

Thomas Jahnke

## Teaching to the Test – Erfahrungen aus den USA

In dem vierzehnteiligen Beschluss der Kultusministerkonferenz zum Bildungsmonitoring vom 02.02.2006, in dem die Teilnahme an Pisa bis 2018 festgeschrieben und diese Untersuchungen als Referenzrahmen bezeichnen werden, wird auf Seite 13 unter der Zwischenüberschrift „Weiterentwicklung der Bildung, aber kein Teaching to the Test“ auf diese Problematik – wie folgt – kurz eingegangen:

Neben der Funktion der Beschreibung von Leistungsanforderungen und der Leistungsmessung dienen die Bildungsstandards primär der Weiterentwicklung des Unterrichts und vor allem der individuellen Förderung aller Schülerinnen und Schüler. Die Länder sind sich darin einig, dass mit der Setzung der Bildungsstandards als übergreifenden Referenzrahmen eine Entwicklung hin zum „teaching to the test“ oder eine Verengung des Unterrichts aus die Anforderungen der Standards verhindert werden muss.

Diese Kürze ist trotz der beschworenen Einigkeit der Länder erstaunlich. Es liegt nahe, wenn man die ‚Weiterentwicklung des Unterrichts und vor allem der individuellen Förderungen der Schülerinnen und Schüler‘ durch Bildungsstandards befördern oder anordnen will, deren Erreichen im Wesentlichen durch Tests überprüft wird, die Erfahrungen von Ländern und darunter insbesondere der USA zu rezipieren, die seit Jahren oder Jahrzehnten eine solche Politik verfolgen.

For several decades, some measurement experts have warned that high-stakes testing could lead to inappropriate forms of test preparation and score inflation, which we define as a gain in scores that substantially overstates the improvement in learning it implies. (p. 99)

leitet Daniel Koretz, Erziehungswissenschaftler an der Harvard-Universität und assoziierter Direktor des Center of Research, Standards, and Student Testing (CRESST), seinen Aufsatz *Alignment, High Stakes, and the Inflation of Test Scores* ein und beschreibt einen Ausgangspunkt, über den eine öffentliche Diskussion in Deutschland bisher kaum hinausgekommen ist:

On common response to this problem has been to seek “tests worth teaching to”. The search for such tests has led reformers in several directions over the years, but currently, many argue that tests well aligned with standards meet this criterion. If tests are aligned with standards, the arguments runs, they test material deemed important, and teaching to the test therefore teaches what is important. If students are being taught what is important, how can the resulting score gains be misleading? (p. 99)

Koretz, D.: *Alignment, High Stakes, and the Inflation of Test Scores*. Yearbook of the National Society for the Study of Education (2005) 104 (2), 99–118.

Koretz begründet seinen Widerspruch gegen solche Naivität theoretisch und empirisch unter anderem eindrücklich mit Sägezahnkurven (“sawtooth pattern“) für die gemessenen Leistungen der gleichen oder einer vergleichbaren Population, die sich in verschiedenen Erhebungen je nach den verwendeten Tests in unterschiedlichster Weise ergaben. Auch der Hoffnung, solche Effekte seien allein der Testkonstruktion und den Testumständen zuzuschreiben, widerspricht er:

The problem is not confined to commercial, off-the-shelf, multiple-choice tests. It has appeared as well with standards-based tests and with tests using no multiple-choice items. (p. 106)

Die Vorstellung, Schülerleistungen ließen sich in einem Test objektiv oder mit angebbaren Fehlermargen – gleichsam physikalisch messen, ist schlicht (und) irreführend. Folgerungen aus solcher Vorstellung mehr als fragwürdig. Wird dies in Abrede gestellt, verschwiegen oder das Gegenteil präntiert, liegen in aller Regel massive Erkenntnisinteressen der Auftraggeber oder -nehmer der Testungen vor.

Auch die Auswirkungen von Testungen auf den Unterricht werden in den USA seit Jahrzehnten untersucht. Koretz zum Beispiel beschreibt und charakterisiert in dem zitierten Papier *Reallocation, Alignment* und *Coaching*:

*Reallocation.* Reallocation refers to shifts in instructional resources among the elements of performance. Research has shown that when scores on a test are important to teachers, many of them will reallocate their instructional time to focus more on the material emphasized by the test. (. . .) Many observers believe that reallocation is among the most important factors causing the sawtooth pattern (. . .).

*Alignment.* Content and performance standards comprise material – performance elements, in the terminology used here – that someone (not necessarily the ultimate user of scores) has decided are important. If the material is emphasized in the standards, that implies that users should give this material substantial weight in the interference they draw about student performance. Alignment gives this same material high weights in the test as well. (. . .)

*Coaching.* The term “coaching” is used in a variety of different ways in writings about test preparation. Here it is used to refer to two specific, related types of test preparation, called substantive and non-substantive coaching. *Substantive coaching* is an emphasis on narrow, substantive aspects of a test that capitalizes on the particular style or emphasis of test items. The aspects of the tests that are emphasized may be either intended or unintended by the test designers. For example, in one study of the author’s, a teacher noted that the state’s test always used regular polygons in test items and suggested that teachers should focus solely on those and ignore irregular polygons. The intended interferences, however, were about polygons, not specifically regular polygons. (. . .) *Nonsubstantive coaching* refers to the same process when focused on nonsubstantive aspects of a test, such as characteristics of distracters (incorrect answers to multiple-choice items), substantively unimportant aspects of scoring rubrics, and so on. Teaching test-taking tricks (process of elimination, plug-in, etc.) can also be seen as nonsubstantive coaching. In some cases – for example, when first introducing young children to the op-scan answer sheets used with multiple-choice tests – a modest amount of certain types of nonsubstantive coaching can increase scores and improve validity by removing irrelevant barriers to performance. In most cases, however, it either wastes time or inflates scores. (p. 110-112)

An anderer Stelle findet sich ähnliche Kritik. So fasst Brian M. Stecher sein Kapitel 4 *Consequences of large-scale, high-stakes testing on school and classroom practices* in dem von ihm mit herausgegebenen Buch *Making Sense of Test-Based Accountability in Education* folgendermaßen zusammen:

The net effect of high-stakes testing on policy and practice is uncertain. Researchers have not documented the desirable consequences of testing – providing more instruction, working harder, and working more effectively – as clearly as the undesirable ones – such as negative reallocation, negative alignment of classroom time to emphasize topics covered by a test, excessive coaching, and cheating. More important, researchers have not generally measured the extent or magnitude of the shifts in practice that the identified as a result of high-stakes testing. Overall, the evidence suggests that large-scale high-stakes testing has been a relatively potent policy in terms of bringing about changes within schools and classrooms. Many of these changes appear to diminish students’ exposure to curriculum, which undermines the meaning of the test scores. (p. 99/100)

Stecher, B. M.: Consequences of large-scale, high-stakes testing on school and classroom practices. In L. S. Hamilton, B. M. Stecher, and S. P. Klein (Eds.): *Making Sense of Test-Based Accountability in Education*. RAND. Santa Monica 2002. P. 79-100

Der im letzten Absatz angesprochene Antagonismus scheint der deutschen Kultusministerkonferenz möglicherweise von ihren Beratern vorenthalten worden zu sein. Das Gleiche gilt vermutlich für das *Position Statement on High Stakes Testing in PreK-12 Education* der American Evaluation Association (AEA), in dem es heißt:

High stakes testing leads to under-serving or mis-serving all students, especially the most needy and vulnerable, thereby violating the principle of “do no harm.” The American Evaluation Association opposes the use of tests as the sole or primary criterion for making decisions with serious negative consequences for students, educators, and schools. The AEA supports systems of assessment and accountability that help education. Recent years have seen an increased reliance on high stakes testing (the use of tests to make critical decisions about students, teachers, and schools) without full validation

throughout the United States. The rationale for increased uses of testing is often based on a need for solid information to help policy makers shape policies and practices to insure the academic success of all students. Our reading of the accumulated evidence over the past two decades indicates that high stakes testing does not lead to better educational policies and practices. There is evidence that such testing often leads to educationally unjust consequences and unsound practices, even though it occasionally upgrades teaching and learning conditions in some classrooms and schools. The consequences that concern us most are increased drop out rates, teacher and administrator deprofessionalization, loss of curricular integrity, increased cultural insensitivity, and disproportionate allocation of educational resources into testing programs and not into hiring qualified teachers and providing sound educational programs. The deleterious effects of high stakes testing need further study, but the evidence of injury is compelling enough that AEA does not support continuation of the practice. While the shortcomings of contemporary schooling are serious, the simplistic application of single tests or test batteries to make high stakes decisions about individuals and groups impede rather than improve student learning. Comparisons of schools and students based on test scores promote teaching to the test, especially in ways that do not constitute an improvement in teaching and learning. Although used for more than two decades, state mandated high stakes testing has not improved the quality of schools; nor diminished disparities in academic achievement along gender, race or class lines; nor moved the country forward in moral, social, or economic terms. The American Evaluation Association (AEA) is a staunch supporter of accountability, but not test driven accountability. AEA joins many other professional associations in opposing the inappropriate use of tests to make high stakes decisions.

In einer Endnote zu diesem Text wird auf weitere Organisationen verwiesen, die ebenfalls dagegen opponieren, weit reichende Entscheidungen auf Grund von Testergebnissen zu fällen.

AEA joins many other professional associations, teacher unions, parent advocacy groups in opposing the inappropriate use of tests to make high stakes decisions. These include, but are not limited to the American Educational Research Association, the National Council for Teachers of English, the National Council for Teachers of Mathematics, the International Reading Association, the College and University Faculty Assembly of the National Council for the Social Studies, and the National Education Association.

American Evaluation Association (AEA): Position Statement on HIGH STAKES TESTING  
In PreK-12 Education. 2002 (Online unter: <http://www.eval.org/hst3.htm>)

Für den deutschen Betrachter ist es kaum nachvollziehbar, mit welchen Besserungs- wenn nicht gar Heilserwartungen gleich in welcher Richtung die hiesige Bildungspolitik umfangreichste Testprogramme einführt, während der sicherlich nicht zimperliche angelsächsische Evaluations-Pragmatismus sich in kaum zu übertreffender Deutlichkeit nach mehr als zwanzigjähriger Erfahrung von solchen Bestrebungen distanziert.

---

I.W. handelt es sich um einen Textauszug aus: Jahnke, Th.: Deutsche Pisa-Folgen. In: Stefan T Hopmann / Gertrude Brinek / Martin Retzl (Hrsg.): PISA zufolge PISA/PISA According to PISA. Wien / Berlin (LIT-Verlag) 2007, S. 305 -320