Julia Claire Prieß-Buchheit

Testfolgen im Bildungsbereich

Aktionen und Reaktionen im deutsch-amerikanischen Vergleich

WAXMANN

Julia Claire Prieß-Buchheit

Testfolgen im Bildungsbereich

Aktionen und Reaktionen im deutsch-amerikanischen Vergleich

Waxmann 2016
Münster • New York

Bibliografische Informationen der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

Internationale Hochschulschriften, Bd. 635

Die Reihe für Habilitationen und sehr gute und ausgezeichnete Dissertationen

ISSN 0932-4763
Print-ISBN 978-3-8309-3512-4
E-Book-ISBN 978-3-8309-8512-9

www.waxmann.com
info@waxmann.com

Umschlaggestaltung: Anne Breitenbach, Münster
Umschlagabbildung: © alexlmx – Fotolia.com
Satz: Stoddart Satz & Layout, Münster
Druck: Hubert & Co, Göttingen

Gedruckt auf alterungsbeständigem Papier,
säurefrei gemäß ISO 9706

Printed in Germany

Danksagung

Bedanken möchte ich mich bei all jenen, die mich beim Anfertigen der Studie unterstützt haben, mir mit ihrer Kritik neue Möglichkeiten eröffnet haben und mich finanziell in der Zeit abgesichert haben. Ermöglicht wurde die Studie durch das Professorinnen-Programm des Bundesministeriums für Bildung und Forschung, in dem mir durch die Frauenförderung der Christian-Albrechts-Universität zu Kiel ein Stipendium gewährt wurde. Die Idee zur Ausführung der Studie entstand während meines Aufenthalts am Soziologischen Institut der Boston University, USA. Durch die dortigen konstruktiven Diskussionen konnte sich der Erkenntnisweg zur Studie öffnen. Bemerkenswerte Unterstützung habe ich von Prof. Dr. Jongebloed und Prof. Dr. Krope erhalten. Ihre wissenschaftlichen Kolloquien waren stets eine Plattform, auf der sich meine Forschungsschritte behaupten mussten. Als externes Institut ist das TestDaF zu nennen, im besonderen PD Dr. Eckes, der mir mit seinem Ratschlag die Option der manifesten und latenten Testfolgen nahegelegt hat. Formal überarbeitet wurde der Text von Dr. Vera Schanz. Ohne sie wäre der Text nie zum Abschluss gekommen. Ihnen allen sei an dieser Stelle herzlich gedankt!

Zum Geleit

Um dem Erkenntnisinteresse der vorliegenden Studie verstehend nachfolgen zu können, erweist es sich als vorteilhaft, sich jene empirische Wende zu vergegenwärtigen, die die deutsche Pädagogik seit den 60er Jahren paradigmatisch und höchst einflussreich zu bestimmen vermag.

Vor dieser Zeit waren Urteile über Leistungen, die von Lehrenden im Bildungsbereich gegenüber Lernenden gesprochen wurden, gleichermaßen selbstverständlich, wie dies heute der Fall ist, jedoch standen sie nicht im Fokus einer streng messtheoretisch begründeten Testtheorie.

Erst mit der empirischen Wende gerieten auch diese Urteile in den Fokus strenger testtheoretischer Fundierung, was zu dem didaktischen Spezialgebiet der Pädagogischen Diagnostik führte, die heute – betrachtet man das Geschehen der modernen empirischen Bildungsforschung – geradezu ein Kerngebiet pädagogischen Erkenntnisinteresses auszumachen scheint.

Natürlich hatten auch zuvor gesprochene Lehrerurteile Folgen für die Betroffenen, die aber jeweils im individuellen Betroffenheitsfokus derjenigen blieben, an die sie gerichtet waren – eine Testfolge, die in der vorliegenden Studie unter der Kategorie „ontisch" erwähnt, für relevant erachtet, aber nicht weiter problematisiert wird.

Mit zunehmender wissenschaftlicher Ausformung der Testtheorie, die sich weitgehend an den Methoden der empirischen Sozialforschung orientierte, erreichten Testfolgen eine ganz neue, erheblich weitreichendere Bedeutung, ohne jedoch ernsthaft als Bestandteil pädagogisch-diagnostischen Erkenntnisinteresses wahrgenommen zu werden.

Wer jedoch die fast schon für jedermann nachzuvollziehenden Änderungen im Bildungssystem seit Beginn des neuen Jahrtausends, die sich mit den Stichworten TIMSS, PISA oder auch IGLU verknüpfen, verfolgt hat, muss konstatieren, dass diese Änderungen sich zumeist als Folgen eben dieser genannten Teststudien erweisen.

Es liegt also auf der Hand, dass eine derartige Testentwicklung deutlichen Einfluss auf das Bildungsgeschehen des Staates, einzelner Länder, von Schulformen aber auch einzelner Lehrerinnen und Lehrer sowie Schülerinnen und Schüler nimmt, ohne dass kasuistisch wie strukturell die damit verbundenen Konsequenzen konzeptionell in ihrer diagnostischen Wirkung intersubjektiv erkennbar sind.

Genau hier setzt nun die Studie von Frau Prieß-Buchheit an, deren Erkenntnisinteresse darauf fokussiert, die durch Testfolgen entstehenden pädagogisch-diagnostischen Einflüsse und Transformationen bildungstheoretisch und bildungspolitisch nachvollziehbar zu evaluieren; denn, obwohl durchaus schon Ansätze zur Untersuchung der Wirkung von Testfolgen vorfindlich sind, wie sich im sechsten Kapitel dieser Untersuchung nachvollziehen lässt, muss dennoch konstatiert wer-

den, dass ein wissenschaftlich systematischer Zugriff auf dieses Problemfeld bisher aussteht.

Die von Frau Prieß-Buchheit angegangene Fragestellung kann daher als außerordentlich neu angesehen werden, indem sie versucht, ein Raster des Zugriffs auf Testfolgen zu entwickeln, um diese als relevanten Bestandteil pädagogisch-diagnostischen Handelns von Anfang an in Rechnung zu stellen.

Für dieses Unterfangen stehen bisher methodische Zugriffe und wissenschaftssystematische Anleitungen im Grunde nicht zur Verfügung, sodass die vorliegende Studie von ihrem wissenschaftstheoretischen Status her als ein erster explorativer Ansatz gesehen werden muss, das Problemfeld der Testfolgen wissenschaftlich zu erschließen.

Frau Prieß-Buchheit wählt dazu methodische Zugänge, die sich zum einen darin ausdrücken, dass sie eine empirische Rückversicherung für die Relevanz der sich einstellenden Folgen anstrebt, während zugleich die Notwendigkeit besteht, das empirische Material datenrelevant aufzubereiten und in gewisser Weise – soweit möglich – theoretisch rückzubinden und einzuordnen.

Da es – zumindest zu diesem Zeitpunkt des Forschungsstandes – nicht möglich erscheint, einen direkten empirischen Zugriff durch übliche Erhebungsmethoden einzuschlagen, werden als empirische Basis Zeitungsartikel gewählt, die sich im weitesten Sinne mit Testfolgen befassen, die das Bildungssystem und das Bildungsgeschehen betreffen. Gleichwohl ergibt sich eine Problematik der angemessenen Samplebildung, die hier in klassisch explorativem Sinne gelöst wird, indem jene Zeitungen ausgewählt werden, die sich überhaupt mit der in Rede stehenden Problematik befassen, und zugleich eine angemessen hohe Auflage haben.

Da in Bezug auf eine Testfolgenwirkungsforschung die USA eine Vorreiterrolle einnehmen und die empirische Bildungsforschung insgesamt eine stark angloamerikanische Akzentuierung aufweist, wird für die USA die New York Times und für Deutschland die Süddeutsche Zeitung ausgewählt, womit sich zugleich die Möglichkeit zu komparativen Erkenntnissen eröffnet.

So wenig vor dem Hintergrund des gegebenen Forschungsstandes gegen die Auswahl des empirischen Datenmaterials dem Grunde nach Einwände erhoben werden können, so sehr muss an dieser Stelle dennoch gesagt werden, dass – ganz unabhängig von der Auflagenhöhe – mit der Süddeutschen Zeitung gleichwohl eine journalistische Position verbunden ist, die eben eine bestimmte bildungspolitische Tendenz vertritt, so dass es nicht verwundert, dass in dem für die Studie entwickelten Raster bestimmte, eher einschränkende Konstellationen häufiger vorkommen als freistellende, wenn man im Sinne der rasterinternen Organisationstypen, hier Organisationstyp Z, argumentiert.

Die Auswertung des Textmaterials orientiert sich an der von Bohnsack u.a. entwickelten, als qualitatives Auswertungsverfahren methodologisch jedoch nicht weiter problematisierten dokumentarischen Methode der exemplarischen Textinterpretation, mit der es möglich wird, die unterschiedlichen, durchaus reichlichen Textauszüge aus Sicht des gewählten Erkenntnisinteresses zu vergleichen.

Dieser Umgang mit den für den Forschungsprozess herangezogenen Methoden – und das gilt auch für die Rückgriffe auf die Sprechakttheorie im Sinne Austins und die Rechtfertigungstheorie im Sinne Toulmins sowie im Sinne des Erlanger Konstruktivismus – ist kennzeichnend für die vorliegende Studie insgesamt, die insoweit eher in der Tradition angloamerikanisch-pragmatischer Forschungspraxis steht als in der cartesianischen Tradition tiefgehender Analytizität. Und auch der hier und da eher großzügig geratene Umgang mit begrifflicher Strenge mag darauf zurückzuführen sein, dass die Arbeit stets unter dem komparativen Anspruch steht, auch amerikanische Textstellen miteinzubeziehen.

Das für die Studie erarbeitete Systematisierungsraster als solches ist gleichwohl geprägt von einem analytischen Systematisierungsanspruch, der in einem gewissen Sinne einen vierdimensionalen Raum aufspannt, in dem die Ergebnisse evaluativ verortet werden.

Diese vier Dimensionen, die aufgezeigt werden, bestehen darin, dass zunächst eine Perspektive der Anlässe für Testfolgen strukturiert wird, die zwischen „Äußerungen über Test", „Tests als solchen" und „Testergebnissen" unterscheidet, wobei es nur um jene Testfolgen geht, denen sich ein deontischer Charakter zuordnen lässt, also um jene, bei denen jemand aufgefordert wird, etwas zu tun.

Dieser Dimension der Anlässe ordnet sich eine Dimension der Testfolgentypen zu, von denen sich vier unterscheiden lassen. Testfolgentyp A „plant und überlegt hypothetisch, wie es möglich ist durch Testfolgen die Handlungsspielräume im Bildungsbereich zu erweitern". Beim Testfolgentyp U geht es um Folgen, die für viele Personen für einen längeren Zeitraum gelten, während es beim Testfolgentyp G darum geht, dass sie durch Gesetzgebung festgeschrieben werden. Mit den Typ S verbinden sich all jene Testfolgen, die explizit aus Gründen der Steuerung eingeführt werden.

Die dritte Dimension konstruiert nun Aktionsformen, die sich mit Testfolgen verbinden, von denen fünf unterschieden werden: „Durchführen", „Einführen", „Einhalten", „Abändern", „Umgehen".

Die vierte Dimension schließlich grenzt vier Operationstypen voneinander ab, von denen der Operationstyp Z (Normtyp) zwischen „Freistellen" und „Einschränken" differenziert, und die Operationstypen N (Norm) sich als „Vollziehen", „Planen" und „Hypothetisch betrachten" zeigen. Die Operationstypen W (Wirkungsgrad) differenzieren zwischen „Lokal" und „Umfassend" und die Operationstypen T (Territorium) zwischen „auf Lehr-Lernprozesse abgestimmt" und „auf andere Gebiete abgestimmt".

Auf diese Art und Weise versucht die Studie der außerordentlich hohen empirischen Komplexität und Differenziertheit systematisch Herr zu werden, und das Produkt der denkbaren Zuordnungsfelder zeigt an, wie groß der Korpus der relevanten Möglichkeiten anzunehmen ist.

Diesem ersten Zugriff auf eine systematische Strukturierung von Testfolgen, die es eigentlich unter dem Anspruch Pädagogischer Diagnostik stets zu gewärtigen

gäbe, ordnet sich nun noch ein Umfeld zu, das unter dem analytischen Ausdruck „soziopolitischer Interpretationsraum von Testfolgen" untersucht wird.

Es zeigt sich nämlich, dass, je nachdem, in welcher soziopolitischen Struktur die Testfolge auftritt, die Wahrscheinlichkeiten und Möglichkeiten, nach bestimmten Aktionsformen zu handeln, unterschiedlich sind. Hierin drückt sich aus, dass unabhängig von der rein analytischen Auseinandersetzung mit Testfolgen die jeweiligen ideologischen Rahmenbedingungen zu beachten sind, die einerseits dazu geführt haben können, dass Testfolgen überhaupt auftreten und andererseits bestimmte Testfolgen präferieren oder konterkarieren.

Um die Auseinandersetzung mit Testfolgen und um Testfolgen kommunikativ abzuwägen, werden die als Sprechakte aufzufassenden Texte zurückgebunden an die Sprechakttheorie und an die Rechtfertigungs- und Begründungstheorien zu Aussagen von Toulmin.

Dabei geht es vor allen Dingen um die Rollen von Proponent und Opponent und um die damit verbundenen Ansprüche an Validität und Legitimität. Aus diesem Grunde setzt sich Frau Prieß-Buchheit mit dem Problem der Validität von Tests auseinander, die jedoch eigentlich mehr ein Problem der Testkonstrukteure ist als derjenigen, die den Testfolgen ausgesetzt sind.

Das Phänomen der Testvalidität, das in seinen Varianten „Kontentvalidität", „Kriteriumsvalidität" und „Konstruktvalidität" auftaucht, erweist sich dabei als Eigenschaft, die unabhängig vom Rechtfertigungs- und Begründungsdiskurs insofern eine Sonderstellung einnimmt, als sie von den Herstellern der Tests zu verantworten ist, während das Legitimitätsproblem stets zurückgeworfen bleibt auf diejenigen, die die begründeten Testfolgen denjenigen gegenüber rechtfertigen müssen, die sie betreffen.

Dieses Problem erfährt eine besondere Schärfe für Testfolgen, die in sogenannten High-Stake-Tests auftritt, da diese Tests sich dadurch auszeichnen, dass Testfolgen z.B. in Form von Entlassungen oder Entzug der Mittel unmittelbar nach den Testergebnissen eintreten. Diese Form spielt jedoch im angloamerikanischen Raum eine wesentlich größere Rolle als in Deutschland, obwohl High-Stake-Tendenzen durchaus auch hier zu erkennen sind.

Betrachtet man die gesamte Ausarbeitung, die in vielen Teilen und Einzelergebnissen noch viel differenzierter und detaillierter ist, insgesamt, dann zeigt sich, auf was für ein komplexes Feld das Erkenntnisinteresse der vorliegenden Studie gerichtet ist, und wie notwendig es ist, dieser Problematik in systematischer Weise nachzugehen.

Das in dieser Studie entwickelte Systematisierungsraster kann deswegen zumindest aus deutscher Perspektive als ein erster Versuch gewertet werden, dieser gleichwohl eminent empirisch relevanten Problematik ordnende Struktur abzugewinnen, obwohl – wie im letzten Kapitel ausgeführt – damit allenfalls ein Anfang gefunden wurde.

Auf jeden Fall ist zu konstatieren, dass die Pädagogische Diagnostik mit der in dieser Untersuchung eröffneten Perspektive auf die Testfolgen eine vollständig

neue, wenngleich unabdingbar notwendige Erweiterung erfährt, die den Umgang mit Äußerungen über Tests, Tests als solchen und Testergebnissen in ein neues pädagogisches Licht rückt.

Wie notwendig das Angehen der hier aufgerufenen Problematik ist, zeigt die große Beliebigkeit der bildungspolitischen Konsequenzen, die sich mit TIMSS und PISA in Deutschland bislang verbunden haben und die sich unter Beachtung der hier vorgeschlagenen Rechtfertigungs- und Legitimierungsdiskurse hätten vermeiden lassen. Insoweit versprechen die erarbeiteten und gewonnenen Ergebnisse und Erkenntnisse nicht nur wissenschaftliche, sondern auch bildungspolitische Innovationen.

Es wäre gleichwohl zu wünschen, dass die Auseinandersetzung beispielsweise mit dem von Frau Prieß-Buchheit vorgestellten Testwirkungsvaliditätskonzept besonders mit Bezug auf die Unterscheidung zwischen latenten und manifesten Testfolgen auch eine stringentere und im Umgang mit Begrifflichkeit und theoretischer Sicherung präzisere Strenge gewonnen werden könnte, die auch für die Rechtfertigungs- und Legitimitätsdiskurse bildungspolitischer Art eine höhere Gewissheit ermöglichen könnten.

Vor dem Hintergrund des bisherigen Forschungsstandes jedoch ist die hier vorgelegte Studie auch für die empirische Bildungsforschung insgesamt ein großer Schritt nach vorn, der das Bemühen des Bildungssystems um die Sicherung der Kompetenzen seiner Bürger sowie Vergleichbarkeit und Chancengerechtigkeit für diese ermöglichen hilft.

Der Studie ist deshalb ein großer Leserkreis zu wünschen!

Nortorf, Kiel im Juni 2016
Prof. Dr. Hans-Carl Jongebloed

Inhalt

1. Einleitung

In Zeiten, in denen standardisierte Tests und Evaluationen eine immer größere Rolle in der Bildungslandschaft spielen, sind verschiedenste Folgen dieser Verfahren – teilweise absichtlich und teilweise unabsichtlich herbeigeführt – bemerkbar. Leistungssteigerung, Prüfungsangst, *Teaching to the Test* und *Narrowing the Curricula* sind Schlagwörter, die regelmäßig bei Diskussionen um Testfolgen vorgetragen werden. Dass es Testfolgen gibt, ist nicht zu bestreiten. So schreibt z.B. die New York Times am 28. Dezember 2002: „Rigorous testing that decides whether students graduate, teachers win bonuses and schools are shuttered, an approach already in place in more than half the nation, does little to improve achievement and may actually worsen academic performance and dropout rates" (Winter 2002).

In einem anonymen Interview zum Thema: „What do you think about standardized testing", antwortet ein US-amerikanischer Experte aus der Bildungsforschung:

> „[…] So what that does, that becomes a systems changing mechanism by a backdoor via an inducement. […] For example, in *Race to the Top* one of the things […] one of the elements that has been great controversy is, should students standardize […] should students' test scores be part of the evaluation of teachers? […] This goes back to Tennessee back in 1991[.]" (Interview geführt am 2. November 2010, Code 2A, Prieß-Buchheit).

Der Begriff Testen beschreibt die Tätigkeit, jemanden zu beurteilen, jemanden zu bewerten oder – im schulischen Rahmen – ihm eine Note zu geben. Tests verändern das Bildungssystem. Im Zuge der Tests werden neue Techniken und Verfahren kontrovers diskutiert und eingeführt. Diese Umsetzungen von Verfahren und Techniken sind Testfolgen. Sie sind Antworten auf Tests.

Die vielen verschiedenen Testfolgen, wie zum Beispiel Motivationssteigerung, Lernzuwachs, Testbetrug und Ressourcenumverteilung, lassen sich in zwei unterschiedliche Typen einteilen. Zum einen gibt es Folgen auf Tests, die eine Verhaltensreaktion sind. Ähnlich einem Reflex reagieren betroffene Personen auf Tests und Testsituationen. So eine Verhaltensreaktion ist zum Beispiel das Aufkommen von Prüfungsangst, wenn der Test ausgefüllt wird. Zum anderen gibt es im Gegensatz dazu Personen, die aufgrund eines standardisierten Testverfahrens beginnen, ihre persönlichen Handlungen darauf auszurichten. So zum Beispiel die Abänderung der Unterrichtsvorbereitung einer Lehrerin aufgrund eines externen Tests. Diese Testfolgen sind Handlungsreaktionen auf Tests. So wie man auf eine Frage seines Gesprächspartners mit einer Antwort (Handlungsreaktion) reagiert, so reagieren Personen handelnd auf Testverfahren.

Bei diesen Testfolgen des zweiten Typs werden individuelle Entscheidungen getroffen, neue Techniken entwickelt, Gewohnheiten umgestellt etc. Ein Beispiel für individuelle Entscheidungen im Zuge eines Tests ist die Umstrukturierung von

Lerngewohnheiten. Aufgrund von Testergebnissen werden Lerngewohnheiten bestätigt oder bisherige Lerngewohnheiten in Frage gestellt und verändert. Für die Entwicklung einer neuen Technik als Testfolge ist das Beispiel von Verfahrensänderungen bei Lehrereinstellungen bzw. -entlassungen ein Prototyp. Administrative Verträge werden nach dem Auftreten von Faktorausprägungen beurteilt. Anhand von (Schüler-)Testergebnissen wird abgelesen, ob Lehrer weiter zu beschäftigen oder zu entlassen sind. Auch darüber berichtet der Experte:

> „[…] about the issue of using students' scores for teacher evaluation, I've been asked to join a state committee to try to develop something. […] look at that issue and obviously, since I am the chair, I want my institution to be at that table and looking at major policy. […] We need to shape it, […] what that really looks like and also to take a look on what's the back-side of this? […] If we say, that teacher can be dismissed on the bases of student scores, […] then you have the issue […] can you hold a teacher accountable?" (Interview geführt am 2. November 2010, Code 2A, Prieß-Buchheit).

Diese Technik der Lehrerbeurteilung wird durch das ständige Testen der Schüler ermöglicht. So bedingen standardisierte Lernstandserhebungen von Schülern (administrative) Techniken. Die Technik bzw. Testfolge wird eingeführt, indem eine administrative Weisung gegeben wird. Ist diese Weisung, Lehrer aufgrund von Testergebnissen weiter zu beschäftigen oder zu entlassen, berechtigt bzw. legitim? Manche Argumente sprechen dagegen, sagt der Experte:

> „If your teachers are dismissed, when there are other factors that come into play. It's like a physician. If I'm your patient and you're my physician, then you say: ‚XXX you have to take this pill every six hours.' And I don't take it and bad things happen to me, my health deteriorates. And I try […] I sue you for malpractice […] I didn't do what I was supposed to do, I didn't take the medication, that I was supposed to, on time. […] Just like the student didn't study and here she was supposed to […] didn't turn in the questions, […] didn't write the papers, they were supposed to do […] It goes a lot deeper than the policy discussion has been[.]" (Interview geführt am 2. November 2010, Code 2A, Prieß-Buchheit 2010).

Es lohnt, sich mit den jeweiligen Testfolgen auseinanderzusetzen: In einem Einigungs- und Durchsetzungsprozess werden manche Testfolgen umgesetzt, andere nicht. Wann eine Testfolge berechtigt bzw. legitim ist und umgesetzt wird, bleibt zu klären.

Die Gruppe der Testfolgen differenziert sich weiter in Reaktionen, die eine Person selber betreffen und in Reaktionen, in welchen eine Person andere zu Handlungen auffordert. Ein Beispiel für die letztere Gruppe ist die Testfolge, dass Lerngewohnheiten einer ganzen Klasse umstrukturiert werden, weil der Lehrer anhand des Testergebnisses die bisherigen Lernstrategien seiner Schüler für ineffizient hält.

Hier ändern sich aufgrund des Tests nicht nur die Handlungszusammenhänge des Lehrers, sondern ebenfalls die der Schüler (Herman & Golan 1993).

Solche Testfolgen haben das besondere Merkmal, dass sie in Form eines Rats oder Tipps, einer Aufforderung bzw. Anweisung, eines Gebots bzw. Verbots oder eines Gesetzes etc. auftreten. Von besonderem Interesse sind diese Testfolgen deshalb, weil sie eine große Wirkung bei vielen Personen erreichen. Viele Personen sind von solchen Testfolgen betroffen. Ferner sind viele Personen im Umsetzungsprozess der Testfolge beteiligt und haben – je nach Umständen – auch an der Entscheidung für und gegen die Testfolge teil. Eine solche Testfolge mit großer Wirkung wird zum Beispiel in diesem Bericht dargelegt: „Die Schüler schnitten im internationalen Vergleich schlecht ab. Im Lesen, Rechnen und in den Naturwissenschaften landeten sie auf hinteren Plätzen. Die Kulturminister [sic!] reagierten darauf mit dem Erlass von Bildungsstandards“ (o.V. 2010).

Die vorliegende Untersuchung *Testfolgen im Bildungsbereich* (TeFoBi) untersucht diese Art von Testfolgen. Analysiert werden Handlungsreaktionen auf Tests, die andere zu etwas verpflichten. In diesem Feld soll die TeFoBi-Studie Antworten auf folgende Fragen liefern: Was für (handlungsreaktive) Folgen werden durch standardisiertes Testen bewirkt? Welche Form haben diese Testfolgen? Was sind typische Testfolgen? Und wie kann man zukünftig mit Tests und Testfolgen umgehen?

Neben einem instrumentellen und konzeptuellen Nutzen werden Tests ebenfalls legitimatorisch für Veränderungen im Bildungssystem eingesetzt. Deshalb untersucht die hier ausgearbeitete TeFoBi-Studie das Abwägen und Beschließen einer Testfolge im sozialen Miteinander. Das Kernstück der Untersuchung bilden Quellen über Diskussionen, ob und wann welche Testfolgen umgesetzt werden. Es werden Aufforderungen bzw. Weisungen analysiert, die aufgrund von Tests andere dazu bringen (wollen), etwas zu tun.

Argumente für und wider Testfolgen werden in ihrer Argumentationsstruktur und in ihrer Verwendungsweise erschlossen. Auf zwei Ebenen werden dafür Sprechsequenzen im gesellschaftlichen Miteinander untersucht. Die Erkenntnisse werden zum einen über sprechakttheoretische Zusammenhänge und zum anderen über argumentations- und rechtfertigungstheoretische Reflexionen erarbeitet. Sowohl die erste methodologische Ebene (die Ebene der Sprechakttheorie) als auch die zweite (die Ebene der Argumentations- und Rechtfertigungstheorie) begrenzen die Erkenntnismöglichkeiten und schließen ontologische und psychologische Erklärungen für Testfolgen aus. Die methodologischen Ebenen zielen auf Erkenntnisse in Mitten von Sprache und Sprechen ab. In diesem Sinne folgt die TeFoBi-Studie folgendem Grundsatz: Um zu Testen und Testfolgen zu bestimmen, wird gesprochen. Welche Formen und Typen werden dabei verwendet?

2. Forschungsgegenstand der TeFoBi-Studie: Praktische Probleme im Bildungsbereich

Testfolgen erscheinen an vielen Orten. In Lehrerzimmern wird diskutiert, ob es legitim ist, aufgrund der anstehenden standardisierten Leistungstests die Lernziele für eine Klasse abzuändern. In den USA werden Schulen für ihr gutes Abschneiden bei standardisierten Leistungstests prämiert und andere für ihr schlechtes Abschneiden sanktioniert. Zwischenzeitlich werden in der Praxis Gegenstimmen laut, und Zweifel sowie Fragen nach der Bedeutsamkeit dieser Tests und ihrer Folgen werden erörtert. Der Titel „A Tyranny of standardized Tests" (Botstein) eines Artikels in der New York Times belegt bereits im Jahr 2000 Stimmen dieser Gegenbewegung.

Tests und Testfolgen sind im Alltag von Bildungseinrichtungen nicht mehr wegzudenken. Und die Praxis zeigt, dass Tests und vor allem ihre Testfolgen praktische Probleme im pädagogischen Alltag lösen, gleichzeitig jedoch auch Probleme mit sich bringen. Es wird gestritten, ob eine bestimmte Testfolge und nicht eine andere in den Bildungseinrichtungen umgesetzt werden soll oder nicht (siehe z.B. Schneider 2008). Eine genauere Betrachtung dieser praktischen Probleme führt zu Fragen wie: Was sind Testfolgen eigentlich? Welche Testfolgen gibt es? Wie ist die Validität bzw. Legitimität der Testfolgen? Die vorliegende TeFoBi-Studie ist ein erster Versuch, diese Fragen zu klären.

Tests im Bildungsbereich werden realisiert, indem Kenntnisse der Testtheorie umgesetzt werden. Hingegen können Forschungsfragen wie: „Welche Folgen werden durch Tests bewirkt und wodurch werden sie hervorgerufen?" mit testtheoretischen Methoden nicht beantwortet werden. Methoden der Testtheorie, die Rost (2004) als „Verfahren zur Erfassung psychischer Eigenschaften oder Merkmale von Personen" (S. 17) bezeichnet, sind bei solchen Problemen nicht erkenntnisleitend.

Das Forschungsinteresse der TeFoBi-Studie stützt sich inhaltlich auf das Praxisgebiet der Testtheorie, ist methodisch allerdings hauptsächlich auf qualitative Methoden der Sozialwissenschaften angewiesen. Daher ist die Studie transdisziplinär. Antworten auf die Fragen, welche Folgen durch Tests hervorgerufen werden und wie sich diese interpretieren lassen, werden mit Hilfe qualitativer Forschungsmethoden und analysierenden sowie systematisierender Techniken erforscht. Die TeFoBi-Studie nimmt Testfolgen in den Blick und schafft es, mittels einer qualitativen Erhebung, interpretativer Verfahren und Analysen von Sprach- sowie Sprechzusammenhängen, die vielschichtigen Formen und Typen von Testfolgen zu beobachten und zu verstehen.

Mit Testfolgen und den daraus resultierenden Problemen sind unterschiedliche wissenschaftliche Disziplinen konfrontiert. Unter anderem stehen die Disziplinen Soziologie, Psychologie und Pädagogik vor Problemfeldern. Zum Beispiel lassen der rapide Anstieg von Tests Mechanismen des Testens und Reaktionen darauf zu einem soziokulturellen Phänomen werden, in dem Ziffern zu Wahrheiten stilisiert werden. Ein weiteres Problemfeld sind die wiederkehrenden Herausforderun-

gen, sich in bestimmten Lebenslagen Tests stellen zu müssen, was wiederum individuelle Ängste fördert. Und ein Problemfeld der Pädagogik ist die Schwierigkeit, inwieweit Getestete im Bereich empirische Bildungsforschung ausgebildet sein sollen und wie dies umgesetzt werden kann.

Das Gebiet Testen und Testfolgen findet in vielen wissenschaftlichen Disziplinen und in unterschiedlichen Gebieten der Gesellschaft Anwendung. Eben deshalb wird es durch verschiedene Forschungsgebiete analysiert, evaluiert und neu ausgerichtet. Infolgedessen sind zentrale Begriffe des Gebiets aus vielen Perspektiven heraus belegt, in ihrer Terminologie jedoch uneindeutig verwendet. Unter dem Stichwort *Test* ist in der Enzyklopädie *Philosophie und Wissenschaftstheorie* (2004) zu lesen: Test ist eine „in der Umgangs- und Wissenschaftssprache allgemein gebräuchliche Bezeichnung für Prüfverfahren, z.B. der Leistung einer Person, der Funktionsfähigkeit eines Geräts oder der Richtigkeit einer Behauptung" (Heister und Schröder-Heister 2004, Band 4, S. 240). Um das weite Gebiet einzugrenzen, wird in der TeFoBi-Studie der Begriff *Testen* explizit für ein regelgeleitetes Vorgehen verwendet, um neue, zahlenbasierte Informationen zu gewinnen. Diese werden anhand großflächig angelegter externer Prüfverfahren gewonnen. Wenn im Folgenden der Begriff *Testen* angewendet wird, dann ist damit immer solch ein Verfahren gemeint.

Die Verbindung zwischen Test und Testfolge kann zunächst beschrieben werden als Zusammenhang zwischen Aktion und Reaktion. Üblicherweise implizieren bestimmte Prämissen, Bedingungen und Gründe eine Folge (vgl. Kambartel 2004, Band 1, S. 654). Im Falle von Testfolgen sind ebenfalls Prämissen, Bedingungen und Gründe ausschlaggebend. In der TeFoBi-Studie wird exakt darauf eingegangen, und es werden Faktoren untersucht, die Testfolgen bestimmen. Vielen Testfolgen voraus geht eine argumentative Auseinandersetzung darüber, welche Testfolgen Testergebnisse implizieren. Das heißt, bevor Testfolgen umgesetzt werden, werden diese erst einmal aus Tests und den Testergebnissen *gefolgert* und im privaten wie öffentlichen kommunikativen Handeln benannt.

Zum Beispiel diskutieren Klieme und Prenzel in der Zeitung Die Zeit: „Die jüngste Pisa-Studie hat den Vorteil einer kontinuierlichen Beobachtung des Bildungswesens besonders deutlich gemacht, weil sie die Entwicklung eines Jahrzehnts nüchtern bilanzieren konnte. […] Wo war das System erfolgreich, und wo liegen nach wie vor Probleme? […] Wie werden in Deutschland Lese- und Sprachförderung genutzt? […] [N]ach wie vor ist die Benachteiligung vor allem in Migrantengruppen sehr groß. Es mangelt an Fördermaßnahmen, gerade im sprachlichen Bereich" (Klieme und Prenzel 2011, S. 68; Unterstreichung von der Autorin hinzugefügt). Die Testergebnisse der internationalen Erhebung werden hier angeführt, um Fördermaßnahmen im sprachlichen Bereich zu fordern. Sie dienen als Argument für eine Veränderung im Bildungssystem. Mit der Aussage, es mangele an Fördermaßnahmen im sprachlichen Bereich, wird der Wunsch oder das Begehren ausgedrückt, Änderungen einzuführen und das Bildungssystem dementspre-

chend zu verändern. Es wird dazu aufgefordert, einen Missstand mittels bestimmter Verfahren zu beheben.

Die meisten Testfolgen werden im kommunikativen Handeln durch Aussagen dieser Art eingefordert und eingeleitet. Doch nicht alle Vorschläge dieser Art führen zu tatsächlichen Folgen. Nur manche Testfolgen werden faktisch zur Umsetzung angeordnet, andere werden untersagt oder gar verboten. Wieder andere werden zunächst in Frage gestellt. Und die Vielfalt ist noch längst nicht erschöpft. Testfolgen werden ebenso zugelassen, genehmigt oder sie werden verwehrt etc.

In der TeFoBi-Studie wird das Testen in seiner Funktion betrachtet, Ziele zu erreichen und Folgen einzuleiten. Parallel dazu wird die Art und Weise der legitimatorisch sprachlichen Funktion des Testens betrachtet, um zu bestimmen, ob etwas berechtigterweise ausgeführt werden darf oder nicht. Ebenfalls werden die Funktionen des Testens, wie zum Beispiel über Missstände aufzuklären, etwas zu belegen, oder eine Entscheidung zu treffen bzw. sie zu untermauern, in Abgrenzung zu anderen Testfunktionen (zum Beispiel Zustände zu beschreiben oder etwas zu messen) diskutiert.

Vom einzelnen Schüler hin zu Schuladministratoren, Eltern und Lehrern: Viele Menschen sind von Testfolgen im Bildungsbereich betroffen. Testfolgen sind Veränderungen im theoretischen und sozialen Raum und betreffen somit unterschiedlichste Personen bzw. Personengruppen. Aus Tests können bestimmte Schlüsse gezogen werden.

Wer bestimmt Testfolgen? Sind sie steuerbar? Wer leitet Testfolgen aus Tests ab? Die TeFoBi-Studie zielt darauf ab, auch Antworten auf diese Fragen zu finden. Es werden sowohl Diskussionen untersucht, die aufgrund von Tests zukünftige Verfahren abwägen, als auch Umsetzungen dieser Diskussionen, die tatsächliche Testfolgen darstellen. Diese Testfolgen werden in Sprechsequenzen ausgehandelt. Innerhalb des sprachlichen Kontexts sind folgende Fragen aus Sicht der Tester interessant: Was tun wir, wenn wir Testergebnisse aussprechen? Und was tun wir, indem wir sie aussprechen? (Vgl. dazu die sprechakttheoretischen Unterschiede dieser Aussagen bei Austin 2002, S. 35).

Aus testtheoretischer Sicht stellen Tester durch die Bekanntgabe der Testergebnisse falsifizierbare Aussagen auf. Aus sprechakttheoretischer Sicht äußern sie Behauptungen, die dadurch, dass sie geäußert wurden, Zustände in der Welt verändern. Beide Varianten sind wissenschaftlich von Bedeutung. Die TeFoBi-Untersuchung wendet sich dem zweiten Fall zu. In den Fokus wird die performative Kraft der Tests genommen, denn diese löst Testfolgen aus.

Ebenfalls ins Blickfeld rücken in der TeFoBi-Studie die Test*konsumenten*. Test*konsumenten* sind Personen, die Tests verwenden (im übertragenen Sinne verbrauchen), indem sie Tests als Argumente anbringen, um Handlungstrategien vorzuschlagen. Der Fokus darauf führt zu diesen Fragen: Wie werden Tests konsumiert? Wie reagieren die Konsumenten auf Tests? Wie gehen sie mit Testergebnissen um? Die TeFoBi-Studie dokumentiert und analysiert, welche Reaktionsschemata bzw. welche Formen und Typen von Testfolgen auftreten. Indem in der TeFoBi-Studie

Berichtsequenzen gesammelt und kategorisiert werden, können typische Schemata bei Testfolgen entdeckt werden. Sie beschreiben, wie Personen in der Öffentlichkeit über Testfolgen diskutieren und wie vorgeschlagene Testfolgen (schließlich) umgesetzt werden.

Um zu klären, was Testfolgen sind, wie man sie interpretieren kann und wie man zukünftig damit umgehen soll, wird in der TeFoBi-Studie eine gesellschaftliche Perspektive eingenommen. Einzelne Personen, Gruppen oder Institutionen werden dabei beobachtet, wie sie Tests konsumieren. Die erkenntnisleitenden Fragen der TeFoBi-Untersuchung lauten: Wie verändern Testfolgen das Bildungssystem? Wer verändert das Bildungssystem mit Hilfe von Tests?

Die skizzierten Praxisprobleme können folgendermaßen zusammengefasst werden: Testvorgänge verändern Bildungsprozesse. Daher wird diesem Phänomen in der TeFoBi-Studie nachgegangen. Neben dem Interviewbeitrag des Experten trifft man auf viele Äußerungen ähnlicher Art aus unterschiedlichsten Quellen. Brügelmann titelt am 13. Januar 2011 in Die Zeit: „Pisa macht die Schulen nicht besser". Und am 21. Januar 2010 fordert Baumert in einem Interview in derselbigen: „In erster Linie muss den schwächsten Schülern ein Mindestmaß an Bildung vermittelt werden" (Interview, Baumert 2010). Beide Äußerungen sind Forderungen aufgrund von Testergebnissen. Sowohl der in der Einleitung zu Wort gekommene US-amerikanische Experte als auch die deutschen Experten beobachten und fordern Veränderungen und beziehen die Folgen dabei auf Testresultate. Meist werden Veränderungen gefordert, weil Testergebnisse dem Bildungssystem Mängel nachweisen oder weil wiederholte Tests zeigen, dass andere Bildungssysteme besser seien.

Berichte wie die obigen sind erste Belege, dass auf verschiedene Art und Weise mit Testfolgen umgegangen wird. Personen tauschen sich über Testergebnisse aus und diskutieren, wie mit ihnen umgegangen werden soll, bzw. was sich aus ihnen folgern lässt. Aufgrund von Tests werden Aufforderungen in gesellschaftlichen Argumentations- und Rechtfertigungsdiskussionen formuliert, um etwas zu gestalten. In den Diskussionsbeiträgen wird geäußert, dass Bildungsgegebenheiten verändert bzw. behalten, eingestellt oder hervorgebracht werden sollen. Die Verwendung von Testergebnissen als Referenzrahmen im öffentlichen oder privaten kommunikativen Handeln belegt im praktischen Vollzug, dass Tests einem Zweck dienen oder einem Zweck dienlich gemacht werden. Sie werden eingesetzt, um etwas zu erreichen. Doch wann erreichen sie etwas? Wann führen sie zu einer Veränderung? Und wann zählen sie als triftiges Argument? Diese Fragen repräsentieren das Erkenntnisinteresse, in dem die Zusammenhänge und Übergänge zwischen Test, Testfolge und Steuerung skizziert sind.