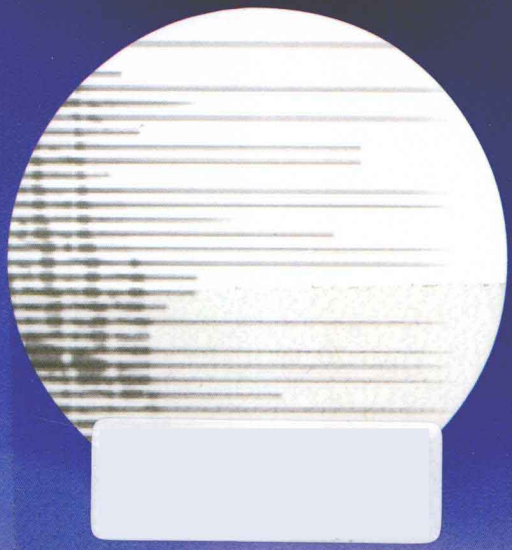
Achievement Assessment in an Educational Reform Era

杨向东　黄小瑞　主编

教育改革时代的学业测量与评价



华东师范大学出版社
全国百佳图书出版单位

杨向东　黄小瑞　主编

教育改革时代的学业测量与评价

Achievement Assessment in an Educational Reform Era

华东师范大学出版社

图书在版编目(CIP)数据

教育改革时代的学业测量与评价/杨向东，黄小瑞主编. —上海：华东师范大学出版社，2013.6
ISBN 978-7-5675-0929-0

Ⅰ.①教… Ⅱ.①杨…②黄… Ⅲ.①基础教育—教学评估—研究—中国 Ⅳ.①G632.0

中国版本图书馆 CIP 数据核字(2013)第 140858 号

教育改革时代的学业测量与评价

主　　编　杨向东　黄小瑞
策划编辑　彭呈军
审读编辑　韩秀秀
责任校对　王　卫
装帧设计　高　山

出版发行　华东师范大学出版社
社　　址　上海市中山北路 3663 号　邮编 200062
网　　址　www.ecnupress.com.cn
电　　话　021-60821666　行政传真 021-62572105
客服电话　021-62865537　门市(邮购) 电话 021-62869887
地　　址　上海市中山北路 3663 号华东师范大学校内先锋路口
网　　店　http://hdsdcbs.tmall.com

印 刷 者　上海商务联西印刷有限公司
开　　本　787×1092　16 开
印　　张　18.75
字　　数　373 千字
版　　次　2013 年 10 月第 1 版
印　　次　2013 年 10 月第 1 次
书　　号　ISBN 978-7-5675-0929-0/G·6645
定　　价　38.00 元

出 版 人　朱杰人

目录

01 大规模学业测评的理论与技术

02 课堂评价:如何与课程、教学相整合

03 新颖与创新型学业评价的研究进展

04 学校及区域经验分享

01

大规模学业测评的理论与技术

National Testing: Promises and Pitfalls — The NZ Perspective

Gavin T L Brown

The University of Auckland

Abstract: Tests and examinations are frequently used to evaluate both students and schools/teachers. This is easily done and intuitive — good schools produce students who do well on examinations. However, there are problems with this simplistic approach. Tests and examinations generally report achievement using total scores (often percentages or letter grades) that are used to determine how well a candidate did compared to others. Unfortunately, this information does not help teachers, students, parents, or even employers determine the strengths and weaknesses of a learner or what the next thing to learn is. Clearly, assessments must provide richer diagnostic information so as to contribute to the improvement agenda. In New Zealand, there is wide-spread use of tests for improvement rather than evaluation or certification within a context that prioritises assessment for learning. This paper will illustrate the New Zealand approach by focusing on the Assessment Tools for Teaching and Learning system which was designed at the University of Auckland as a means of informing teachers and school leaders in Years 5 to 12 about which children needed to be taught which part of reading, writing, or mathematics curriculum.

Keywords: formative assessment; diagnostic testing; standardised testing; asTTle; assessment policy

Politicians, public policy, and parents want to know that students in schools are learning what is expected. Those expectations are normally laid out in curriculum documents which indicate content and difficulty aspects. Further, society wants to know if and how much value schools are adding to student learning, given that not every school starts with the same quality of students (Butterfield, Williams, & Marr, 1999; Firestone, Mayrowetz, & Fairman, 1998; Smith, Heinecke, & Noble, 1999). Society has a valid concern for information about student academic outcomes because these contribute to an individual student's life chances as well as overall societal development (i.e., educated individuals make life better for all).

Unfortunately, not every school or teacher is as good as another. However, it is difficult, time-consuming, and expensive to improve teachers. Provision of curriculum documents is not enough and so many societies turn to educational testing to evaluate schools and teachers.

Using tests to evaluate is easy (Linn, 2000). Tests are relatively cheap and quick to create; as well, the process of calling for new tests, implementing them, and using them to evaluate is reasonably quick — usually quicker than the time it takes to train a new cohort of teachers. Unfortunately, test scores can be inflated without any real learning taking place. Known techniques (Cannell, 1989; Hamilton, 2003) include: 1) teaching to the test, 2) using the same test year after year so that norms are out of date, 3) teachers learning what is on the test, 4) getting low performing students not to come to school on testing days, and 5) prompting students during testing, 6) correcting student test responses before submission. Despite these threats to the validity of tests, many societies believe that examination and test scores are a robust way of telling which teachers and schools did well, acceptably, or unacceptably.

Since there is a commitment to use tests and exams, let us consider what they can tell us. Obviously, tests provide a total score usually in the form of a percentage or grade and sometimes in terms of a rank order score (e. g., place in class, percentile, or stanine). However, we need to ask whether this type of information is enough to guide improved quality of educational outcomes. Grades can be ambiguous. They might mean place in rank order (i. e., A=Top few, B=Middle some, C=Bottom many) Grades might indicate proportion of questions or tasks completed correctly (i. e., A=85%+correct, B=65−84% correct, C=50−64% correct). Grades might indicate the quality of performance (i. e., A=Excellent, B=Good, C=Acceptable). A study with American families found that parents were happy if their children brought home C grades because they understood that to mean Satisfactory/Average; whereas, the teachers meant C=Unacceptable (Waltman & Frisbie, 1994). So it is important, if grades are to be used, for everyone to have a clear and common understanding of what grades mean.

Rank scores are independent of the actual quality of performance, since rank depends on who else is in the group not the actual ability of the people. If grades reflect quality rather than rank, it is possible for no student to have done as well as the required standard for excellence, despite being top in the class; likewise, it is possible for all students in a well-taught class to meet expectations for an A grade. Another problem with norm-referenced scores is that teachers tend to normalise on their own sample (i. e., the best in my class must be as good as

the population; the worst in my cohort must be really bad). This is clearly a problem when students are pre-assigned to schools or classes according to prior performance — a good student in a weak class is not really a good student, even if he is 1st in class and a strong student in the top class is still a good student even if she is 9th in class. When classes and schools are not equal because of streaming or tracking, a high score does not mean the teacher assigned to the top class has done a good job. Instead, under a value-added approach, a teacher with a weak class to start with may have made a large difference, despite the class still being below average. And all teachers know this, and this way of thinking discredits the examination system.

What if we were to report the percentage correct, instead? First off, many important skills, knowledge, and understandings can NOT be scored discretely, they must be judged as to the quality of the whole work, not piece-by-piece (e. g. , essays; reflective writing; reports, course work assignments, projects; portfolios, discussions, presentations/forums, scrap books, critical comments or reviews). These tasks could be assigned a percentage score but, given the unreliability of human scoring (Brown, 2009), a quality grade is probably more appropriate, albeit less precise. Furthermore, setting a certain percentage as the pass mark (e. g. , 50% in NZ or 40% in HK) is not a valid approach; if the test were very easy, that mark would be far too low and may be too high on an exceptionally difficult examination. Furthermore, in many high-risk domains (e. g. , flying passenger aircraft), the accuracy we should demand for passing ought to be much higher than 50 or 40%. Hence the percentage correct needed for each grade or standard (e. g. , competent or highly accomplished) should be set by a process that takes into account the difficulty of the items. Furthermore, the percentage correct approach assumes that all items with the same mark value are equally difficult. Psychometric statistical analysis of items and test-takers shows that this is simply not so (Embretson & Reise, 2000). Using item response theory approaches to creating a test score means that higher scores are awarded to students who get hard questions right, not just lots of easy questions. Nonetheless, even if scores are calculated using the most advanced statistical techniques, the total score will not tell us who needs to be taught what next, which is the real purpose of educational use of tests (Popham, 2000).

If tests and exams evaluate teachers and students and systems also expect teachers to improve outcomes, what do teachers think about the goals or purposes of assessment? Since, teachers deliver the curriculum and are going to be evaluated, it makes sense to take into account their point of view. In a series of survey studies, it has been clearly shown that teachers believe in using assessment to improve their teaching and student learning (Brown, 2011;

Brown, Lake, & Matters, 2011; Brown & Michaelides, 2011). That's the good news! A recent study of teachers in the PRC and HK (Brown, Hui, Yu, & Kennedy, 2011) showed that there was a positive correlation between assessments (including examinations) for accountability purposes and assessment being irrelevant; while assessment for improvement was seen as NOT irrelevant. This suggests that even Chinese teachers have qualms about the consequences and effects of examinations.

Fundamentally, as long as tests and examinations only give limited information that is at the end of the teaching process, they cannot inform teachers about how to improve their teaching and student learning. What is needed then is a testing system that systematically aligns to the goals of teaching (*curriculum*) and gives information to instructors early enough to make a difference (*formative*) and that tells teachers what they need to know in order to make a difference to student learning (*diagnostic*) (Brown & Hattie, 2012; Hattie & Brown, 2010). Without timeliness and detailed information, standardised tests and public examinations are doomed to repeat the cycle of rewarding most of all the children of privilege. It is this fundamental set of presuppositions that has been adopted in New Zealand as the basis for educational testing.

New Zealand has adopted a formative assessment policy, committed resources to enabling teachers to implement the policy, kept consequences for schools and teachers relatively low, and safe-guarded the professionalism of its teachers. New Zealand has a national curriculum framework of 8 levels of progress containing ordered achievement objectives within strands and covers the whole of schooling from Year 1 to Year 13. The National Curriculum requires that teaching, assessment, reporting, school qualifications, and school evaluation are aligned with these curriculum-based objectives. Specifically the NZ assessment policy focuses on improvement rather than evaluation and embeds assessment within teaching:

> Assessment for the purpose of improving student learning is best understood as an ongoing process that arises out of the interaction between teaching and learning. It involves the focused and timely gathering, analysis, interpretation, and use of information that can provide evidence of student progress. Much of this evidence is 'of the moment'. Analysis and interpretation often take place in the mind of the teacher, who then uses the insights gained to shape their actions as they continue to work with their students. Ministry of Education, 2007, p. 39

Nonetheless, New Zealand requires teachers to make use of high-quality, norm-referenced standardised tests as an important adjunct to their professional judgements about student learning strengths and needs. To that end, the NZ government has commissioned or supported the development and deployment of a range of test systems. One of those systems — Assessment Tools for Teaching and Learning — is one I was proud to work on from 2000 to 2005. The development of asTTle took place within a framework committed to 1) aligning assessment with curriculum, 2) giving teachers and administrators choice about information, 3) giving teachers and administrators control over assessment practices, and 4) improving the quality and accuracy of communication from the test system to teachers and administrators (Hattie, Brown, & Keegan, 2003). These positive goals were complemented by a strong commitment from government NOT to 1) exercise central control or reporting of testing or data and 2) make use of the system compulsory. Nonetheless, the project was funded by the NZ Ministry of Education to meet the improvement agenda and provide evaluative information about the quality of student learning and, by implication, the quality of school teaching.

To meet these goals, the research and development team at the University of Auckland led by Professor John Hattie created a bank of IRT-calibrated test items for Levels 2 - 6 of the reading, writing, and mathematics curricula in both English and Maori (Hattie & Brown, 2008). This test bank was given a computer-assisted interface that allowed teachers and administrators to devise and administer tests, analyse and report performance, and seek additional teaching resources for identified student needs. The system currently allows completely paper-based testing or completely on-screen administration with an option for computer adaptive testing.

Choice. The system gives teachers choice over curriculum content & difficulty through an interface in which teachers select strands, levels, and test constraints (i. e. , length, style). Choice is further permitted by allowing teachers and administrators the ability to select reports that provide information about 1) individuals or groups, 2) performance relative to norms or objectives, 3) performance over time or current status, or 4) seek resources to enrich current teaching practices. In accordance with best practice, no one report attempts to meet all objectives simultaneously (Hattie, 2010).

Control. By allowing teachers to see the test, greater validity of the test for the actual class can be attained. Decisions about when, who, and what to test are made at the school and classroom level, ensuring that the testing process is determined by local and national priorities. Since the data belong to the school and can be evaluated promptly (the computer creates the

reports immediately upon completion of data entry or online testing). This means that schools see the results (with their successes and disappointments) very quickly and don't have to wait for analysis from an external agency. Speed of results also means that changes to instruction can be implemented and evaluated within a school year to see if value is being added. Control within the school, fundamentally, gives teachers confidence to look in the dark corners — who's not learning, what are the students not getting, and so on — instead of placing the blame for not learning on the learners.

Calibration. The calibration processes used anchor the bank of items, and thus tests, to curriculum objectives and levels, as well as the performance of large samples of students in grades 5 to 12. Thanks to IRT calibration, students do not need to be given the same test to compare their performances; in test-retest conditions, this mitigates the tendency for score inflation through practice. More importantly, it means that, while all students are tested, tests can be customised (i. e., hard tests for able students and easier tests for weaker students) while still assuring comparability of information. And since the scores are determined by the difficulty of the items, teachers do not get false impressions of their class's ability by choosing an easy test — such tests generate low scores.

Communication. While testing traditionally provides a table of numbers, the asTTle system has focused on improving the communication of test information to test-users by developing graphical reporting mechanisms that allow users to see the information rather than have to studiously examine detailed tables or values. These visual reports, in accordance with best practice (Hattie, 2010), have been rigorously trialled with users to ensure that the correct interpretations are made. Furthermore, to help close the curriculum, teaching, and assessment cycle, teachers are guided to a website that catalogues high-quality materials and resources to support their teaching of achievement objectives, strands, and levels.

While this model of national assessment is currently working effectively in New Zealand there are constant policy challenges. It is clear that high-stakes will elicit negative effects (especially, in such an open system, cheating is feasible) and so consequences must be kept low. Educationally this is a simple decision. The people who are closest to educational problems are the teachers and school leaders. It must be safe for them to discover learning problems within their classes and schools before it is too late to do anything about it. Just as important, test systems must inform teachers with rich diagnostic analysis of who needs to be taught what next so that they can plan changes in their teaching early enough to make a difference. It is clear also that no one test would ever be enough. Nonetheless, since testing and examining are so

important in so many countries, it behoves national policies and ministries to give educators both the professional responsibility and the tools to understand what changes are needed to improve outcomes. Furthermore, change of this sort is slow, since it requires professional educators to use assessment to analyse the effect of their own work so as to evaluate it AND it requires strong policy commitment to using assessment diagnostically, not just evaluatively. This is a big step from using total scores to judge students, teachers, and schools. But it is a step that can be achieved, even in China.

References

Brown, G. T. L. (2009). The reliability of essay scores: The necessity of rubrics and moderation. In L. H. Meyer, S. Davidson, H. Anderson, R. Fletcher, P. M. Johnston & M. Rees (Eds.), *Tertiary assessment and higher education student outcomes: Policy, practice and research* (pp. 40 - 48). Wellington, NZ: Ako Aotearoa.

Brown, G. T. L. (2011). Teachers' conceptions of assessment: Comparing primary and secondary teachers in New Zealand. *Assessment Matters*, *3*, 45 - 70.

Brown, G. T. L., & Hattie, J. A. (2012). The benefits of regular standardized assessment in childhood education: Guiding improved instruction and learning. In S. Suggate & E. Reese (Eds.) *Contemporary debates in child development and education* (pp. 287 - 292). London: Routledge.

Brown, G. T. L., Hui, S. K. F., Yu, W. M., & Kennedy, K. J. (2011). Teachers' conceptions of assessment in Chinese contexts: A tripartite model of accountability, improvement, and irrelevance. *International Journal of Educational Research*, *50*(5 - 6), 307 - 320. doi: 10. 1016/j. ijer. 2011. 10. 003

Brown, G. T. L., Lake, R., & Matters, G. (2011). Queensland teachers' conceptions of assessment: The impact of policy priorities on teacher attitudes. *Teaching and Teacher Education*, *27*(1), 210 - 220. doi:10. 1016/j. tate. 2010. 08. 003

Brown, G. T. L., & Michaelides, M. (2011). Ecological rationality in teachers' conceptions of assessment across samples from Cyprus and New Zealand. *European Journal of Psychology of Education*, *26*(3), 319 - 337. doi: 10. 1007/s10212 - 010 - 0052 - 3.

Butterfield, S., Williams, A., & Marr, A. (1999). Talking about assessment: mentor-student dialogues about pupil assessment in initial teacher training. *Assessment in Education*, *6*(2), 225 - 246.

Cannell, J. J. (1989). *How public educators cheat on standardized achievement tests*. Albuquerque, NM: Friends for Education.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: LEA.

Firestone, W. A., Mayrowetz, D., & Fairman, J. (1998). Performance-based assessment and instructional change: The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis*, *20*(2), 95 - 113.

Hamilton, L. (2003). Assessment as a policy tool. *Review of Research in Education*, *27*(1), 25 - 68.

Hattie, J. A. C. (2010). The validity of reports. *Online Educational Research Journal*. Retrieved from http://www.oerj.org/View? action=viewPaper&paper=6.

Hattie, J. A. C., & Brown, G. T. L. (2008). Technology for school-based assessment and assessment for learning: Development principles from New Zealand. *Journal of Educational Technology Systems*, *36*(2), 189 - 201.

Hattie, J. A., & Brown, G. T. L. (2010). Assessment and evaluation. In C. Rubie-Davies (Ed.), *Educational psychology: Concepts, research and challenges* (pp. 102 - 117). Abingdon, UK Routledge.

Hattie, J. A. C., Brown, G. T. L., & Keegan, P. J. (2003). A national teacher-managed, curriculum-based assessment system: Assessment Tools for Teaching & Learning (asTTle). *International Journal of Learning*, *10*(771 - 778).

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

Ministry of Education. (2007). *The New Zealand Curriculum for English-medium teaching and learning in years 1 - 13*. Wellington, NZ: Learning Media.

Popham, W. J. (2000). *Modern educational measurement: Practical guidelines for educational leaders* (6th ed.). Boston: Allyn & Bacon.

Smith, M. L., Heinecke, W., & Noble, A. J. (1999). Assessment policy and political spectacle. *Teachers College Record*, *101*(2), 157 - 191.

国家测试：承诺与陷阱

——新西兰的视角

Gavin T L Brown

【摘要】测验和考试经常用来评价学生、学校或老师。这种评价很容易做，并且很直观——好学校培养善于考试的学生。但是，这种简单的方法也有一些问题。考试和测验通常用总分数（一般是百分制或等级制）来报告学生的学业成绩，并以此来评判考生的好坏。不幸的是，这种信息并不能帮助教师、学生、家长，或雇主来确定学习者的优缺点，或不能用来指导学习者下一步应该学习什么。显然，评价必须提供更丰富的诊断性信息，从而帮助提高学生的学习日程。在新西兰，测验普遍作为提高学生学习的手段，而不是用来作为评价或证书优先考虑学生的

学习。这篇文章将通过聚焦于教学和学习系统的测量工具来展示新西兰的测评方法。这个工具是由奥克兰大学设计的，是用来告知5—12年级的教师和学校领导哪些学生需要学习阅读、写作或数学课程中的哪部分内容。

【关键词】形成性评价；诊断性测验；标准化测试；asTTle；评价政策

【作者简介】Gavin T L Brown/新西兰奥克兰大学副教授

Learning from Large Scale International Assessments: Benefits of Secondary Analysis for Research and Policy[①]

Kerry J. Kennedy[②]

Abstract: International large scale assessments such as the PISA, TIMSS, ICCS and PIRLS are now a regular part of the education landscape. Governments and policymakers in different countries have increasingly come to value such studies because of the feedback they provide on students' learning performance. Changes to curriculum and classroom practice are often undertaken within countries based on the results of these international assessments (Ringarp & Rothland, 2010) There are also criticism of these studies, their reliability and the assumptions on which they are based (Hopmann, 2008). There is little doubt, however, that we can expect to see large scale assessments playing an increasingly important role in shaping future policy and practice related to curriculum and pedagogy.

Given the role of such assessments, the purpose of this presentation is to highlight the benefits of secondary data analysis. It will show secondary analysis of large scale assessment data can be a powerful tool for understanding more about student learning than the usual portrayal of results might suggest. This presentation will also question the usefulness of traditional analyses that rely on a single country scale score to represent the level of learning of that construct in each participating country. It will highlight those forms of analysis that recognize the importance of observed differences such as gender, SES, immigrant status etc but also those that can detect unobserved differences that are masked when too much reliance is placed in single scale scores.

Recognizing the importance of secondary data analysis has the potential to provide insight and understanding of important data sets. It can be accessible to policymakers, universities and

---

① This paper has been prepared as part of a General Research Fund Project funded by the Hong Kong Research Grants Council, *Asian Students' Conceptions of Citizenship: Constructing Indigenous Views of Citizens, Citizenship Education and the State*. [HKIEd 842211]. The views expressed here are those of the author and not the funding body.

② Professor Kennedy is Chair Professor of Curriculum Studies, Dean of the Faculty of Education and Human Development and Director of the Centre for Governance and Citizenship at the Hong Kong Institute of Education. He is also the Associate Vice-President (Quality Assurance).