David W Fletcher, July 2001

REVIEW OF JOURNAL ARTICLES ABOUT EVALUATION IN HIGHER EDUCATION

Bibliography:

Baresic, Jean, and David Gilman. "How Does the Pendulum Swing on Standardized Testing?" *Education Digest* 66 (Jan/2001): 12-16.

Baumert, Jurgen, and Olaf Koller. "National and International School Performance Studies: What Can They Do, What Are Their Limits?" *European Education* 32 (Fall/2000): 37-49.

Drake, Frederick D., and Lawrence W. McBride. "The Summative Teaching Portfolio and the Reflective Practitioner of History." *The History Teacher* 34 (Nov/2000): 41-60.

Goldhaber, Dan D., and Dominic J. Brewer. "Does Teacher Certification Matter? High School Teacher Certification Status and Student Achievement." *Educational Evaluation and Policy Analysis* 22 (Summer/2000): 129-145.

Hancock, Dawson R. "Effects of Test Anxiety and Evaluative Threat on Students' Achievement and Motivation." *The Journal of Educational Research* 94 (May-Jun/2001): 284-290.

Horner, Wolfgang. "'Europe' as a Challenge for Comparative Education–Reflections on the Political Function of a Pedagogical Discipline." *European Education* 32 (Summer/2000): 22-36.

Kaplan, Robert M., and Dennis P. Saccuzzo. Chapter 10, "Theories of Intelligence and the Binet Scale" in *Psychology Testing: Principles, Applications, and Issues*, 5th ed. Belmont, CA: Wadsworth/Thomson Learning (2001): 253-277.

Kohn, Alfie. "High-Stakes Testing as Educational Ethnic Cleansing." *Education Digest* 66 (Dec/ 2000): 13-18.

Midgley, Carol, Avi Kaplan, and Michael Middleton. "Performance-Approach Goals: Good For What, For Whom, Under What Circumstances, and At What Cost?" *Journal of Educational Psychology* 93 (No. 1/2001): 77-86.

Schiller, Kathryn S., and Chandra Muller. "External Examinations and Accountability, Educational Expectations, and High School Graduation." *American Journal of Education* 108 (Feb/2000): 73-102.

Baresic, Jean, and David Gilman. "How Does the Pendulum Swing on Standardized Testing?" *Education Digest* 66 (Jan/2001): 12-16.

Baresic and Gilman present an interesting but brief survey of the history of standardized testing for the last half century, i.e., since World War II. They assert, "Standardized tests originated to answer the need of teachers and principals for a way to ascertain how their students were doing," but this is only partly true (for example, in France standardization of evaluation for children originated with the French government's perceived need to identify *idiots*, *imbeciles*, and *morons* for the express purpose of removing them from the general educational system in order to institutionalize and specialize). Nevertheless, the authors very nicely highlight the ups and downs, the pros and cons, and the beliefs of those who favor versus those who reject the practice of standardization. They note that "over the years, standardized testing has received mixed reviews. At one end of the spectrum are those who view teachers and schools as the problem and tests as the solution. Supporters of this point of view believe that tests are the only way to hold schools and teachers accountable, as well as the only way to determine whether schools, students, and teachers really are successful. . . . At the other end of the spectrum are those who consider schools and students to be the solution, and standardized tests to be the problem. Although these educators are not necessarily opposed to testing, they believe that the tests and the time spent preparing students for them get in the way of learning. They believe that the test, rather than learning, becomes the objective." According to the authors, between these two ideological poles the pendulum swings back and forth.

From the outset of standardization, many viewed the tests as a "necessary evil" or even a proverbial "bad boy" of education, something akin to corporal punishment. This continued after criterion-referenced testing was introduced in the mid-1960s, and especially during the height of

the anti-testing movement in the early 1970s. During the latter, the National Education Association (NEA) tried to abolish all standardized tests, and with some limited success, especially in New York City schools, some tests became difficult to market (i.e., they were criticized as biased, elitist, immoral, and evil). However, modest gains in the agenda of the anti-testing movement soon suffered a major and irreversible setback. "In an 18-month period between late 1976 and early 1978, 39 states passed laws or regulations requiring students to take statewide minimum-competency tests," and many of the states aligned these competency tests with graduation or promotion.

Since the mid-1970s, the market for standardized tests has increased steadily, and publishers have quietly reintroduced the tests that they previously discontinued, often with just semantic changes like "tests of cognitive ability" instead of "intelligence tests." For the last quarter of the twentieth-century, "mandatory statewide testing has been the rule rather than the exception," but the authors see this development as "a political rather than an educational solution" to assessment and standards. They highlight the notion that "educators now have begun to realize that statewide testing does no more to improve students' learning than measuring humidity improves the weather, or measuring blood pressure improves health." This belief results from the difficulties that educators have with standardization: the problem of equity or fairness, the problem of accurate assessment, and the pressure on administrators and educators to cheat or skew results.

As more states mandate higher standards and more tests, quality instruction gives way to intense cram sessions, test-coaching programs, and pull-out remediation classes. Students who are low-achievers or special-education are encouraged to stay home. The emphasis on results

puts enormous pressure on administrators and school personnel and thereby increases the likelihood that educators will "bend the rules" in order to comply with bureaucratic standards. Protests of standardization have helped raise the issue to the forefront of public consciousness but have produced little change (e.g., the Organized Students of Chicago, 1999). More action is needed to swing the pendulum back in the other direction away from standardization. While most teachers desire this "move beyond the current fixation on test scores," the reality is that "in their zeal to score well on statewide mandated tests, many educators have resorted to a massive teach-to-the-test strategy that is diverting efforts to improve student achievement, schools, teacher education, and public opinion." So, "as a result, quality education in the public schools continues to plummet, while charter and private schools increase in both number and popularity."

With this assessment, the authors conclude, "Only time will tell when and how the testing pendulum will swing. But if it is to swing back toward a more reasonable role for testing, it ought to do so soon, because the pendulums for charter and private schools are swinging, too–in the opposite direction."

Baumert, Jurgen, and Olaf Koller. "National and International School Performance Studies: What Can They Do, What Are Their Limits?" *European Education* 32 (Fall/2000): 37-49.

Baumert and Koller lament "no tradition in the Federal Republic of Germany of a permanent monitoring of the results of institutionalized education," but they applaud efforts of educational science over the last two decades toward "the development and testing of models to optimize work in individual schools and to work out didactic models and introduce them into classroom practice." Addressing European and primarily German concerns, they focus on two studies that began in the 1990s–the "Educational Careers and Psychosocial Development in Youth" (ECPDY) and the "Third International Mathematics and Science Study" (TIMSS). First, they summarize some of the findings from both of these important studies. Then second, they relate a few of the problem areas and offer some plausible solutions.

ECPDY began in Germany when their "school system was transformed in the new states to a tiered school system and the cognitive and psychosocial development of youth and young adults [was] studied as a function of the various family and institutional conditions. The study comprised performance in several disciplines; a broad range of cognitive, sociocognitive, motivational, and other personality features; the institutional conditions of learning and development; and quality attributes of the classroom process." The TIMSS project was carried out by the International Association for the Evaluation of Educational Achievement (IEA), which is a research organization composed of several member nation states. In their study, which is primarily descriptive and is on-going since the intent is "to gather information for a long-term observation of school systems," the IEA studied simultaneously for the first time "mathematics

and science achievements of key core cohorts in the primary school and the first and second level of secondary schools."

Based upon the research conducted by these projects, Baumert and Koller offer what they call "a few well-founded conjectures on cause-and-effect relationship." First, "there is no single causal factor running throughout that would explain the sometimes immense differences . . . as regards pupils' achievement." Second, "the findings of the TIMSS yield no structural argument for or against an integral school form." Third, "systematic explanations for differences in achievement are probably to be found in the following areas: (1) in the general evaluation of education and school learning and the associated willingness to invest personal resources; (2) in the specific learning culture of a school system and in the significance attributed to a continued acquisition of knowledge and the associated effort and endurance; (3) in how society and the school evaluate certain disciplines; and (4) in the quality of instruction itself." As far as the ECPDY study goes, the authors mention the following central points: (1) "schools in the unified system of the former GDR tended to better achievement results"; (2) "the gymnasium is socially always the most homogeneous, as regards achievement"; (3) "differences in the level of achievement among pupils can be largely attributed to institutional markers–the particular school type, the specific school, and, to a lesser degree, the particular class"; and (4) "differences in achievement over time are for the most part institutionally related."

The authors then discuss criteria for the reliability of results from international achievement studies–i.e., "the test assignments chosen, curricula and classroom validity, the appropriate representation of the abilities being determined, the format of the answers, and a formulation suitable for the cohort being studied." Points of criticism for these criteria include

the problem of curricular validity, the problem of answer format, and the cognitive skills and abilities being tested. They offer some salient criticisms of evaluation of achievement via testing. First, "a pupil can only master material that is also dealt with in the classroom. In school achievement studies there is almost always the dilemma that no special achievement tests can be constructed for the classroom instruction on individual classes. Achievement tests are therefore oriented to the curriculum." This problem of curricular validity will almost always skew results. Second, "the problems in school achievement tests usually have multiple-choice items," whereas "scientific criticism has long held that multiple-choice problems essentially determine only reproductive achievements, not productive skills–inferential thinking, complex operations, or even problem solving. What is more, multiple-choice problems also do not allow for systematic analyses of mistakes. Open-answer formats are considered preferable." But the authors do challenge the accuracy of the assumption in this argument: "Open problems are generally only somewhat more difficult, since there is no possibility of an intelligent guess. It is a mistake to believe that challenging problems cannot be developed in the multiple-choice format–problems that give a picture of independent thinking, methodological skills–as in experiments in natural science–or an understanding of the problem." Third, school achievement studies help broaden the acquisition of knowledge to "the other learning" or "cross-subject thinking, application, research, problem solving, and self-organization of learning" by highlighting these contexts as testable. The authors remark:

> The best precondition for a cumulative learning process and independent successful continued learning is not formal key qualifications but a solid and well-organized basic knowledge in each school subject, by which we mean not knowledge acquired piecemeal and by rote, but an intelligently ordered, interlinked knowledge tested in various situations and flexibly adaptable. This applies equally to facts, concepts, theories, methods, and processes. As the cognitive demands of tasks and problems

> increase in difficulty and complexity, the importance of specific prior knowledge for working successfully with them increases as well. The accumulation of intelligent knowledge is usually a process of many long years requiring intensive practice; it demands not only effort and endurance but at the same time systematic training in the elements of "the other learning"–application, transfer, restructuring, and integrating. Conversely, cross-subject thinking, application, research, and methods or even the self-organization of learning are hard to imagine without a solid foundation in knowledge.

Baumert and Koller conclude by discussing the problem of how to determine social behavior in the classroom and the implications for student achievement. They also mention analyses of the different types of schools (in Germany, the gymnasium, the *Realschule*, the comprehensive school, and the *Hauptschule*). While "no differences in achievement were found in a comparison of the *Hauptschule* and the comprehensive school," the authors do report that "a comparison between the *Realschule* and the comprehensive school shows that in the *Realschule* the pattern of achievement is more favorable according to tests of cognitive and social starting variables." They argue that the influence of social stratum in achievement differentiation "is relatively slight or statistically not demonstrable according to the test of cognitive preconditions," therefore "influence of social origin on the achievement pattern *within* a type of school is usually overestimated in comparison to its importance in the selection process for the next stage of school." Rather, they assert "the development of prosocial motives as an example of differences in development in the social-cognitive domain." Since "prosocial motivation shows a typical pattern of variation during the course of adolescence," the egoism (decreasing with age) / altruism (increasing with age) construct, although burdened with weighty presuppositions, may help explain such an achievement differentiation.

All in all, the authors realize that the ECPDY and TIMSS studies are limited and not well suited for "the development and revising of didactic measures." But they do indicate positively

that “the importance of school for cognitive and social development can now be separated from the effects of different starting conditions on the part of pupils. However, there are limits here in the micro-analysis of individual learning and developmental processes.”

Drake, Frederick D., and Lawrence W. McBride. "The Summative Teaching Portfolio and the Reflective Practitioner of History." *The History Teacher* 34 (Nov/2000): 41-60.

Drake and McBride outline for beginning and experienced teachers of history both rationale and strategy for planning, constructing, and presenting a summative teaching portfolio. This use of an individualized tool for growth and reflection intends to help not only teachers of history but also administrators in their assessment of performance. To these ends, Drake and McBride provide the following pragmatics: a checklist of documents, criteria for assessment at different stages of a teacher's career, a chart integrating education and history standards, and an analytic rubric for formal review of the portfolio. The rationale for the creation of a summative teaching portfolio includes the following: to encourage teacher self-reflection and the search for deeper insights, to instill practitioner confidence especially in light of a competitive employment market, to increase administrator confidence in making hiring decisions, to chart and document goal accomplishment, and to demonstrate professional growth and development. A further important rationale involves the growing consensus towards standards-based documentation of teacher performance at both state and national levels.

Drake and McBride offer a very practical guide for compilation and use of the summative teaching portfolio. Based upon Dewey's distinction between "unreflective" and "reflective" teaching (reflective teaching "combines interesting content and sound pedagogical practices . . . adapts content and methodology according to experiential levels and interests of students . . . exercises responsibilities toward students and community"), they integrate the five traditions of reflective practice as defined by Zeichner and Liston: academic, social efficiency, developmentalist, social reconstructionist, and generic. These approaches can equally inform each of the three broad categories of items in the portfolio: (1) materials that document the

teacher's performance as a scholar, (2) materials that document the teacher's philosophy of education and his / her performance as a teacher, and (3) materials that document the teacher's professional qualifications and personal achievements.

Drake and McBride recommend that the portfolio be self-explanatory, i.e., that it can stand alone without explanation or additional comment, as this would help administrators easily assess the teacher's *knowledge* of content and pedagogy (the cognitive domain), *performance* in the classroom (the psycho-motor domain), and *disposition* towards the profession (the affective domain). This assessment could be done according to six levels of proficiency–Master, Distinguished, Accomplished, Proficient, Apprentice, and Novice. In order to help administrators evaluate systematically, the authors offer a six level rubric, with guidelines for its effective use, by which to compare and rank teacher portfolios.

Overall, application and appreciation for the content of history is assumed by Drake and McBride, and they add this important disclaimer: "the reliability of a summative portfolio as a performance of a teacher's knowledge, teaching, and disposition will depend on the portfolio's comparison with other teaching performances *including direct observation* [emphasis mine, DWF]. Thus, the pragmatic value of the summative teaching portfolio remains limited and must yield to more empirical, existential evidence! Further, Drake and McBride include too many details in the assessment parts, i.e., criteria and rubrics, so one wonders whether they have ever functioned as administrators with many, time-consumptive tasks!

Goldhaber, Dan D., and Dominic J. Brewer. "Does Teacher Certification Matter? High School Teacher Certification Status and Student Achievement." *Educational Evaluation and Policy Analysis* 22 (Summer/2000): 129-145.

Goldhaber and Brewer examine the impact of high school teacher certification on student achievement. The authors note the importance of the issue for two reasons: (1) the increase of teachers due to greater enrollment and unprecedented teacher retirements (i.e., according to the National Commission for Teaching and America's Future [NCTAF], "more new teachers will be hired in the next decade than in any previous decade in our history"); and (2) the fact that "very little research exists on the effectiveness of the teacher licensure system, in terms of how well teachers subsequently teach and what works to promote positive student outcomes." Since "much of the educational establishment takes for granted that licensure is an important and effective screen on the quality of teachers," a rigorous study looking at performance of teachers with standard certification versus teachers with temporary or provisional license is justified. Further, since licensure remains a function of each individual state, a comparison of differences in requirements state-by-state in relation to student achievement is warranted.

Goldhaber and Brewer give some important background information before they relate their findings. They raise the question, "What does licensure do?" in order to highlight a significant lacunae in educational research, i.e., "to our knowledge, there have been no studies that use national data to examine the relationship between teacher licensure and student outcomes." Today, "a relatively standard path by which teachers obtain the credentials necessary to teach in the system" exists. Earlier "normal" schools (late 19th century) and apprenticeship programs (early 20th century) have been replaced by the typical college or university baccalaureate degree program that adheres to state guidelines for accreditation of teachers. The

authors agree that most educators assume the value of such programs. But "it is not entirely clear how much is really known about the components of an effective licensure program" (*a la* the degree of variables from state to state, and even from program to program within a state), and there is little reliable evidence about the impact of these programs (i.e., "only a few quantitative studies have explicitly analyzed the link between licensure and student performance"). Accordingly, "the very concept of teacher certification has been criticized by some on the grounds that there is no concrete evidence backing the claim that teacher certification policies result in more qualified teachers." True, some evidence–based on the use of standardized tests, minimum grade point average, and licensure exams for screening prospective teachers–links such certification criteria to teacher performance, but the existence of different types of certification in addition to "standard" certification, plus the fact that state licensure policies fail to screen out poor candidates for the teaching profession (along with the problem of tenure which makes the dismissal of poor teachers very difficult), minimize the relevancy of this evidence about the relationship between these credentials and student outcomes. In their study, Goldhaber and Brewer focus on this relationship.

For methodology, the authors use a multiple regression framework to analyze variable determinants of individual performance of high school students, primarily in mathematics and science. They reference the *National Educational Longitudinal Study of 1988*, "a nationally representative survey of about 24,000 8th-grade students," for a primary source of data, while their sample set consists of 3,786 students in mathematics and 2,524 students in science, all in public schools in the 12th grade. Classifications of certification status include "regular" or "standard," "probationary," "emergency," "private school certification," and "not certified."

They also consider the socioeconomic situation of students, and they conclude, "These results provide prima facie evidence that students are not randomly distributed across teachers by type of certification." Goldhaber and Brewer look at some key state licensure features, which are: (1) "whether a state requires prospective teachers to take a test prior to entering a school of education"; (2) "whether a state requires new teacher graduates to take a test prior to licensing"; (3) "the minimum cutoff score on the National Teachers Exam and the Pre-Professional Skills Test that teachers much achieve prior to obtaining a license"; (4) "the percentage of teachers in each state who pass the state licensure exams"; (5) "whether states require field experience prior to student teaching"; and (6) "the number of weeks of full-time student teaching required prior to licensure." Finally, in order to summarize their results they use four sets of general explanatory variables: "individual and family background variables, school variables, teacher variables, and class variables."

Basically, this study by Goldhaber and Brewer discovers little evidence to support the belief of a "strong relationship between state certification policies and student outcomes." Some evidence indicates higher test scores from mathematics students who have teachers with a degree in mathematics versus students whose teachers have out-of-subject degrees. Other evidence shows no variation in science based on the subject-specific degree of the teacher. Yet, one quite interesting variation reveals a *negative* impact on mathematics scores of students when the teacher holds a degree in education! The authors do not seem surprised by this, since "most college students selecting education majors tend to be drawn from the lower part of the ability distribution" [they cite Hanushek & Pace (1995) and Henke, et al. (1996)]. They also conclude that students of teachers with provisional types of licensure perform no worse than students who

have teachers with standard certification! In their research, they uncover little evidence of any advantage in outcomes from teachers with regular credentials versus teachers with emergency certification. Further, they find no data to support differentiation in student achievement based on licensing criteria for individual states relative to each other. They note that the presumption that equates higher certification requirements with higher quality teachers is not well founded.

What are some possible explanations for this lack of evidence supporting a correlation between teacher certification and student achievement? Goldhaber and Brewer offer four generalizations: (1) "state policy variables may not be properly linked with teachers"; (2) "the decision by states to impose or strengthen certification requirements may not be random, i.e., it is possible that states with poor student performance are systematically more likely to change certification policies"; (3) "it is possible that the state requirements do not represent a significant barrier to becoming a teacher"; and (4) "it is conceivable that these requirements actually act to screen out individuals who might have been successful teachers." From the perspective of this reviewer, these latter two judgments seem highly likely, since most post-baccalaureate teacher licensure programs treat seasoned educators as if they were raw teenage recruits. The work of Goldhaber and Brewer not only calls for more research on this important connection between certification and outcomes, but also for a push to overhaul existing certification programs with a view toward application to real world outcomes.

Hancock, Dawson R. "Effects of Test Anxiety and Evaluative Threat on Students' Achievement and Motivation." *The Journal of Educational Research* 94 (May-Jun/2001): 284-290.

Hancock seeks to extend to a specific postsecondary setting the work of other researchers on the link between test anxiety / evaluative threat and student achievement / motivation. He defines test anxiety as a personality trait that "prompts an individual to react to threatening situations with sometimes debilitating psychological, physiological, and behavioral responses." He reviews the substantial empirical evidence for the "negative effects of test anxiety on academic performance." He also highlights both positive and negative impact of evaluations by teachers on student performance. To evaluate student motivation, Hancock follows a model offered by Vroom (1964) called *expectancy theory* which suggests three perceptual relationships causative to "the amount of effort an individual ultimately will exert": "(a) *expectancy*–an individual's subjective estimation of the likelihood of successfully performing a particular behavior; (b) *instrumentality*–a person's subjective estimation of the likelihood that a particular behavior will be rewarded; and (c) *valence*–the positive or negative value that a person places on a reward."

In this study, Hancock used the Test Anxiety Inventory (TAI; Spielberger, 1980) and the Competition and Teacher Control subscales of the Classroom Environment Scale (CES; Moos & Trickett, 1986) for his independent variables, since both have established reliability and validity. His dependent variables were measured by "a professor-made, criterion-referenced test in a graduate-level course in research methods" (i.e., achievement level) and "a version of Vroom's expectancy theory model" (i.e., motivation level), but both have limited evidence of reliability and validity. Hancock's methods involved subjecting 61 postbaccalaureate and graduate students, who were taking an Educational Research Methods course at a large university in the

South, to varying levels of anxiety and threats (i.e., high-threat and low-threat conditions). Further, in order to facilitate differentiation from conditions explored by other researchers, he followed a quasi-experimental design in this study.

Hancock's results parallel that of previous researchers, so little is to be gained from this particular study. Generally, "students with a predisposition toward test anxiety do not achieve at a lower level than do students who are not naturally test anxious," but they are "significantly more sensitive to environments in which competition is emphasized and teacher control is evident." Also, the author's results show that "all students, regardless of their tendencies toward test anxiety, achieve more poorly under conditions of high evaluative threat." Similarly, motivation of students rises or falls correspondingly with feelings of anxiety and perceptions of threat. While the conclusion that "to help students master course content and remain motivated to learn, university and college professors might lessen their control over classroom procedures and the extent to which students perceive the need to compete with each other" is well taken, Hancock's study, which basically mimics previous research on academic motivation, relies upon minimal data with a narrow representation. In order to give his work more merit, he would do well to broaden his questions to include cultural, economic, and social factors that possibly relate to the link between anxiety / threat and achievement / motivation. To improve his research significantly, he would also need to pool a much larger, more diversified student population.

Horner, Wolfgang. "'Europe' as a Challenge for Comparative Education–Reflections on the Political Function of a Pedagogical Discipline." *European Education* 32 (Summer/2000): 22-36.

In his introduction, Horner carefully defines the pedagogical relevancy of his topic, and then continues his discussion by dabbling in semantics (comparative education as "an international conceptual setting"). Ideologically, the move from a multiple nation-state Europe toward a united Europe means that "Europe" is "not just a geographical concept but a *political process*." Pragmatically, evidence of relevancy for higher education curricula is seen, for example, in a mandatory module in the M.A. program in education at the University of Leipzig: "Comparative Education with Special Consideration of the Process of European Unification." Overall, "the presumption is that the process of European unification is not just a politically managed process but that it is understood as a process of social consciousness that perhaps has some feedback effects on the object, namely, comparative education." According to the author, that this process of social consciousness involves a "challenge" implies "a set criterion for the individual limits of performance," and such a "performance of social subsystems" presumes that "concrete ideas exist as to the manner in which the discipline concerned may be expected to 'perform'."

Horner gives the greatest attention to "concrete ideas" rather than any "set criterion" that define performance. He surveys the historic development of the German concept *Bildung* (along with its many hybrids), the French expression *education compare*, and the English phrases *comparative education* and *pedagogical science*. He notes that these variations in "the logic of the language" cast doubt on whether the discipline can be narrowly reduced to "a heuristic procedure alone" as versus the object of the discipline (i.e., education "elsewhere" as implied by

the terminology "international education"). He then raises the important question, "What should this discipline accomplish?"

A necessary function of the discipline, which is implied in the terminology, is the role of comparison. But how is such defined? Horner offers "four functions of comparative research . . . derived by hybridization of the two pairs of opposite (theoretical versus practical) interest, and a 'particularity and universality'." First, the *ideographic function* [theoretical and particular] utilizes classic historical methodology to describe and explain "that which is special and unique in the phenomena being studied." From this perspective, pedagogical processes are seen as closely connected with their cultural and social milieu. Second, the *ameliorative function* [practical and particular] searches for a better way to do things by comparing the work of others, i.e., "makes positive use of the experience of others." Third, the *evolutionary function* [practical and universal] looks for general trends in development, "a definite, unique pattern in the dynamic of movement." Quite often this perspective discerns "a secondary intention bearing on educational policy," so that "the quest for one's own position on this implicit evolutionary scale imparts a 'cryptonormative' function to the developmental dynamic." Fourth, the *experimental function* [theoretical and universal] compares similar phenomena in different contexts and thereby recreates the experimental situation. This perspective originates in the work of French sociologist Emile Durkheim who suggested that "comparison in the social sciences should perform the role that experiment performs in the natural sciences, since the creation of artificial experimental situations to isolate variables is for practical and ethical reasons not directly possible in the social sciences." Horner then applies these functions of comparison to the epistemological problem of transferability, and he concludes that "the possibility in principle of

transfer is the source of much of the political attractiveness of comparative educational research." The author likewise notes an important *didactic function* to comparative education, as well as a *practical-pedagogical function* and an *intercultural function*. Each of these latter three aspects, especially the didactic role, impact both evaluation and testing in the international [European] setting.

Finally, the author addresses the problem of European integration from the viewpoint of diverse nation-state *educational systems*. In this regard, the Treaty of Maastricht provides for the development of top-quality academic and vocational education "with strict observance of the responsibility of the member states for the content and the structure of the educational system" and "to the exclusion of any harmonization of the legal and administrative rules and regulations of the member states." But with very little fleshed out by way of practical details, Horner laments that "unity of European education remains strictly an ideal." So what is the role of comparative education for this epochal situation? Horner himself asks, "Can comparative education have any function at all in this established context? Is not this assumption an enormous self-exaggeration?"

With this query, he finds a valid connect, a pragmatic function of comparative education, between "the development of criteria for evaluating formal achievement" and "degree and diploma equivalents." Furthermore, "what society expects from comparative education goes clearly beyond a mere idiographic function toward providing political know-how, for example, to determine functional equivalents for formal certifications defined in terms of different criteria." For Horner and others, this task is an urgent one. What the author suggests, therefore, is "an objectifying and relatively aloof comparative view from without . . . teachers working locally

with pupil exchange . . . to achieve a more precise understanding of [other] systems and school cultures." In reality, Horner need look no further than across the Atlantic Ocean to the practice of comparative education within the United States of America! But when he looks to the diversity of educational theory and practice within the different states in this country, he will likely be just as bewildered about what constitutes valid "criteria for evaluating formal achievement!"

Kaplan, Robert M., and Dennis P. Saccuzzo. Chapter 10, "Theories of Intelligence and the Binet Scale" in *Psychology Testing: Principles, Applications, and Issues*, 5th ed. Belmont, CA: Wadsworth/Thomson Learning (2001): 253-277.

Kaplan and Saccuzzo begin by saying that "of all the major concepts in the field of testing, intelligence is among the most elusive," and then they quote sundry definitions from Binet (1916), Spearman (1923), Freeman (1955), Das (1973), Gardner (1983), and Sternberg (1986). They identify, from T. R. Taylor (1994), three independent research traditions on human intelligence: the *psychometric*, which examines the "fundamental structure" of a test; the *information-processing*, that examines the processes humans use to learn and solve problems; and the *cognitive*, which analyzes how humans adapt to real-world demands. Of these three, the psychometric is the oldest and is the focus of the authors who discuss in depth the history of the Binet Scale. Standardized intelligence tests, which ironically were initially developed to eliminate subjectivity in the evaluation of children, but consistently have exhibited a correlation between scores and socioeconomic backgrounds (i.e., children of college graduates average ten points above the mean), have been criticized since their inception.

In 1904, Alfred Binet, a member of a commission appointed by the French minister in charge of education to recommend a procedure for identifying "subnormal" (i.e., intellectually limited) children, worked with few guideposts apart from his earlier research on human abilities. He defined intelligence as "the capacity (1) to find and maintain a definite direction or purpose, (2) to make necessary adaptations–strategic adjustments–to achieve that purpose, and (3) for self-criticism so that necessary adjustments in strategy can be made." Along with his colleague, T. Simon, Binet originally developed a thirty item test, arranged in increasing order of difficulty,

that measured judgment, attention, and reasoning according to two major concepts–age differentiation and general mental ability.

In Great Britain, Charles Spearman paralleled Binet's work with his notion of a general mental ability factor–*psychometric g*–underlying all intelligent behavior. Spearman advanced the phenomenon called *positive manifold* (i.e., positive correlations to results from diverse ability tests), as well as a statistical technique called *factor analysis* (i.e., to reduce variables to a smaller number of hypothetical factors). Spearman's idea of a single intelligence influenced psychologists through the twentieth century until theories of multiple intelligences appeared quite recently (i.e., the *gf-gc theory* of a *fluid* and a *crystallized* intelligence).

The early Binet Scale (1905) identified three levels of intellectual deficiency (i.e., idiot, imbecile, and moron), but it lacked adequate measuring units, normative data, and verification bases. The 1908 scale, however, did add an *age scale*, whereby items were grouped according to age level rather than by difficulty, although the test remained heavily weighted on language, reading, and verbal skills (something not corrected until the 1986 revision). Additional revisions of what was called the Stanford-Binet Intelligence Scale in 1916, 1937, and 1960 added the Intelligent Quotient (IQ) and improved standardization samples, but made no provisions for restandardization or *normative sampling*.

The modern Stanford-Binet (1986, 4th ed.) continues the "tradition of innovation and incorporation of central psychometric and theoretical concepts" with the benefits of new research in cognitive psychology, adjustments for social and cultural changes, and improved subtests, summary scores, and procedures for administration. The modern Binet uses a three-level hierarchical model of intelligence as its basis–*crystallized abilities* or learning, the realization of

original potential through experience; *fluid-analytic abilities* or original potential, the basic capabilities that a person uses to acquire crystallized abilities; and *short-term memory* or memory during short intervals, the amount of information a person can retain briefly after a single, short presentation.

Overall, the modern Binet compares well with other tools used to evaluate intellectual capability. While cumbersome to administer and still showing a consistent trend in correlation of mean score increases with socioeconomic factors, the modern version of Binet remains "a well-constructed instrument [that] meets the highest standards for a modern psychological test." As such, the Stanford-Binet Intelligence Scale can be used effectively as an accurate and fair indicator of cognitive / analytic achievement of students, albeit with a recognition of the scale's limitations. Kaplan and Saccuzzo have done an excellent job of tracing both the history of the Binet scales and their modern significance.

Kohn, Alfie. "High-Stakes Testing as Educational Ethnic Cleansing." *Education Digest* 66 (Dec/ 2000): 13-18.

Kohn's hard-hitting indictment of standardized testing begins with eight "indisputable facts":

> — Fact 1. Our children are tested to an extent unprecedented in our history and unparalleled in the world.
> — Fact 2. Noninstructional factors explain most variance among test scores when schools or districts are compared.
> — Fact 3. Norm-referenced tests were never intended to measure quality of learning or teaching.
> — Fact 4. Standardized-test scores often measure superficial thinking.
> — Fact 5. Virtually all specialists condemn giving standardized tests to children under 8 or 9 years old.
> — Fact 6. Virtually all relevant experts and organizations condemn basing important decisions, such as graduation or promotion, on the results of a single test.
> — Fact 7. The time, energy, and money devoted to preparing students for standardized tests have to come from somewhere.
> — Fact 8. Many educators are leaving the field because of what is being done to schools in the name of "accountability" and "tougher standards."

Kohn develops each of these facts briefly, then affirms that the supposed cure, i.e., standardized testing, to many educational ills is much worse than the disease. He quotes Senator Paul Wellstone (D-Minnesota), "Making students accountable for test scores works well on a bumper sticker, and it allows many politicians to look good by saying that they will not tolerate failure. But it represents a hollow promise. Far from improving education, high-stakes testing marks a major retreat from fairness, from accuracy, from quality, and from equity."

Kohn affirms potent reasons why he believes the statement from Senator Wellstone is true, and each of these is based in socioeconomic realities. First, the tests may be biased in favor of students with affluent, well-educated parents. Second, affluent families, schools, and districts can more readily afford people, products, and processes that enhance test preparation. Third, since "standardized tests tend to measure the temporary acquisition of facts and skills, including

the skill of test-taking itself," quality instruction declines for those who need it the most. Fourth, these tests tend to focus on "standards" alone, while other concerns (good faculty, adequate facilities, and wholesome environments) remain ignored. Fifth, those who are supposed to be helped (e.g., minority / low-income students) will be excluded, since they are "disproportionately affected by the incessant pressure on teachers to raise scores." Kohn charges, "Unless we act to stop this, we will be facing a scenario that might be described without exaggeration as an educational ethnic cleansing."

In conclusion, the author offers an interesting hypothetical scenario by way of support for his belief that the use of high-stakes testing, i.e., standardized testing, is nothing less than about sorting. He writes:

> Imagine that almost all the students in a given state met the standards and passed the tests. What would be the reaction from most politicians, business people, and pundits? Would they concede our public schools are terrific–or would they take this to show that the standards were too low and the tests too easy? The phrase "high standards" by *definition* means standards that everyone won't be able to meet.

Rather than play the game of testing, which many students of color cannot win, Kohn suggests an alternative approach toward school reform: the elimination of tracking, more equitable funding, and the use of more sophisticated pedagogical methods. Contrarily, "standardized testing, while bad news across the board, is especially hurtful to students who need our help the most." Kohn has made an important point, perhaps overstated with the use of the "ethnic cleansing" metaphor, but he has offered a crisp, compact argument that calls for a rejoinder with pertinent empirical data.

Midgley, Carol, Avi Kaplan, and Michael Middleton. "Performance-Approach Goals: Good For What, For Whom, Under What Circumstances, and At What Cost?" *Journal of Educational Psychology* 93 (No. 1/2001): 77-86.

Midgley, Kaplan, and Middleton, like most recent achievement goal theorists, focus on *why* individuals are motivated, rather than concentrating on individuals as either possessing or lacking motivation. They define achievement goals as "purposes for behavior that are perceived or pursued in a competence-relevant setting," and while they do not purport to distinguish research on *pursued* goals versus research on *perceived* goals, they do focus on understandings about *performance goals* in contrast to *mastery goals*. The latter describes "the goal to *develop* ability," whereas the former relates "the goal to *demonstrate* ability or to *avoid the demonstration* of lack of ability." Another difference is that mastery goals tend to "focus the individual on the task at hand and relate especially to developing competency and gaining understanding and insight," but performance goals tend to "focus the individual on the self and relate especially to how ability is judged and how one performs, especially relative to others." Since theorists often depict these performance goals in terms of an orientation to demonstrate one's ability or *approach*, as well as an orientation to avoid showing a lack of ability or *avoidance*, they are termed performance-approach (PA) goals (to highlight their more positive aspect). The aim of the authors is to survey relevant research on this type of goal orientation, to indicate themes and questions that help explain the nature of PA goals, and to provide a framework for future research on their effects.

While Midgley, Kaplan, and Middleton cite studies that show a positive correlation between PA goals and adaptive patterns of learning, they intentionally fail to cite studies that indicate no correlation or a negative effect of PA goals on the same outcomes. Hence, from the

very start, they limit the scope of their investigations in favor of an uncritical and unbalanced methodology. For this reason, their conclusions will remain tentative, skewed, and open to criticism.

On the whole, the study presents a confusing array of summaries of findings from quite a host of researchers. However, some broad conclusions seem to be consistent with a certain clarity, e.g., the effect of PA goals on adaptive patterns of learning; a positive link between PA goals and effort; the value of PA goals for cognitive, metacognitive, and self-regulatory strategies; and positive association of PA goals with achievement. However, other outcomes for PA goals are not so certain, such as their relation to "deep processing" (undefined by the authors), "long-term retention" (facilitated best by mastery goals), and intrinsic motivation (there is no evidence of linkage with PA goals, but how do the authors know for sure if they do not evaluate studies highlighting negative influences or outcomes for these type goals?).

The authors raise interesting questions for PA goals, namely, good for whom, and under what circumstances? In reply, the authors feel certain of some benefit for boys versus girls, older students versus younger students, and when used in conjunction with other goals versus a single goal-type orientation. But the evidence again is not conclusive, since research on goal orientation for perceived competence, gender and ethnicity, age and context, and multiple goals is sporadic.

One promising area for the sure effect of PA goals, however, focuses on the question: goals at what cost? Certain self-handicapping strategies, used by students to protect self-worth when they pursue PA goals, have been found to be consistent, especially when students hold little confidence in their abilities. Types of self-handicapping identified are "the avoidance of novelty

and challenge, the avoidance of help seeking, the use of cheating, and reluctance to cooperate with peers."

Ultimately, the authors do not call for a new or revised goal theory, since the evidence for positive effects of PA goal theory remains unverifiable. But the authors have failed to push forward understanding of the overall effects of PA goals by omitting critical negative evidence. While this indeed supports one of the aims of the authors in promoting further research on the topic, it definitely invalidates their own contribution and renders it superfluous.

In the opinion of this reviewer, little of pragmatic value is to be gained by this particular work of Midgley, Kaplan, and Middleton.

Schiller, Kathryn S., and Chandra Muller. "External Examinations and Accountability, Educational Expectations, and High School Graduation." *American Journal of Education* 108 (Feb/2000): 73-102.

Schiller and Muller give a balanced assessment of the value of external examinations for school accountability, the expectations of educators, and high school graduation. They note that "proponents argue that state-mandated assessments of student performance will encourage cooperation between teachers and students leading to greater academic success, while opponents argue they will increase inequality by structurally limiting the opportunities of disadvantaged students." The authors are absolutely correct in their understanding that the current debate over external examinations "is a controversy over which students–motivated and bright, or racially and economically advantaged–will succeed in school if state or national assessments are implemented." This shows that the current dispute is highly political, i.e., "about how policy will structure opportunities" and "whether such policies will improve schools by providing incentives for teachers and students to work cooperatively toward the goal of academic progress."

In their study, the authors assess this ongoing debate by looking at the correlation of policy concerning external assessments and accountability to a particular negative academic consequence, i.e., failure to complete high school. They utilize theory and empirical research about assessment and accountability systems, and they develop a structural analysis to link state testing programs to variations in student achievement of a high school diploma. They also employ hierarchical linear modeling (HLM) to study two national data sets–the National Education Longitudinal Study (NELS, 1988-1992) and the National Longitudinal Study of Schools (NLSS, early 1990s). In order to focus on their particular interest of state variations in assessment policy and effects on students earning a diploma, Schiller and Muller organize their

research around "three well-established predictors of academic success": (1) the social background of students; (2) the motivation exhibited by students as measured by their academic aspirations; and (3) the "gatekeeping" function of educational expectations of teachers.

As introduction to their work, Schiller and Muller rehearse the standard background information about external assessments as incentive systems, and this includes the scope of testing, consequences for students, and accountability for schools. [In my opinion, this is actually the best part of the article.] But, to this reader, the results gleaned from their "analytic plan" as applied to "the data and the variables" seems a bit confused and overall unimpressive. While their analytical technique and statistical reporting certainly merit reflection, the reported results do not match exactly the intended goals of the researchers. Their findings show "that students' probabilities of graduating from high school are associated with states' testing policies, [so that] widespread and frequent testing is positively associated with students' greater likelihood of earning a high school diploma," but this says relatively nothing about their particular interest in "between-state variations in the associations of earning a diploma." No differentiation, no comparison, and no contrast in the data and its interpretation from individual states is mentioned anywhere in the study. This is a serious lacunae that registers as a disconnect to invalidate their conclusions.

Schiller and Muller juggle a lot of data, but they do not make any unqualified assertion. Perhaps the closest they come to a definite affirmation is: "The extensiveness of the states' testing programs is significantly associated with the probability a student will graduate from high school, which increases as the number of academic tests and frequency of testing increases." But this is a generalization, as no particular state is mentioned, and the authors use the qualifying

limitation–"probability." In their discussion (conclusion), the two professors do offer some links to their three guiding principles or predictors of academic success, but the argument remains fuzzy, i.e., "some states," "states with more consequences for schools," "states with more extensive testing," and so forth. The problem, in the view of this reviewer, is that no specific state is ever mentioned anywhere, so the goal sought by the authors remains unapparent. A rewrite of the entire article to clarify these issues would be welcomed.