SPRINGER BRIEFS IN EDUCATION

Jaap Scheerens *Editor*

Opportunity to Learn, Curriculum Alignment and Test Preparation

A Research Review

Springer

SpringerBriefs in Education

More information about this series at http://www.springer.com/series/8914

Jaap Scheerens
Editor

Opportunity to Learn, Curriculum Alignment and Test Preparation

A Research Review

With Contributions from:
Marloes Lamain
Hans Luyten
Peter Noort

Springer

Editor
Jaap Scheerens
Oberon Research and Consultancy
Utrecht
The Netherlands

ISSN 2211-1921 ISSN 2211-193X (electronic)
SpringerBriefs in Education
ISBN 978-3-319-43109-3 ISBN 978-3-319-43110-9 (eBook)
DOI 10.1007/978-3-319-43110-9

Library of Congress Control Number: 2016946940

© The Author(s) 2017
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Acknowledgments

This study was carried out as cooperation between Oberon Research and Consultancy and the Department of Research Methodology, Measurement and Data Collection of the University of Twente. The work was supported by a grant from NRO, the Netherlands Initiative for Educational Research in The Hague.

Contents

About the Book

Alignment between educational goals, intended and implemented curricula, and educational outcomes is considered as a characteristic of effective education. The expectation is that better alignment leads to better student performance. The concept of opportunity to learn, abbreviated as OTL, is commonly used to compare content covered, as part of the implemented curriculum, with student achievement. As such it is to be seen as a facet of the broader concept of "alignment." One of the aims of this study is to further clarify these concepts, identify how they have been used in research studies, and are employed in practice; the other major aim is to assess the state of the art on OTL effects. This is done on the base of review of meta-analyses, review of 51 research studies carried out during the last twenty years, and a secondary analysis of TIMSS and PISA results. The results of the study indicate a modest, but educationally significant effect size for OTL. Legitimate forms of test preparation are seen as a promising approach to optimize OTL. Enhancing the curricular validity of high-stakes tests and closer monitoring of alignment chains are seen as relevant interventions for educational policy aimed at improving OTL.

Chapter 1
Focus and Design of the Review Study

Jaap Scheerens

Abstract Alignment between educational goals, intended and implemented curricula and educational outcomes is considered as a characteristic of effective education. The expectation is that better alignment leads to better student performance. The concept of Opportunity to Learn, abbreviated as OTL, is commonly used to compare content covered, as part of the implemented curriculum, with student achievement. As such it is to be seen as a facet of the broader concept of "alignment". One of the aims of this study was to further clarify these concepts, identify how they have been used in research studies, and are employed in practice. A second major aim of this study was to review the state of the art on research on OTL effects by means of a search for meta-analyses and recent primary studies. The following research questions were addressed: (1) Which facets are to be distinguished in clarifying the overall concept of OTL, and how are these to be placed as part of more general models of educational alignment and systemic reform? (2) What is the average effect size of OTL (association of OTL with student achievement outcomes), as evident from available meta-analyses, review studies, secondary analyses of international data-sets and (recent) primary research studies? (3) What are the implications of the results on (1) and (2) for educational policy and practice?

Study Aims and Research Questions

Alignment between educational goals, intended and implemented curricula and educational outcomes is considered as a characteristic of effective education. The expectation is that better alignment leads to better student performance. The concept

J. Scheerens (✉)
University of Twente, Enschede, The Netherlands
e-mail: j.scheerens@utwente.nl

J. Scheerens
Oberon, Utrecht, The Netherlands

© The Author(s) 2017

J. Scheerens (ed.), *Opportunity to Learn, Curriculum Alignment and Test Preparation*, SpringerBriefs in Education,
DOI 10.1007/978-3-319-43110-9_1

of Opportunity to Learn, abbreviated as OTL, is commonly used to compare content covered, as part of the implemented curriculum, with student achievement. As such it is to be seen as a facet of the broader concept of "alignment". One of the aims of this study is to further clarify these concepts, identify how they have been used in research studies, and are employed in practice. Although opportunity to learn was originally studied within the context of curriculum research, it has also obtained a place in educational effectiveness research. Within this research orientation OTL is seen as "an effectiveness enhancing condition" and can be compared with other such factors for its influence on student achievement. As a matter of fact, results of meta-analyses would suggest that OTL has a relatively substantial average-effect size when it is compared to other effectiveness enhancing conditions, such as learning time and instructional leadership (Scheerens 2016). Yet, the number of meta-analyses and review studies on the effects of OTL is rather limited. This study seeks to make a step towards updating the state of the art, by means of a search for meta-analyses and recent primary studies.

Several recent trends in the ongoing global efforts to improve the quality of education provide further perspective to assessing the state of the art on OTL, these are *alignment within the context of systemic reform*, *the accountability movement*, and *task related cooperation between teachers.*

Alignment Within the Context of Systemic Reform

In influential reports by the OECD and McKinsey the quality of educational systems is considered in systemic terms, as a whole of impulses and mechanisms at system, school and classroom level (OECD 2010; McKinsey & Company 2010). Alignment between levels in various functional domains is a key concept in finding out why certain educational systems do better than others. The expectation is that systems do better when aims, objectives, curricula and assessment programs are well-aligned. The conceptual analysis in this report, starting out from OTL intends to further clarify the complexity of alignment between "curricular elements" and opens up discussion on alternative interpretations, for example by comparing proactive structuring and retroactive planning.

Accountability and Its Influence on Teaching

As indicated in the above, OTL originates from curriculum theory and research. According to a pro-active logic, aims are operationalized to standards, worked out as intended curricula, which are expected to be implemented with a certain fidelity, and finally evaluated and assessed, by means of examinations and formative and summative assessment. This is still a valid logic, although developments in the

direction of greater school and teacher autonomy may give rise to a different orientation. More curricular autonomy that goes together with a more prominent role of "high stakes" testing might lead to situations where teaching gets more direction from alignment to the assessment programs than from references to rather global and "open" curricula. A negative interpretation from this phenomenon is "teaching to the test". A more positive interpretation is described by terms like "exam preparation" and "instructional alignment" (Popham 2003; Sturman 2011; Polikoff and Porter 2014). One of the challenges of this study is to provide suggestions for legitimate test preparation, while avoiding harmful interpretations in "teaching to the test".

Task Related Cooperation Between Teachers

The teacher has a key role in realizing "opportunity to learn"; the choice and use of textbooks may be one issue in how this plays out. Another medium is teacher training and professional development of teachers. Recent studies in the realm of teacher training effectiveness underline the importance of teacher content knowledge and pedagogical content knowledge (Baumert et al. 2010; Blömeke et al. 2014; Scheerens and Bloemeke 2016). Within the context of continuous "on the job" professional development teacher cooperation and "peer learning" have obtained high profile (e.g. Thurlings and den Brok 2014). Results of meta-analyses underline the importance of task related work in order to make teacher cooperation effective (Lomos et al. 2011). The results of this study will be used to provide suggestions for placing OTL and instructional alignment on the agenda of task related teacher cooperation.

The general objectives of the review study are to create more clarity about the conceptualization of OTL within a broader framework of educational alignment and to assess the available research evidence about OTL effectiveness. For this latter objective the focus is on the positive significance of OTL effects, effect sizes, and the degree to which OTL effect are related to contextual conditions, such as subject matter area, grade level and national context where the study was conducted.

More specifically the following research questions are addressed:

(1) Which facets are to be distinguished in clarifying the overall concept of OTL, and how are these to be placed as part of more general models of educational alignment and systemic reform?
(2) What is the average effect size of OTL (association of OTL with student achievement outcomes), as evident from available meta-analyses, review studies, secondary analyses of international data-sets and (recent) primary research studies?
(3) What are the implications of the results on (1) and (2) for educational policy and practice?

Methods

The study approach consists of a conceptual analysis, based on literature review. Review of research literature: meta-analyses, research reviews and primary research studies, and secondary analyses on data from international assessment studies, TIMSS and PISA.

Conceptual Analysis

The following issues are addressed

- The definition of OTL. OTL will be defined from the perspective of three research traditions: curriculum research, educational effectiveness research and (international) student assessment.
- Embedding OTL in a broader framework of "educational alignment".
- Alternative ways to measure OTL (in terms of research methods, respondents, content focus and/or focus on psychological operations that students are expected to master).
- The role of teachers in realizing OTL.

Literature Search

First of all, an inventory will be made of available meta-analyses with respect to OTL and instructional alignment. Next, from the available international assessment study reports one or two examples will be selected for secondary analyses of OTL effects. Finally, a systematic search of recent primary OTL effectiveness studies will be carried out. A set of explicit selection criteria will be used to arrive at a set of relevant studies with sufficient research quality.

Analyses of Research Literature and Available (International) Data Sets

A narrative review will provide a summary of the identified review studies and meta-analyses on OTL effectiveness. Average effect sizes from these meta-analyses will be compared with similar results for other effectiveness enhancing conditions, like learning time, educational leadership, teacher cooperation and evaluation at school level. The data from international assessment studies yield descriptions of the way OTL was measured in these studies, as well as effect sizes (OTL associated

with student achievement) within and between countries. The individual research studies identified by means of the systematic searches, and application of the selection criteria will be schematically summarized. Basic analyses of the tabulated descriptions provide information about the proportion of studies in which OTL had a positive and significant effect on student achievement (a so called vote-count analysis), grade-levels addressed in the studies, subject matter area in which OTL was measured and nationality of the study. Vote counts found for OTL in this study are compared to vote counts for other effectiveness enhancing variables, computed in other review studies.

Exploration on How OTL and Educational Alignment Are Addressed in the Practice of Dutch Primary Education

This exploration is based on a limited number of interviews with experts and officials in the areas of curriculum development, educational testing, and textbook production. Preliminary results will be discussed with a panel of teachers.

Structure of the Report

In the second chapter a conceptual analysis of Opportunity to Learn (OTL) is given, covering also related terms, such as instructional alignment and test preparation. The OTL issue is highlighted from three educational research traditions: educational effectiveness research, curriculum research and achievement test development. The conceptual analysis leads to pinpointing OTL as a specific type of alignment in educational systems; a taxonomy of alignment forms is presented. Next, different facets of the way OTL is measured empirically are discussed. The conceptual analysis is given further theoretical depth, by discussing De Groot's (1986) integrative model of didactic and evaluative operationalization. Reflecting on this model brings the alignment issue in a systemic perspective, leading up to the conjecture that alignment, OTL and test preparation aim for integration in organizational structures that are often to be characterized as loosely coupled.

In the third chapter and inventory of meta-analyses of OTL effects (association between measures of OTL and instructional alignment with cognitive achievement outcomes) is presented. This leads to a first impression of the average magnitude of OTL effects. Next, seven case-study descriptions of illustrative OTL research studies are given, spanning four decades of research. The illustrative studies provide an impression of the diversity in emphasis, with exposure of content taught, and alignment between different curriculum elements (like standards, textbooks, taught content and tested content) as two different kinds of independent variables. One of the meta-studies is more specifically oriented to implications of high stakes test for content selection in teaching.

In the fourth chapter an overview is given of about 50 primary studies, conducted during the last twenty years. Schematic summary descriptions are put together in a table. Although quantitative meta-analysis of these studies is beyond the scope of this review study, some basic summary tables are produced to provide an overall orientation on how OTL has been researched and what can be concluded about its effectiveness.

In the fifth chapter secondary analyses based on data from international studies are presented.

In the sixth and concluding chapter conclusions are drawn, and the relevance for educational science and policy and practice is considered. Illustrations will be provided that are drawn from the exploration of policy and practices in the Netherlands. The chapter leads up to recommendations for educational policy planners and teachers.

References

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., et al. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*, 133–180.

Blömeke, S., Hsieh, F.-J., Kaiser, G., & Schmidt, W. H. (2014). International perspectives on teacher knowledge, beliefs and opportunities to learn. Teachers Education and Development Study in Mathmeatics, (TEDS-M). Dordtrecht: Springer.

De Groot, A. D. (1986). *Begrip van evalueren*. 's-Gravenhage: Vuga.

Lomos, C., Hofland, R., & Bosker, R. (2011). Professional communities and learning achievement —A meta-analysis. *School Effectiveness and School Improvement, 22*, 121–148.

McKinsey & Company. (2010). *How the world's most improved school systems keep getting better*. McKinsey & Company.

OECD. (2010). *Strong Performers and successful reformers in education: Lessons from PISA for the United States*. Paris: OECD.

Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis, 36*, 399–416.

Popham, W. J. (2003). *Test better, teach better: The instructional role of assessment*. Alexandria, Virginia: ACSD.

Scheerens, J. (2016) Educational Effectiveness and Ineffectiveness. *A critical review of the knowledge base*. Dordrecht, Heidelberg, New-York, London: Springer.

Scheerens, J., & Bloemeke, S. (2016). Integrating teacher education effectiveness research into educational effectiveness models (Submitted).

Sturman, L. (2011). Test prepatation: Valid and valuable, or wasteful? *The Journal of the Imagination of Language Learning, 9*, 31–37.

Thurlings, M., & den Brok, P. (2014). *Leraren leren als gelijken: Wat werkt?*. Eindhoven: Eindhoven School of Education, Technische Universiteit Eindhoven.

Chapter 2
Conceptualization

Jaap Scheerens

Abstract In the second chapter a conceptual analysis of Opportunity to Learn (OTL) is given, covering also related terms, such as instructional alignment and test preparation. The OTL issue is highlighted from three educational research traditions: educational effectiveness research, curriculum research and achievement test development. The conceptual analysis leads to pinpointing OTL as a specific type of alignment in educational systems; a taxonomy of alignment forms is presented. Next, different facets of the way OTL is measured empirically were discussed. The conceptual analysis is given further theoretical depth, by discussing De Groot's (Begrip van evalueren. Vuga, 's-Gravenhage, 1986) integrative model of didactic and evaluative operationalization. Reflecting on this model brought the alignment issue in a systemic perspective, leading up to the conjecture that alignment, OTL and test preparation aim for integration in organizational structures that are often characterized as loosely coupled, which might explain sub-optimal effects of these policies.

Introduction

At first glance "opportunity to learn" would seem to be one of the few concepts in educational science that you could clarify to you mother or grandmother in two minutes. It addresses the expectation that students will do better on educational content tested when that content has actually been taught, which almost sounds like a truism. Throughout this report we will remain close to this basic clarification, since we are not here to complicate matters unnecessarily. As this study is a

J. Scheerens (✉)
University of Twente, Enschede, The Netherlands
e-mail: j.scheerens@utwente.nl

J. Scheerens
Oberon, Utrecht, The Netherlands

© The Author(s) 2017

J. Scheerens (ed.), *Opportunity to Learn, Curriculum Alignment and Test Preparation*, SpringerBriefs in Education,
DOI 10.1007/978-3-319-43110-9_2

research review we shall also encounter the basic and simple conception of opportunity to learn (abbreviated as OTL) in the empirical studies that will be analyzed. The correlation between a measure of OTL and cognitive achievement in school subjects, like mathematics, science and mother tongue langue, is the key parameter of investigation. Nevertheless the exploration of the literature shows complexity that goes beyond the basic definition. The OTL issue can be situated in at least three different traditions of educational research and development, with corresponding differences in research orientation, shows considerable variability in the way it has been defined and operationalized, and has different contexts of practical application as well (e.g. national educational policy and school level teaching). In this chapter a "conceptual map" of OTL will be presented.

Building Blocks for a Conceptual Framework

OTL is a construct that depends on the *alignment* of educational goals or standards, actual teaching and assessment (measurement of student achievement). These elements can be positioned in a basic system model of education, which is often used to define educational effectiveness.

The elementary design of educational effectiveness research is the association of hypothetical effectiveness enhancing conditions (OTL being one of these) and output measures, mostly student achievement. The basic model from systems theory, shown in Fig. 2.1, is helpful in clarifying this design. The major task of educational effectiveness research is to reveal the impact of relevant input characteristics on output and to "break open" the black box in order to show which process or throughput factors "work", next to the impact of contextual conditions. The model, shown in Fig. 2.1, can be used at different levels of aggregation. In the figure this is indicated by mentioning three levels in the central box of the model: the level of a national educational system, the school level and the level of the instructional setting, often indicated as the classroom level. The three levels are nested, in the sense that schools function within an educational system at national level and classrooms function within schools.

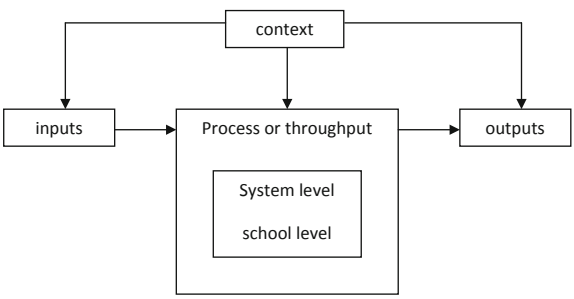

Fig. 2.1 A basic systems model on the functioning of education

In terms of this model the alignment between standards, actual teaching and student achievement can be seen as the association of inputs (e.g. national standards), processes (teaching) and outputs (student achievement). Accordingly, OTL can be characterized as the alignment between inputs and teaching processes, as the alignment between teaching processes and student achievement, or as the alignment of standards and output measures, mediated by teaching processes. An example of a relevant context variable is the degree of centralization of an educational system. To the degree that the educational system is centralized, national standards, or a national curriculum, are likely to be more detailed and prescriptive in the way they are supposed to be applied by schools. When an educational system is more decentralized national curricula might be just rudimentary, consist of quite general goals, or even be totally absent.

The educational effectiveness perspective is just one of three research and development orientations in which OTL is approached in a specific manner. The other two perspectives are the logic of curriculum research and test preparation. In this study our emphasis will be on the educational effectiveness perspective; so this context for OTL research will be explained first.

OTL in the Context of Educational Effectiveness Research

In educational effectiveness research OTL is one of a series of malleable, effectiveness enhancing conditions at national system, school and classroom level that are expected to be positively associated with student achievement, also when outcomes are adjusted for student characteristics like previous achievement, scholastic aptitude and socio economic status. Other malleable variables addressed in educational effectiveness research are indicated in the overview presented in Table 2.1, cited from Scheerens (2014).

Table 2.1 Summary of effectiveness enhancing teaching variables by Muys et al. (2014), adapted from Scheerens (2014)

Teaching effectiveness, Muys et al. (2014)
Opportunity to learn
Time
Classroom management
Structuring and scaffolding, including feedback
Productive classroom climate
Clarity of presentation
Enhancing self-regulated learning
Teaching meta-cognitive strategies
Teaching modelling
More sophisticated diagnosis
Importance of prior knowledge

In educational effectiveness research OTL has been used as an independent variable defined mostly at school and classroom level. With the development of international assessment studies country level definitions of OTL have also been used. A defining characteristic of OTL studied from an educational effectiveness perspective is that measured student achievement is the dependent variable.

OTL in Curriculum Research

Curriculum research, rather than educational effectiveness research, forms the intellectual heritage of OTL. OTL in curriculum research shares the systemic perspective with the more recent multi-level studies in educational effectiveness (Scheerens 2016). The research orientation in curriculum research is broader than in educational effectiveness research. In the curriculum context alignment is addressed in its broadest sense, including sometimes "alignment" with measured student achievement, interpreted as the "realized curriculum", but not limited to that. The building blocks for our conceptual framework on OTL that were previously mentioned (standards, actual teaching and student outcomes) have specific terminology in curriculum research, where one speaks of the intended, implemented and realized curriculum. In curriculum research alignment between national curriculum standards (intended curriculum), intermediary elements, such as school level standards and textbook content, and taught content is studied in its own right, without necessarily involving student outcomes.

Alignment between standards, intermediary elements, teaching and assessment of student outcomes has different interpretations when considering the association between pairs of elements.

(1) *Alignment between curriculum goals or standards and intermediary elements* such as school standards or textbooks could be considered in terms of construct validity; the key question is whether textbooks or school standards provide a "covering" representation of the national standards. Assuming that national goals or standards are likely to be defined in more general terms than are the content elements in school standards and textbooks, the analogy with construct validity seems more appropriate than content validity, which would presuppose matching of elements from two sources of comparable specificity. In the case of construct validity expert panels would be needed to decide whether school standards or textbook content could be seen as adequate representations of the more general goals, or national standards.

(2) *Alignment between curriculum goals and standards and teaching* (i.e. *the implemented curriculum*). Here the same reasoning could be applied as in case 1, in the above. The actual feasibility of assessing this kind of alignment would strongly depend on the national standards being sufficiently specific; in addition empirical methods to observe or otherwise measure teaching behavior in practice would be required.

(3) *Alignment between intermediary elements* (*school standards and textbooks*) would allow for a more straightforward consistency check, based on content analysis of the school standards and textbooks and matching with measures of content covered by teachers. Here the practical reason for carrying out a consistency check could be the choice of the most suitable textbook, given school standards
(4) *Alignment between intermediary elements and assessment instruments*. Here content elements of the intermediary elements would be matched with the content elements that make up the assessment instrument. This might be done either based on actual test items or on (more general) content elements derived from the analytic frameworks on which the test is based. Such frameworks usually consist of two dimensional matrices, specifying cognitive operations required in relationship to content elements. The context of application might be test validation or analyzing the opportunities to learn that are stimulated by school standards or textbooks.
(5) *Alignment between teaching* (*implemented curriculum*) *and assessment content*. This would have essentially the same interpretation and contexts of application as described with point 4.
(6) *Alignment between general goals or national standards and assessment content.* As with the alignment discussed in point 2, the feasibility of this approach would strongly depend on the specificity of the national standards.
(7) "*Alignment*" *of any other of the main elements to student achievement outcomes* (*the realized curriculum*). This kind of alignment refers to the most common definition of OTL. The term alignment is questionable in this case, because the association, although mostly just measured by means of correlational measures is likely to be interpreted as causal. The most frequent application is the one between content covered in teaching and achievement results. Contexts of application are: establishing the predictive validity of OTL measures and assessing school or teaching effectiveness. In the latter sense curriculum research and educational effectiveness research overlap.
(8) More complex models of alignment, where intermediary elements may be studied as mediators of higher level elements (examples will be provided in Chap. 3).

In this taxonomy of alignment types, when national standards, intermediary elements (school standards and textbooks) and assessment instruments and measures are the basic elements, the emphasis has been on matching and consistency. Pelgrum (1989), presents a conceptual framework in which mismatches and deficiencies, next to matches, are given explicit attention. His work took place in the context of international comparative assessment studies by the IEA (International Association for the Evaluation of Educational Achievement), in which variability between countries in the way the international assessment test corresponded to national intended and implemented curricula, was scrutinized from the perspective of "fair" comparison.

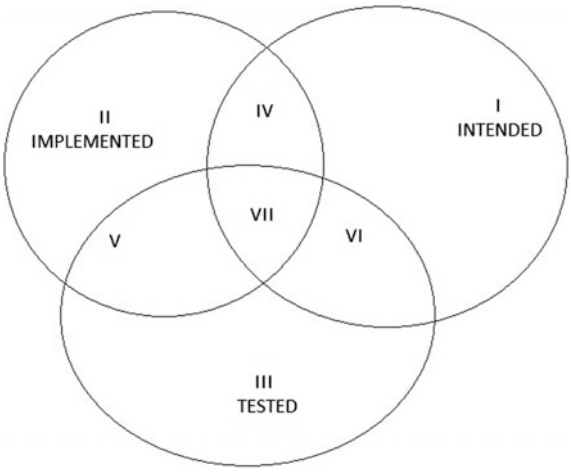

Fig. 2.2 Venn diagram of intended, implemented and tested curriculum, from Pelgrum (1989)

The presentation of Pelgrum's model is cited in Fig. 2.2 (Pelgrum 1989, p. 17). The numbered areas in Fig. 2.3 are described as follows:

"I + IV + VI + VII: what students should learn.
II + IV + V + VII: what is actually taught at school.
III + V + VI + VII: what is tested.
I: what students should learn, but is actually not taught at school, and not tested.
II: what actually is taught at school, but not tested and not part or what students are supposed to learn.
III: what is tested, actually not taught at schools, and not part of what students are supposed to learn.
IV: what students should learn and what is actually taught at school, but not tested.
V: what actually is taught at school and tested, but is not part of what should be learned.
VI: what students should learn and is tested, but is actually not taught at schools.
VII: what students should learn, what is tested and taught." (ibid., 17).

The theoretical principle behind these analyses of consistency between the various facets of curriculum can be indicated with the term "coupling". Analyses that tend to underline deficiencies could be seen as manifestations of "loose coupling" in educational organizations (Weick 1976); the positive alternative of good integration between the curriculum facets can be indicated with the term "alignment". Successful OTL is an example of alignment, fallible OTL can be seen as a manifestation of loose coupling.

OTL as Test Preparation

In this section we shall start out with on orientation on the process of educational achievement test development. As we shall see test development follows the same kind of specification process, from general goal descriptions to test items, as were encountered in curriculum analysis and development. When comparing curriculum development and test construction we can establish, first of all that they have the first and last step of the development process in common, the first being a national curriculum with general goals and national standards, and the last step being the assessment instruments. Most interesting are intermediary "products". In curriculum research we encountered school standards, textbooks and implemented curriculum as intermediary steps. Analyzing test construction shows other kinds of intermediary products. When discussing test preparation as an interpretation of OTL these intermediary products are quite interesting. Holcome's: "taxonomy of score inflation", illustrates what is meant by intermediary products in test development (Holcombe, 2011).

The term "score inflation" refers to the context of application of this taxonomy, which is teaching to the test. Each of the decisions in test design (like specifying subsets of standards and material to be covered within standards) is seen as narrowing the domain for testing and creating opportunities for teaching to the test. The subsequent decisions in test development concern content specification but also choice of representations, such as item formats.

The specification process in test development for a particular subject could be seen solely from a content perspective. The deductive steps then go from major domains of a discipline, to subdomains, to more specific topics and ultimately to item content. However, at the more detailed levels, the level of topics and test items, a second dimension is usually added, in the form of the cognitive demand of the topic or item. Topics are thus defined as a combination of the specification of a content element and a particular psychological operation. The cognitive demand dimension can be arranged from simple to more complex cognitive operations. See Porter et al. (2011, 104).

In the test frameworks for PISA the cognitive operation dimension is further sub-divided in terms of process categories and cognitive demand. Next, a context dimension and a response type dimension are distinguished to further characterize test domains and test items. The context dimension consists of a personal, societal, occupational and scientific sub dimension.

So what does it mean that intermediary specification levels in curriculum and test design have quite similar analytic structures consisting of specification of content and psychological operations with a certain demand or difficulty level? Obviously this facilitates empirical research on different types of curriculum alignment, see for example Porter et al. (2011). Perhaps more interesting is to further reflect on implications for OTL optimization. Here the attention would go particularly to the association between teaching content and test content. The question is whether "test preparation" can be seen as a constructive and "legitimate" way to optimize OTL.

Traditionally this kind of alignment has the unfavorable connotation of "teaching to the test". But, perhaps, when certain technical requirements of tests are met, specific ways to direct teaching to these tests are not so bad. We shall return to these questions after having further analyzed the communalities and differences between OTL from the curriculum perspective and test preparation facilitated by test characteristics. In the next section an integrative model of "didactic and evaluative specification" (De Groot 1986) will be discussed to try and make further progress on these issues.

An Overarching Model of "Didactic and Evaluative Specification of Educational Goals"

De Groot (1986) describes the development of curriculum programs as the result of a process in which policy goals, background characteristics of students and societal demands are the key inputs to choose general goals, and create an overall vision of how to attain these goals. In a subsequent step of specification, goals are formulated as attainable end-terms (effects); "standards" in more contemporary terminology.

In Fig. 2.3, these steps are represented with A, B and C, in the upper part of the figure. Next the specification process splits up in two directions: "didactic operationalization" (D) and "evaluative operationalization (E)". The didactic operationalization leads to a concrete plan in the form of school standards and teaching methods, which in a next step is brought into practice (the implemented curriculum). The evaluative specification leads to the design of test instruments and ultimately to test items, norms, and decision rules about success or failure. All relationships in the figure, A through H, are indicated as "coverage problems"; the total of specifications at a lower level should cover the main themes of a higher level. Because higher level descriptions are in broader terms De Groot prefers the analogy of construct validation to judge the success of coverage of goals by curriculum elements and test frameworks at lower levels to content validity. Content validity would be theoretically adequate if the higher level goal formulation would be a precise collection of elements, and a test a representative sample from those elements. However, according to De Groot, educational goals at higher level are more than collections of content elements, because they may also refer to general skills, like problem solving or social skills. And this means that, ultimately, expert judgment is required to assess the content validity of lower level elements, like textbooks, test frameworks and tests. Relationship H in Fig. 2.3 is crucial, it refers to our basic definition of OTL: the degree to which the content tested has actually been taught.

De Groot's framework underlines the analogy between curriculum and test design, and offers criteria to determine the quality and alignment of these two construction processes. In the recognition of vertical coverage in the didactic and the evaluative column, and "horizontal consistency" between the two columns in

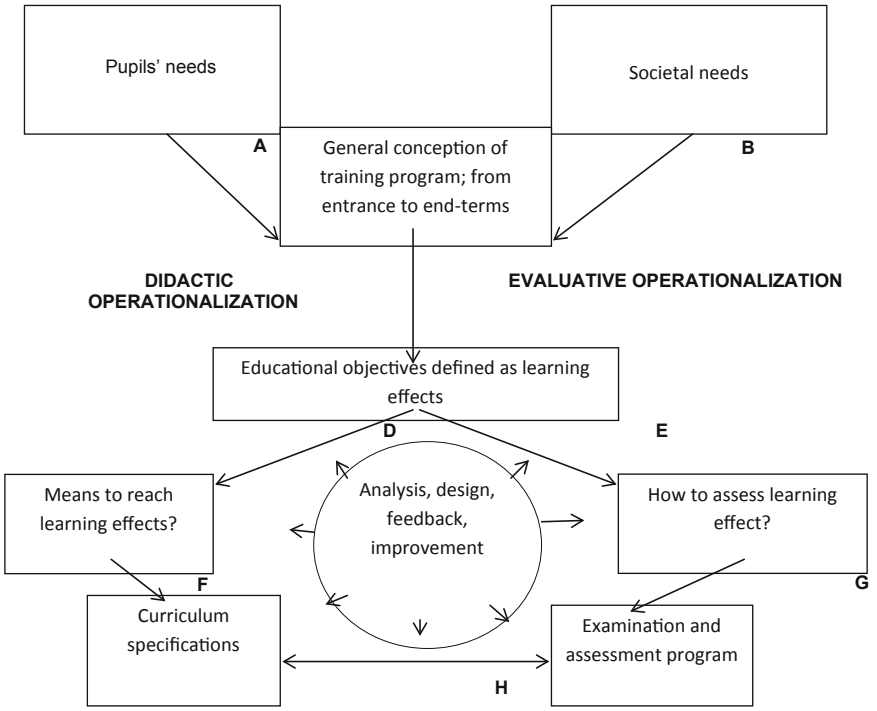

Fig. 2.3 Two kinds of operationalizations of educational goals, adapted from De Groot (1986)

the form of OTL. A few more practical issues inspired by this integrative theoretical framework concern the way the two processes should be organized over time (which should be done first?), whether didactic and evaluative operationalization should be carried out by different, independent organizations and finally, whether alignment would not be served by streamlining the organizational conditions.

Should Evaluative Specification Precede Didactic Specification or Vice Versa?

De Groot argues that evaluative operationalization should happen first, because curriculum design needs verifiable learning effects to adequately resolve issues of instrumentality, in other words constructing means that are adequate to reach goals and intended effects. If the evaluative specification would follow the didactic specification there would be too big a chance of pressure to adapt tests to preferred methods (goal displacement).

Should Evaluative Specification and Didactic Specification be Carried Out in Different Organizations?

According to De Groot evaluative and didactic specification should be independent. His main argument is that curriculum developers lack the know- how to carry out test construction. Next, in high stakes evaluative applications, like examinations, independence is required to guarantee objectivity. In actual practice various organizational units are often involved in specific phases of curriculum development and test development. Developers of teaching methods and textbooks are often independent firms operating outside the public sector; the same may apply to test developers.

From De Groot's analytic model, but also from our earlier presentation of alignment issues, it seems that what we have are two operationalization processes that are quite similar. From a theoretical perspective, but also from the point of practical application it is therefore challenging to think further about a more efficient approach in organizational structures that might be more aligned and less "loosely coupled".

How Feasible is Optimizing Alignment in a Leaner Organizational Structure?

To a degree, alignment in educational systems, as discussed so far, is a remedy to a self-created problem. Particularly in the curriculum development column in Fig. 2.3 organizational units at various levels are involved in the process of what De Groot describes as "didactic operationalization". At national level priorities, general goals and overall time tables are established by either the central administration or national institutes that operate closely to the central administration. Depending on the degree of centralization of the system and the specificity of the national curriculum, at intermediary level various organizational units may have a prominent role in the process of didactic operationalization as well: textbook writers, educational support organizations, school districts and school boards. These intermediary units develop "intermediary curricular elements", like district standards, textbook content coverage, and school standards, creating as many areas where alignment becomes an issue. Again, depending on the specificity of the intermediary elements and the autonomy of teachers, the teachers will have more or less space to make their own choices in what is actually taught. So, in summary, there is a lot of need for vertical coordination in the didactic specification column. Looking once more at the question of horizontal alignment, i.e. the correspondence between elements in the didactic specification column and elements in the evaluative specification column, we saw that De Groot argues for a leading role of test development. Evaluative specification should precede curriculum specification because concrete and specific end terms (i.e. ultimately collections of test items) are needed to guide

the curriculum development process. It is rather questionable whether such kind of interplay and coordination between test development and curriculum development is frequently realized in practice. If it is not realized another alignment issue arises, creating, most probably important discrepancies between what is taught and tested; in other words limited OTL. It is important to realize that the quest for alignment in educational systems, of which OTL is a specific issue, happens in a context where structural arrangements tend to be loosely coupled.

The question is how matters could be improved, first of all "in theory" and secondly in practice, when all kinds of contextual conditions of a structural and cultural nature should be taken into consideration.

Hypothetical Solutions to the Alignment Problem in Educational Systems

Two ideal type scenarios will be addressed in the next section: centralism and synoptic planning, and retroactive planning, combined with high stakes accountability.

Centralism and Synoptic Planning

Although, during the last two decades, there has been a strong tendency in many countries to decentralize educational systems and provide more autonomy to lower levels (schools in particular), some previously decentralized countries like the UK and the USA have gone in the other direction. In the UK national programs for numeracy and literacy were developed and implemented, and in the USA Common Core Standards have been developed. Explicit national curriculum standards provide clear direction for both didactic and evaluative operationalization. At the very least it opens the possibility to empirically verify the alignment between, for example, the national standards and the contents of assessment instruments.

Rational, synoptic planning can be seen as the theoretical background of national curriculum planning.

The ideal of "synoptic" planning is to conceptualise a broad spectrum of long term goals and possible means to attain these goals. Scientific knowledge about instrumental relationships is thought to play an important role in the selection of alternative ways to realize these goals.

The main characteristics of synoptic planning as a prescriptive principle conducive to effective (in the sense of productive) organizational functioning, as applied to education, are:

- "proactive" statement of goals, careful deduction of concrete goals, operational objectives and assessment instruments;

- decomposition of subject-matter, creating sequences in a way that intermediate and ultimate objectives are approached systematically;
- alignment of teaching methods (design of didactical situations) to subject-matter segments;
- monitoring of the learning progress of students, preferably by means of objective tests.

According to this model curriculum development is seen as a deductive process of operationalizing general goals, ultimately also in terms of achievement test items. Developing achievement tests is the last step in this deductive process.

There are many obstacles to apply this model: resistance against national standards and "state pedagogy", incomplete knowledge about instrumental relationships, lack of vertical coordination between the central administration, intermediary organizations and schools, technical problems in getting from general goals to more operational formulations, and resistance by schools against externally developed guidelines and programs. Finally, the linear sequence from general goals to test items implies that didactic specification precedes evaluative specification and this is probably less efficient (see De Groot's argumentation in favour of the opposite position in which evaluative specification precedes didactic specification).

Retroactive Planning, Combined with High Stakes Accountability

A less demanding type of planning than synoptic planning is the practice of using evaluative information as a basis for corrective or improvement-oriented action, sometimes indicated as "retroactive planning" (Scheerens et al. 2003). In that case planning is likely to have a more "step by step", incremental orientation, and "goals" or expectations get the function of standards for interpreting evaluative information. The discrepancy between actual achievement and expectations creates the dynamics that could eventually lead to more effectiveness. In cybernetics the cycle of assessment, feedback and corrective action is one of the central principles.

Evaluation—feedback—corrective action and learning cycles comprise four phases:

- measurement and assessment of performance;
- evaluative interpretation based on "given" or newly created norms;
- communication or feedback of this information to units that have the capacity to take corrective action;
- actual and sustained use (learning) of this information to improve organizational performance.

This model resembles approaching alignment by given precedence to what De Groot indicates as "evaluative specification".

When evaluative specification proceeds curriculum and didactic specification, it could also be seen as "taking the lead" in a more substantive way. Substantively processes of curriculum specification and test construction have much in common. This is particularly the case if we focus on learning tasks and assessable educational objectives. Admittedly, curriculum design has a broader scope, in also needing to address the choice and development of means (teaching strategies, classroom organization, and application of educational resources) apart from just having to select and ultimately implement subject matter based *content*. When recognizing the thoroughness of achievement test development one might wonder why a parallel process of specification and a parallel intermediary structure would be required in the didactic specification column. All that would be required might be an additional unit in a test development agency which proposes evidence based teaching strategies in relationship to content elements and educational objectives. Next, specific technical issues should be met.

Firstly, construction teams should have multi-disciplinary expertise with subject matter specialists in the lead, supported by didactical experts and test development experts.

Secondly, tests should be curriculum valid, relative to national standards and criterion referenced.

Thirdly, "half products" of test development like test frameworks and the specification of sub-domains should be made publicly available; for example to advice textbook writers.

Fourthly, calibrated item banks should be publicly available as well, in order to allow targeted test preparation by schools (van der Linden 1985).

Finally, moderate or high stakes accountability regimes would give schools a motivational impulse to align their teaching with educational objectives, standards and tests.

Particularly the fourth condition, calibrated item banks, allowing for legitimate test preparation would, in principle, be a strong step forward in attaining content alignment between what is taught and tested.

Ways to Empirically Assess OTL

Empirical procedures to measure OTL vary according to the *scope of the OTL definition*, the *data source*, the *level of the curricular unit envisaged,* and whether *exposure* or *alignment* is measured.

Scope of the OTL Definition

The basic definition of OTL refers to educational content. Further elaboration of this basic orientation considers qualitatively different cognitive operations in

association with each content element, often also expressing accumulating complexity (see the example from PISA 2012, presented in an earlier section). A next step in enlarging the scope is to add an indication of the time students were exposed to the specific content elements. Sometimes the theoretical option to include *quality of deliverance* to the OTL rating is considered as well. This option will be disregarded here as it is seen as stretching the OTL definition to a degree that it approaches a multi-dimensional measure of overall instructional quality.

Data source

OTL measures may be based on teacher judgments or student judgments. A third option is to consider unusual scoring patterns as instances of OTL differences.

The level of the Curricular Unit Envisaged

Curricular sub-domains, more specific topics, or test-items represent different levels at which OTL may be assessed.

Exposure or Alignment

The independent variable in assessing the impact of OTL on achievement can be a measure of exposure (has this content element been taught or not, or with a certain frequency) or an alignment measure. An example of an alignment measure as the independent variable is the correspondence between a measure of exposure and the contents of standards or assessment instruments. So in the latter case alignment is first assessed by means of content analyses methods, and then correlated with student achievement. An example is provided in the study by Polikoff and Porter (2014) which will be discussed in more detail in the next chapter.

Since basically all these dimensions on which OTL measure may differ can be crossed with one another, it follows that there is a broad range of ways to empirically assess OTL. This diversity could be seen as a problem when the objective would be to conduct meta-analyses of OTL effectiveness research.

Conclusion

What seemed like a relatively simple concept, at second sight, OTL proves to be rather complex. From the perspective of curriculum research, but also in fairly recent systemic modeling of educational effectiveness research (Scheerens 2016), OTL is part of a range of alignment issues, usually involving national standards, prescriptive formulations at intermediary level, like school standards and test frameworks, actual teaching and ultimate achievement measurement. Operational definition and measurement of OTL is also complex, in the sense that many options are possible, depending on the scope of the operational measure, measurement source, the curricular unit used to define OTL and the question whether OTL is operationalized as exposure or alignment.

In the final sections of the chapter, optimizing OTL was connected to the way educational systems are organized, particularly with respect to those facets of the system created to play a role in curriculum and test development. A preliminary conclusions was that alignment is an ideal of strong matching and coupling, projected in an actual organizational context that is usually "loosely coupled".

The option to give precedence to what De Groot calls "evaluative operationalization" puts the spotlight on test-preparation, which in its turn opens up the question about legitimate and dysfunctional applications (teaching to the test). We shall turn back to all these issues in the last chapter of the report, in which we shall also formulate recommendations for educational practice and policy. But this will be done after we have taken a thorough look at the research evidence, concentrating on OTL effects on student achievement, which is the main issue of this study.

References

De Groot, A. D. (1986). *Begrip van evalueren*. 's-Gravenhage: Vuga.

Holcombe, R. W. (2011). *Can the 10th Grade Mathematics MCAS Test be Gamed? Opportunities for Score Inflation.* Unpublished manuscript. Harvard University, Cambridge, Massachusetts.

Muys, D., Creemers, B., Kyriakides, L., Van der Werf, G., Timperley, H., & Earl, L. (2014). Teaching Effectiveness. A state of the art review. *School Effectiveness and School Improvement., 25*, 231–257.

OECD. (2014). *PISA 2012 Results, Volume I, What students know and can do. Student performance in mathematics, reading and science.* Paris: OECD Publishing.

Pelgrum, W. J. (1989). *Educational assessment. Monitoring, evaluation and the curriculum.* Enschede: Febo Printers.

Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis, 20*, 1–18.

Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common core standards: The new U.S. intended curriculum. *Educational Researcher, 40*(3), 103–116.

Scheerens, J., Glas, C. and Thomas S. M. (2003) *Educational evaluation, assessment and monitoring. A systemic approach*. Lisse: Swets & Zeitlinger.

Scheerens, J. (2014). School, teaching, and system effectiveness: some comments on three state-of-the-art reviews. *School Effectiveness and School Improvement, 25*(2), 282–290. ISSN 0924-3453.

Scheerens, J. (2016). *Educational Effectiveness and Ineffectiveness. A critical review of the knowledge base.* Dordrecht, Heidelberg, London; New York: Springer. http://www.springer.com/gp/book/9789401774574

van der Linden, W. J. (1985). *Van standaardtest tot itembank. Inaugural Lecture*. Enschede: University of Twente.

Weick, K. E. (1976). Educational organizations as loosely coupled systems. *Administrative Science Quarterly, 21*, 1–19.

Chapter 3
Meta-Analyses and Descriptions of Illustrative Studies

Jaap Scheerens

Abstract In the third chapter an inventory of meta-analyses of OTL effects (association between measures of OTL and instructional alignment with cognitive achievement outcomes) is presented. This leads to a first impression of the average magnitude of OTL effects. The evidence from meta-studies that reviewed OTL effects appeared to be less solid than was expected. Leaving out a strongly outlying result, the OTL effect-size (in terms of the *d*-coefficient) compares to other relatively strong effectiveness enhancing conditions at school level, at about 0.30. One of the meta-studies was more specifically oriented to implications of high stakes tests for content selection in teaching. Next, seven case-study descriptions of illustrative OTL research studies are given, spanning four decades of research. The illustrative studies provide an impression of the diversity in emphasis, with exposure of content taught, and alignment between different curriculum elements (like standards, textbooks, taught content and tested content) as two different kinds of independent variables.

Introduction

In this chapter research evidence about OTL will be reviewed by means of an overview of meta-analyses and descriptions of illustrative primary studies.

J. Scheerens (✉)
University of Twente, Enschede, The Netherlands
e-mail: j.scheerens@utwente.nl

J. Scheerens
Oberon, Utrecht, The Netherlands

© The Author(s) 2017

J. Scheerens (ed.), *Opportunity to Learn, Curriculum Alignment and Test Preparation*, SpringerBriefs in Education,
DOI 10.1007/978-3-319-43110-9_3

Meta-Analyses of OTL Effects

Opportunity to learn has been incorporated in several meta-analyses of educational effectiveness (Scheerens and Bosker 1997; Marzano 2003; Scheerens et al. 2007; Hattie 2009). Results from these studies are summarized in Table 3.1. Apart from opportunity to learn, other independent variables are included, to allow for a comparison with the effect sizes for OTL.

Several remarks about this table should be made.

First of all, the results from Marzano are citations from Scheerens and Bosker, for all variables apart from opportunity to learn. Secondly, the study by Scheerens et al. (2007) is an update of the analyses by Scheerens and Bosker, by including studies carried out between 1995 and 2005, roughly doubling the total number of effect sizes that were analyzed. However, the considerable overlap between both data-sets should be recognized and explains the close correspondence in results between 1997 and 2007.

Thirdly, the result cited from Hattie's meta-analysis of meta-analyses, as far as OTL are concerned, is based on a somewhat remote "proxy" of OTL, namely "enrichment programs for gifted children". It is quite striking, given the comprehensiveness of Hattie's study, that opportunity to learn is not directly addressed in his analyses of school and teaching variables.

Table 3.1 Rank ordering of school effectiveness variables according to the average effect sizes (d-coefficient) reported in three reviews/meta-analyses

	Scheerens and Bosker (1997)	Marzano (2003)	Scheerens et al. (2007)	Hattie (2009)	Average effect size
Opportunity to learn	0.18	0.88	0.30	0.39*	0.44
Instruction time	0.39	0.39	0.30	0.38	0.37
Monitoring	0.30	0.30	0.12	0.64	0.34
Achievement pressure	0.27	0.27	0.28	0.43**	0.31
Parental involvement	0.26	0.26	0.18	0.50	0.30
School climate	0.22	0.22	0.26	0.34	0.26
School leadership	0.10	0.10	0.10	0.36	0.17
Cooperation	0.06	0.06	0.04	0.18***	0.09

*Operationalized as "enrichment programs for gifted children"
**Operationalized as "teacher expectations"
***Operationalized as "team teaching"

Table 3.2 Summary of studies that had produced OTL effects, based on results presented by Creemers (1994), and cited by Marzano (2003)

Study	Effect size (*d*)
Husen (1976)	0.68
Horn and Walberg (1984)	1.36
Pelgrum et al. (1983)	0.45
Bruggencate et al. (1986)	1.07
Average effect size	0.88

As matters stand, the results presented in Table 3.1 have several dependencies among the studies that are cited. As far as OTL is concerned, it is interesting to trace Marzano's coefficient of 0.88 a bit further. Where all other coefficients of the variables were cited from Scheerens and Bosker (1997), the coefficient for OTL was constructed on the basis of data published by Creemers (1994). These results are cited in Table 3.2.

Marzano explains that for OTL he preferred to cite Creemers, and not Scheerens and Bosker, because Creemers used a definition that was more specifically targeted at OTL in the sense of congruence between the coverage of content in teaching and the test. Scheerens and Bosker had used a more composite interpretation of OTL and curriculum quality, based on:

(1) The school having a well-articulated curriculum
(2) Choice of methods and text-books
(3) Application of methods and textbooks
(4) Opportunity to learn
(5) Satisfaction with the curriculum (Scheerens and Bosker 1997, 135).

I think that Marzano's choice is quite defensible, however the evidence that leads to an average effect size of 0.88 is just based on four early studies, and influenced by two values, that are exceptionally high (Horn and Walberg's 1.63 and Ten Bruggencate's 1.07). All in all considerable reservation is in place with respect to the relatively high average effect size shown for OTL in Table 3.1.

The results in Table 3.1 mostly concern independent variables interpreted at school level. Scheerens et al. (2007), and Seidel and Shavelson (2007) also addressed classroom level teaching interpretations of OTL. Both studies were conducted on largely converging data sets, originally developed by Seidel and Steen (2005), with both studies using slightly different analyses techniques (ibid., 2007). Scheerens et al. found an effect size of r = 0.12 for OTL at teaching level. Seidel and Shavelson computed a composite indicator in which time and opportunity to learn and homework are integrated. The effect size they report for this indicator is r = 0.04. Both studies also included variables about teaching subject matter related learning strategies. Scheerens et al. label a composite of these variables as "Teaching learning strategies", Seidel and Shavelson use the term "domain specific processing". In both meta-analyses the effect size for this composite comes out as a relatively high r = 0.22. It could be argued that this variable has something in common with OTL, in the sense that the focus is on domain related cognitive

operation facets (see the conceptual analysis in Chap. 1, in which both learning tasks and test items are defined two-dimensionally, namely on content elements and cognitive operations). Still this indicator is closer to pedagogy than to content covered, which is the basic characteristic of our OTL definition.

A literature search directed at research studies that had assessed OTL effects on student achievement, conducted for this study in the autumn of 2015, and which will be described in Chap. 4, included a search for meta-analytic studies on OTL The initial search yielded a disappointingly small number of quantitative meta-analyses: Kablan et al. (2013), Kyriakides et al. (2013), Schroeder et al. (2007) and Spada and Tomita (2010). The study by Au (2007) qualifies itself as a qualitative meta-synthesis. The study by Kablan et al. (2013) had an independent variable defined as "use of educational material"; which is a too general concept to be seen as OTL.

The meta-analysis by Kyriakides et al. (2013) considers the teaching level factors of the "Dynamic model of educational effectiveness", developed by Creemers and Kyriakides (2008). These teaching factors are: orientation (goal setting), structuring, questioning, teaching modeling, application, the classroom as a learning environment, management of time and assessment. "Application" is the factor that might be seen as having common elements with OTL. Application is a composite of—opportunities to practice a skill or a procedure presented in the lesson,—opportunities to apply a formula to solve a problem, and opportunities to transfer knowledge to solve everyday problems. Since the connection with content covered and content aligned to assessment measures is not explicitly addressed, this meta-analysis has no great relevance for this study. The effect size (correlation) that was found for application was 0.18; actually the smallest of the whole set of factors (with effect sizes of the other factors ranging from 0.34 to 0.45).

The study by Schroeder et al. (2007) consisted of a meta-analysis of U.S. research published from 1980 to 2004 on the effect of specific science teaching strategies on student achievement. The following eight categories of teaching strategies were revealed during analysis of the studies (effect sizes –*d*—coefficients—in parentheses): Questioning Strategies (0.74); Manipulation Strategies (0.57); Enhanced Material Strategies (0.29); Assessment Strategies (0.51); Inquiry Strategies (0.65); Enhanced Context Strategies (1.48); Instructional Technology (IT) Strategies (0.48); and Collaborative Learning Strategies (0.95). A total of 61 studies were analyzed. The independent variable in this meta-analysis that has some resemblance to OTL is "Enhanced material strategies". This factor is described in terms of teachers' modifying instructional materials (e.g., rewriting or annotating text materials, tape recording directions and simplifying laboratory apparatus). OTL resemblance of this factor would depend on closer alignment to assessment being a strong rationale for adapting texts and other materials. Since there is no trace of this actually being the case, the results of this meta-analysis are of limited relevance to this study.

The meta-analysis by Spada and Tomita (2010) was conducted to investigate the effects of explicit and implicit instruction on the acquisition of simple and complex grammatical features in English as a second language. The results indicated larger effect sizes for explicit over implicit instruction for simple and complex features. The findings also suggested that explicit instruction positively contributes to

learners' controlled knowledge and spontaneous use of complex and simple forms. The theoretical background of the study is the claim by some researchers that whereas easy rules can be taught, hard rules are by their very nature too complex to be successfully taught and thus difficult to learn through traditional explanation—and—practice pedagogy. Hard rules are thought to be best learned implicitly, embedded in meaning-based practice. Although "explicit teaching" might be compared to teaching content that is well matched to what is assessed, in other words OTL, the connection is considered rather weak, because there is no specific attention for content coverage and test content and therefore no further attention is given to this meta-analysis.

As expressed in the titled, the study by Au (2007) "High-stakes testing and curricular control: a qualitative meta-synthesis" is not a quantitative meta-analysis yielding a numerical average effect size across studies. Instead, a rubricating exercise is followed, in which a total of 49 qualitative studies is sorted according to a template that consists of 6 categories that express a certain kind of implication of high-stakes testing for teaching content, knowledge form and pedagogy.

In most cases changes in content (83.7 %), knowledge form (69.4 %) and pedagogy (77.6 %) were observed as an alignment response to high stakes accountability. In a majority of cases content became narrower and more fractured, and pedagogy became more teacher-centered. Yet, in a relative minority of studies content was expanded to deliberately teach extra material over and above the test content, and the way the knowledge was offered became more integrated.

The study is an example of a retro-active interpretation of content alignment, namely the (implemented) curriculum adapting to the test content, and not the other way around (the test adapted to the curriculum). An interesting observation by the author is that in the cases where the respondents (teachers) were positive about the test, they had a favorable judgment about the curricular alignment as well. The "triplet" content contraction, more fractured knowledge and more teacher centered pedagogy was the dominant combination (75 %) of the cases in which all three were measured. The opposite triplet, content expansion, more integrated knowledge and more student-centered teaching occurred in 24 % of the cases.

When making up de balance about what meta-analyses tell us about OTL-effects the yield appears to be less than expected. The number of meta-analyses that have explicitly addressed OTL is limited. In several cases OTL has been combined with other characteristics in calculating effect sizes (for example in the meta-analyses by Scheerens and Bosker, Scheerens et al., and Seidel and Shavelson). Marzano's (2003) relatively high coefficient (d = 0.88) appears to depend on just four relatively "old" studies, two of which have exceptionally high effect sizes. More recent meta-analyses Schroeder et al. (2007), Kyriakides et al. (2013), and Spada and Tomita, only studied variables, that for our purposes, are to be seen as relatively remote proxies of OTL. The study by Au (2007) was the only review which considered retroactive alignment (teaching to assessment). We shall return to the assessment of the available empirical knowledge base in the final chapter of this report.

Meta-Analysis of Test Preparation Effects

For substantive and practical reasons the review of studies on test preparation was given lower priority than originally intended. Initial analysis of studies that were identified in a first literature search produced a limited number of studies. When this material was examined it appeared that a large part of the studies had looked at test preparation from the perspective of providing training to students in test taking skills and familiarity with certain types of questions and test items. Another, smaller part had addressed content preparation as well. For our purposes, the latter kind of studies are more relevant, and better defensible as a form of enhancing OTL, whereas preparation in test taking skills, on the other hand, hardly fits this perspective. The usefulness of the studies that were identified was further limited by the fact that earlier studies (roughly before 2000) had mainly studied test preparation for norm referenced tests of general scholastic aptitude. Again this is somewhat remote from test preparation as a form of curricular alignment with criterion referenced tests, based on content standards. Last but not least test preparation is often depicted as "bad" practice, associated with teaching to the test. Although this leads to very interesting discussions on legitimate and illegitimate forms of test preparation, it makes the empirical research results quite heterogeneous. To provide a flavor of the debate, the following citation from Sturman (2003) illustrates the various perspectives well:

> Koretz et al. (2001) note that the term 'test preparation', in common usage has a negative connotation. However, they distinguish seven types of test preparation, three of which, they argue, can produce unambiguous, meaningful gains in test scores.
>
> These are 'teaching more', 'working harder', and 'working more effectively'. In contrast, three strategies ('reallocation of resources', 'alignment' of tests with curricula and 'coaching' of substantive elements) can lead to either meaningful or inflated gains, whilst the seventh strategy ('cheating') can lead only to inflated grades.

In the final chapter, when discussing implications for teachers of the results that were brought together in this report, we shall turn back more extensively to the issue of legitimate and illegitimate test preparation.

As far as empirical research results are concerned we found two meta-analyses. Bangert-Drowns et al. (1983) reported a mean effect size of 0.25 sd, based on analysis of 30 studies on effects of test coaching. Messick (1982) reported an average effect size of lower than one fifth of a standard deviation based on his meta-analysis of 39 studies on test preparation for SAT. These relatively small effects may have been due to the nature of the norm referenced tests that were used as the criterion.

A handful of individual research studies Xie (2013), Farnsworth (2013) and Sturman (2003) report small but positive effects of various forms of coaching and test preparation.

Descriptions of Illustrative Studies

In the second part of this chapter descriptions of illustrative studies addressing OTL or instructional alignment are presented. The descriptions are placed in a chronological order, with the first dating from 1992 and the latest dating from 2015. It is quite striking that all but two of the illustrative studies are linked to international assessment studies, such as TIMSS and PISA, whereas the two exceptions are associated with large scale studies in the USA, namely NAEP and the MET (Measuring Effective Teaching, by the Bill and Melinda Gates Foundation). An overview of the 7 study descriptions is provided in Table 3.3.

Measuring Test-Curriculum Overlap
Dieuwke De Haan
Thesis, University of Twente, 1992

Scope

The main purpose of the study was the development and validation of an instrument to measure "test curriculum overlap" (TCO). Test curriculum overlap is defined as the degree of overlap, measured on the basis of teachers' self-reports, between taught content and content as operationalized in test items. The context of application is to use the OTL measure as a control variable in international assessment studies, in order to make comparisons that could be considered more faire, based on commonly taught content.

Table 3.3 Overview of illustrative studies covering OTL and "alignment"

Study	Setting
De Haan (1992)	The Netherlands, classic OTL
Schmidt et al. (2001)	USA, classic OTL
Schmidt et al. (2011) *Content Coverage Differences across Districts/States: A Persisting Challenge for U.S. Education Policy*	USA, classic OTL
Porter et al. (2011)	USA, alignment, NAEP
Polikoff and Porter (2014)	USA, alignment
OECD (2014)	International, classic OTL
Schmidt et al. (2015)	International, classic OTL

Instrument Development

In IEA studies TCO had been measured by asking teachers whether the content of *test items* had been taught. Reference to Pelgrum et al. (1986) is made to refer to some results of the application of the IEA instrument. These authors concluded on the basis of analyses of SISS (second international science study) and SIMS- (second international mathematics study) data, that in general the validity of the instrument seemed quite high; the predictive validity at between country level appeared to be 0.57 (correlation TCO and cognitive achievement). However, within country analysis for the Netherlands, when comparing scores between classes, showed a considerably lower correlation of 0.22. The authors (ibid., 1986) indicated some limitations of the IEA procedure to measure OTL: lack of knowledge by responding secondary teachers about previous learning experiences, content that students had learned, not from the actual responding teachers but in previous episodes of their school career; teachers' judgements about content taught possibly being confounded by considering item formats, or being arbitrary when an item would cover different content elements (in fact the inter rater reliability of the procedure is being questioned). Next, problems might occur when content is taught not only in the chosen time period teachers were asked to report on, but also previously or later. A further questionable assumption is that all students in a class were taught the same content. And, finally, teachers were asked to make absolute judgements about content being taught or not, with no room for nuances as to the degree of emphasis in teaching a particular content element.

These problematic aspects of the IEA procedure prompted the researcher to develop an alternative instrument, indicated as the "D-TCO instrument". This instrument includes additional questions about when the specific content element is taught, and whether the item format is familiar or not to the students. Because this appeared to be quite a labor intensive procedure, a second alternative instrument was developed "H-TCO" in which the teacher is asked to select those items that would be fit to be included in an assessment test relevant for the content taught until the testing date.

Research Design

In order to study the validity of the D-TCO instrument a pretest-posttest design was chosen (p. 44). The instruments were administered twice. In April 1990, 31 mathematics teachers of different types of secondary education judged a set of items according to the D-TCO instrument and the H-TCO instrument. The items were also judged by their students; for each item they answered the question whether an item was taught before the date of testing. In order to study the predictive validity of the D-TCO instrument, the students also answered the items. The data collection was repeated at the end of the school year, in July 1990. To make it possible to

adapt a textbook analysis for each teacher individually, in the intervening time period, teachers registered which part of the textbook was treated and which homework was given. In order to compare the efficiency of both the teacher based instruments, teachers registered the time needed to judge the item set.

Results: Validity of the TCO Measures

The construct validity of the instruments was computed by comparing teacher administered versions to textbooks and student judgements of whether items had been taught or not. The conclusion was that correspondence between these alternative methods was fair at aggregated level (items and groups of teachers). In an absolute sense percentages taught were higher for the textbook analyses than for the instruments.

Study of the predictive validity pointed out that both the D-TCO and the H-TCO scores correlated "somewhat" with student achievement. More precisely, correlations at the test level, i.e. between average student test score per teacher and percentage of taught items per teacher varied between 0.37 and 0.46. At item level, the correlation between average item scores over all students and percentage of teachers judging an item as taught, correlations varied between 0.32 and 0.50 (p. 91). Correlations for the H-TCO instrument were slightly higher (0.42–0.54).

The difference in student achievement scores between the pre-test and post-test were larger for the items that had been taught in the in-between period, this was also taken as a support of the predictive validity.

Comment: it should be noted that correlations were not adjusted for student background characteristics. The author concludes that "it might be questioned whether the D-TCO measure is a good predictor for student achievement". (91).

Schmidt, W. H., McKnight, C.C., Houang, R.T., Wiley, D.E, Cogan, L.S., and Wolfe, R.G. (2001) Why schools matter. A cross-national comparison of curriculum and learning. San Francisco: Jossey-Bass

Abstract, General Description

This is a comprehensive study on what nowadays would probably be indicated as "curriculum alignment". The curriculum is defined as the intended course of study and sequences of learning opportunities in formal schooling.

The curriculum is described as comprising of four "artefacts" which together express curriculum intentions and implementation:

- Standards (official documents)
- Textbooks

- Teachers' content goals (the proportion of teachers in a country who say they have taught the content; score per topic)
- Duration of (actual, implemented) content coverage (national average of how long-time—a topic was taught, during a study year).

The dependent variable in the study is Learning Gains in Math and Science, as measured in TIMSS 1995, at 8 grade level. Gain constructed on the basis of students being either in grade 7 or grade 8. The study includes an interesting discussion on tests measuring general competencies (more prone to be influenced by background and given student characteristics) and tests that are sensitive to curriculum differences. (Issue of norm referenced and criterion referenced tests).

The focus of the study is on the content required to answer test items correctly.

Attention is given to the context of national differences, by referring to national culture and institutions, in the form of national goals and purposes that reflect cultural beliefs and values (distinction of general and fine grained goals), distribution of authority (locus of decision making), the articulation of curricular areas and topics, preferences concerning achievement assessment, formal or informal.

The contents of the book consists of, an introductory chapter (Chap. 1): a part on measuring the facets of an overall conceptual model (Chaps. 2 and 3); a part focused at relationships comprising intentions and implementations of the curriculum, as well as the variability of these, within and between countries (Chaps. 4–6, and two chapters focused at the relationship between the implemented curriculum (textbook use and teacher implementation) and student achievement (Chaps. 9 and 10).

Conceptual Model and Measures

The study is based on a conceptual model at three levels of abstraction: national cultures—curriculum areas and types of achievement assessment, intended, implemented and attained curriculum and the operational level: content standards, textbooks, teacher content covered and student learning.

The TIMSS math and science frameworks consist of content units (topics) and performances expected of students (other authors have referred to the latter as cognitive operations). The TIMSS framework was used as a metric to compare national curricula, in terms of topics addressed proportional to the whole range of topics in the TIMSS framework. These analyses yield measures of national curricula or national content standards. In measuring curriculum documents, like textbooks, "blocks" were the fundamental units that were coded, counted and analyzed. In the measures of the implemented curriculum, based on teacher questionnaires, "lessons", or "instructional periods" were the unit. Classroom instruction was measured as both the percentage of teachers within a country addressing a topic, as well as the mean percent of instructional time they reported as teaching the topic. For all curricular artefacts the same metric was used, so that intended,

implemented and attained curriculum could be matched. The dependent variable consisted of achievement on a subset of TIMSS math and science items (actually 22 of the total of 44 topics in the Framework) covering a specific sub set of topics from the framework. "The use of a common framework for both analyzing test items and elements of curriculum content allow careful matching of test content to curriculum manifestations." (27) To quote the authors: "Only specificity has a reasonable chance to reveal relationships between curriculum and achievement, and then only when achievement gains on specific content is related to curriculum efforts in that specific period related to that same content".

Results Concerning Curriculum Variability and Alignment Between Intensions and Implementations

International comparison of curriculum goals and content standards show a lot of variation. The common core of a World Curriculum, which would be determined by the international structure of the discipline, would appear to be quite narrow. This core curriculum consists mostly of algebra and geometry.

"Among those countries planning to cover the fewest of the tested topics, several (Czech Republic, Japan, Korea) were among the top 7 performers on the 8 grade math tests" (p. 86). Possible explanation: these topics were taught at an earlier grade level. Question for international comparisons: how fair is a test if the topics were not covered?

When considering textbook coverage of topics, countries varied from 15 framework topics (Japan) to 44 (the USA). A result illustrating the degrees to which textbooks covered a certain topic was that 50 % of the countries have at most 20 % of their textbooks devoted to that topic. Also large within country variation of content covered in textbooks was found. "The large ranges for most countries suggest considerable variation in complex performance expectations within each country (100).

When considering teacher coverage it was established that 5 topics were emphasized by each of the four indicators of curriculum (content standards, textbooks, teacher coverage and achievement measurement) across all TIMSS countries (area of equations and formula and two-dimensional geometry). From this point begins the divergence among the indicators. Often a breach was noted between intended and implemented content, this was interpreted as a split between the worlds of policy and practice. Chapter 6 provides further details on the association between the intended and implemented curriculum. Correlations were mostly positive significant. Structural relationship were significant for math but not for science (169). At general aggregate level the median correspondence between any two curricular indicators was very small across all topics, essentially 0. (176) Per topic (math) correlations varied between 0 and 0.20. "The coefficients of determination show textbook space accounting for around 10–40 % of the variance in

both instructional time and proportion of teachers covering a topic". One relationship appeared to be constant, namely a direct relationship of textbook space to teacher implementation. Varying patterns of association between different curriculum indicators and achievement were found. For example, in the Netherlands, math achievement assessment was positively associated with textbook use, but not with any other curriculum indicator. Quite surprisingly all path coefficients between curriculum indicators were positive in the NL for science. No direct influence of curricular centralization was noted, but there was an indirect effect through the path between content standards and teacher implementation as mediated by textbook space.

Association of Curriculum Aspects and Measured Achievement Gain Across Countries

The first type of analysis that was conducted was a pair-wise comparison between topic coverage of a particular curriculum indicator (e.g. textbook) and the achievement results for that topic.

"The pairwise coefficients at the interaction level (topics x countries) are all positive for both mathematics and science. This indicates that more curriculum coverage of a topic area—no matter whether manifested as emphasis in content standards, as proportion of textbook space, or as measured by either teacher implementation variable—is related to larger gains in that same topic area" (261). These are analyses at country level; the results show that the nature of these general relationships is not the same for all countries.

Next, a two-way analysis was carried out, by means of testing a structural model. In this structural model, the effect of a particular curriculum aspect was estimated while controlling for the effect of the other aspects of curriculum.

The results of the complete structural model analyses (264,265) are summarized as follows:

> The conception proposed by and represented in the structural model is supported by the data. In general, content standards in mathematics were related to teacher implementation both directly and indirectly through textbook coverage. Teacher implementation was in turn related to achievement gain (although only marginally so for instructional time per topic as the dependent variable). The results for science were more complex, with positive direct effects for content standards (sign 0.016) and both teacher implementation variables (significance at 0.002 and 0.003 respectively).

The relationships described were generally not uniform across countries. Main exception was the relationship between textbook coverage and learning, which appeared to be generalizable across all TIMSS countries; for math the relationship was positive (R square was 0.31, modestly strong), for science negative. The regression coefficient for pair-wise uncontrolled association between textbook coverage and achievement was 0.75, for content standards is was 0.72.

The report presents interesting comparisons of country patterns, e.g. between Japan, Singapore, and the US. However, in 20 of the 30 countries not a single path coefficient reached statistical significance. For the uncontrolled analyses instructional time had a positive significant coefficient in only 2 of the 30 countries. For textbook coverage this was the case for 3 of the 30 countries.

Finally within country associations between instructional time and achievement gain were analyzed, controlling for SES and some instructional variables. Results showed that in the USA there were significant relationships for thirteen of twenty topic areas (ignoring marginally significant relationships). In France, there were no significant topic areas with significant relationships. Canada was closest to the USA in results, with nine of twenty areas with significant relationships, and Japan had five. The effect size in the USA was rendered as follows: an increase of 3 % OTL (one week of instruction devoted to a topic) would give and effect ranging from 3 to 24 % increase in achievement gain on that topic.

William H. Schmidt, Leland S. Cogan, Richard T. Houang, and Curtis C. McKnight Source: American Journal of Education, Vol. 117, No. 3 (May 2011), pp. 399–427 Content Coverage Differences across Districts/States: A Persisting Challenge for U.S. Education Policy

This article utilizes the 1999 TIMSS-R data from U.S. states and districts to explore the consequences of variation in opportunities to learn specific mathematics content. Analyses explore the relationship between classroom mathematics content coverage and student achievement as measured by the TIMSS-R international mathematics scaled score. District/state-level socioeconomic status indicators demonstrated significant relationships with the dependent variable, mathematics achievement, and the classroom-level measure of content coverage. A three level hierarchical linear model demonstrated a significant effect of classroom content coverage on achievement while controlling for student background at the student level and SES at all three levels, documenting significant differences in mathematics learning opportunities, which were interpreted as a function of the U.S. education system structure.

Scope

> We use the term "opportunity to learn" (OTL) or "educational opportunity" specifically with reference to content coverage, that is, the specific mathematics topics covered in classroom instruction. This reflects both the narrow curricular sense in which the concept was originally developed by Carroll and the International Association for the Evaluation of Educational Achievement (IEA) international studies.

> Our focus is the seminal definition of OTL as content for two reasons: (1) the provision of content is the fundamental rationale of schooling and the education system, and (2) this is an aspect of schooling that both reflects education policy and is amenable to education policy reform.

The study has a systemic perspective: "what many may not realize is the extent to which differences in students' learning opportunities are embedded in and are a function of the very structure of the U.S. education system" (ibid., 400). Even with the presence of well-defined state standards and, more recently, the increasing presence of corresponding state assessments, local districts still maintain de facto control of their curriculum. The American educational governance system is a system of shared responsibility among the states and the more than 15,000 local school districts.

Individual local districts make choices about content coverage with, at best, indirect and more global state control. "American children simply are not likely to have *equal educational opportunities* as defined at the most basic level of equivalent content coverage (ibid., 400)".

Conceptual Framework and Measurements

Reference is made to Caroll and time related interpretations of OTL, time to study a task.

A similar vein of OTL research developed somewhat independently from Carroll. Specifically, the International Study of Achievement in Mathematics (later called FIMS) framed OTL as a content coverage variable without specific regard to allocated time. In the Third International Mathematics and Science Study (TIMSS) Carroll's notion of time was incorporated, yielding a more refined definition of OTL —in terms of the number of periods a specific topic was taught—that became a central focus of the study and emerged as a major factor in making sense of cross-country achievement differences.

The authors consider the estimated strong relationship between SES and OTL as unique for the USA. The common U.S. practice of tracking provides students in the same school with different content opportunities.

> For example, suppose that solving linear equations (the simplest kinds of equations familiar from a first course in algebra and even before) is a learning content goal at eighth grade. Suppose that it is something that all children should know. If so, then exposure to this part of mathematics is central to providing equal educational opportunity for all eighth-grade students in mathematics. (ibid., 405)

Achievement may be seen as an interactive function of the actual content covered and delivery effectiveness.

This study focuses on equality of content coverage across two organizational entities, that is, districts and states. It examines the consistency of educational opportunities with respect to specific mathematics topics together with the associated student academic achievement. To explore this, 13 districts and 9 states that replicated the TIMSS mathematics study in 1999 were used.

The internationally scaled total test score in eighth-grade mathematics for TIMSS-R (Third International Mathematics and Science Study-Repeat) was used as the dependent variable.

Important to the development of the teacher content questionnaires and the tests was the mathematics content framework, which spelled out in detail the specific content covered across the TIMSS world in school mathematics. A hierarchical array of 44 specific mathematics topics within 10 broad areas was developed to cover the full range of K–12 mathematics.

Using the TIMSS cross-national curriculum data, an index of difficulty for each topic was developed. The scale was grade-related (1–12) and referred to as the "international grade placement" index, (IGP), for a topic.

The data came from the teacher questionnaire in which teachers indicated the number of periods of coverage associated with each of a set of topics ("taught before this year," "taught 1–5 periods," "taught more than 5 periods"). Each of the 34 teacher questionnaire topics contained one or more of the 44 mathematics framework topics. The proportion of the school year's instruction in each of the 34 topics was calculated from the questionnaire, creating a profile of mathematics content coverage for the classroom of students the teacher taught. These estimated teacher content profiles were then weighted by the corresponding IGP values and summed across all topics. This produced a single value that was an estimate of the rigor or content-related difficulty of the implemented mathematics curriculum for each teacher.

This measure is described as "the weighted content coverage index", which is a multifaceted measure based on three distinct aspects of OTL: (1) the mathematics content itself (topic coverage—yes/no), (2) instruction time for each topic, and (3) rigor or content difficulty (as estimated from international curriculum data). Therefore, the IGP measure of the mathematics taught in the classroom was seen as a measure of content-specific OTL defined at the classroom level, which can be unambiguously related to classroom achievement.

Types of Results

District-Level Variation in OTL

The percentage of eighth-grade students in the district who were in mathematics classes that focused mainly on the coverage of algebra and geometry. That percentage ranges across the districts, from 14 % in one district to 95 % in another. The IGP index varied from 6.05 to 6.88—almost one complete grade-level difference across the districts.

Relationship of OTL and Achievement at the District Level

R square OTL and achievement, not-controlling for SES was 0.67.

After controlling for SES: Even after controlling for district-level SES, the effect size was estimated at around three-fourths of a standard deviation.

Relationship of SES and OTL

Negative relationship Rsquare = 0.51.

Differentiating Between Classroom and District-Level Relationships

Data were available to estimate the teacher variation in content coverage and its effect on achievement and, as a result, to adjust for it in the statistical model.

> The consequences of decisions made by system-entity components (i.e., districts or states), the focus of this article, can only be believed to exist if OTL differences persist even after controlling for the teacher variation just discussed (ibid., 419).

Controlling for SES and prior achievement, the IGP measure was significant ($p < 0.001$). The coefficient indicated that for a one grade-level difference in IGP, the increase in mean achievement at the classroom level was 0.15 of a standard deviation (419). Teacher knowledge was also significant, but had a relatively small effect relative to OTL.

The authors conclude with the following observation:

> In school mathematics, at least, the United States is sadly not the "land of opportunity" for any student, regardless of wealth or social class. It is the land of the lucky few and the unlucky many in which educational opportunity depends on a fluke (or on other societal factors)—that is, in which school district an education is sought. It depends on factors that cannot be wholly overcome by student ability and effort. (Ibid., 423)

Common Core Standards: The New U.S. Intended Curriculum
Andrew Porter, Jennifer McMaken, Jun Hwang, and Rui Yang Educational Researcher, (2011) Vol. 40, No. 3, pp. 103–116

Abstract

The Common Core standards were released in 2010 for English language arts and mathematics and had been adopted by dozens of states, when this article was published. The central research question that was addressed looked at how much change these new standards represented, and what the nature of that change was In this study the Common Core standards were compared with current state standards and assessments and with standards in top-performing countries, as well as with reports from a sample of teachers from across the country describing their own practices.

Scope

The Common Core State Standards Initiative developed these standards as a state-led effort to establish consensus on expectations for student knowledge and skills that should be developed in Grades K–12. By late 2010 36 states had adopted the standards. Standards pertain to "the content of the intended curriculum," NOT on how that content is to be taught (curriculum, pedagogy). The standards are grade specific. Both (math and language) standards intend to influence the assessed and enacted curricula. Ambitions: shared expectations, focus (benchmarked to high performing countries), efficiency (not each state re-inventing the wheel), quality of assessment (electronical and adaptive). The question was also asked how current state assessments and the National Assessment of Educational Progress (NAEP) compare to the Common Core standards. The Common Core was also benchmarked to standards and assessments from selected other countries. Finally, the authors compared estimates of the enacted curriculum with the Common Core.

Conceptual Framework

The underlying framework for the study was a matrix of topics and cognitive operations: memorize, perform procedures, demonstrate understanding, conjecture, generalize, proof, sole non routine problems. It employs a two-dimensional framework defining content at the intersections of topics and cognitive demands. The topic dimension is divided into general areas: 16 for mathematics and 14 for ELAR (English language and reading). Each general area is further divided into 4 to 19 topics, for a total of 217 topics in mathematics and 163 topics in ELAR. The second dimension consists of five levels of cognitive demand, which differ by subject. Thus, for mathematics, there are 1085 distinct types of content contained in the categories; for ELAR, there are 815.

The alignment index assesses the extent to which two documents have the same content message, based on the extent to which the cell proportions (topics by cognitive demand) are equal cell by cell across two documents.

The index ranges from 0 to 1, with 1 indicating perfect alignment (i.e., having 100 % of the content in common). The value of the index can be thought of as the proportion of content in common across the two documents.

Results from previous research shows that there is great interstate variability for both standards and assessments. Correlation between National assessment and state standards is moderate.

Types of Results

Common Core standards compared to state standards averages an alignment index of 0.30 for ELAR and 0.25 for Math, when aggregated to higher content level dimensions the index rose from 0.25 to 38 for math and from 0.30 to 0.41 for ELAR.

What changes from state standards to common core standards?
For mathematics the Common Core standards represent a modest shift toward higher levels of cognitive demand than currently represented in state standards.

For ELAR, the Common Core standards would shift the content even more strongly than they would for mathematics toward higher levels of cognitive demand (but, in both cases, not to the highest level of cognitive demand).

A lot of differences between topics addressed in Common standards and state standards for both subjects.

Does the Common Core represent greater focus than is currently represented in state content standards?
Focus was addressed by investigating how many cells were needed in the content matrix of topics by cognitive demand to capture 80 % of the total content; the fewer the cells, the greater the focus. The average for state content standards represented greater focus than was seen in the Common Core for ELAR, but the Common Core for mathematics is still more focused. The Common Core has more focus than some states' standards and less focus than other states' standards, both for mathematics and for ELAR.

Comparing Common Core Standards with State Assessments
Across Grades 3–12 in math, the average alignment of state assessments to Common Core standards is 0.19, compared with 0.25 for state standards to the Common Core (Table 6). In ELAR, the average alignment of assessments to the Common Core standards is 0.17, compared with 0.30 for state standards (see Table 7).

For math, the alignment index ranges from 0.10 to 0.31 across states and grades for assessments. For ELAR, the alignment index ranges from 0.07 to 0.32.

The degree to which NAEP assessments align with the Common Core standards: NAEP's alignment with the Common Core standards was 0.28 for fourth grade and 0.21 for eighth grade; the average alignment of state math assessments to the Common Core standards was 0.20 in both grades. In ELAR, however, NAEP has a higher alignment than the average of state assessments in both fourth and eighth grades. NAEP's. Alignment to the Common Core standards is 0.25 in fourth grade and 0.24 in eighth grade, compared with an average alignment of 0.17 for state assessments in both grades.

Benchmarking the Common Core Against Massachusetts (Highest performing state).
The alignment between Massachusetts and the Common Core for mathematics at Grade 7 was 0.19—less than the state average of 0.23 and considerably less than the figure for the most aligned state, 0.34. In ELAR, Massachusetts' alignment was 0.13, again less than the state average of 0.32 and substantially less than the state maximum of 0.43.

International Benchmarking (*Finland*, *Japan and Singapore*)
All three of these countries have higher eighth-grade mathematics achievement levels than does the United States. The content differences that lead to these low levels of alignment for cognitive demand are, for all three countries, a much greater emphasis on "perform procedures" than found in the U.S. Common Core standards. For each country, approximately 75 % of the content involves "perform procedures," whereas in Common Core standards, the percentage for procedures is 38 %.

Clearly, these three benchmarking countries with high student achievement do not have standards that emphasize higher levels of cognitive demand than does the Common Core. (There was a strong common core focus on solving non routine problems.)

Comparing Common Core Curriculum with the Current Enacted Curriculum
For mathematics, the average alignment across teachers to Common Core standards was 0.22, with a standard deviation of 0.042, a minimum alignment of 0.00 and a maximum alignment of 0.33. For ELAR, the mean alignment was 0.27, with a standard deviation of 0.071, a minimum alignment of 0.001, and a maximum alignment of 0.398. Again, generally low levels of alignment were found.

Comment

Focus of this article is not on OTL/achievement correlation, nor on strategies to enhance instructional alignment. Results show that a huge effort needs to be made to align assessments, textbooks and instruction to the new common core standards.

Instructional alignment as a measure of teaching quality
Polikoff, M.S. and Porter, A.C. (2014)
***Educational Evaluation and Policy Analysis,* 36, 4, pp 399–416**

Abstract

> This article is the first to explore the extent to which teachers' instructional alignment is associated with their contribution to student learning and their effectiveness on new composite evaluation measures using data from the Bill and Melinda Gates Foundation's Measurement of Effective Teaching (MET) study. Finding surprisingly weak associations, we discuss potential research and policy implication for both streams of policy (p. 399).

The following research questions were addressed:

(1) To what extent is the alignment of teachers' reported instruction with the content of standards and assessments associated with value-added to student achievement (note; included are two alignment measures, instruction with standards, and instruction with assessments).
(2) To what extent does pedagogical quality moderate the relationship between instructional alignment and value added to student achievement?
(3) To what extent is the alignment of teachers' reported instruction associated with a composite measure of teacher effectiveness?

Conceptual Framework

> It is thought that providing teachers with more consistent messages through content standards and aligned assessments, curriculum materials, and professional development will lead them to align their instruction with the standards, and student knowledge of standards content will improve" (400). "There is abundant evidence from the literature that OTL (including instructional alignment and pedagogical quality) affects student achievement." "In some cases, the effects of OTL are so large that there is no statistically significant residual between-course level (e.g. between general and honors levels) variation in achievement after controlling for OTL (ibid., 401)

Main dimensions of teachers' instruction are: instructional time, content and pedagogical quality.

Ways of capturing these dimensions are self-reports (content), observations and student surveys (pedagogical quality).

The assessment sensitivity depends on the proximity of assessment to instruction: immediate, proximal and distal. State assessments- distal to instruction- differ in their sensitivity to instruction. One reasonable hypothesis is that the more sensitive assessments are more tightly aligned to the content taught by teachers.

> Final elements of this study are the use of classroom-level VAM (value-added) measures and relating instructional alignment to a composite measure of teacher effectiveness (based on MET, observation, student reporting and past performance in terms of VAM).

Method

Investigation of the association of instructional alignment with other measures of teacher effectiveness in two subjects, Math and ELA English Language Arts) and two grade levels, 4th and 8th grade. Earlier research found that the relationship between assessment-instruction alignment and achievement was a correlation of 0.45 (source Gamoran et al. 1997). Given 81teachers per grade/subject cell, the total target sample was 324 teachers. Online survey of the content of their instruction, using the Surveys of Enacted Curriculum (SEC) content taxonomies. Teachers were asked to think about a target MET class when answering the survey. The response-rate was 39 % (p. 402)

Instruments:
SEC (Surveys of the enacted curriculum); the surveys define content at the intersection of specific topics and levels of cognitive demand; there are 183 fine-grained topics in mathematics and 133 in ELA. Cognitive demand varies from memorization to application or proof. Instruction to respondents: first decide which topic were taught or not (in a school year); for those taught indicate a) the number of lessons spent on each topic and b) the level of cognitive demand (cell = topic by cognitive demand combination).

Content analyses to compare teachers' survey responses with SEC content analyses of standards and tests, to estimate instructional alignment to the documents. The raters were content-area specialists with advanced degrees. Fine-grained topics, objects (from the standards) and test items. Test items might be placed in up to 3 cells, while objectives may be placed in up to 6 cells. Test items are weighted for their occurrence in more than one cell. Results: set of proportions, indicating the percentage of total standards or test items in each cell of the SEC. Ergo: a standard profile, an assessment profile, and a teacher self-rating "behavioral", instructional profile (JS).

Validity evidence contained, among others correlations of nearly 0.5 between instruction to test alignment and student achievement gains.

The alignment index is defined as 1—Sum of cell differences between two documents divided by 2. The sum of the cell-by-cell minima can be used to compute a proportion ranging from 0 to 1. Used as a teacher level independent or dependent variable.

Pedagogical quality measures were based on student surveys, video based observations, analyzed in classroom environment sub-scales: environment of respect and rapport, establishing a culture of learning, managing classroom procedures and managing student behavior. Next, responses were used to compute instructional sub-scales; communicating with students, using questioning and discussion techniques, engaging students in learning and using assessment in instruction. All averaged to obtain a composite score.

Achievement was measured by means of VAM (value added measures) (text cited or paraphrased from pp. 403–406).

Results

The alignment of teachers' instruction with state standards and state and alternate assessments was low; with a mean of 0.20 in mathematics and 0.28 in ELA. (Note: these are consistency measures based on content analysis.)

When it came to the zero-order correlations of VAM scores with SEC instructional alignment indices, most of the correlation were not significant. Three correlations with VAM were analyzed: (1) the alignment between instruction and state standards, and (2) the alignment between instruction and state or (3) alternative tests. In those grade, district, subject combinations where the correlations were significant the average was 0.16 for math and 0.14 for ELA. A further result of the correlational analysis was that "Overall, the correlations do not show evidence of strong relationships between alignment of pedagogical quality and VAM scores" (408).

Fixed effects regression pointed out that there was just one coefficient (among some 30 others) that was significant at the 0.10 level, namely the coefficient for the relationship of instruction-alternate test alignment, with alternate test VAM in ELA (408) This significant coefficient indicated that a 1 standard deviation increase in instruction-alternative test alignment with alternate test VAM is associated with a 0.05 unit (approximately 0.2 standard deviation increase in alternate test VAM. Concerning the climate and pedagogical quality variables, there were no significant correlations of any of the three variables with any of the VAM outcomes (in the full study data set of the MET significant relationship were found, but the sizes of the relationships were small).

Conducting a series of interaction models, to see if SEC interacted with pedagogical quality in influencing VAM scores, one significant interaction effects out of 6 was established; effect sizes in the order of 0.3 sd. "Together these results provide, at best, modest evidence of an interactive effect of alignment and pedagogical quality, though the results are in the expected direction that the effects of pedagogical quality is positive when alignment is stronger but not when alignment is weaker" (410). Finally it was established that there was no evidence of relationships between alignment and a composite measure of effectiveness (namely the composite used in the MET, based on VAM and two different observation schedules).

Discussion

The authors conclude as follows:

> We found modest evidence through zero order correlations and regressions that alignment indices were related to VAM scores. These relationships went away when controlling for pedagogical quality. Only one significant interaction effect in the expected direction (out of 6). No evidence of associations of instructional alignment with a composite measure of teaching effectiveness. Correlations were much smaller than expected, the design

> anticipated an increase in Rsquare of 0.10, suggesting a correlation of greater than 0.30 (ibid., 410).

Hypotheses for small effects: poor data (ruled out); other indicators of alignment, e.g. just the content that teachers covered without comparison to standards or assessment were ruled out; limited variation in either the independent or dependent variables leading to attenuated correlations; comparison to the larger MET study; there was indeed reduced variance. "Overall the results are disappointing" (411). Authors find it hard to accept that there would not be a fair OTL effect, "given the volume of literature that links OTL to achievement" (414) Conclusion could be that the decades of previous research have not identified what really matters in instruction. Plausible explanation: the tests were not sufficiently sensitive to detect differences in the quality and content of instruction; but that would be a troubling interpretation. The results suggest challenges to the effective use of VAM data. "It is essential that the research community develops a better understanding of how state tests reflect differences in instructional content and quality" (414).

OECD (2014) ***PISA 2012 Results: What students know and can do. Student performance in mathematics, reading and science. Volume 1. Paris: OECD Publishing. Chapter 3: Measuring opportunities to learn mathematics.***

The way OTL was Measured in PISA 2012

The data were collected by means of the student questionnaire, to cover both the content and time aspects of students' opportunity to learn.

> Four of the questions focused on the degree to which students encountered various types of mathematics problems or tasks during their schooling, which all form part of the PISA mathematics framework and assessment. Some of the tasks included in those questions involved formal mathematics content, such as solving an equation or calculating the volume of a box. Others involved using mathematics in a real-world applied context. Still another type of task required using mathematics in its own context, such as using geometric theorems to determine the height of a pyramid (see Question 5 at the end of this chapter). The last type of task involved formal mathematics, but situated in a word problem like those typically found in textbooks where it is obvious to students what mathematics knowledge and skills are needed to solve them. Students were asked to indicate how frequently they encountered similar tasks in their mathematics lessons using a four-point scale: never, rarely, sometimes, or frequently.
>
> In another question, students were asked how familiar they were with certain formal mathematics content, including such topics as quadratic functions, radicals and the cosine of an angle Responses to these tasks were recorded on a five-point scale indicating the degree to which students had heard of the topic. Having heard of a topic more often was assumed to reflect a greater degree of opportunity to learn. In addition, a question asked students to indicate, on a four-point scale, how frequently they had been taught to solve eight specific mathematics tasks. These tasks included both formal and applied

> mathematics. All but the last question were used to create three indices: "formal mathematics", "word problems", and "applied mathematics". Values of these indices range from 0 to 3, indicating the degree of exposure to opportunity to learn, with 0 corresponding to no exposure and 3 to frequent exposure. When interpreting the data, it needs to be borne in mind that the 15-year-olds assessed by PISA are, in some countries, dispersed over a range of grades and mathematical programmes and will therefore be exposed to a range of mathematical content (p. 146).

On the basis of the student response on the 6 items, three opportunity to learn indices were constructed. The index of formal mathematics was computed as the average of three scales. One scale was based on the degree to which students said they had heard of a particular topic in formal mathematics, e.g. "radicals"; a second scale was based on the frequency codes of being exposed to exponential functions, quadratic functions and linear equations as an indicator of "familiarity with algebra". The third scale derived from the item where students indicated how often they had been confronted (in their lessons and in tests) with problems defined as formal mathematics.

Results

On average, 15-year-olds in OECD countries indicated that they encounter applied mathematics tasks and word problems "sometimes" and formal mathematics tasks somewhat less frequently.

To examine the overall relationship between opportunity to learn and achievement, a three-level model was fitted to the data showing that at all three levels—country, school and student—there was a statistically significant relationship between opportunity to learn and student performance. Therefore, examinations of the relationship between opportunity to learn and achievement can be made at student, school and country levels simultaneously (150).

In the sequel of this summary only the results of opportunity to learn formal mathematics will be considered. It appeared that the relationship between OTL in formal mathematics and mathematics achievement was linear, and the relationship was positive and significant at student, school and country levels.

> Within each country the relationship between opportunity to learn and performance can be observed at both the school and student levels. These relationships were analyzed using a two-level model. Of the 64 countries and economies that participated in PISA 2012 with available data for the index of opportunity to learn formal mathematics, all but Albania and Liechtenstein show a positive and statistically significant relationship between exposure to formal mathematics and performance at both the student and school levels (Figure I.3.3). Among the OECD countries, the average impact of the degree of exposure to algebra and geometry topics on performance is around 50 points at the student level (i.e. increase in PISA mathematics score associated with one unit increase in the index of exposure to formal mathematics). (150)

The OECD report concludes that: "It is noteworthy that in the high-performing East Asian countries and economies on the PISA assessment—Shanghai-China, Singapore, Hong Kong-China, Chinese Taipei, Korea, Macao-China and Japan—the exposure to formal mathematics is significantly stronger than in the remaining PISA participating countries and economies (2.1 vs. 1.7)." (155) and "At the student level, the estimated effect of a greater degree of familiarity with such content on performance is almost 50 points. The results could indicate that students exposed to advanced mathematics content are also good at applying that content to PISA tasks. Alternatively, the results could indicate that high-performing students attend mathematics classes that offer more advanced mathematics content. Exposure to word problems, which are usually designed by textbook writers as applications of mathematics, are also related to performance, but not as strongly" (155). In terms of policy relevance of these findings, the report concludes that the findings suggest that policy makers can learn through PISA how their decisions about curricula are ultimately reflected in student performance.

Comments

Basically students were asked about specific mathematics content, how frequently they had been confronted with that content in their lessons and their tests. It should be noted that content was related to their regular tests at school and not to the PISA mathematics test. In the summary above only the results concerning formal mathematics were discussed, and this was the facet of OTL that correlated highest with achievement (associations for word problems and applied mathematics were smaller). From the way the results are reported it appears that the relationship between OTL in formal mathematics and mathematics achievement were not adjusted for student background characteristics; so the associations should be considered as raw correlations. This means that the effect sizes are probably inflated. Still, when comparing the OTL effects to other malleable variables in PISA, the effects compare quite favorably, both in effect size and number of countries in which the relationship between OTL and mathematics achievement is significant (Scheerens 2016).

Note: for a further treatment of these results from PISA, see the secondary analyses, described in Chap. 5 of this book.

Schmidt, W.H., Burroughs, N.A., Zoido, P., and Houang, R.T. (2015) The Role of schooling in perpetuating educational inequality: An international perspective. Educational Researcher Vol XX, No. X, pages 1–16.[1] http://er.aera.net

[1]This study is a further analysis of the overall effect of OTL in PISA 2012 (OECD 2014) as described in the previous project summary.

General Description

Schmidt et al. (2015) carried out a secondary analysis on the PISA 2012 data set, in which they focused on OTL effects in relationship to socio economic status (SES). The study yields international comparative results on the effect of OTL and SES on mathematical literacy performance as measured in PISA. Analyses were carried out at country, between school and within school level. The main focus of analysis is the way OTL and SES jointly influence student achievement. This joint effect was analyzed in terms of interaction effects (random effects regression model) and in terms of direct and indirect effects (path analysis). Indirect effects are interpreted in the sense of OTL mediating the SES effect. The total SES effect (at country, between school and within school level) was expressed as the sum of the direct and indirect effect of SES, where the indirect SES effect expresses scak the mediating influence of OTL. The main substantive issue is the hypothesis, supported by earlier research that is cited in the article, that the influence of SES on performance is "boosted" by the effect of less rigorous OTL provision for low SES students. In other words socio economic inequalities are enforced by OTL inequalities.

Research Questions

The research questions addressed in this study are the following ones:

1. *What is the joint relationship of SES and OTL to PISA literacy at both the between- and within-school levels*?
2. *What is the relationship of both between- and within school inequalities in OTL to the corresponding between- and within-school inequalities in SES and how those inequalities relate to differences in achievement*?
3. *To what degree does content coverage (OTL) function as a meditator of SES in its relationship to achievement*?

Data and OTL Measurement

The 2012 version of PISA surveyed students about the intensity of their exposure to selected mathematics topics. Data from 33 OECD and 29 non-OECD, were used in the analyses. The mathematics domain that was concentrated on was formal mathematics.

> Two separate scales were constructed using the item asking for the degree of the student's familiarity with 7 of the 13 mathematics content areas (Question 2). The five response categories reflecting the degree to which they had heard of the topic were scaled 0 to 4 with 0 representing "never heard of it" 4 representing they "knew it well". [As stated in another

section of the report, "having heard of a topic more often was assumed to reflect a greater degree of opportunity to learn" (OECD 2014, p. 146).] The frequency codes for the three topics—exponential functions, quadratic functions, and linear equations—were averaged to define familiarity with algebra. Similarly, the average of four topics defined a geometry scale, including vectors, polygons, congruent figures, and cosines. The third scale was derived from the item where students indicated how often they had been confronted with problems defined as formal mathematics (Question 4). The frequency categories were coded as "frequently", "sometimes", and "rarely" equaling 1 and "never" equal to 0, resulting in a dichotomous variable. The algebra, geometry and formal mathematics tasks were averaged to form the index "formal mathematics", which ranged in values from 0 to 3, similar to the other three indices. (OECD 2014, pp. 172–173) (ibid., p 2)

Main Results

Using a three-level model with individual SES and OTL centered on the school mean and school average SES and OTL centered on the country mean, we found that student-, school-, and country-level indicators of SES and OTL had a statistically significant relationship to student mathematics performance on the PISA scale. The inclusion of both variables into a single model reduced the size of the student-level SES coefficient by 32 %, but the positive coefficient for the student-level OTL variable was essentially the same being reduced by only 5 %.

Similarly, two-level (school and individual) analyses within each country indicated that these relationships were quite consistent across PISA participants. In the full model, SES continued to have a statistically significant relationship with student mathematics outcomes in most countries (57 of 62), whereas OTL was significant in all but one (Sweden). In comparison with the SES-only model, the size of student- and school-level SES coefficients was reduced for 62 countries, with an average of about one third in the OECD (ibid., 4).

The hypothesis that SES and OTL have an interactive effect on student mathematics literacy received partial support at the student level, with a statistically significant relationship in the full model. The interaction effects were stronger at the school than at the individual level (ibid., 4).

A second strand of analyses in this study was based on the definition of SES and OTL gaps (average differences between the top and bottom quartiles) defined and analyzed at country, between schools and within schools levels). The results showed that Substantial SES within school achievement gaps existed in most countries, with an average 44-point difference in PISA scores in OECD countries (approaching half a standard deviation), ranging from New Zealand's 74 points to Slovenia's mere 7 points. (ibid., p. 6). Within-school achievement gaps appeared to be smaller than between-school gaps in every country but Sweden and Finland but were still appreciable, with an average of 42 % the size of the mean between-school gap of 105 points (ibid., 6/7) The data also showed similar inequalities in OTL, with a 0.27 average OECD difference in OTL within school and 0.46 between schools, with substantial variation across countries.

In a next series of analyses OTL gaps were used as the independent variable and SES achievement gaps as the dependent variable. The results showed that countries with larger average differences in OTL between high- and low-income students within the same school tend to have larger average differences in performance. The results of the pooled within country analyses suggested that a one-unit increase in a school's OTL gap is associated with a 31-point increase in the SES achievement gap (about a third of a standard deviation).

"Turning to the ordinary least squares regressions for each country separately, the association between within-school OTL and within-school performance inequalities is positive for every OECD country (and 59 of the 62 in the PISA sample) and statistically significant for 23 of the 33 OECD countries and 20 of the 28 non-OECD systems" (ibid., 6). A notable result was also the finding that once the OTL gap was controlled for, variables like tracking and streaming did no longer show a significant effect. The authors conclude that OTL accounts for most of the tracking effects.

As a third strand of analysis path-analysis was used to test a conceptual model, in which SES and OTL are the main independent variables, including direct effects of these two variables and an indirect effect, which pictures OTL as a mediator of SES.

The estimated path coefficients among SES, OTL, and performance as hypothesized by the model were statistically significant at the 0.05 level across 32 of 33 OECD countries. The magnitude of the direct relationship between OTL and PISA performance controlling for SES had an OECD average of 60 points on the PISA mathematics scale, suggesting an effect size of three fifths of a standard deviation. Three countries stood out for their somewhat extreme values. Korea and Japan had estimated values of around 100 (a full standard deviation), and Sweden's estimated path coefficient was only 5 (5 % of a standard deviation).

The estimated total effect of SES on PISA performance for each country, further subdivided by the relative contribution that is indirect versus direct, showed a large variation across the 33 countries, with an average total SES effect size of 39, ranging from 19 in Mexico to 58 in France. The average proportion of that total effect that was attributable to the indirect effect was one third but varied appreciably, ranging from 1 % in Sweden to 58 % in the Netherlands (ibid., 8).

The size of the indirect effect of SES in many countries appeared to be a relatively large contributor to the total SES effect. The authors conclude that this suggests 'that the perceived role of schooling as the "great equalizer" may well be a myth and that the reality is better characterized as the "exacerbater." And they point at Australia, Korea, and the Netherlands, where over one half of the total estimated effect contributed by the relationship of SES to performance is mediated by OTL (ibid., 9).

Within-school relationships showed statistically significant associations along all three paths for nearly all countries, with an OECD average OTL effect of 45 and an SES total effect of 19 (13 direct, 6 indirect). SES was positively related to OTL in all 62 educational systems in the PISA sample. As with the overall results, the SES-OTL relationship accounted for roughly a third of the total SES effect on

performance, with sizable variation across countries, ranging from the indirect effect accounting for nearly three quarters in (Japan) to less than 10 % (Iceland and Sweden).

The results of the path analysis for differentiation *between schools* again showed positive and statistically significant but highly variable relationships among SES, OTL, and student performance Among OECD countries, there was a large positive average total SES effect (100), with indirect effects constituting a large share of that relationship (average 43 points in the OECD, statistically significant in 29 of 33 systems). OTL was also strongly related to PISA performance (average coefficient of 90, statistically significant in 29 countries). In all 62 PISA educational systems, SES is significantly related to OTL (OECD average 0.44). Between-school relationships varied appreciably across countries, with the indirect effect related to OTL having the strongest relationship in Korea, the Netherlands, and Japan and the smallest in Sweden and Estonia. Further, the relationships between SES and OTL were greatest in a group of European countries: the Netherlands, Austria, Switzerland, and Germany.

In the Netherlands and Korea, all of the very large total SES effect was derived from the SES-OTL relationship, together with the largest effect sizes for OTL at the between-school level (p. 10)

The study concludes that: (a) OTL has a strong direct relationship to student achievement, (b) high-SES students tend to receive more rigorous OTL, and (c) a substantial share of the total relationship of SES to literacy occurs through its association with OTL. The authors say that the implication of these findings is that any serious effort to reduce educational inequalities must address unequal content coverage within schools.

Conclusions

The evidence from meta-studies that reviewed OTL effects appears to be less solid than was expected, given the relatively high expectations about OTL effects expressed by various authors, like Porter, Schmidt and Polikoff. The number of meta-analyses was limited, and further analyses revealed that not all meta-studies listed as such were independent from one another (e.g. Marzano 2003 largely quoting the results presented by Scheerens and Bosker 1997). Leaving out the outlying results from Marzano, the OTL effect-size (in terms of the *d*-coefficient) compares to other relatively strong effectiveness enhancing conditions at school level, at about 0.30. A sophisticated recent study (Polikoff et al. 2014) suggests that effect sizes may be lower when adjustments are made for other variables.

The review of 7 illustrative studies showed considerable diversity in the way OTL is measured. An important difference, introduced in Chap. 1, is between studies that associate an empirical measure of exposure to content to achievement, as compared to studies that related an alignment index to achievement (as was the main emphasis in the studies by Porter et al., and Polikoff et al. 2014). The results

from PISA 2012 are striking, in the sense that OTL effects are higher, and more generalizable across countries than any of the other school/teaching variables that are usually analyzed as background variables in PISA.

References

Au, W. (2007). High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis *Educational Researcher, 36*(5), 258–267.

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1983). Effects of coaching programs on achievement test performance. *Review of Educational Research, 53,*571–586.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., et al. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*, 133–180.

Bruggencate, C. G., Pelgrum, W. J., & Plomp, T. (1986). First results of the Second IEA Science Study in the Netherlands. In W. J. Nijhof & E. Warries (Eds.), *Outcomes of education and training*. Lisse: Swets & Zeitlinger.

Creemers, B. P. M. (1994). *The effective classroom*. London: Cassell.

Creemers, B. P. M., and Kyriakydes, L. (2008). *The dynamics of eduational effectieness.* London and New York: Routledge.

De Haan, D. M. (1992). *Measuring test-curriculum overlap*. Enschede: Universiteit Twente (Dissertatie).

Farnsworth, T. (2013). Effects of targeted test preparation on scores of two tests of oral English as a second language. *Tesol Quarterly, 47*, 1.

Gamoran, A., Porter, A. C., Smithoins, J., & White, P. A. (1997). Upgrading high school mathematics instruction: improving learning opportunities for low-achieving, low-income youth. *Educational Evaluation and Policy Analysis, 19*, 325–338.

Hattie, J. (2009). *Visible learning*. Abingdon: Routledge.

Horn, A., & Walberg, H. J. (1984). Achievement and interest as a function of quantity and quality of instruction. *Journal of Educational Research, 77*, 227–237.

Husen, T. (1967). *International study of achievement in mathematics: A comparison of twelve countries*. New York: Wiley.

Kablan, Z., Toblan, B., & Erkan, B. (2013). The effectiveness level of material use in classroom instruction: A meta-analysis Study. *Educational Sciences: Theory & Practice, 13*(3), 1638–1644.

Koretz, D. M., McCaffrey, D. F., & Hamilton, L. S. (2001). Towards a framework for validating under high-stakes conditions. CSE Technical Report, 551. (CRESST) http://www.cse.ucla.edu/CRESST/reports?TR551.pdf

Kyriakides, L., Christoforou, C., & Charalambous, C. I. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teacher and Teacher Education, 36*(2013), 143–152.

Marzano R. J. (2003). *What works in schools. Translating research into action.* Alexandria, VA: Association for Supervision and Curriculum Development.

Messick, S. (1982). Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing practice. *Educational Psychologist, 17*, 67–91.

OECD. (2014). *PISA 2012 results: What students know and can do*. Volume 1. Paris: OECD.

Pelgrum, W. J., Eggen, T. J. H. M., & Plomp, T. (1983). *The second mathematics study: Results.* Enschede: Twente University.

Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis, 36*, 399–416.

Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). *Common Core Standards: The New U.S. Intended Curriculum*. *Educational Researcher, 40*(3), 103–116

Scheerens, J. (2016). *Educational Effectiveness and Ineffectiveness. A critical review of the knowledge base* (p. 389). Dordrecht, Heidelberg, London, New York: Springer. http://www.springer.com/gp/book/9789401774574

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford: Elsevier Science Ltd.

Scheerens, J., Luyten, H., Steen, R., & Luyten-de Thouars, Y. (2007). *Review and meta-analyses of school and teaching effectiveness*. Enschede: University of Twente, Department of Educational Organisation and Management.

Schmidt, W. H., McKnight, C. C., Houang, R. T., Wiley, D. E., Cogan, L. S., & Wolfe, R. G. (2001). *Why schools matter. A cross-national comparison of curriculum and learning*. San Francisco: Jossey-Bass.

Schmidt, W. B., Cogan, L. S., Houang, R. T., & MCKnight, C. C. (2011, May). Content coverage across countries/states: A persisting challenge for US educational policy. *American Journal of Education, 117*, 399–427.

Schmidt, W. H., Burroughs, N. A., Zoido, P., & Houang, R. T. (2015). The Role of schooling in perpetuating educational inequality: An international perspective. *Educational Researcher, 44*, 1–16. http://er.aera.net

Schroeder, C. M., Scott, T., Tolson, H., Huang, Tse-Yang, & Lee, Yi-Hsuan. (2007). A meta-analysis of national research: Effects of teaching strategies on student achievement in science in the United States. *Journal of research in science teaching., 44*(10), 1436–1460.

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: the role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454–499.

Seidel, T., & Steen, R. (2005). The indicators on teaching and learning compared to the review of recent research articles on school and instructional effectiveness. In J. Scheerens, T. Seidel, B. Witziers, M. Hendriks, & G. Doornekamp (Eds.), *Positioning the supervision frameworks for primary and secondary education of the Dutch Educational Inspectorate in current educational discourse and validating core indicators against the knowledge base of educational effectiveness research*. Enschede: University of Twente; Kiel: Institute for Science Education (IPN).

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning, 60*(2), 263–308.

Sturman, L. (2003). Test prepatation: Valid and valuable, or wasteful? *The Journal of the Imagination of Language Learning, 9*, 31–37.

Xie, Q. (2013). Does test preparation work? Implications for score validity. *Language Assessment Quarterly, 10*(2), 196–218. doi:10.1080/15434303.2012.721423

Chapter 4
Review and "Vote Count" Analysis of OTL-Effect Studies

Marloes Lamain, Jaap Scheerens and Peter Noort

Abstract In the fourth chapter an overview is given of 51 primary studies, conducted during the last twenty years. Schematic summary descriptions are put together in a large summary table. Although quantitative meta-analysis of these studies was beyond the scope of this review study, some basic summary tables were produced to provide an overall orientation on how OTL had been researched and what can be concluded about its effectiveness. It is concluded that the vote count measure of OTL, (i.e. the percentage of effect sizes that were statistically significant and positive) established in this study, and which was 44 %, is of comparable size to other effectiveness enhancing conditions like achievement orientation, learning time and parental involvement, but dramatically higher than vote count measures for variables like cooperation and educational leadership. What should be considered is that vote counting is a rather crude procedure and that comparison of quantitative effect sizes is more informative.

Introduction

In this chapter 51 empirical studies, in which opportunity to learn was quantitatively associated with student performance, are schematically summarized. Results are synthesized by means of a so called "vote count" procedure, where studies are categorized for showing positive or negative statistically significant or insignificant associations between OTL and student performance. Such analyses provide a crude impression about the empirical support for the assumption that opportunity to learn is positively associated with student performance in a meaningful way. The

M. Lamain · J. Scheerens (✉) · P. Noort
University of Twente, Enschede, The Netherlands
e-mail: j.scheerens@utwente.nl

J. Scheerens
Oberon, Utrecht, The Netherlands

© The Author(s) 2017

J. Scheerens (ed.), *Opportunity to Learn, Curriculum Alignment and Test Preparation*, SpringerBriefs in Education,
DOI 10.1007/978-3-319-43110-9_4

advantage of this way of synthesizing research outcomes is that they allow for comparison with similar results for other independent variables in schooling and teaching that are expected to enhance effectiveness. As such we shall provide comparative data based on earlier studies with respect to: evaluation, achievement orientation, learning time, parental involvement, staff consensus and cohesion and educational leadership.

Methods

Identification and Collection of Studies

The identification of studies was carried out by means of explicit selection and inclusion criteria and the choice of relevant data bases.

We were interested in selecting primary empirical research studies as well as meta-analyses, review-studies, best-evidence syntheses and research syntheses. The following inclusion criteria were used:

- OTL as independent variable.
- Cognitive achievement in basic school subjects like vernacular (mother tongue) language, mathematics, arithmetic and science, preferably adjusted for relevant student background characteristics, as dependent variable(s) of the study.
- Studies carried out in primary and secondary education (age groups 6–18).
- Studies carried out in regular education (excluding special education).
- Studies reported in Dutch, English or German.
- Research reports containing sufficient quantitative information to potentially calculate effect sizes (like correlations or *d*-coefficients). In the case of meta-analyses we required at least mean effects-sizes across the analyzed primary studies.
- Studies that were published between 1995 and 2015.

The following data bases were used: Eric, PsychARTICALS, PsychBOOKS, Psych INFO and Psychology and Behavioral Sciences collection. In the Annex to this chapter further details are provided about the search that was conducted. In addition to this systematic search of data-bases, the reference lists of a limited number of key articles (see Chap. 3) were scrutinized for relevant studies as well.

Description Categories for Basic Description of the Identified Studies

The core of this chapter is the schematic summary description of the 51 studies that were identified; see the next section. The description categories address the focus of

the study, the definition of the OTL variable that was used, the number of respondents, the dependent variable, the effect measures that were computed, and a category for additional comments. In the *Focus* category the general orientation of the study is summarized, next, an indication is given of the study being (quasi) experimental or correlational. The category for the *OTL measure*, describes from which respondents (e.g. teachers, students) data were collected and by means of which method (e.g. questionnaires or content analyses of relevant documents). *Respondents* are identified as far as the measurement of the independent and dependent variables are concerned; numbers of respondents are included. The *Dependent variable*(*s*) mention the type of student achievement measures that were used, identified according to subject matter category. The *Comment* category is focused at facets of the methodological quality of the study, particularly with respect to adjustment of the achievement results for student background characteristics.

Results

The Identified Studies

From an initial total of 6006 studies, 51 met the inclusion criteria. Subsequent tables provide an overview of the following study characteristics: the year of publication, the nationality (including a category for international studies), and the subject matter area that was addressed in the OTL and student achievement measures (Tables 4.1, 4.2, 4.3, 4.4 and 4.5).

Table 4.1 Number of studies per time interval

Year of publication	Number of articles
Before 1995	3
1995–2000	5
2000–2005	3
2005–2010	16
2010–2015	24
Total	51

Table 4.2 Number of studies by geographical area

Article conducted in….	Table 1
North-America	37
Europe	2
Africa	4
South-America	2
Meta-analyses/reviews excluded	6

Table 4.3 Number of articles per age category

Grade	Number of articles
Kindergarten	3
First grade	2
Second grade	2
Third grade	2
Fourth grade	3
Fifth grade	4
Sixth grade	8
Seventh grade	3
Eighth grade	20
Ninth grade	1
Tenth grade	1
Eleventh grade	0
Twelfth grade	2
Unknown	1

Table 4.4 Number of articles by subject matter area

Subject	Number of articles
Mathematics (only Algebra)	39 (4)
English Language Arts (only reading/only writing)	10 (5/1)
Science	3
History	1

Table 4.5 Number of articles per control variable

Control variables	Number of articles
Prior achievement	6
General ability	1
SES	4
Prior achievement and FRL	5
Prior achievement and SES	15
Prior achievement and parental education	2
Prior achievement and family income	2
Parental education and FRL	1
None	9
Inapplicable	5
Unknown	1

The results indicate that the large majority of studies were conducted between 2005 and 2015, that no less than 37 of the 51 studies were carried out in the USA, that studies in which eighth grade students were the respondents predominated (20

out of 51), and that most studies addressed mathematics, with language as the second important area. Interestingly, the number of studies that addressed OTL effects increased over time, with almost half of the identified studies being conducted between 2010 and 2015.

Schematic Summary Description

In Table 4.6 the 50 studies that were selected are described according to the categories defined in the methods section.

Table 4.6 Schematic description of studies

Aguirre-Muñoz and Boscardin (2008)	Focus	This investigation examined the impact of opportunity to learn content and skills targeted by a writing assessment on the achievement of English learners (ELs), including the potential for differential impact of increased exposure to literary analysis and writing instruction (p. 186). This study is characterized as correlational research and is conducted in the USA
	OTL measure	The six OTL constructs measured by the teacher survey that were included in the validity analysis were expertise, literary analysis content coverage, writing content coverage, classroom processes, assessment practices, and LAPA (Language Arts Performance Assignment) preparation (p. 191)
	Respondents	Teachers (N = 27) and 6th grade students (N = 1.038)
	Dependent variable	Language arts performance assessment score
	Effect size	The ordinal logistic hierarchical linear modelling analyses indicated that **of the nine OTL variables, only two showed significant effects on student performance**: literacy analysis coverage ($\beta = 0.46$, $p = 0.03$) and writing coverage ($\beta = 0.54$, $p = 0.02$) (pp. 196/197)
	Comments	Results controlled for all of the predictors at the grand mean level, including gender and language proficiency status (p. 196)
Au (2007)	Focus	This study analyses 49 qualitative studies to interrogate how high-stakes testing affects curriculum, defined here as embodying content, knowledge form, and pedagogy (p. 258). See Chap. 3: Meta-analyses

(continued)

Table 4.6 (continued)

Boscardin et al. (2005)	Focus	Examination of the impact of OTL variables on student performance on English and algebra assessments. This study showed that content coverage was positively correlated with student performance in English and algebra (p. 307). This is a correlational study conducted in the USA
	OTL measure	The findings were based on the Teacher OTL Survey developed by the UCLA National Center for Research on Evaluation, Standards, and Student Testing (CRESST) in collaboration with content experts. The survey had five major sections corresponding to critical aspects of OTL: (a) teaching experience, (b) teacher expertise in content topics, (c) topic coverage, (d) classroom activities, and (e) assessment strategies and preparation
	Respondents	Students (N = 4.715 English and N = 4.724 Algebra) and Teachers (N = 118 English and N = 124 Algebra)
	Dependent variable	English and Algebra achievement
	Effect size	**Initial analyses revealed that of the five OTL variables, only two were significant predicators**. The other variables were not considered in the final model. The proportion of variance accounted for in the final model by the OTL variables (teacher expertise in content topics and topic coverage) was 0.28 for the algebra test and 0.35 for the English test. This means that including measures of teacher expertise, content coverage, and mean student SES reduced the variance by 28 % for the algebra test and 35 % for the English test (p. 324)
	Comments	Results controlled for differences in the average SES of the students and individual differences among students, like initial course grades
Cai et al. (2011)	Focus	The impact of a standards-based or reform mathematics curriculum (called CMP) and traditional mathematics curricula (called non-CMP) on students' learning of algebra using various outcome measures (p. 117). This is an experimental study conducted in the USA
	OTL measure	Influence of differences in the ways that the CMP curriculum and the non-CMP curriculum define variables, define equation solving, introduce equation solving and use mathematical problems to develop algebraic thinking on algebra achievement. Conceptual and procedural emphases as a classroom variable is used to examine the impact of curriculum on students' learning
	Respondents	6th–8th grade students (N = 1.284)
	Dependent variable	Algebra achievement

(continued)

Table 4.6 (continued)

	Effect size	Findings suggest that the use of the CMP curriculum is associated with a significantly greater gain in conceptual understanding than is associated with the use of non-CMP curricula. So, **in favour of the CMP curriculum, there is a positive significant treatment effect**
	Comments	This article particularly examines students' achievement gains across the three middle school years while controlling for the conceptual and procedural emphases in classroom instruction (p. 120). Results are controlled for prior achievement
Calhoon and Petscher (2013)	Focus	Examination of group- and individual-level responses by struggling adolescent readers to three different modalities of the same reading program, Reading Achievement Multi-Component Program. The three modalities differ in the combination of reading components (phonological decoding, spelling, fluency and comprehension) that are taught and their organization. Latent change scores were used to examine changes in phonological decoding, fluency, and comprehension for each modality at the group level. Individual students were classified as gainers versus non-gainers so that characteristics of gainers and differential sensitivity to instructional modality could be investigated (p. 565). This is an experimental research conducted in the USA
	OTL measure	Effects of modality differences (alternating, integrated and additive) on reading achievement (phonological decoding, fluency, and comprehension). Eight reading measures were used, but two of the tests were not given in all three studies. Each student was tested individually in a single session that occurred within 1 week of the beginning and end of the intervention period (p. 572)
	Respondents	6th–8th grade students (N = 155)
	Dependent variable	Reading skills
	Effect size	Mean latent change scores in standardized units (z units) *Decoding*: Alternating, 0.93; Integrated, 1.09; Additive, 1.24 *Fluency*: Alternating, 0.49; Integrated, 0.83; Additive, 0.84 *Comprehension*: Alternating, 1.39; Integrated, 1.17; Additive, 1.66 **Overall, the additive modality produced significantly more gain than the other modalities** (p. 584)

(continued)

Table 4.6 (continued)

	Comments	The extent to which modality differences existed on the pre-intervention assessments were examined prior to the main analyses. When data from the three studies were concatenated, a series of ANOVAs were run to test for initial fall differences across the conditions, and a Linear Step-Up was applied to control for the false discovery rate (p. 573)
Carnoy and Arends (2012)	Focus	The purpose of the study is to test whether and how classroom and school factors contribute to student gains in mathematics learning. From a classroom perspective, the emphasis is on teacher mathematics knowledge, classroom pedagogy and opportunity to learn in of sample of grade 6 classrooms (p. 453). This is a correlational study conducted in South Africa and Botswana
	OTL measure	Total lessons on topic and total topics taught, measured by analysing the contents of the three best student notebooks each classroom
	Respondents	6th grade students (N = 5.500) and teachers (N = 126)
	Dependent variable	Mathematics achievement
	Effect size	The OTL measures are both significantly related to student learning gains in South Africa with a significance level of $p < 0.10$ instead of $p < 0.05$ (**total lessons on topic: 0.0012, $p < 0.10$; total topics taught: 0.0021, $p < 0.10$**), but not in Botswana (**total lessons on topic: 0.0000, $p > 0.10$; total topics taught: 0.0002, $p > 0.10$**) (p. 465)
	Comments	Results are controlled for initial student achievement, several teacher quality variables, characteristics of individual students and students' families, average classroom SES, observed class size and school conditions
Claessens et al. (2012)	Focus	Examination of how reading and mathematics content coverage in kindergarten is associated with the maintenance of preschool skills advantages. Results suggest that increased exposure to advanced content could help maintain preschool skill advantages while promoting the skills of children who did not attend preschool (p. 1). This is a correlational study conducted in the USA
	OTL measure	Content exposure: the influence of four measures of kindergarten academic content—basic mathematics, advanced mathematics, basic reading and advanced reading—on math and reading achievement test scores. Teachers were surveyed about classroom activities and content. Distinction is being made between children who attended Center Care, children who went to a funded childhood program, like Head Start, and children with Other Care

(continued)

Table 4.6 (continued)

	Respondents	Kindergarten students (N = 17.981) and teachers (N = 3.038)
	Dependent variable	Mathematics and reading achievement
	Effect size	Coefficients from regressions predicting spring mathematics/reading achievement with mathematics/reading content measures by children's preschool experience: Center Care: Basic math (−0.0463, $p < 0.01$), Advanced math (0.066, $p < 0.01$), Basic reading (−0.0192, $p < 0.05$) and Advanced reading (0.0611, $p < 0.01$) (**4/4 significant**) Head Start: Basic math (−0.0215, $p > 0.1$), Advanced math (0.0416, $p < 0.01$), Basic reading (0.00424, $p > 0.1$) and Advanced reading (0.0251, $p < 0.1$) (**2/4 significant**) Other Care: Basic math (−0.0446, $p < 0.01$), Advanced math (0.0531, $p < 0.01$), Basic reading (−0.0086, $p > 0.1$) and Advanced reading (0.0530, $p < 0.01$) (**3/4 significant**) (pp. 41–42)
	Comments	The OTL measure is based on specific content categories Analyses control for observable characteristics of teachers and classrooms, a variety of child characteristics and home environment factors that might be correlated with both content measures and student initial achievement (pp. 15–17) Effect net of co-variables and other independent variables
Cogan et al. (2001)	Focus	This article examines the range of eighth-grade mathematics learning opportunities in the U.S. drawing on data gathered for the Third International Mathematics and Science Study (TIMSS). Comparison of students' learning opportunities includes consideration of the specific course in which they were enrolled, the type of textbook employed for the course, and the proportion of time teachers devoted to teaching specific topics (p. 1). This is a correlational study conducted in the USA
	OTL measure	Variation in 8th grade mathematics in the USA: the opportunity students get to study mathematics. The OTL variables are topic and course-text difficulty. Their influence on mathematics score is being measured by students surveys and teacher questionnaires
	Respondents	8th grade students (N = over 13.000) and their teachers
	Dependent variable	Mathematics achievement

(continued)

Table 4.6 (continued)

	Effect size	**Class' topic difficulty and course-text challenge are both significant predictors explaining nearly 40 % of the variance in mathematics score across classrooms. Topic coverage has a coefficient of 23.2 with $p < 0.001$ and Course Challenge 13.8 with $p < 0.001$**. Classes exposed to more challenging topics tended to have higher TIMSS scores—on average, 23 points higher for every year increase in the class' international topic difficulty. Each increase in a class's course-text challenge rank was associated with nearly a 14 points increase on TIMSS score (p. 20)
	Comments	Results controlled for school's location (urban, rural or suburban), size and percent of minority enrolment. Results are not controlled for initial achievement scores because the study was cross-sectional
Cueto et al. (2014)	Focus	This paper explores the relationship between SES measured at age one, OTL and achievement in mathematics ten years later. (p. 50). This is a correlational study conducted in Peru
	OTL measure	Four OTL variables were measured: hours of class per year, curriculum coverage, quality of teachers' feedback and level of cognitive demands. Each variable was measured based on exercises found in the notebooks and workbooks, except hours of class per year, which was reported by the head teacher (p. 53)
	Respondents	4th grade students (N = 104)
	Dependent variable	Mathematics achievement
	Effect size	Curriculum coverage (0.44) and the level of cognitive demand (0.29) indicate a positive and significant association with achievement in mathematics. **With several covariates added, only curriculum coverage remained as a significant predictor of achievement**
	Comments	Net effect of each variable, results controlled for several covariates, like students' prior abilities
Cueto et al. (2006)	Focus	Opportunities to learn mathematics of sixth grade students from 22 public schools in Lima, Peru. Where OTL is defined as curriculum coverage, cognitive demand of the tasks posed to the students, percent of mathematical exercises that were correct and quality of feedback. OTL is positively associated with achievement (p. 25). This is a correlational study conducted in Peru

(continued)

Table 4.6 (continued)

	OTL measure	Curriculum coverage, cognitive demand of the tasks posed to the students, percent of mathematical exercises that were correct and quality of feedback. These variables were coded in the workbooks and notebooks of the students, which were gathered at the end of the school year (p. 25)
	Respondents	Sixth grade students (N = 369)
	Dependent variable	Mathematics achievement
	Effect size	**Two of the three variables are positively and significantly related to achievement**, namely cognitive demands and adequate feedback. When the three OTL variables were included in a factor analysis with varimax rotation one factor resulted, which accounted for 68 % of the total variance (p. 43)
	Comments	Different independent variables next to the OTL variables, i.e. gender, age, attendance rate during the school year, whether math is the preferred subject for the student, number of persons living at home, the SES score and type of school. Results are not controlled for students' initial achievement level
D'agostino et al. (2007)	Focus	The purpose of this study was to examine the instructional sensitivity of Arizona's fifth-grade mathematics standards-based assessment (p. 6). For this study, a new method was developed for capturing the alignment between how teachers bring standards to life in their classrooms and how the standards are defined on a test (p. 1). This study is characterized as correlational research and is conducted in the USA
	OTL measure	Two curriculum experts judged the alignment between how teachers brought the objectives to life in their classrooms and how the objectives were operationalized on the state test (p. 1). Achievement was measured by the fifth-grade mathematics AIMS (Arizona Instrument to Measure Standards). The AIMS math test was designed to measure the Arizona Academic Mathematics standards for Grades 3 through 5. At that time, the standards consisted of six strands: (a) number sense, (b) data analysis and probability, (c) patterns, algebra, and functions, (d) geometry, (e) measurement and discrete mathematics, and (f) mathematical structure/logic (p. 9)
	Respondents	5th grade students (N = 1.003) and teachers (N = 52)
	Dependent variable	Mathematics achievement

(continued)

Table 4.6 (continued)

	Effect size	To interpret the magnitude of the effect, one can consider a teacher who is one standard deviation above the mean on Alignment and on the Emphasis × Alignment interaction variable. On average, students in the teacher's classroom would be expected to score about 11 scale score points [5.17 points for Alignment (***p*** < **0.05**) and 5.75 points for the interaction (***p*** < **0.05**)] higher than students, on average, in classrooms at the grand mean of both predictors, which was about a one-fifth standard deviation difference (from Table 4.1, the standard deviation for the outcome was 55.46). Notice that Emphasis alone has a negative coefficient, −2.00 (***p*** > **0.05**) (p. 17)
	Comments	Results controlled for initial achievement differences between classrooms and student socioeconomic status
Desimone et al. (2013)	Focus	This study examines relationships between teachers' participation in professional development and changes in instruction, and between instruction and student achievement growth (p. 4). This is a correlational study conducted in the USA
	OTL measure	Minutes per day spent on mathematics, topic focus (basic math and advanced math) and cognitive demands (memorize facts and solve novel problems) are the five OTL variables which are related to the initial status of students achievement and to achievement growth Reports of professional development and instruction (time spent on mathematics instruction, topic focus, type of learning required or cognitive demands) are taken from teacher's self-report surveys. Student achievement was measured by a special administration of a set of open-ended questions from the Stanford Achievement Test, Ninth Edition, which assessed problem solving and procedures (p. 6)
	Respondents	3th–5th grade students (N = 4.803) and teachers (N = 457)
	Dependent variable	Mathematics achievement

(continued)

Table 4.6 (continued)

	Effect size	Minutes per day spent on mathematics did not significantly predict either initial achievement status or growth. Increased emphasis on Memorizing Facts was associated with slower than average growth in achievement (b = −6.02, $p < 0.037$). Emphasis on Solving Novel Problems was associated with extremely modest achievement growth (b = 0.69, $p < 0.041$) (p. 30, 31). The correlation between Focus on Basic Math Topics and achievement growth is −0.042, with $p = 0.036$. For Focus on Advanced Math Topics b = 0.061, with $p = 0.043$ (p. 55) **Looking at the initial status only 1 out of the 5 variables is significant (Focus on advanced math topics). When it comes to growth 4 out of 5 are significant of which two are positively and two are negatively related**
	Comments	Results controlled for teacher, school and student characteristics, like teacher's years of experience, school enrolment and initial achievement level
Elliott (1998)	Focus	This article illuminates both the relationship between spending practices and students' achievement and the specific components of OTL in the classroom that affect students' outcomes. Moreover, it indicates how financial resources indirectly affect students' achievement by creating differential access to OTL (p. 223). This is a correlational study conducted in the USA
	OTL measure	Key links between expenditures and achievement: the effect of expenditures on teachers' effectiveness and the effect of expenditures on classroom resources (p. 226) Achievement was measured by the 10th grade IRT theta scores, a mathematical transformation of the standardized test score is designed to reflect over time
	Respondents	8th–10th grade students (N = 14.868)
	Dependent variable	Mathematics and science achievement
	Effect size	Expenditures correlate significantly with most measures of OTL. 8 out of 9 OTL variables correlate significantly with expenditures for math students (ranging from −0.126 to 0.142), and 6 out of 7 OTL variables correlate significantly with expenditures for science students (ranging from −0.139 to 0.191) **9 out of 9** OTL variables correlate significantly with students' IRT math test score (ranging from −0.059 to 0.330) and **6 out of 7** OTL variables correlate significantly with students' IRT science test score (ranging from −0.066 to 0.186)

(continued)

Table 4.6 (continued)

	Comments	Results controlled for student background characteristics (e.g. SES, racial background and gender) and school characteristics. The 8th-grade IRT theta score was controlled in all analysis such that the true outcome was actually gains in math or science achievement between the 8th and 10th grade (p. 229)
Engel et al. (2013)	Focus	This study explored the relationship between students' school-entry math skills, classroom content coverage, and end-of-kindergarten math achievement (p. 157). This is a correlational study conducted in the USA
	OTL measure	Exposure to specific mathematics content (the OTL variables: basic counting and shapes, patterns and measurement, place value and currency, and addition and subtraction) and children's early math skills, measured by teacher reports
	Respondents	Students in kindergarten (N = 11.517) and teachers (N = 2.176)
	Dependent variable	Mathematics achievement
	Effect size	Devoting additional days per month to Basic Counting and Shapes was negatively associated with the end-of-kindergarten mathematics test scores (−0.02 SD). For Patterns and Measurement there was no statistically significant association. For Place Value and Currency there was an increase (0.03 SD), and for Addition and Substraction as well (0.04 SD). **3 out of 4 OTL variables are significantly correlated to student achievement**
	Comments	Effect net of co-variables and other independent variables. Results are controlled for initial reading and math skills and cognitive ability (p. 164)
Gamoran (1987)	Focus	This paper argues that research on school effects has come to regard instruction as the core of the schooling process. Two aspects of instruction, the use of time and the coverage of curricular content, are discussed in detail (p. 1). This is a correlational study conducted in the USA
	OTL measure	Content coverage and instructional time (explanation, review and discipline time) assessed by IEA study '81/'82 through teacher reports
	Respondents	8th grade students (N > 3.995)
	Dependent variable	Mathematics achievement

(continued)

Table 4.6 (continued)

	Effect size	The effects of content coverage are statistically significant but substantively small. The coefficient of b = 0.046 means that a teacher would need to cover about 20 more items to raise achievement one point The effects of the three types of instructional time are quite small. The coefficient for review time is significant (b = 0.007, $p < 0.01$) and indicates that an increase of about 140 min per week, would be needed to raise achievement by a single point. The effect of discipline time is a little bigger (b = −0.019, $p < 0.001$), each additional 50 min weekly, reduces achievement by one point (pp. 29, 37). The coefficient for explanation time is 0.001 and not significant. **3 out of 4 OTL variables are significant**
	Comments	Organizational constraints, class characteristics, social designations and teacher and student attributes are taken into account. Results are controlled for prior achievement
Gamoran et al. (1997)	Focus	In this article, the authors evaluate the success of "transition" math courses in California and New York, which are designed to bridge the gap between elementary and college-preparatory mathematics and to provide access to more challenging and meaningful mathematics for students who enter high school with poor skills (p. 325). This study is conducted in the USA
	OTL measure	The extent to which mathematical content and cognitive demands were included in sample classes. Content coverage reflects both the proportion of instructional time that was spent covering tested content and the match of relative emphases of types of content between instruction and the test (19 content areas). Teacher questionnaires provided information on the extent to which the topics on our tests were covered in the sample classes and whether the cognitive demands made on the test also occurred in mathematics instruction (p. 330)
	Respondents	7th grade students (N = 882)
	Dependent variable	Mathematics achievement
	Effect size	More rigorous content coverage accounts for much of the achievement advantage of college-preparatory classes (p. 325). **The correlation between content coverage and instructional effects is 0.11 with $p < 0.10$** (p. 334)
	Comments	Results are controlled for prior achievement and other student characteristics

(continued)

Table 4.6 (continued)

Gau (1997)	Focus	The focus of this paper is further understanding of the distribution and the effects of an expanded conception of OTL on student mathematics achievement. In addition to descriptive statistics, a set of two-level hierarchical linear models was employed to analyse a subset of the restricted-use National Education Longitudinal Study of 1988 database. The results revealed that on different scales, various kinds of opportunities to learn mathematics are associated with student mathematics achievement, and opportunities are unequally distributed among different categories of schools (p. 3). This is a correlational study conducted in the USA
	OTL measure	Content and level of instruction (high achievement group, textbook coverage, instructional time and weekly homework). The variables are measured by students' surveys and teachers' questionnaires. This study cites resources and teachers' mathematical knowledge also as OTL variables
	Respondents	8th grade students (N = 9.702) and their teachers
	Dependent variable	Mathematics achievement
	Effect size	The results of the content and level of instruction analyses are mixed. **Three of the four OTL variables are statistically significant in a positive direction, while the other is significant but negative** (p. 15) The effects of teachers' mathematical knowledge are significant, but the effects of school mathematical resources are not significant
	Comments	Results controlled for teachers' mathematical knowledge, content and level of instruction, school mathematical resources, gender, race, SES, prior achievement, school sector, minority concentration, community type and school average student SES (pp. 3, 4)
Grouws et al. (2013)	Focus	This study examined the effect of 2 types of mathematics content organization on high school students' mathematics learning while taking account of curriculum implementation and student prior achievement. Approximately ½ of the students studied from an integrated curriculum and ½ studied from a subject-specific curriculum (p. 416). This is an experimental study conducted in the USA

(continued)

Table 4.6 (continued)

	OTL measure	Table of contents (TOC) records, three indices from TOC records capture the nature and extent of textbook use. (1) The Opportunity to Learn (OTL) index, derived from the TOC, represents the percentage of content in the textbook that students were provided an opportunity to learn. (2) The Extent of Textbook Implementation (ETI) index is a weighted average of students' opportunity to learn from their textbook. (3) A Textbook Content Taught (TCT) index, which differs from the ETI index by considering only lessons taught, thereby ignoring content that students were not given the opportunity to learn. (p. 427)
	Respondents	8th grade students (N = 2.161)
	Dependent variable	Mathematics achievement
	Effect size	The group of teachers of the subject specific curriculum have a higher mean score for OTL, the teacher of the integrated curriculum have higher mean scores for ETI and TCT (p. 440). Although the mean indices by curriculum type were not statistically different, Levene F-tests showed significantly more variability among teachers of the integrated curriculum materials than teachers of the subject-specific curriculum material for the ETI index ($p = 0.055$) and the OTL index ($p < 0.001$). However, the Levene F-test showed significantly more variation in the TCT index ($p = 0.001$) for teachers of the subject-specific curriculum (p. 437) On the Test of Common Objectives, the Problem Solving and Reasoning Test and the Iowa Test of Educational Development, students of teachers teaching from the integrated textbook outperformed students in classrooms in which the teacher taught from a subject-specific textbook, with effect sizes of 0.31, 0.45 and 0.17, respectively (p. 451). So, **in favour of the integrated textbook there is a positive significant treatment effect**
	Comments	Results controlled for covariates, a measure of prior learning and demographics like SES
Heafner and Fitchett (2015)	Focus	The authors examine National Assessment of Educational Progress in U.S. History (NAEP-USH) assessment data in order to better understand the relationship between classroom- and student-level variables associated with historical knowledge as measured in the 12th grade. Findings document that instructional exposure (OTL) is a factor associated with learning outcomes (p. 226). This is a correlational study conducted in the USA

(continued)

Table 4.6 (continued)

	OTL measure	Two categories of instructional exposure: Multimodel Instruction (based on work on group project, give presentation to the class, write a report, use books or computers in library for schoolwork, listen to information presented online, go on field trips or have outside speakers and watch movies or videos) and Text-Dependent Instruction (based on frequency of report writing, discussing material studied, reading extra material, read material from textbook, use publications of historical people, write short answer to questions). OTL variables measured by student surveys
	Respondents	12th grade students (N = 8.610)
	Dependent variable	History achievement
	Effect size	Analysis of the exposure to instruction factors indicated that for each standard deviation increase in text-dependent instruction, NAEP-USH scores increased by 8.61 points (***p* < 0.001**, SE 0.38) where the mean is 250. Conversely, each standard deviation increase in exposure to multimodel instruction was associated with a decrease of 7.48 (***p* < 0.001**, SE 0.49) (pp. 236/237)
	Comments	The OTL measure is rather global, not based on more specific content categories Effect net of co-variables and other independent variables
Herman and Abedi (2004)	Focus	Exploration of two complimentary approaches for exploring English Language Learners' (ELL) opportunity to learn Algebra 1, representing opposite ends of the cost continuum (p. 6). This is a correlational study conducted in the USA
	OTL measure	Content coverage measured by surveys of teachers and student, 28 content areas are listed. Teacher-student interactions details through observation
	Respondents	Survey study: 8th grade students (N = 602) and teachers (N = 9) Observation phase: nine classes of students (N = 271) and their teachers
	Dependent variable	Algebra achievement
	Effect size	Results suggest that OTL is a more determining factor in algebra achievement for ELL students than for the non-ELL group (p. 13) **Results show that the classroom-level OTL measure has significant effects on the outcome variable. However, after accounting for the classroom-level OTL measure, the student-level preparation/OTL factor had no significant effect** (p. 16)

(continued)

Table 4.6 (continued)

	Comments	Three multiple regression models, for all students, for ELLs and for non-ELLs. Each model is controlled for prior math ability and prior student preparation (p. 13)
Holtzman (2009)	Focus	This dissertation addresses the following questions: (1) To what extent is the content of instruction aligned with the California content standards and with the blueprint for the California Standards Test (CST)? (2) How do instruction, the standards, and the CST blueprint compare with one another in the topics covered and the levels of cognitive demand emphasized? (3) To what extent is the alignment of instruction with either the standards or the CST blueprint related to student achievement on the CST? (p. iv). The last question is addressed separately for each school-level (grades 3–6 or grades 6–8) and subject-area (ELA or maths) combinations. This is a correlational study conducted in the USA
	OTL measure	Topic coverage and cognitive demand emphases in classroom instruction; The data were from a survey of middle school teachers in San Diego City Schools (SDCS). The survey presents teachers with a list of highly detailed topics. For each of the specific topics, teachers first fill in the amount of time spent on the topic by their class during the past school year, and then indicate the proportion of the total time spent on the topic designed to help students meet expectations in each of five different categories of cognitive demand Student achievement data were provided by SDCS. Scaled scores on the CST in ELA and math for years 2002–03, 2003–04, and 2004–05 are used (pp. 41, 42) OTL variables: (1) Alignment with Standards: Overall, (2) Alignment with Standards: Topic, (3) Alignment with Standards: Cognitive Demand, (4) Alignment with CST Blueprint: Overall, (5) Alignment with Blueprint: Topic and (6) Alignment with CST blueprint: Cognitive Demand
	Respondents	Teachers (N = 724), ELA students grades 3–6 (N = 2715), Math students grades 3–6 (N = 2946), ELA students grades 6–8 (N = 1753), and Math students grades 6–8 (N = 2556)
	Dependent variable	English language arts (ELA) and mathematics achievement

(continued)

Table 4.6 (continued)

	Effect size	Elementary ELA results: **All the correlations are negative and not statistically significant** ($p < 0.05$) Elementary Math results: 4 out of 6 correlations are negative and not significant ($p < 0.05$). **Both positive correlations are statistically significant** Middle School ELA results: 4 out of 6 variables are positive of which **3 are statistically significant** ($p < 0.05$). **One of the 2 negative correlations is significant** Middle School Math results: 4 out of 6 correlations are positive of which **only one is statistically significant**. The other 2 correlations are negative and not statistically significant ($p < 0.05$) (p. 138)
	Comments	Results controlled for student prior achievement, student demographics, and teacher characteristics
Kablan et al. (2013)	Focus	The aim of this study was to combine the results obtained in independent studies aiming to determine the effectiveness of material use. The main questions of the study is: "Does material use in classroom instruction improve students' academic achievements?" 57 experimental studies are included in this meta-analysis. See Chap. 3: Meta-analyses
Kurz et al. (2014)	Focus	This study provides initial evidence supporting intended score interpretations for the purpose of assessing OTL via an online teacher log. MyiLOGS yields 5 scores related to instructional time, content and quality. Agreements between log data from teachers and independent observers were comparable to agreements reported in similar studies. Moreover, several OTL scores exhibited moderate correlations with achievement and virtually nonexistent correlations with a curricular alignment index (p. 159). This is a correlational study conducted in the USA
	OTL measure	The extent to which a teacher dedicates instructional time to cover the content prescribed by intended standards using a range of cognitive processes, instructional practices and grouping formats. MyiLOGS scores are designed to allow interpretations about time spent on academic standards, content coverage of academic standards, emphases along a range of cognitive processes, emphases along a range of instructional practices and emphases along a range of instructional grouping formats (pp. 165, 166). Each teacher received the standard professional development on the use of MyiLOGS and each teacher participant was observed at least once during his or her logging period (p. 171)

(continued)

Table 4.6 (continued)

	Respondents	General and special education teachers (N = 38) and 8th grade students (N = 56)
	Dependent variable	Mathematics and reading achievement
	Effect size	Three out of five OTL variables are significantly related to average class achievement. The correlation between the yearly summary score for Time on Standards and class achievement was **r = 0.56**, ***p* < 0.05**, accounting for about 31 % of the variance in average class achievement. The correlation between the yearly summary score for Cognitive Processes and class achievement was **r = 0.64**, ***p* < 0.05**, accounting for about 41 % of the variance in average class achievement. Last, the correlation between the yearly summary score for Grouping Formats and class achievement was **r = 0.71**, ***p* < 0.05**, accounting for about 50 % of the variance in average class achievement (p. 177)
	Comments	Results controlled for state and subject and not for students' prior achievement
Kurz et al. (2010)	Focus	Examination of the content of the planned and enacted eighth-grade mathematics curriculum for 18 general and special education teachers and the curricula's alignment to state standards via the Surveys of the Enacted Curriculum (SEC). The relation between alignment and student achievement was analyzed for three formative assessments and the corresponding state test within a school year (p. 131). This is a correlational study conducted in the USA
	OTL measure	Measurement of students' OTL the enacted curriculum and qualification of the alignment of the enacted curriculum to state standards by using SEC as traditional end of year surveys. In addition the surveys were administered midyear to allow for reporting across a shorter period of time. To supplement the standard use of the SEC, the SEC was employed as a prospective survey to measure teachers' planned curriculum at the beginning of the school year. Last, the SEC's alignment statistics were used to examine the presume relation between alignment and achievement (p. 134)
	Respondents	8th grade students (N = 238) and teachers (N = 18)
	Dependent variable	Mathematics achievement
	Effect size	Significant correlations between student achievement averages and teacher alignment indices were equal to or greater than 0.48 (**significant 10 out of 15**). When teacher groups were examined separately, the relation between alignment and achievement remained significant only for special education, with correlations equal to or greater than 0.75 (p. 131)

(continued)

Table 4.6 (continued)

	Comments	Results not controlled for other independent variables or co variables (including prior achievement). Only the distinction between special and regular education is being made
Kyriakides et al. (2013)	Focus	Factors of effective teaching with the indicators orientation, questioning, structuring, application, management of time, assessment, the classroom and learning environment and teaching modeling. See Chap. 3: Meta-analyses
Marsha (2008)	Focus	This exploration includes multiple measures of classroom instruction to evaluate the instructional sensitivity of multiple measures of math achievement and applies an analytic method that makes it possible to relate student-level outcomes to teacher-level measures of instruction (p. 23). This is a correlational study conducted in the USA
	OTL measure	Instructional sensitivity, a link between instructional opportunities and performance on particular assessment items, by measuring two different performance levels, proximal and distal, with students assessments, teacher assessments and teacher interviews
	Respondents	Third grade students (N = 486) and third grade teachers (N = 24)
	Dependent variable	Mathematics and algebraic reasoning
	Effect size	The correlation between prior student achievement and the outcome measure was highest for the distal items, r = 0.61, $p < 0.01$ (proximal items: r = 0.30, $p < 0.01$). The correlation between OTL and performance on the proximal items was r = 0.28, $p > 0.01$ and on the distal items r = 0.05, $p > 0.01$. **No statistically significant effects for OTL on achievement**
	Comments	General measures of student prior achievement collected at the end of the previous school year were used as covariates in the multilevel analyses (p. 31)
Mo et al. (2013)	Focus	This study examined the individual, class, and school level variability of the students' science achievement. And it makes a contribution to a better understanding of the OTL variables at classroom and school level in students' science achievement (p. 3). This is a correlational study conducted in the USA
	OTL measure	OTL was measured as a classroom-level factor. Operationally, it included two factors: an indicator of teacher quality (science certification) and an indicator of instructional practice in terms of topic coverage (p. 4). Data from TIMSS 2003 is used

(continued)

Table 4.6 (continued)

	Respondents	8th grade students (N = 8.544)
	Dependent variable	Science achievement
	Effect size	**The two-class-level OTL variables significantly influenced the class-mean science achievement**. The percentage of variance in student science achievement explained by OTL at the class level (Level 2) was 23.32 %
	Comments	Study includes individual- (students' science- and classroom engagement and students' interests), teacher- (teacher quality and topic coverage), and school-level factors (availability of remedial and enriched courses and the SES of the school (p. 4). Results are not controlled for initial achievement scores
Niemi et al. (2007)	Focus	This study investigates the instructional sensitivity of a standards-based ninth grade performance assessment that requires students to write an essay about conflict in a literary work. Students were randomly assigned to one of three instructional groups: literary analysis, organization of writing and teacher selected instruction (p. 215). Experimental testing of an assessment's sensitivity to construct-focused instruction is likely to provide stronger validation evidence than OTL data alone (p. 217). This is an experimental study conducted in the USA
	OTL measure	Sensitivity of a ninth-grade writing performance assessment to different types of standards-based instruction: the differential effects of instruction focused on the organization of writing, literary analysis, or teacher selected goals, controlling for student background variables (p. 218). Sensitivity is measured by data from the district's ninth grade language arts performance assessment made by the students after 8 days of a certain type of instruction
	Respondents	9th grade students (N = 886) and teachers (N = 25)
	Dependent variable	Writing performance
	Effect size	The overall performance assessment score shows an advantage of 0.22 points for the literary analysis group versus the teacher choice group, after controlling for SAT-9 reading scores and language scores, and **this difference is significant**. Scores for students in the writing group were **not significantly different** from scores from students in the teacher choice group (p. 226)
	Comments	Results controlled for students' Grade 8 SAT-9 language scores, SAT-9 reading scores, free or reduced-price lunch program status and English language proficiency levels (pp. 223, 224)

(continued)

Table 4.6 (continued)

Oketch et al. (2012)	Focus	The primary concern in this paper is to understand some of the classroom-school factors that may explain the persistent differences in achievement between the top and bottom schools. The focus is on time-on-task and curriculum content and whether this explains the difference in performance (p. 19). This is a correlational study conducted in Kenya
	OTL measure	The effect of active teaching and content coverage on student achievement between low and high performing schools. To conduct this analysis a two-level multilevel model is fitted to evaluate to what degree content coverage, proportion of lesson time spent on active teaching influence student achievement (p. 23). Content coverage is measured through the analysis of classroom observation videos. Item response theory was used to calculate test scores, it generated 40 items in each test (p. 22)
	Respondents	6th grade students (N = 2.437) and teachers (N = 72)
	Dependent variable	Mathematics achievement
	Effect size	In the final model shows the effect OTL and time on active teaching on pupil IRT gain score, it controls for pupil, school and teacher characteristics. **Proportion of topic covered (OTL) is positive though not significant**. The proportion of time on active teaching is negative and not significant (pp. 29, 31)
	Comments	Model controls for pupil, school and teacher variables. Achievement is tested at two points in time
Ottmar et al. (2013)	Focus	Examination of the extent to which exposure to content and instructional practice contributes to mathematics achievement in fifth grade. Result suggest that more exposure to content beyond numbers and operations (i.e., geometry, algebra, measurement, and data analysis) contribute to student mathematics achievement, but there is no main effect for increased exposure on developing numbers and operations (p. 345). This is a correlational study conducted in the USA
	OTL measure	Contribution of exposure to specific mathematical content and instructional practice (i.e., geometry, algebra, measurement, data analysis) to mathematics achievement scores. Teachers of sampled children were asked to respond to 24 instructional practice and content items taken from the revised child-level fifth-grade mathematics teacher questionnaire. The fifth-grade mathematics assessment was administered to children using workbooks with open-ended questions (pp. 348, 349)

(continued)

Table 4.6 (continued)

	Respondents	5th grade students (N = 5.181), teachers and parents	
	Dependent variable	Mathematics achievement	
	Effect size	Results indicate that greater exposure to content beyond numbers and operations contributed to higher achievement, $\boldsymbol{p < 0.01}$. More exposure to numbers and operations or instructional practices **did not significantly contribute** to achievement growth, all p's > 0.05 (p. 351)	
	Comments	Results controlled for child and teacher/classroom variables, like students' SES but not for previous student achievement	
Plewis (1998)	Focus	This paper looks at between teacher differences in pupils' mathematics progress from two correlational studies in London schools. It was found that the more of the mathematics curriculum was covered by teachers, the greater the progress made by pupils in those classrooms (p. 97). Both studies are correlational and conducted in England	
	OTL measure	Effects of curriculum coverage and classroom grouping. The method of measuring curriculum coverage was essentially the same in the two studies. Each teacher completed a checklist for each pupil in the class, the checklist consisting of separate items put into groups such as addition, money, etc., which the teachers ticked if they had covered that item during the year with a particular pupil. Thus, we measured coverage of the curriculum experienced by the pupils but reported by their teachers. Each teacher was interviewed about their grouping practices at the end of Year 2 (p. 101)	
	Respondents	First grade students (N = 776)	Second grade students (N = about 550) and teachers (N = 28)
	Dependent variable	Mathematics achievement	
	Effect size	**The effect size for mean curriculum coverage is 0.11 SD units** accounting for 15 % of between teacher variance ($\boldsymbol{p < 0.02}$). The effects of classroom grouping on achievement were small. There was some benefit in being in a grouped classroom, with pupils making 0.18 SD units	**The effect size for mean curriculum coverage is 0.18 SD units**, accounting for 65 % of the between teacher variance ($\boldsymbol{p < 0.001}$). In contrast to study 1, content coverage was lowest for the 'grouped instruction' group. The effect on progress of being in the 'grouped instruction'

(continued)

Table 4.6 (continued)

		more progress than pupils in the other two types (whole class and individual instruction) of classroom after allowing for the effect of curriculum coverage at the pupil level. The differences in mean curriculum coverage across these three groups were not statistically significant (pp. 103, 104)	category was very small and negative. Differences between whole class and individual instruction were not significant (pp. 104, 105)
	Comments	Unknown	Results at least controlled for gender and ethnicity proportions in classrooms
Polikoff and Porter (2014)	Focus	This article is the first to explore the extent to which teachers' instructional alignment is associated with their contribution to student learning and their effectiveness on new composite evaluation measures using data from the Bill and Melinda Gates Foundation's Measurement of Effective Teaching (MET) study (p. 1). This is a correlational study conducted in the USA	
	OTL measure	The Surveys of Enacted Curriculum (SEC); The surveys define content at the intersection of specific topics and levels of cognitive demand; there are 183 fine-grained topics in mathematics and 133 in ELA. Cognitive demand varies from memorization to application or proof. Application: first decide which topics were taught or not (in a school year); for those taught indicate (a) the number of lessons spent on each topic and (b) the level of cognitive demand (cell = topic by cognitive demand combination)	
	Respondents	4th and 8th grade teachers (N = 701, 327 completed surveys)	
	Dependent variable	Value added measurement in Math and ELA	
	Effect size	When it comes to the zero-order correlations of VAM scores with SEC instructional alignment indices, most of the correlations are not significant. Three correlations with VAM were analyzed: the alignment between instruction and state standards, and the alignment between instruction and state or alternate test **In those grade, district, subject combinations where the correlations were significant the average was 0.16 for math and 0.14 for ELA**	

(continued)

Table 4.6 (continued)

	Comments	Most of the zero order correlations were not significant It should be noted that the independent variable was not the enacted curriculum, but various alignment indicators, e.g. the consistency between SEC and the contents of assessment tests. State and Alternate Assessment VAM (value-added models) scores were used
Ramírez (2006)	Focus	This study compared Chile to three countries and one large school system that had similar economic conditions but superior mathematics performance and examined how important characteristics of the Chilean education system could account for poor student achievement in mathematics. One of the results: the Chilean mathematics curriculum covered less content and fewer cognitive skills (p. 102). This study is correlational and is conducted in the USA
	OTL measure	Content coverage measured by students and teachers' self-reported questionnaires. This study used TIMSS 1998/99 data from Chile, South Korea, Malaysia, the Slovak Republic and Miami Dade Country Public Schools
	Respondents	8th grade students (N between 1.356 and 6.114 per country) and their Mathematics teachers
	Dependent variable	Mathematics achievement
	Effect size	In Chile, 73 % of the students were taught by teachers who emphasized basic mathematics content. In the comparison jurisdiction, this proportion was substantially smaller (6, 12, 19 and 33 %). In Chile, **content coverage was significantly related to mathematics performance**. This relationship held true after controlling statistically for schools' socio-economic index and type of administration, 4.8, $p < 0.05$)
	Comments	Results are controlled for schools' socio-economic index and type of administration (public/private), but not for prior achievement
Reeves (2005)	Focus	This thesis investigates whether the existing South African policy approach is supported through research, or whether, in accordance with the international evidence, 'Opportunity-to-Learn' (curriculum content and skills actually made available to learners in classrooms) has a greater effect on achievement (than 'type of pedagogy') and is therefore a policy variable worth taking more seriously for narrowing the gap in achievement between South African learners on different socio-economic backgrounds (p. iii). This is a correlational research conducted in South Africa

(continued)

Table 4.6 (continued)

	OTL measure	Four OTL dimensions: content coverage by cognitive demand, content exposure, curricular coherence and curricular pacing, measured by lesson observations, teacher survey interviews, teachers' year or term plans and students questionnaires (partly items from TIMSS) and students' workbooks and reports
	Respondents	6th grade students (N = 1.001) and their mathematics teachers
	Dependent variable	Mathematics achievement
	Effect size	The study's findings do not confirm the assumption that in relation to achievement gain, OTL is more important than 'type of pedagogy'. The results show that OTL and pedagogy variables both significantly affect achievement (p. 230). The variable that had the highest correlation with achievement gain was the level of cognitive demand (a correlation co-efficient of 0.28)
	Comments	Results controlled for individual learner background variables. This study uses achievement gain scores
Reeves and Major (2012)	Focus	Research has shown that rural high school students in the United States have lower academic achievement than their non-rural counterparts. The evidence for why this inequality exists is unclear, however. The present study takes up this issue with a narrowing of the focus. Using the database of the Educational Longitudinal Study of 2002–2004, the author investigates reasons for the rural achievement gap in mathematics during the last 2 years of high school. His approach focuses on the geographic disparities in the opportunity to learn advanced math (p. 887). This is a correlational study conducted in the USA
	OTL measure	The supply-side factors of OTL, such as school offerings of advanced math units, restriction on student admission to advanced math courses, or the quality of advanced math instruction. This will be measured using survey regression models focused on comparative effects of family SES on course taking in different geographic locations and separate regression models of math achievement gain will be estimated for each type of school location
	Respondents	10th grade (and two years later, 12th grade) students (N = 11.170)
	Dependent variable	Mathematics achievement

(continued)

Table 4.6 (continued)

	Effect size	In Model 3, we find that the addition of the opportunity-to-learn variable—total advanced math units taken—not only has a large effect on the math achievement gain, but it also accounts for more than two third of the residual rural gap and reduces the remaining gap to non-significance (p. 901)
	Comments	Results controlled for 10th grade achievement, student demographics, private school attendance, school size, family SES, and friends' educational engagement and aspirations (p. 899)
Reeves et al. (2013)	Focus	This paper estimates the effect of OTL on students' academic performance using rich data we gathered on the teaching process in a large number of South African and Botswana Grade 6 classrooms (p. 426). This is a correlational study conducted in Africa
	OTL measure	Curriculum coverage, including: content coverage, content exposure and content emphasis). The data comes from student notebooks and videotaped mathematics lessons
	Respondents	6th grade students (N > 5.000) and teachers (N = 116)
	Dependent variable	Mathematics achievement
	Effect size	The study's estimates suggest that in many of the South African classrooms the relation of additional lessons on test items to test score gains, although positive, is **not statistically significant**. The test score gain on items in Botswana classrooms is generally **negatively related** to the number of lessons given by teachers on each test item (p. 432)
	Comments	Results are controlled for pre-test scores, but not for students' SES
Roncagliolo (2013)	Focus	This study aims to explore and understand differences between the implemented curriculum in public and private schools in 4th, 5th, and 6th grades in the Dominican Republic, specifically with respect to the instructional time allocated by teachers in Mathematics Activities and Mathematics Contents. And if these differences do in fact exist, then do they help to explain differences in student mathematics achievement between public and private institutions (public rural, public urban, and accredited private schools)? (p. iv). This is a correlational study conducted in the USA

(continued)

Table 4.6 (continued)

	OTL measure	Allocated time based on teachers' questionnaires. The first section is made up of general information. The specific mathematics questions seek to map out different aspects of opportunities, such as time allocated for teaching, homework, use of instructional materials, teaching and learning activities regarding specific aspects of the Dominican curriculum, the number of lessons in specific contents, and specific questions related to the curriculum covered (p. 50). The section of Mathematics Content is composed of 70 variables (p. 80) Math performance is measured by 2005–2007 EERC applications, 156 common items in five areas: whole numbers, fractions and decimals, geometry, measurement and statistics (p. 49)
	Respondents	**Students** 2005: 3th grade (N = 6954), 4th grade (N = 7238), and 5th grade (N = 7746). 2006: 4th grade (N = 6354), 5th grade (N = 6942), and 6th grade (N = 7214). 2007: 5th grade (N = 6639), 6th grade (N = 6960), and 7th grade (N = 6999) **Teachers** 2005 (N = 29), 2006 (N270), and 2007 (210)
	Dependent variable	Mathematics achievement
	Effect size	The main curricular differences are more concentrated in 4th and 5th grade than they are in 6th grade; and this is especially the case in public rural and private schools (p. 102). The variance explained by Mathematics Content in grade 4 is 12.64, in grade 5 it is 17.67 and in grade 6 it is 17.81 (p. 149) A multilevel analysis carried out in this study did not show consistent effects of mathematics contents and mathematics activities predictors on mathematics achievement. Only one predictor, coverage of the mathematics content of addition, was found to be statistically significant (p. 164)
	Comments	The primary control variable is the type of institution: public urban, public rural or accredited private. Another control variable is instructional resources, such as computer and classroom mathematics learning materials. Besides, poverty is a control variable. Achievement is measured in three consecutive years
Schmidt (2009)	Focus	Exploration of the relationship of tracking in eighth grade to what mathematics topics are studied during eighth grade (content exposure) and to what is learned during the year as well as to what is achieved by the end of eighth grade (p. 6). This is an experimental study conducted in the USA

(continued)

Table 4.6 (continued)

	OTL measure	Effect of tracking in different types of courses: regular, pre-algebra and algebra. Each of the sampled school defines a track in the sense of providing different content opportunities to learn mathematics. TIMSS surveyed the mathematics teachers of the sampled classes
	Respondents	7th grade students (N = 3.886), 8th grade students (N = 7.087), 7th grade teachers (N = 127) and 8th grade teachers (N = 241)
	Dependent variable	Mathematics achievement
	Effect size	In tracked schools, the algebra track was statistically significantly different from the other two tracks ($p < 0.0001$), it covered content slightly over one grade level higher (1.09) than the regular track and almost one (0.92) than the pre-algebra track (p. 16) For algebra classes the 70-point difference in mean achievement between those in tracked schools versus non-tracked schools is significant (***p* < 0.003**), but the differences in mean achievement for the other two types of courses are not significant. Across the non-tracked schools there were no significant differences in eighth grade achievement for the three different type of courses ($p < 0.38$) (p. 21)
	Comments	The track designation was included as a dummy variable at the classroom level. The model also included several covariates at each of the levels in the design. The student-level included racial identity and SES, the class-level included 7th grade pre-measure, mean SES and track, the school-level included the school-level mean SES, percent minority enrolment, location and size of the school (p. 23)
Schmidt et al. (2009)	Focus	Analyses that explores the relationship between classroom coverage of specific mathematics content and student achievement as measured by the TIMSS-R international mathematics scaled score (p. i). This is a correlational study conducted in the USA
	OTL measure	Variation in content coverage across a set of districts and states and relating it to cross-district/state variation in achievement. The study uses IGP, "international grade placement", that provides an indication of the conceptual complexity for each topic. The data came from the teacher questionnaire in which they indicated the number of periods of coverage associated with each of a set of topics
	Respondents	8th grade students (N = 36.654) and their mathematics teachers

(continued)

Table 4.6 (continued)

	Dependent variable	Mathematics achievement
	Effect size	Districts that had a higher average value on the IGP index also had a correspondingly higher mean achievement (R2 = 67 %, $\boldsymbol{p < 0.01}$). To test the effect of OTL on achievement controlling for SES, both variables were included in the same district level regression model. Both were related to achievement (R2 = 82 %, $\boldsymbol{p < 0.0002}$).
	Comments	Results controlled for SES at all three levels and prior achievement
Schroeder et al. (2007)	Focus	Eight categories of teaching strategies: Questioning, manipulation, enhanced material, assessment, inquiry, enhanced context, instructional technology and collaborative learning. See Chap. 3: Meta-analyses
Snow-Renner (2001)	Focus	Academic achievement and opportunity to learn were studied using data from the 195 TIMSS for Colorado students at the elementary level. The study used a comprehensive definition of OTL that includes content coverage, curricular focus, duration of instruction, and instructional strategies. The implications for using large-scale measures to indicate how fairly educational opportunities were distributed were studied in a context of comparative accountability measures (p. 1). This is a correlational study conducted in the USA
	OTL measure	Content coverage measured in terms of curricular focus (number of topics taught b teachers) an topic coverage. The measures from student achievement scores and teacher surveys are mapped specifically onto six different subtopic: whole number; fractions and proportionality; measurement, estimation, and number sense; data representation, analysis and probability; geometry; and pattern, relations, and functions. Due to lack of a reasonable level of consistency reliability, the geometry and patterns subscales were omitted from the remainder of the study (pp. 8, 9)
	Respondents	Third and 4th grade students (N = 2.163) and their teachers
	Dependent variable	Mathematics achievement

(continued)

Table 4.6 (continued)

	Effect size	The only variable that correlates significantly and positively across grade levels with all four achievement subscales is the curricular focus variable **(8/8)**. For the other variables concerning topic coverage, correlations are inconsistent by grade level. **No fourth grade classes showed any significant relationships between achievement and topic coverage. The third grade classes showed significant correlations for 6 of the 12 variables, all positive (overall 6/24 significant)**. In contrast, fourth grade achievement correlated most highly and significantly with variables measuring instructional practices rather than topic coverage (pp. 13, 14)
	Comments	Results not controlled for other independent variables or co variables
Squires (2012)	Focus	This article reviewed the research around curriculum alignment in terms of the taught curriculum, the tested curriculum, and the written curriculum. The research demonstrates that school districts can improve student achievement by paying attention and aligning their written, taught, and tested curriculum (p. 134). See Chap. 3: Meta-analyses
Tarr et al. (2008)	Focus	Examination of student achievement in relation to the implementation of textbooks developed with funding from the National Science Foundation or published-developed textbooks (p. 247). This is an experimental study conducted in the USA
	OTL measure	Influence of curriculum type: NSF funded versus publisher developed. The study uses teacher surveys, the appendix of textbooks, textbook-use diary, observations, table-of-contents implementation record, existing school records, the TerraNova Survey (TNS) and the Balanced Assessment in Mathematics (BAM)
	Respondents	6th, 7th and 8th grade students (N = 2.533) and their teachers
	Dependent variable	Mathematics achievement
	Effect size	**The study detected no significant differences between the two groups of teachers on any singular variable** in year 1 and year 2, with one exception: teachers using NSF-funded curricula reported a significantly higher frequency of use of textbooks (p. 264) With regard to predicting student achievement from curriculum type there was **no main effect of curriculum type found to be statistically significant**

(continued)

Table 4.6 (continued)

	Comments	The design of this study took into account "treatment integrity" by including teachers' use of available curricular materials and their provision of key instructional practices associated with *Standards*-based instruction (p. 250). Prior achievement as a covariate
Tarr et al. (2010)	Focus	American curricula seems more skills oriented, more repetitive and less conceptually deep than those of nations that score better than America on TIMSS. This research-study focuses on the question whether there are differences in mathematical learning when students study from an integrated approach textbook and when they study from an subject-specific textbook. And what are the relationships among curriculum type, fidelity of implementation and student learning (pp. 1, 2). This is a correlational study conducted in the USA
	OTL measure	Influence of curriculum type: integrated approach textbook versus subject-specific textbook. The study uses classroom visits, teacher surveys, textbook diaries, project developed tests and standardized tests Another factor is what they called OTL, including the percentage of textbook lessons taught by the teacher during the year, the Extent of Textbook Implementation index, the seating arrangement of observed lessons and the dominant level of student engagement in observed lessons Lastly, the factor implementation fidelity, including Textbook Content Taught index, Content fidelity rating and the Extent of Textbook Implementation index (p. 19)
	Respondents	8th grade students (N = 2.621) and teachers (N = 43)
	Dependent variable	Mathematics achievement
	Effect size	Curriculum type is positively related in the three different test when no other variable is disregarded, but only 2 of the 3 are significant (**$r = 0.304$, $p < 0.05$; $r = 0.518$, $p < 0.001$; $r = 0.264$, $p > 0.05$**). OTL, is in all three tests positively and significantly related to student outcomes (**$r = 0.388$, $p < 0.01$; $r = 0.370$, $p < 0.05$; $r = 0.291$, $p < 0.05$**). Fidelity, is in non of the test significantly related to student outcomes (**$r = -0.189$, $p > 0.05$; $r = -0.085$, $p > 0.05$; $r = -0.115$, $p > 0.05$**) (p. 23)
	Comments	Results controlled for prior achievement. Correlations between student outcomes and ten other variables are measured partialling out the other variables one at a time

(continued)

Table 4.6 (continued)

Törnroos (2005)	Focus	Relation between OTL and mathematics achievement in which OTL is approached in three ways. Firstly, it was measured as the proportion of textbooks dedicated to different topics. The second approach was based on the data given by teachers in TIMSS 1999. The third approach involved an item-based analysis of the textbooks (p. 320). This is a correlational study conducted in Finland
	OTL measure	Content coverage divided into three variables: Proportion of textbooks dedicated to topics (INBOOKx), what has been taught by teachers (TAUGHTx) and the proportional analysis of the textbook content (CONTENTx)
	Respondents	7th grade students, teachers and textbooks (N = 9)
	Dependent variable	Mathematic achievement
	Effect size	For textbook K only the variable TAUGHT had a statistically significant correlation with achievement **(1/3)** Textbook P showed no statistically significant correlations between OTL and achievement **(0/3)** For textbook MM the variable INBOOK had what were clearly the highest correlations with achievement **(1/3)** (p. 321)
	Comments	This analysis was based on students' actual achievement instead of achievement gains over a specific time period (p. 321). Results are not controlled for other variables. However, a distinction is being made between raw and standardized scores
Wang (1998)	Focus	This study investigated the relationship between students' OTL and their science achievement. Hierarchical linear modelling was used to analyze OTL variables at two levels of instructional processes: the classroom level and the student level (p. 137). This is a correlational study conducted in the USA
	OTL measure	Eight OTL variables covered by four constructs: content coverage, content exposure, content emphasis, and quality of instructional delivery. The latter is rather broad including also for example teacher preparation and equipment use. Science achievement is measured in both a written test and a hands-on test. In addition, teachers were interviewed about content coverage, activities and their prediction of how well their students would do on post-test. The teachers also provided copies of all the material they used as well as student daily attendance lists (p. 141)
	Respondents	8th grade students (N = 623) and science teachers (N = 6)
	Dependent variable	Science achievement

(continued)

Table 4.6 (continued)

	Effect size	It was found that OTL variables were significant predictors of both written and hands-on test scores even after students' general ability level, ethnicity, and gender were controlled. Content exposure was the most significant predictor of students' written test scores, and quality of instructional delivery was the most significant predictor of the hands-on test scores (p. 137). Written tests: Content Exposure (β = 11.1, SE = 5.4), Content Coverage (β = 10.6, SE = 9.9) and Quality of Instructional Delivery (β = 5.8, SE = 4.0) Hands-on tests: Content Exposure (β = 14.2, SE = 7.0), Content Coverage (β = 25.4, SE = 12.8) and Quality of Instructional Delivery (β = 10.8, SE = 4.9) (p. 149)
	Comments	Results are controlled for students' general ability level, ethnicity, and gender, but not for students' SES Content emphasis was omitted from the analyses because of its high correlation coefficients with content coverage, content exposure, and quality of instructional delivery (p. 152)
Wang (2009)	Focus	This study empirically examined a subset of children from low-income families to determine whether African American and Caucasian students have differential opportunity to learn mathematics and the extent to which opportunities to learn predict gains in mathematics achievement at kindergarten (p. 295). This is a correlational study conducted in the USA
	OTL measure	OTL variables representing maths instructional time, maths instructional method (three variables), and maths instructional emphasis (two variables) Students were assessed in maths skills and knowledge both kindergarten entry and exit, and teachers were asked to complete a survey that included 48 items relating to maths OTL (p. 297)
	Respondents	Kindergarten students who lived below the poverty line (N = 1.721)
	Dependent variable	Mathematics achievement
	Effect size	OTL was found to predict maths achievement of African American and Caucasian kindergartners from low-income families. Both groups showed only 1 statistically positive significant correlation with achievement, which is the OTL variable 'Emphasis: Telling time, estimating quantities and coin values accurately 1–2 times per week'. For African American children the variable 'Method: Used math manipulatives at least 1–2 times per week' is negatively but statistically significantly related to math achievement (b = −1.23) (**Total: 2/6 significant correlations**). For the Caucasian students there a no other significant correlations (**Total: 1/6 significant correlations**)

(continued)

Table 4.6 (continued)

	Comments	Results controlled for mathematics achievement at kindergarten entry, student age, student gender, and full-day versus half-day kindergarten programs
Winfield (1987)	Focus	This study investigates the relation between first grade Chap. 1 students' test content coverage and performance on a standardized reading achievement test. It provides a strategy for obtaining teachers' estimates of test content coverage and an instructional context in which to assess students' opportunity to learn and their assessment-test performance (p. 436). This is an experimental study conducted in the USA
	OTL measure	Content coverage; To assess the relation between test content covered and student achievement a situation in which students receive supplementary services is useful. Chapter 1 students are students who receive supplementary services in reading. First grade classroom teachers and Chap. 1 teachers who instructed the same students within a school were surveyed to assess how much content of a first-grade standardized reading achievement test students had covered (pp. 440, 441). Average composite scores for each teacher group were used to categorize items into four groups; (1) Items receiving greater coverage by both Chap. 1 and regular first grade teachers, (2) Items receiving greater coverage by regular classroom teachers, (3) Items receiving greater coverage by Chap. 1 teachers and (4) Items receiving low coverage by both groups of teachers (p. 445)
	Respondents	First grade students (N = 105) and teachers (N = 19)
	Dependent variable	Reading achievement
	Effect size	The absolute difference between Chap. 1 and classroom teachers' ratings was small for all the categories (p. 448). Group 1 items: Chap. 1 teachers' ratings were significantly higher than those of regular classroom teachers (**$z = -2.37$, $p < 0.01$**), Chap. 1 students' performance was slightly but significantly lower (**$t = -3.12$, $p < 0.01$**). Group 2 items: Chap. 1 teachers' ratings were significantly lower than those of regular teachers (**$z = -2.02$, $p < 0.04$**), performance of Chap. 1 students was significantly lower than the performance of students in the National Reference Group. Group 3 items: Chap. 1 teachers' ratings were significantly higher than those of regular classroom teachers (**$z = -3.17$, $p < 0.001$**), student from the National Reference Group scores significantly higher than the Chap. 1 students (**$t = -5.23$, $p < 0.001$**). Group 4 items: The difference in rating between both groups of teachers was not significant (**$z = -1.33$, $p < 0.18$**), performance of Chap. 1 students was much lower than the performance of the students in the National Reference group (**$t = -7.86$, $p < 0.001$**) (p. 446)

(continued)

Table 4.6 (continued)

	Comments	Results are not controlled for other independent variables It is hard to draw conclusions because both groups of students are not comparable when it comes to the amount of instruction and their prior achievement level. The latter is the reason why they are receiving the extra support
Wonder-McDowell et al. (2011)	Focus	The purpose of this study was to explore the effects of aligning classroom core reading instruction with the supplementary reading instruction provided to 133 struggling grade 2 readers. A 2-group, pre-posttest true experimental design was employed in this study conducted in the USA (p. 259)
	OTL measure	Influence of aligned and unaligned supplementary reading instruction after a maximum of 20 weeks. Effect is measured by pre- and posttest with a focus on reading fluency, word identification, word attack and reading comprehension
	Respondents	Second grade students (N = 133) and teachers (N = 12)
	Dependent variable	Reading achievement
	Effect size	Struggling readers in both the aligned and unaligned supplementary reading instruction groups made significant growth across all measures from pretest to posttest during the treatment period. The eta-squared effect size indicated **for all four variables a small but statistically significant positive effect** of aligning supplementary reading instruction on students growth. The effect size for reading fluency is 0.17 ($p < 0.001$), for word identification 0.08 ($p < 011$), for word attack 0.13 ($p < 0.001$) and for reading comprehension the effect size is 0.18 ($p < 0.001$) (p. 272)
	Comments	Demographic variables of gender, reading achievement, ethnicity, English learner status, and free and reduced-price meals qualification are taken into account, there were no significant differences between both groups
Yoon et al. (1990)	Focus	The purpose of this study was to investigates the degree of consistency of teachers' content coverage reports with logical expectations about the contents of a course with a given title for two consecutive years and to detect the effects of content coverage by comparing student performance patterns (for students lower than Pre Algebra, math A and math B, Pre Algebra, Algebra 1 and Geometry) associated with teachers' reports of content coverage for 1988 and 1989 (p. 1). This is a correlational study conducted in the USA

(continued)

Table 4.6 (continued)

	OTL measure	Content coverage; The data were collected from teachers who volunteered to participate in the Mathematics Diagnostic Testing Program (MDTP). Under this project a series of four diagnostic tests have been developed (Algebra Readiness, Elementary Algebra were used in this study). Teachers are presented with different math topics and are asked to indicate how these topics are covered in each mathematics course they teach (new, extended, review, assumed, taught later, not in curriculum and don't know) (p. 3). The teacher topic coverage response data is related to student performance. Depending on the course in which students are enrolled, they will have taken either the MDTP Algebra Readiness or the Elementary Algebra tests and one of the six randomly assigned forms of the SIMS (Second International Mathematics Study) Benchmark test (p. 6)
	Respondents	8th grade students and teachers
	Dependent variable	Mathematics achievement
	Effect size	The pattern of performance on items from the MDTP tests classified according to topic and specific teachers' reports of content coverage agree with expectations, but are somewhat uneven. For example, for both Algebra Readiness and SIMS Benchmark, *p*-values were highest when topics were indicated as 'Taught as New' with 'Assumed as Prerequisite' taken out of consideration. *P*-values were lowest when topics were indicated as 'Not in Curriculum', 'Don't Know' and 'No Response' in MDTP Algebra Readiness both years. For the SIMS Benchmark items, the simple rank ordering of average p-values appears confusing because of high *p*-values for 'Taught Later', 'Not in Curriculum', and 'Don't Know' for both years (p. 8)
	Comments	Results are not controlled for other variables
Yoon et al. (1991)	Focus	Investigation of the validity of teachers' reports of students' instructional experiences (content exposure or coverage) and content validity of a given course. And examination of the sensitivity of the test to instruction by linking student performance patterns to instructional experiences of students as possible corroborating evidence of their relationship (p. 2). This is a correlational study conducted in the USA

(continued)

Table 4.6 (continued)

	OTL measure	Sensitivity of the test to instruction. Achievement scores are based on data from the Algebra Readiness and Elementary Algebra examinations for different courses (lower than Pre Algebra, math A and math B, Pre Algebra, Algebra 1 and Geometry). Content coverage was measured by teacher questionnaires about their coverage of mathematics topics (new, extended, review, assumed, taught later, not in curriculum and don't know)
	Respondents	8th grade students (N = approx. 2000) and teachers (N = approx. 20)
	Dependent variable	Mathematics achievement
	Effect size	Results show the evidence of content validity of test items by analyzing what was taught at secondary school mathematics and what was tested. Content coverage of test item topics was related to students' performance on the Algebra Readiness Test and Elementary Algebra Test. When *p*-value differences were considered for each topic, some topics were relatively more sensitive to content coverage than others. For example, the topics 'exponents with integral exponent', 'order and comparison of fractions', and 'perimeter and area of triangles and squares' showed relatively large *p*-value differences greater than 0.20. These topics were taught as CORE in 1988 and as PRIOR in 1989. Results are shown per course level (pp. 9, 10)
	Comments	Results are not controlled for other variables

Results presented on the nature of the OTL measure show considerable diversity. The most common reference is to content covered as indicated by teachers. Only incidentally are students asked to indicate whether content has been taught. Alternative operational definitions used in the studies are "program content modalities", "difficulty level of mathematics content", "topic and course text difficulty", "topic focus, in terms of basic and advanced math", "textbook coverage", "topic coverage and cognitive demand", "instruction time per intended content standards", "the enacted curriculum and its alignment with state standards", "instructional opportunities" "content coverage in terms of topic coverage, topic emphasis and topic exposure", "cognitive complexity per topic", "The quality of teaching a particular topic" "Aligned and unaligned exposure to reading instruction", "curriculum type". From these descriptions it appears that considerable

heterogeneity exist in the way researchers employ operational definitions of OTL. Additional, more minute content analyses would be needed to decipher to what extent alternative labels still represent the "core idea" of OTL. These results relate to the initial analyses of OTL conceptualization and measurement in Chap. 2, and will be taken up further in the final chapter of this report.

As far as the schematic overview in Table 4.6 provides an impression of the overall quality of the studies, the large majority have used student background adjustment of achievement measurements (in about 10 studies there was no adjustment, or it could not be inferred from the publication). In terms of research design 7 studies used an experimental or quasi experimental design, while the overlarge majority of studies was correlational.

Proportions of Significant and Insignificant Effects (Vote Counting)

Table 4.7 provides an overview of the number of effect sizes computed per study, whether OTL was positively or negatively associated with student achievement, and whether the association was statistically significant (5 % level).

Table 4.7 Significant and insignificant OTL effects

Article	Number of OTL effects	Number of statistical significant effects ($p < 0.05$)	Number of statistical insignificant effects ($p < 0.05$)	Number of statistical significant positive effects ($p < 0.05$)	Number of statistical significant negative effects ($p < 0.05$)
Aguirre-Muñoz and Boscardin (2008)	2	2	0	2	0
Boscardin et al. (2005)	1	1	0	1	0
Cai et al. (2011)	1	1	0	1	0
Calhoon and Petscher (2013)	1	1	0	1	0
Carnoy and Arends (2012)	2	0	2	0	0
Claessens et al. (2012)	12	8	4	5	3
Cogan et al. (2001)	2	2	0	2	0
Cueto et al. (2014)	1	1	0	1	0
Cueto et al. (2006)	2	1	1	1	0
D'agostino et al. (2007)	3	2	1	2	0

(continued)

Table 4.7 (continued)

Article	Number of OTL effects	Number of statistical significant effects ($p < 0.05$)	Number of statistical insignificant effects ($p < 0.05$)	Number of statistical significant positive effects ($p < 0.05$)	Number of statistical significant negative effects ($p < 0.05$)
Desimone et al. (2013)	4	4	0	2	2
Elliott (1998)	6	6	0	2	4
Engel et al. (2013)	4	3	1	2	1
Gamoran (1987)	1	1	0	1	0
Gamoran et al. (1997)	1	0	0	0	0
Gau (1997)	1	1	0	1	0
Grouws et al. (2013)	1	1	0	1	0
Heafner and Fitchett (2015)	2	2	0	1	1
Herman and Abedi (2004)	1	1	0	1	0
Holtzman (2009)	24	7	17	6	1
Kurz et al. (2014)	1	0	1	0	0
Kurz et al. (2010)	15	10	5	10	0
Marsha (2008)	2	0	2	0	0
Mo et al. (2013)	1	1	0	1	0
Niemi et al. (2007)	2	1	1	1	0
Oketch et al. (2012)	1	0	1	0	0
Ottmar et al. (2013)	2	1	1	1	0
Plewis (1998)	2	2	0	2	0
Polikoff and Porter (2014)	6	2	4	2	0
Ramirez (2006)	1	1	0	1	0
Reeves (2005)	4	1	3	1	0
Reeves and Major (2012)	1	1	0	1	0
Reeves et al. (2013)	2	1	1	0	1
Roncagliolo (2013)	8	1	7	1	0
Schmidt (2009)	3	1	2	1	0
Schmidt et al. (2009)	1	1	0	1	0
Snow-renner (2001)	32	14	18	14	0
Tarr et al. (2008)	1	0	1	0	0
Tarr et al. (2010)	9	5	4	5	0
Törnroos (2005)	9	2	7	2	0

(continued)

Table 4.7 (continued)

Article	Number of OTL effects	Number of statistical significant effects ($p < 0.05$)	Number of statistical insignificant effects ($p < 0.05$)	Number of statistical significant positive effects ($p < 0.05$)	Number of statistical significant negative effects ($p < 0.05$)
Wang (2009)	12	3	9	2	1
Winfield (1987)	1	1	0	1	0
Wonder-McDowell et al. (2011)	4	4	0	4	0
	Total effects	Total significant effects	Total insignificant effects	Total significant positive	Total significant negative
	192	98	93	84	14

Please note that the total number of studies is 43. Out of the 51 studies that were selected, 8 were left out because they were meta-analyses, or, on second notice, were considered as not addressing OTL effects

The results in Table 4.7 show that slightly more than half of the OTL effects are statistically significant and the other half is statistically insignificant. The most relevant indicator is the proportion of statistically significant positive associations, and this proportion is 84/192 (43.75 %). In order to put this proportion of positive significant effect in proportion Table 4.8 shows comparable indicators for other effectiveness enhancing conditions: consensus and cohesion between school staff, educational leadership, parental involvement, frequent evaluation and achievement orientation. Results of an earlier vote count on an OTL related variable, namely "curriculum quality and opportunity to learn" from Scheerens et al. are also included in the table.

Table 4.8 OTL percentage positive significant compared to similar indicators on other variables that are expected to enhance effectiveness from other research reviews

Variable	Percentage positive significant (%)	Source
OTL	42	This study
Curriculum quality and OTL	24	Scheerens et al. (2007)
Evaluation	28	Hendriks (2014)
Achievement orientation	41	Scheerens et al. (2007)
Learning time	36	Scheerens et al. (2007)
Parental involvement	34	Scheerens et al. (2007)
Staff consensus and cohesion	2	Scheerens et al. (2007)
Educational leadership	2	Scheerens et al. (2007)

It appears that the vote count measure of OTL, established in this study, (44 %) is of comparable size to other conditions like achievement orientation, learning time and parental involvement. What should be considered is that vote counting is a rather crude procedure and that comparison of quantitative effect sizes is more informative (compare the results of quantitative meta-analyses summarized in Chap. 3). The overview in Table 4.8 shows more dramatic outcomes when comparing organizational conditions like leadership and staff consensus with the other variables that are closer to the learning environment and the primary process of teaching and learning.

Annex: Descriptors Used in the Literature Search

Database: ERIC, PsycARTICLES, Psychology and Behavioral Sciences Collection, PsycINFO
Publication date: 1995–2015

"opportunity to learn" OR "curricul* align*" OR "learn* what is expected" OR "access to instruction" OR "curricul* exposure" OR "test preparat*" OR "exam* preparat*" OR "instruction* align*" OR "instructional sensitivity" OR "enacted curricul*" OR "curricul* cover*" OR "content cover*" OR "curricul* implement*" OR "curriculum teaching" OR "curricul* differen*" OR "curricul* coherence" OR "topic cover*"
AND
"Effectiveness" OR "achievement" OR "outcome" OR "success" OR "influence" OR "added-value" OR "grade"
NOT: ICT
NOT: disab* OR disadvantage*
NOT: material*
NOT: higher education
NOT: business
NOT: special.

References

Aguirre-Muñoz, Z., & Boscardin, C. K. (2008). Opportunity to learn and English learner achievement: Is increased content exposure beneficial? *Journal of Latinos and Education, 7*(3), 186–205.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher, 36*(5), 258–267.

Boscardin, C. K., Aguirre-Muñoz, Z., Stoker, G., Kim, J., Kim, M., & Lee, J. (2005). Relationship between opportunity to learn and student performance on English and Algebra assessments. *Educational Assessment, 10*(4), 307–332.

Cai, J., Wang, N., Moyer, J. C., Wang, C., & Nie, B. (2011). Longitudinal investigation of the curricular effect: An analysis of student learning outcomes from the LieCal Project in the United States. *International Journal of Educational Research, 50*(2), 117–136.

Calhoon, M. B., & Petscher, Y. (2013). Individual and group sensitivity to remedial reading program design: Examining reading gains across three middle school reading projects. *Reading and Writing, 26*(4), 565–592.

Carnoy, M., & Arends, F. (2012). Explaining mathematics achievement gains in Botswana and South Africa. *Prospects, 42*(4), 453–468.

Claessens, A., Engel, M., & Curran, F. C. (2012). Academic content, student learning, and the persistence of preschool effects. *American Educational Research Journal, 51*(2), 403–434.

Cogan, L. S., Schmidt, W. H., & Wiley, D. E. (2001). Who takes what math and in which track? Using TIMSS to characterize U.S. students' eighth grade mathematics learning opportunities. *Educational Evaluation and Policy Analysis, 23*(4), 323–341.

Cueto, S., Ramirez, C., & Leon, J. (2006). Opportunities to learn and achievement in mathematics in a sample of sixth grade students in Lima, Peru. *Educational Studies in Mathematics, 62*(1), 25–55.

Cueto, S., Guerrero, G., Leon, J., Zapata, M., & Freire, S. (2014). The relationship between socioeconomic status at age one, opportunities to learn and achievement in mathematics in fourth grade in Peru. *Oxford Review of Education, 40*(1), 50–72.

D'agostino, J., Welsh, M. E., & Nina, M. C. (2007). Instructional sensitivity of a state's standards-based assessment. *Educational Assessment, 12*(1), 1–22.

Desimone, L. M., Smith, T. M., & Phillips, K. (2013). Linking student achievement growth to professional development participation and changes in instruction: A longitudinal study elementary students and teachers in title I schools. *Teachers College Records, 115*(5), 1–46.

Elliott, M. (1998). School finance and opportunities to learn: Does money well spent enhance students' achievement? *Sociology of Education, 71*(3), 223–245.

Engel, M., Claessens, A., & Finch, M. A. (2013). Teaching students what they already know? The (mis)alignment between mathematics instructional content and student knowledge in kindergarten. *Educational Evaluation and Policy Analysis, 35*(2), 157–178.

Gamoran, A. (1987). *Instruction and the effects of schooling*. Paper Presented at the Annual Meetings of the American Sociological Association.

Gamoran, A., Porter, A. C., Smithson, J., & White, P. A. (1997). Upgrading high school mathematics instruction: Improving learning opportunities for low-achieving, low-income youth. *Educational Evaluation and Policy Analysis, 19*(4), 325–338.

Gau, S.-J. (1997). *The distribution and the effects of opportunity to learn on mathematics achievement*. Paper Presented at the Annual Meeting of the American Educational Research Association, Chicago, IL, timeMarch. (ERIC Document Reproduction Service No. ED407231).

Grouws, D. A., Tarr, J. E., Chávez, O., Sears, R., Soria, V. M., & Taylan, R. D. (2013). Curriculum and implementation effects on high school students' mathematics learning from curricula representing subject-specific and integrated content organizations. *Journal for Research in Mathematics Education, 44*(2), 416–463.

Heafner, T. L., & Fitchett, P. G. (2015). An opportunity to learn US history: What NAEP data suggest regarding the opportunity gap. *The High School Journal, 98*(3), 226–249.

Hendriks, M.A. (2014) *The influence on school size, leadership, evaluation, and time on student outcomes*. Enschede: University of Twente, Doctoral Thesis.

Herman, J. L. & Abedi, J. (2004). *Issues in assessing English language learners' opportunity to learn mathematics* (CSE Report No. 633). Los Angeles: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing.

Holtzman, D. J. (2009). *Relationships among content standards, instruction, and student achievement.* Retrieved from ProQuest Dissertations and Theses.

Kablan, Z., Topan, B., & Erkan, B. (2013). The effectiveness level of material use in classroom instruction: A meta-analysis study. *Educational Sciences: Theory and Practice, 13*(3), 1638–1644.

Kurz, A., Elliott, S. N., Wehby, J. H., & Smithson, J. L. (2010). Alignment of the intended, planned, and enacted curriculum in general and special education and its relation to student achievement. *The Journal of Special Education, 44*(3), 131–145.

Kurz, A., Elliott, S. N., Kettler, R. J., & Yel, N. (2014). Assessing students' opportunity to learn the intended curriculum using an online teacher log: Initial validity evidence. *Educational Assessment, 19*(3), 159–184.

Kyriakides, L., Christoforou, C., & Charalambous, C. Y. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teaching and Teacher Education, 36*, 143–152.

Marsha, I. (2008). Using instructional sensitivity and instructional opportunities to interpret students' mathematics performance. *Journal of Educational Research and Policy Studies, 8*(1), 23–43.

Mo, Y., Singh, K., & Chang, M. (2013). Opportunity to learn and student engagement A HLM study on eighth grade science achievement. *Educational Research for Policy and Practice, 12* (1), 3–19.

Niemi, D., Wang, J., Steinberg, D. H., Baker, E. L., & Wang, H. (2007). Instructional sensitivity of a complex language arts performance assessment. *Educational Assessment, 12*(3&4), 215–237.

Oketch, M., Mutisya, M., Sagwe, J., Musyoka, P., & Ngware, M. W. (2012). The effect of active teaching and subject content coverage on students' achievement: Evidence from primary schools in Kenya. *London Review of Education, 10*(1), 19–33.

Ottmar, E. R., Grissmer, D. W., Konold, T. R., Cameron, C. E., & Berry, R. Q. (2013). Increasing equity and achievement in fifth grade mathematics: The contribution of content exposure. *School Science and Mathematics, 133*(7), 345–355.

Plewis, I. (1998). Curriculum coverage and classroom grouping as explanations of between teacher differences in pupils' mathematics progress. *Educational Research and Evaluation, 4*(2), 97–107.

Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis, 20*, 1–18.

Ramírez, M.-J. (2006). Understanding the low mathematics achievement of Chilean students: A cross-national analysis using TIMSS data. *International Journal of Educational Research, 45* (3), 102–116.

Reeves, C. A. (2005). *The effect of 'opportunity-to-learn' and classroom pedagogy on mathematics achievement in schools serving low socio-economic status communities in the Cape Peninsula.* PhD thesis, School of Education, Faculty of Humanities, Cape Town, University of Cape Town.

Reeves, C., & Major, T. (2012). Using student notebooks to measure opportunity to learn in Botswana and South African classrooms. *Prospects, 42*(4), 403–413.

Reeves, C., Carnoy, M., & Addy, N. (2013). Comparing opportunity to learn and student achievement gains in southern African primary schools: A new approach. *International Journal of Educational Development, 33*(5), 426–435.

Roncagliolo, R. (2013). *Time to learn mathematics in public and private schools: Understanding difference in aspects of the implemented curriculum in the Dominican Republic* (Dissertation). Retrieved from ProQuest Dissertations and Theses.

Scheerens, J., Luyten, H., Steen, R., & Luyten-De Thouars, Y. (2007). Review and meta-analyses of school and teaching effectiveness. Enschede: University of Twente.

Schmidt, W. H. (2009). *Exploring the relationship between content coverage and achievement: Unpacking the meaning of tracking in eighth grade mathematics.* Education Policy Center, East Lansing, MI: Michigan State University.

Schmidt, W. H., Cogan, L. S., Houang, R. T. & McKnight, C. (2009). *Equality of educational opportunity: A myth or reality in U.S. schooling.* Lansing, MI: The Education Policy Center at Michigan State University.

Schroeder, C. M., Scott, T. P., Tolson, H., Huang, T.-Y., & Lee, Y. H. (2007). A meta-analysis of national research: Effects of teaching strategies on student achievement in science in the United States. *Journal of Research in Science Teaching, 44*(10), 1436–1460.

Snow-Renner, R. (2001). *What is the promise of large-scale classroom practice measures for informing us about equity in student opportunities-to-learn? An example using the Colorado TIMSS*. Paper Presented at the Annual Meeting of the American Educational Research Association. Seattle, WA, 10–14 April 2001.

Squires, D. (2012). Curriculum alignment research suggests that alignment can improve student achievement. *The Clearing House, 85*(4), 129–135.

Tarr, J. E., Reys, R. E., Reys, B. J., Chávez, Ó., Shih, J., & Osterlind, S. J. (2008). The impact of middle-grade mathematics curricula and the classroom learning environment on student achievement. *Journal for Research in Mathematics Education, 39*(3), 247–280.

Tarr, J. E., Ross, D. J., McNaught, M. D., Chávez, O., Grouws, D. A., & Reys, R. E. (2010). *Identification of student- and teacher-level variables in modelling variation of mathematics achievement data.* Paper Presented at the Annual Meeting of the American Educational Research Association, Denver, CO.

Törnroos, J. (2005). Mathematics textbooks, opportunity to learn and student achievement. *Studies in Educational Evaluation, 31*(4), 315–327.

Wang, J. (1998). Opportunity to learn: The impacts and policy implications. *Educational Evaluation and Policy Analysis, 20*(3), 137–156.

Wang, A. H. (2009). Optimizing early mathematics experiences for children from low-income families: A study on opportunity to learn mathematics. *Early Childhood Education Journal, 37*(4), 295–302.

Winfield, L. F. (1987). Teachers' estimates of test content covered in class and first-grade students' reading achievement. *The Elementary School Journal, 87*(4), 436–454.

Wonder-McDowell, C., Reutzel, D. R., & Smith, J. A. (2011). Does instructional alignment matter? Effects on struggling second graders' reading achievement. *The Elementary School Journal, 112*(2), 259–279.

Yoon, B., Burnstein, L., Chen, Z. & Kim, K.-S. (1990). *Patterns in teacher reports of topic coverage and their effects on math achievement: Comparisons across years* (CSE Tech. Rep. No. 309). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Yoon, B, Burnstein, L. & Gold, K. (1991). *Assessing the content validity of teachers' reports of content coverage and its relationship to student achievement* (CSE Rep. No. 328). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chapter 5
Predictive Power of OTL Measures in TIMSS and PISA

Hans Luyten

Abstract In this chapter secondary analyses of international data sets are presented. The analyses are based on data from TIMSS 2011 (grade 4 and grade 8) and PISA 2012, for the 22 countries that participated in both studies. In the analyses on TIMSS data three explanatory variables are taken into account: mathematics OTL, science OTL and number of books at home. In PISA only information on mathematics OTL is available (no information on science OTL was collected). All data were aggregated at the school level. All in all the secondary analyses of these international data sets show a modest effect of OTL for mathematics. An unexpected finding was that math OTL appears to be more strongly related to science achievement, than science OTL. The PISA 2012 results showed relatively high OTL effects, within and between countries. The standardized regression coefficients range from 0.119 in Romania to 0.813 in Qatar. The average effect across the 22 countries in PISA was 0.369. The PISA OTL effects stand out as being much stronger than the mathematics results found in the TIMSS studies. The first hypothetical explanation for this difference that comes to mind is the fact that TIMSS OTL measures were based on teacher responses, and the PISA OTL measures on student responses. The findings leave many questions that would be interesting to take up in future research.

Introduction

The predictive power of the OTL measures in the most recent TIMSS and PISA surveys has been analysed. This was done both at school and country level, controlling for the number of books at home (as reported by the students). The analyses focus on the Netherlands and those 22 countries that participated in the most recent TIMSS survey (data were collected in 2011) with both the grade 4 and grade 8

H. Luyten (✉)
University of Twente, Enschede, The Netherlands

© The Author(s) 2017

J. Scheerens (ed.), *Opportunity to Learn, Curriculum Alignment and Test Preparation*, SpringerBriefs in Education,
DOI 10.1007/978-3-319-43110-9_5

student populations and in the PISA 2012 survey. The Netherlands participated in PISA 2012 and also in TIMSS 2011, but only with the grade 4 student population.

Both TIMSS and PISA are large-scale cross-national comparative surveys focussing on student achievement. The TIMSS-survey is conducted every five years since 1995 and the PISA-survey every three years since 2000. Per country thousands of students participate and take standardized tests on mathematics, science and reading. In addition background information on students, schools, classrooms and teachers is collected. Over sixty countries participated in the most recent TIMSS and PISA surveys. The PISA target population is age-based. All 15-year-old students in a country are part of the target population, irrespective of their grade. TIMSS aims at two target populations, which are both based on grade (grade 4 and grade 8). The grade 4 population contains mostly 8-year-olds and the grade 8 population mostly 14-year-olds.

TIMSS focuses on mathematics and science achievement, whereas PISA covers reading literacy in addition to mathematics and science. The TIMSS study design attempts to align the achievement tests students take as closely as possible to the national curricula of the participating countries. In PISA, test curriculum alignment has never been a goal. The ambition of PISA is to assess the basic cognitive skills that young people need to succeed in later life. For further information on both the TIMSS and PISA project we refer to their respective websites:

http://www.iea.nl/timss_2011.html
http://www.oecd.org/pisa/

Assessing Opportunity to Learn (OTL) in TIMSS and PISA

In both TIMSS 2011 and PISA 2012 opportunity to learn (OTL) has been assessed, thus allowing for a cross-national assessment of the statistical relation between this measure and student achievement. In TIMSS the grade 4 teachers were asked to indicate, for a range of topics within three mathematics domains (Number, Geometric Shapes and Measures and Data Display), which of the following options best described the situation for the students in the TIMSS survey:

- The topic had been mostly taught before this year
- The topic had been mostly taught this year
- The topic had been not yet taught or just introduced.

Examples of the topics included are concepts of whole numbers, including place value and ordering (Number domain), comparing and drawing angles (Geometric Shapes and Measures domain) and reading data from tables, pictographs, bar graphs, or pie charts (Data Display domain).

With regard to science, similar questions were asked for topics within the domains Life Science, Physical Science and Earth Science. Based on these responses, indices were constructed that express on a scale from 0 to 100 to what

extent the various domains had been covered. For the purpose of the analyses that are reported on in this chapter two general OTL measures for mathematics and science were computed. The OTL measure for mathematics is the average of the OTL indices over the three domains. Likewise, the OTL measure for science is the average of the OTL indices over the three science domains.

The same strategy was followed for the TIMSS data that relate to grade 8. In this case there are four mathematics domains (Number, Algebra, Geometry, Data and Chance) and four science domains (Biology, Chemistry, Physics and Earth Science). As a result, for both TIMSS populations there is an OTL measure for mathematics and an OTL measure for science.

In PISA 2012, OTL data was obtained from the students instead of the teachers. Using the student responses, three indices on OTL with regard to mathematics were constructed. The analyses reported here focus on the experience with formal mathematics. Students were asked to indicate for the following mathematics tasks how often they had encountered them during their time at school.

- Solving an equation like $6x^2 + 5 = 29$
- Solving an equation like $2(x + 3) = (x + 3)(x - 3)$
- Solving an equation like $3x + 5 = 17$

The response categories were: frequently—sometimes—rarely—never. The PISA index was constructed by means of item response theory (IRT) scaling, which results in scores that may range from minus infinity to infinity with a zero mean (OECD 2014a, p. 329). Two additional OTL indices were constructed in PISA. The first one relates to experience with applied mathematics (e.g. figuring out from a train schedule how long it would take to get from one place to another) and the second one to familiarity with a range of mathematical concepts (e.g. exponential functions). The index that captures experience with formal mathematics is in the author's opinion most comparable to the OTL measures in TIMSS, which relate to mathematical content in abstract terms. The index on familiarity with mathematical concepts may partly reflect results of learning in addition to genuine OTL (although this cannot be ruled out completely for the measure on experience with formal mathematics). Familiarity with such concepts can also be conceived as a result of teaching in addition to opportunity to learn. Although the index on experience with formal mathematics may seem somewhat restricted, it statistical relation with mathematics achievement has been extensively documented (OECD 2014b, pp. 145–174).

Data Analysis

The relation between OTL and student achievement was assessed by means of a number of regression analyses. In these analyses number of books at home is included as a control variable. Prior analyses on PISA 2012 data revealed a

substantial correlation between OTL and family background (Schmidt et al. 2015). Raw correlations between achievement and OTL may therefore be confounded, due to the joint relation of OTL and achievement with family background. Unfortunately there is little overlap in the questions on family background in the TIMSS and PISA questionnaires, but the question on numbers of books at home is nearly identical in both surveys. The only difference is an additional sixth response category in PISA. In order to realize maximum comparability both the fifth and sixth response category in PISA have been collapsed into a single category. The resulting categories are: 0–10 books—11–25 books—26–100 books—101–200 books—more than 200 books.

As the OTL data in TIMSS are obtained from the teachers, the OTL measures in TIMSS can only account for variation in achievement between classes. In PISA information on classes is not included and it cannot be determined which students are classmates. Therefore it was decided to aggregate the data at the school level. In addition analyses were conducted on country means. Three datasets were analysed: TIMSS, grade 4 students, TIMSS grade 8 students and PISA. In the TIMSS data sets the impact of OTL on both mathematics and science was analysed (controlling for books at home). In these analyses the impact of mathematics and science OTL was analysed. In the PISA data set only the impact of mathematics OTL on mathematics achievement was analysed (controlling for books at home).

All analyses made use of "plausible values". In both the TIMSS and PISA datasets individual student performance is presented by five plausible values instead of a single test score. Plausible values were to deal with assessment situations where the number of items administered is not large enough for precise estimation of their ability. Based on the observed score a distribution is constructed of the student's ability. The plausible values are random draws from this distribution and represent the range of abilities (s)he might reasonably have (Wu and Adams 2002). The use of plausible values prevents researchers from underestimating random error (i.e. overestimating precision) in their findings. Before reporting the findings, descriptive statistics and correlations of country means are presented.

Descriptive Statistics Per Country and Correlations of Country Means

Tables 5.1, 5.2 and 5.3 present for TIMSS (grade 4 and 8) and PISA information per country on the number of schools, student achievement and number of books at home. Some of the countries selected for this study score very high on mathematics and science in both TIMSS and PISA (e.g. Singapore and Korea), but some countries that score far below the international mean (e.g. Tunisia and Qatar) are included as well. Tables 5.1, 5.2 and 5.3 suggest a great deal of consistency between TIMSS grade 4, TIMSS grade 8 and PISA. Countries that show a high average achievement in one survey also tend to score highly in the other surveys.

The consistency between mathematics and science also appears to be (very) strong. More details on the correlations between the country means are provided in Tables 5.4, 5.5 and 5.6. With regard to the numbers of books at home, a similar consistency between surveys can be observed. Korea and Norway show high averages in all three surveys, whereas the opposite goes for Tunisia and Thailand. In these two countries the average score on a scale from 1 to 5 hardly exceeds 2, which suggest a little over 11 books at home on average. In Korea and Norway the average is always well above three. This suggests well over 25 but probably closer to 100 books at home. With regard to OTL the country averages seem less consistent. Countries with a high OTL score on mathematics do not always show high OTL on science. Also across surveys the OTL seems not particularly consistent.

Table 5.1 Descriptive statistics TIMSS, grade 4

Countries	Number of schools	Achievement		OTL		Books at home (5 categories)
		Math	Science	Math	Science	
Australia	516	516	516	88.4	59.5	3.31
Chile	480	462	480	82.9	71.3	2.37
Finland	570	545	570	74.2	56.6	3.29
Hong Kong	535	602	535	81.4	57.3	2.82
Hungary	534	515	534	68.9	69.2	3.01
Italy	524	508	524	79.9	59.1	2.74
Japan	559	585	559	76.7	37.3	2.75
Korea (South)	587	605	587	73.2	50.5	3.86
Lithunia	515	534	515	85.1	81.0	2.57
Norway	494	495	494	67.4	59.0	4.12
New Zealand	497	486	497	76.9	55.5	3.16
Qatar	394	413	394	77.1	64.8	2.77
Romania	505	482	505	76.9	91.9	2.28
Singapore	583	606	583	85.9	39.3	3.08
Slovenia	520	513	520	66.8	63.8	2.98
Sweden	533	504	533	55.7	54.3	3.26
Thailand	472	458	472	78.6	66.8	2.00
Tunesia	346	359	346	52.8	46.6	2.04
Turkey	463	469	463	83.9	74.1	2.30
Taiwan	552	591	552	81.7	57.2	2.90
United Arab Emirates	428	434	428	72.8	64.8	2.62
United States of America	544	541	544	88.6	72.6	2.90
Average		**510**	**507**	**76.2**	**61.5**	**2.87**
Netherlands	128	540	531	63.3	48.2	2.91

Table 5.2 Descriptive statistics TIMSS, grade 8

Countries	Number of schools	Achievement		OTL		Books at home (5 categories)
		Math	Science	Math	Science	
Australia	227	505	519	78.9	59.0	3.27
Chile	184	416	461	71.5	76.7	2.51
Finland	142	514	552	55.6	64.8	3.29
Hong Kong	115	586	535	81.7	55.6	2.69
Hungary	145	505	522	85.9	84.1	3.21
Italy	188	498	501	80.7	76.2	3.03
Japan	138	570	558	89.6	59.4	2.93
Korea (South)	148	613	560	90.9	56.0	3.61
Lithunia	141	502	514	70.2	72.5	2.78
Norway	132	475	494	51.9	41.4	3.36
New Zealand	154	488	512	78.2	49.3	3.14
Qatar	107	410	419	86.2	79.5	2.75
Romania	146	458	465	93.8	96.0	2.50
Singapore	165	611	590	87.8	67.4	2.82
Slovenia	181	505	543	67.4	63.0	2.91
Sweden	142	484	509	60.0	67.1	3.23
Thailand	171	427	451	76.0	75.0	2.08
Tunesia	205	425	439	67.4	37.5	2.17
Turkey	238	452	483	94.7	88.1	2.48
Taiwan	150	609	564	71.2	63.5	3.00
United Arab Emirates	430	456	465	78.9	73.7	2.64
United States of America	384	509	524	90.6	83.8	2.94
Average		**501**	**508**	**77.7**	**67.7**	**2.88**

Table 5.3 Descriptive statistics PISA, 15-year-olds

Countries	Number of schools	Math Achievement	Math OTL	Books at home (5 categories)
Australia	775	504	−0.165	3.40
Chile	221	423	−0.102	2.42
Finland	311	519	0.003	3.33
Hong Kong	148	561	0.149	2.79
Hungary	204	477	0.140	3.46
Italy	1194	485	0.219	3.12
Japan	191	536	0.193	3.35
Korea (South)	156	554	0.428	3.80
Lithunia	216	479	0.133	2.89

(continued)

Table 5.3 (continued)

Countries	Number of schools	Math Achievement	Math OTL	Books at home (5 categories)
Norway	197	489	0.005	3.48
New Zealand	177	500	−0.270	3.33
Qatar	157	376	−0.282	2.81
Romania	178	445	−0.067	2.70
Singapore	172	573	0.331	3.05
Slovenia	338	501	0.199	2.92
Sweden	209	478	−0.251	3.38
Thailand	239	427	−0.090	2.37
Tunesia	153	388	−0.302	2.00
Turkey	170	448	−0.104	2.45
Taiwan	163	560	−0.040	3.13
United Arab Emirates	458	434	−0.097	2.72
United States of America	162	481	0.093	2.83
Average		**484**	**0.006**	**2.99**
Netherlands	179	523	−0.010	

Tables 5.4, 5.5 and 5.6 report the correlations between the country means. All correlations involve the 22 countries that participated in all three surveys. Because the findings for the Netherlands are not included, the figures reported in Tables 5.4, 5.5 and 5.6 all relate to the same 22 countries (the Netherlands did not participate in TIMSS grade 8; still the results hardly change if the Dutch data are included). Table 5.4 shows the correlations between the country achievement means on mathematics and science in the three surveys involved. These figures show a very strong degree of consistency. The correlations between mathematics across the three surveys exceeds 0.90 in each and every case. Countries with high mathematics scores in TIMSS grade 4 also score high in TIMSS grade 8 and in PISA. Within both TIMSS surveys the consistency between mathematics and science consistently exceeds 0.90 as well. The relatively "low" correlations in Table 5.4 (0.800–0.887) all relate to science achievement. The correlation between science in TIMSS grade 4 and grade 8 equals 0.887. The "lowest" correlations refer to science in grade 4 with mathematics in TIMSS grade 8 and PISA. The strongest correlation (0.954) is found for science in TIMSS grade 8 and mathematics in PISA. Table 5.5 shows a strong degree of consistency as well for the country means on number of books at home. The correlations range from 0.873 to 0.954.

Table 5.4 Correlations between country achievement means

Achievement		TIMSS, grade 4		TIMSS, grade 8		PISA
		Math	Science	Math	Science	Math
TIMSS, grade 4	Math	1				
	Science	0.932***	1			
TIMSS, grade 8	Math	0.930***	0.800***	1		
	Science	0.916***	0.887***	0.921***	1	
PISA	Math	0.946***	0.872***	0.951***	0.954***	1

*Significant at 0.001 level (one-tailed)
**Significant at 0.01 level (one-tailed)
***Significant at 0.001 level (one-tailed)
All correlations relate to 22 countries (Netherlands not included)

With regard to OTL (see Table 5.6), the picture is quite different. Only four out of ten correlations are statistically significant (ranging from 0.418 to 0.696). One of the non-significant correlations is even negative (−0.192). The findings clearly show that a country scoring high on OTL in one respect may look very different on OTL in other respects. The significant correlations between OTL all relate to findings from TIMSS:

- Mathematics in TIMSS grade 4 with mathematics in TIMSS grade 8 (0.489)
- Science in TIMSS grade 4 with science in TIMSS grade 8 (0.696)
- Mathematics in TIMSS grade 8 with science in TIMSS grade 8 (0.527)
- Mathematics in TIMSS grade 4 with science in TIMSS grade 8 (0.418).

The country means on OTL in TIMSS, which are based on teacher reports, show no significant correlations with those in PISA, which are based on student reports.

Table 5.5 Correlations between country means books at home

Books at home	TIMSS, grade 4	TIMSS, grade 8	PISA
TIMSS, grade 4	1		
TIMSS, grade 8	0.922***	1	
PISA	0.873***	0.954***	1

*Significant at 0.001 level (one-tailed)
**Significant at 0.01 level (one-tailed)
***Significant at 0.001 level (one-tailed)
All correlations relate to 22 countries (Netherlands not included)

Table 5.6 Correlations between country means OTL

OTL		TIMSS, grade 4		TIMSS, grade 8		PISA
		Math	Science	Math	Science	Math
TIMSS, grade 4	Math	1				
	Science	0.267	1			
TIMSS, grade 8	Math	0.489*	0.153	1		
	Science	0.418*	0.696***	0.527**	1	
PISA	Math	0.277	−0.192	0.280	0.048	1

*Significant at 0.001 level (one-tailed)
**Significant at 0.01 level (one-tailed)
***Significant at 0.001 level (one-tailed)
All correlations relate to 22 countries (Netherlands not included)

Predictive Power of OTL Measures Per Country

This section reports the findings from a series of regression analyses that aim to assess the effect of OTL on mathematics and science achievement controlling for number of books at home. The analyses are based on data from TIMSS 2011 (grade 4 and grade 8) and PISA 2012. In the analyses on TIMSS data three explanatory variables are taken into account: mathematics OTL, science OTL and number of books at home. In PISA only information on mathematics OTL is available (no information on science OTL was collected). All data are aggregated at the school level. This also goes for the data on number of books at home and for the achievement data. The findings can therefore not be interpreted as estimates of the effect of books at home on individual achievement. They reflect the relation between the average background of the school population (measured as books at home) and average achievement per school. A positive coefficient does not necessarily imply that students with many books at home tend to get high test scores (Robinson 1950). It is even conceivable that in schools with high numbers of books at home on average, the students with low numbers of books at home get high scores. With regard to the relation between books at home and academic achievement this may not seem a likely scenario, but in voting studies researchers may find that support for anti-immigrant policies is relatively strong in districts with large percentages of immigrants. In such a case, it is obvious that the relation between immigrant status and political sympathies at the individual level is quite different from the relation at the aggregate level. In the analyses the effects of mathematics OTL and science OTL have been assessed on both outcome measures. The initial expectation was that mathematics achievement is mainly affected by mathematics OTL and science achievement mainly by science OTL.

The findings for TIMSS grade 4 are presented in Table 5.7. The results show that mathematics OTL is significantly related to mathematics achievement in about half of the countries included (12 out of 23). The average effect across the 22 countries that participated both TIMSS surveys and PISA is rather modest (0.074).

A few countries even show negative OTL effects. Finland and the Netherlands show the strongest effects (0.293 and 0.236 respectively). With the exception of Qatar, all countries show a significant effect of books at home on mathematics achievement in grade 4. The average effect of books at home is 0.522. Science OTL hardly shows any effect on mathematics achievement. The average effect across 22 countries is virtually zero and in only one country (New Zealand) a significantly positive effect can be detected. Standardized regression coefficients, like correlations, can range from −1 to +1. If only one explanatory variable is involved, the standardized regression coefficient equals the correlation between the explanatory variable and the dependent variable. In this case the standardized regression coefficient of OTL can be interpreted as the correlation between OTL and student achievement while adjusting for number of books at home.

The findings regarding science in grade 4 are quite surprising. Once again we see significant effects of books at home in each and every country except Qatar. What we also see is that math OTL is much more strongly related to science achievement than science OTL. In nine countries a significant effect of math OTL on science achievement is detected. These are by and large the same countries showing a significant effect of math OTL on math achievement. The average effects of math and science OTL are quite similar to their average effects on mathematics achievement. A significant effect of science OTL on science achievement was found in only one country (United States). This unanticipated finding may possibly be due to a very strong correlation between the school means for mathematics and science.

Table 5.7 Standardized regression coefficients per country in TIMSS grade 4

Dependent variable	Mathematics achievement			Science achievement		
	Math OTL	Science OTL	Books at home	Math OTL	Science OTL	Books at home
Australia	**0.147	−0.003	***0.615	*0.105	0.037	***0.633
Chile	***0.208	−0.142	***0.577	***0.204	−0.128	***0.569
Finland	***0.293	0.005	**0.246	**0.229	−0.009	***0.296
Hong Kong	0.012	−0.126	***0.453	0.021	−0.114	***0.385
Hungary	0.009	0.042	***0.809	−0.003	0.055	***0.808
Italy	*0.137	−0.105	***0.252	0.114	−0.078	***0.304
Japan	−0.119	0.098	***0.560	−0.113	0.073	***0.558
Korea (South)	−0.031	0.003	***0.762	−0.064	−0.002	***0.771
Lithunia	*0.106	−0.043	***0.699	0.090	−0.004	***0.703
Norway	*0.164	0.014	***0.297	*0.146	0.016	***0.380
New Zealand	*0.124	*0.096	***0.627	*0.111	0.084	***0.659
Qatar	−0.023	0.030	0.086	0.012	0.077	0.061
Romania	−0.075	0.126	***0.554	−0.037	0.085	***0.638
Singapore	0.033	0.014	***0.752	0.019	0.023	***0.786
Slovenia	0.027	−0.041	***0.503	0.047	−0.057	***0.532

(continued)

Table 5.7 (continued)

Dependent variable	Mathematics achievement			Science achievement		
	Math OTL	Science OTL	Books at home	Math OTL	Science OTL	Books at home
Sweden	*0.124	0.059	***0.740	0.083	0.060	***0.783
Thailand	−0.123	0.000	***0.267	−0.131	−0.006	***0.263
Tunesia	**0.159	−0.002	***0.416	**0.153	−0.034	***0.506
Turkey	***0.193	−0.089	***0.609	***0.187	−0.081	***0.619
Taiwan	0.006	−0.107	***0.697	0.024	−0.114	***0.712
United Arab Emirates	0.056	−0.049	***0.276	0.073	−0.027	***0.256
United States of America	***0.204	0.047	***0.693	***0.140	*0.084	***0.737
Average	**0.074**	**−0.008**	**0.522**	**0.064**	**−0.003**	**0.544**
Netherlands	**0.236	−0.094	***0.393	***0.273	−0.101	***0.377

*Significant at 0.001 level (one-tailed)
**Significant at 0.01 level (one-tailed)
***Significant at 0.001 level (one-tailed)

The findings for TIMSS grade 8 are presented in Table 5.8. These findings reveal even less convincing evidence for an effect of OTL on mathematics or science achievement. In about one third of the countries included (7 out of 22) math OTL shows a statistically significant relation with mathematics achievement. The average effect across the 22 countries (0.025) is even closer to zero than it is in TIMSS grade 4. Seven countries show negative OTL effects, which is the same amount as those showing significantly positive effects. The strongest negative effect (−0.324; Qatar) is even further away from zero than the strongest positive effect (0.230; New Zealand). All countries show a significant effect of books at home on mathematics achievement in grade 4. The average effect of books at home is 0.677. Science OTL hardly shows any effect on mathematics achievement. The average effect across 22 countries is −0.016 zero and only one country (Singapore) shows a significantly positive effect.

Table 5.8 Standardized regression coefficients per country TIMSS in grade 8

Dependent variable	Mathematics achievement			Science achievement		
	Math OTL	Science OTL	Books at home	Math OTL	Science OTL	Books at home
Australia	**0.143	0.023	***0.717	***0.144	−0.008	***0.777
Chile	−0.039	0.029	***0.776	−0.057	0.040	***0.782
Finland	−0.066	0.009	***0.464	−0.048	0.006	***0.476
Hong Kong	−0.052	−0.149	***0.708	−0.072	−0.122	***0.670
Hungary	0.060	−0.019	***0.826	0.070	−0.006	***0.835
Italy	*0.142	−0.120	***0.516	0.090	−0.060	***0.575
Japan	−0.017	0.029	***0.570	0.009	0.040	***0.608

(continued)

Table 5.8 (continued)

Dependent variable	Mathematics achievement			Science achievement		
	Math OTL	Science OTL	Books at home	Math OTL	Science OTL	Books at home
Korea (South)	0.032	0.024	***0.749	−0.011	0.015	***0.681
Lithunia	−0.113	0.042	***0.689	−0.113	0.042	***0.689
Norway	*0.134	−0.089	***0.603	0.077	−0.060	***0.619
New Zealand	***0.230	0.019	***0.695	***0.189	0.034	***0.763
Qatar	−0.324	−0.056	***0.730	−0.317	−0.094	***0.712
Romania	0.042	0.027	***0.728	0.073	0.023	***0.685
Singapore	0.049	*0.093	***0.746	0.063	*0.100	***0.774
Slovenia	−0.016	−0.098	***0.555	0.022	−0.104	***0.533
Sweden	0.069	−0.008	***0.720	0.057	−0.033	***0.755
Thailand	0.058	−0.096	***0.676	0.066	−0.094	***0.667
Tunesia	−0.045	0.048	***0.698	−0.046	0.063	***0.635
Turkey	*0.096	0.028	***0.709	*0.100	0.030	***0.662
Taiwan	*0.079	−0.048	***0.802	*0.093	−0.065	***0.818
United Arab Emirates	−0.022	−0.036	***0.494	−0.012	−0.024	***0.488
United States of America	***0.116	−0.010	***0.720	**0.083	0.000	***0.750
Average	**0.025**	**−0.016**	**0.677**	**0.021**	**−0.013**	**0.680**

*Significant at 0.001 level (one-tailed)
**Significant at 0.01 level (one-tailed)
***Significant at 0.001 level (one-tailed)

The findings for PISA 2012 are presented in Table 5.9. In this case we find much stronger effects of OTL on mathematics achievement. In each and every country the OTL effect is significant. The standardized regression coefficients range from 0.119 in Romania to 0.813 in Qatar. The average effect across the 22 countries in PISA is 0.369. The effect of books at home on mathematics achievement is significant in all countries as well. This effect ranges from 0.134 in Qatar to 0.743 in Hungary with an average of 0.527.

Table 5.9 Standardized regression coefficients per country in PISA 2012

Dependent variable: Mathematics achievement	Math OTL	Books at home
Australia	***0.368	***0.495
Chile	***0.424	***0.555
Finland	***0.319	***0.383
Hong Kong	***0.546	***0.452
Hungary	***0.196	***0.743
Italy	***0.378	***0.525

(continued)

Table 5.9 (continued)

Dependent variable: Mathematics achievement	Math OTL	Books at home
Japan	***0.561	***0.428
Korea (South)	***0.407	***0.540
Lithunia	***0.368	***0.567
Norway	***0.322	***0.390
New Zealand	***0.368	***0.617
Qatar	***0.813	***0.134
Romania	*0.119	***0.699
Singapore	***0.238	***0.673
Slovenia	***0.219	***0.677
Sweden	***0.204	***0.539
Thailand	***0.301	***0.461
Tunesia	***0.587	***0.360
Turkey	***0.396	***0.516
Taiwan	***0.191	***0.737
United Arab Emirates	***0.508	***0.410
United States of America	***0.291	***0.689
Average	**0.369**	**0.527**
Netherlands	***0.672	***0.307

*Significant at 0.001 level (one-tailed)
**Significant at 0.01 level (one-tailed)
***Significant at 0.001 level (one-tailed)

Predictive Power of OTL Measures Aggregated at Country Level

This section reports the findings on five regression analyses using country means. The same dependent and explanatory variables are used as in the analyses at the school level reported in the previous section, only this time aggregated at country level. These analyses show to what extent countries with a high average OTL across all schools also show high average achievement scores. The results are reported in Table 5.10.

For mathematics the results look fairly similar in PISA and TIMSS (both grades). For TIMSS grade 4 it is found again that math OTL is more strongly related to science achievement than science OTL. In grade 8 no significant effects of either math or science OTL on science achievement are found. The standardized regression coefficients of math OTL on mathematics achievement in TIMSS (both grades) and PISA range from 0.464 to 0.533. The math OTL coefficient on science achievement in grade 4 is 0.430. None of the science OTL coefficients is statistically significant. The coefficients for books at home range from 0.439 to 0.596 and are all significant. Especially with regard to mathematics achievement the results are quite consistent. The effects of OTL is similar in size to that for books at home

Table 5.10 Regression analyses on country means

	TIMSS, grade 4		TIMSS, grade 8		PISA
	Math	Science	Math	Science	Math
OTL Math	**0.533	*0.430	*0.464	0.230	**0.487
OTL Science	−0.289	−0.120	−0.364	−0.181	–
Books at home	*0.439	**0.529	**0.507	**0.596	**0.448
R Square	52.4 %	45.2 %	44.0 %	40.5 %	61.5 %

*Significant at 0.001 level (one-tailed)
**Significant at 0.01 level (one-tailed)
***Significant at 0.001 level (one-tailed)
Findings relate to 22 countries (Netherlands not included)

(both a little below 0.500 on average). Note that these outcomes relate to country means. As such they indicate that in countries where the average number of books at home is relatively high, achievement scores are relative high as well. The same goes for countries with high math OTL on average. Also note that the effect of OTL is assessed controlling for books at home.

Discussion

In this chapter the statistical relation has been assessed between student achievement in science and mathematics and OTL, controlling for number of books at home. Use was made of the cross-national data that were collected in TIMSS 2011 and PISA 2012. The analyses relate to two aggregation levels: the school and country level. An important difference between TIMSS and PISA is the way OTL data have been obtained. In TIMSS the OTL measures are based on teacher responses, whereas in PISA the information is obtained through the students. In addition it should be noted that in comparison to TIMSS the OTL index in PISA seems rather restricted, as it based on only three questions (each referring to highly similar mathematical content).

Taken this into consideration, the modest relations that were found between the OTL measures in TIMSS and average achievement at the school level are striking. Quite remarkable is the finding that math OTL was found to be more strongly related to science achievement than science OTL. The relation between the teacher based OTL measures and student achievement in TIMSS turns out to be much weaker than the relation between the student based OTL measure and achievement in PISA. With regard to the country level, the findings in TIMSS show a stronger relation between mathematics OTL and mathematics achievement. At this level, the relation between math OTL and achievement is quite similar in TIMSS and PISA, although the correlations between the country level OTL measures in PISA and TIMSS are quite modest and statistically non-significant.

These findings call for a closer examination of the validity of the OTL measures in both TIMSS and PISA. The teacher based indices in TIMSS capture information from much more items (about twenty per index) than the student based measure in PISA (three items). Moreover, the items that make up the PISA index of experience with formal mathematics all relate to solving algebraic equations. Still, this seemingly crude measure shows a much stronger relation with achievement than the apparently more sophisticated measures in TIMSS.

There are at least two possibilities that need to be considered. First of all: are the OTL measures in TIMSS lacking in validity or reliability or can confounding variables account for the disappointingly weak relations with student achievement? Second: does the observed relation of the OTL index in PISA with student achievement somehow produce an overestimation of the real relation?

A salient finding is the near zero correlation of science OTL with science achievement. At the same time the relation between math OTL and science achievement is hardly any different from the correlation between math OTL and math achievement. Part of the explanation for this somewhat puzzling outcome may be a very strong correlation between the school means for mathematics and science. Remember that the analyses were conducted on aggregated data and that correlations at an aggregated level are typically stronger than they are at the individual level. In that respect it is not so surprising that the relation between math OTL and science achievement is very similar to the relation between math OTL and math achievement. What remains is the issue that science OTL appears to be virtually unrelated to science achievement, even though the format of the items is very similar to the OTL items for mathematics. The relation between math OTL and math achievement is not particularly strong, but definitely stronger than the relation between science OTL and achievement (and at least positive on average). A substantive interpretation may also apply, as it seems more plausible that mathematics exposure facilitates learning in science than it is plausible that learning science content facilitates math achievement.

It seems that the amount of mathematics topics taught is more strongly related to student achievement than the amount of science topics. Maybe this reflects the central importance of mathematics in school learning. One could also surmise that the weak relation between OTL and achievement in TIMSS results from poor information among teachers about the content that has been taught to their students in previous grades. It is also conceivable that teachers are not always able to assess how effective their instruction has been. Maybe they did teach the topics exactly as they reported, but if their instruction was not very effective, the relation between OTL and student achievement will be strongly weakened. This line of reasoning suggests an interaction effect of OTL and quality of instruction. Only combined with a sufficient instruction quality can we expect substantial effects of OTL. A final possibility is that sometimes teacher provide socially desirable answers to OTL items. Maybe they report what they feel they should have taught rather than what they actually taught.

The validity of the student based OTL index in PISA deserves closer scrutiny as well. It seems possible that OTL as measured in PISA also captures student ability

to some extent and thus produces an overestimation of the real relation between OTL and achievement. Possibly, students are more prone to report that they are familiar with certain topics if they master them. If a teacher did teach certain topics, but failed to get the main points across effectively, then this could make students reluctant to report that these topics have been covered. If any of this applies, the student reports on OTL, do not only reflect OTL but also aspects such as effectiveness of instruction and student aptitude. In that case, there is a serious risk that OTL based on student reports produced inflated correlations with achievement scores.

To sum up, all these possible explanations for the puzzling findings on the relation between OTL and achievement in TIMSS and PISA deserve further study. We argue for study designs to assess the impact of the possibly confounding factors outlined above (and more). It would in any case be useful to collect data on OTL from both student and teachers, so that it can be assessed to what extent they agree on OTL. An in-depth study on the relation between student aptitudes (or prior achievement) and their reports on OTL would be valuable as well, especially if one could also control for OTL as reported by their teachers. With regard to teacher based OTL, their reports should be compared to the instruction actually provided (e.g. as registered in logs or assessed through classroom observations). The possibility that quality of instruction can affect the relation between OTL and achievement also deserves close study. In fact, all the factors that play a part in Caroll's model on school learning (Caroll 1963, 1989) in addition to OTL (student aptitude, quality of instruction, perseverance and ability to understand instruction) should be taken into account when assessing the relation between OTL and student achievement.

Only if we can rule out the confounding factors outlined above, it would make sense to reconsider the theoretical assumptions that stipulate a strong relation between OTL and achievement. For the moment, it makes more sense to focus on possibly confounding factors that may account for the somewhat puzzling findings on the relation between OTL and achievement. One might consider testing the Caroll model in strongly controlled laboratory experiments. For example, a setting in which respondents need to learn relatively simple tasks in a short time span (a few hours). In such settings, one can more easily manipulate aspects like content covered and quality of instruction. Also the monitoring of student perseverance would be relatively straightforward. Even prior knowledge (ability to understand instruction) may be prone to manipulation.

As a final remark the findings on OTL for the Netherlands are highlighted. Two points stand out in this respect. First of all the relation between mathematics OTL and achievement has been found to be relatively strong in the Netherlands (see Tables 5.7 and 5.9). Three regression coefficients of math OTL have been reported with regard to the Netherlands. In two cases there is one country showing a stronger coefficient (math in TIMSS grade 4 and PISA) and in one case the Dutch coefficient is stronger than that of any other country (TIMSS grade 4 with science achievement as the dependent variable). What also stands out are the relatively low scores on OTL in the Netherlands. All three average OTL scores reported for the Netherlands

(see Tables 5.1 and 5.3) are below the international average. This deviation from the international mean is stronger for the teacher reports in TIMSS than it is for the national OTL average in PISA, which is based on student responses. It was noted earlier (Chap. 3) that the result by Schmidt et al. (2015), based on PISA 2012 data, showed that the Netherlands was the only country in which the influence of SES could be attributed for 100 % to OTL.

References

Caroll, J. B. (1963). A model of school learning. *Teachers College Record, 64*, 722–733.

Caroll, J. B. (1989). The Caroll model, a 25-year retrospective and prospective view. *Educational Researcher, 18*, 26–31.

OECD. (2014a). *PISA technical report*. Paris: OECD.

OECD. (2014b). *PISA 2012 results: What student know and can do, student performance in mathematics, reading and science* (Vol. 1). Paris: OECD.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*(3), 351–357.

Schmidt, W. H., Burroughs, N. A., Zoido, P., & Houang, R. H. (2015). The role of schooling in perpetuating educational inequality: an international perspective. *Educational Researcher, 20* (10), 1–16.

Wu, M. & Adams, R. J. (2002). *Plausible values—Why they are important.* Paper presented at the International Objective Measurement Workshop, New Orleans, 6–7 April.

Chapter 6
Recapitalization, Implications for Educational Policy and Practice and Future Research

Jaap Scheerens

Abstract In this concluding chapter conclusions are drawn, and the relevance of the results for educational science and policy and practice are discussed. Illustrations are provided that were drawn from the exploration of policy and practices in the Netherlands. Synthetic answers to the three research questions that guided the study are as follows: The OTL concept is better understood when it is placed in a larger framework of curricular alignment in educational systems. The average effect of OTL, estimated from the various parts of this study, amounts to a modest effect (d coefficient of 0.30, percentage of significant positive associations with achievement results of 44). Implications for educational policy are the recommendations to monitor the quality and curricular validity of high stakes tests, and to actively manage alignment between curricular components. Implications for educational practice in teaching are to consider optimizing OTL in the form of legitimate test preparation practices, and aligning formative and summative tests. Legitimate test preparation procedures are also highlighted as a relevant area for further research.

Summary of Main Findings

In this report OTL was defined as the matching of taught content with tested content. In the conceptual framework it was seen as part of the larger concept of curriculum *alignment* in educational systems.

When national educational systems are seen as multi-level structures, alignment is an issue at each specific level, but also an issue of connectivity between different

J. Scheerens (✉)
University of Twente (Nl), Zandpad 36, 3601 NA Maarssen, The Netherlands
e-mail: j.scheerens@utwente.nl

J. Scheerens
Oberon Research Institute, Utrecht, The Netherlands

© The Author(s) 2017

J. Scheerens (ed.), *Opportunity to Learn, Curriculum Alignment and Test Preparation*, SpringerBriefs in Education,
DOI 10.1007/978-3-319-43110-9_6

layers. General education goals or national standards are defined at the central level. At intermediary level (between the central government and schools) curriculum development, textbook production and test development have their organizational homes. At school level, school curricula or school working plans may be used, and at classroom level, lesson plans and actual teaching are facets of the implemented curriculum. Test taking at individual student level completes the picture. This process of gradual specification of curricula is the domain of curriculum research, with the important distinction between the intended, implemented and realized curriculum, as a core perspective. This perspective is mostly associated with a proactive logic of curriculum planning as an approach that should guarantee a valid operationalization of educational standards into planning documents and implementation in actual teaching.

In decentralized education systems explicit common goals or curriculum standards may be missing, or be of a very general nature. In the particular case when there are no specific central standards, but there is a formal set of examinations, teaching may get direction from being aligned to the contents of the examinations. This perspective could be seen as a "retro-active" orientation to alignment.

In the conceptual part of the report the issue of alignment was further analyzed by comparing proactive processes of curriculum development to test and examination driven approaches, in which accountability might be seen as driving educational improvement and reform. Further reflection on parallel processes in curriculum development on the one hand and test development on the other, led to conjectures about more efficient division of tasks and a discussion about whether one or the other should be leading. More closely related to the basic definition of OTL, the idea of evaluation driven improvement leads to questions about test preparation as an OTL maximizing procedure. These questions will be addressed in a subsequent section of this chapter.

An important realization from the conceptual analysis was the conclusion that *alignment* in multi-level education structures is a complex issue, with quite a few connections in need of being managed. It was noted that the quest for alignment would tend to require connectivity and "tight coupling" under actual conditions of "loose coupling".

The main body of this report was dedicated to assessing the empirical evidence on OTL effects. How consistently was OTL found to be significantly positively associated with student achievement outcomes, what seems to be a reasonable estimate of the quantitative effect size, and how does this compare to effect sizes that were found for other "effectiveness enhancing" school conditions?

The evidence from meta-studies that reviewed OTL effects appeared to be less solid than was expected, given the relatively high expectations about OTL effects expressed by various leading authors, like Porter, Schmidt and Polikoff. The number of meta-analyses was limited, and further analyses revealed that not all meta-studies listed as such were independent from one another. Leaving out the outlying results from Marzano, the OTL effect-size (in terms of the *d*-coefficient) compares to other relatively strong (or rather "relatively less weak") effectiveness enhancing conditions at school level, at about 0.30. A sophisticated recent study

(Polikoff and Porter 2014) suggests that effect sizes may be lower when adjustments are made for other variables.

The review of illustrative studies showed considerable diversity in the way OTL was measured. An important difference exists between studies that associate an empirical measure of exposure to achievement, as compared to studies that related an alignment index to achievement (as was the main emphasis in the studies by Porter et al. and Polikoff et al.). The results from PISA 2012 are considered striking, in the sense that OTL effects are higher and more generalizable across countries than any of the other school/teaching variables that are usually analyzed as background variables in PISA.

The literature search on empirical OTL effect studies yielded 51 studies and 198 effects. It was noted first of all that results presented on the nature of the OTL measure showed considerable diversity. The most common reference was to content covered, as indicated by teachers. Only incidentally were *students* asked to indicate whether content had been taught. Alternative operational definitions used in the studies are "program content modalities", "difficulty level of mathematics content", "topic and course text difficulty", "topic focus, in terms of basic and advanced math", "textbook coverage", "topic coverage and cognitive demand", "instruction time per intended content standards", "the enacted curriculum and its alignment with state standards", "instructional opportunities" "content coverage in terms of topic coverage, topic emphasis and topic exposure", "cognitive complexity per topic", "the quality of teaching a particular topic" "aligned and unaligned exposure to reading instruction", and "curriculum type". From these descriptions it appears that considerable heterogeneity exists in the way researchers employ operational definitions of OTL. Additional, more minute content analyses would be needed to decipher to what extent alternative labels still represent the "core idea" of OTL. As far as research methodology is concerned, the large majority of studies had used student background adjustments of achievement measurements (in about 10 studies there was no adjustment, or it could not be inferred from the publication). In terms of research design 7 studies used an experimental or quasi experimental design, while the overlarge majority of studies was correlational.

It was concluded that the vote count measure of OTL, (i.e. the percentage of effect sizes that were statistically significant and positive) established in this study, and which was 44 %, is of comparable size to other effectiveness enhancing conditions like achievement orientation, learning time and parental involvement, but dramatically higher than vote count measures for variables like cooperation and educational leadership. What should be considered is that vote counting is a rather crude procedure and that comparison of quantitative effect sizes is more informative (compare the results of quantitative meta-analyses summarized in Chap. 3).

The part of this study based on secondary analyses of international data sets is reported in Chap. 5. A series of regression analyses was conducted that aimed to assess the effect of OTL on mathematics and science achievement, controlling for number of books at home. The analyses were based on data from TIMSS 2011 (grade 4 and grade 8) and PISA 2012, for the 22 countries that participated in both studies. In the analyses on TIMSS data three explanatory variables were taken into

account: mathematics OTL, science OTL and number of books at home. In PISA only information on mathematics OTL is available (no information on science OTL was collected). All data were aggregated at the school level.

The findings for TIMSS grade 4 showed that mathematics OTL is significantly related to mathematics achievement in about half of the countries included (12 out of 23). The average effect (standardized regression coefficients, interpretable as correlations) across the 22 countries that participated in both TIMSS surveys and PISA is rather modest (0.074). A few countries even showed negative OTL effects. Finland and the Netherlands had the strongest OTL effects (0.293 and 0.236 respectively).

The findings regarding science in grade 4 are quite surprising. Once again significant effects of books at home were found in each and every country except Qatar. Quite surprisingly *math* OTL was much more strongly related to *science* achievement than *science* OTL. The average effects of math and science OTL were quite similar to their average effects on mathematics achievement. A significant effect of science OTL on science achievement was found in only one country (United States). This unanticipated finding may possibly be due to a very strong correlation between the school means for mathematics and science.

The findings for TIMSS grade 8 revealed even less convincing evidence for an effect of OTL on mathematics or science achievement. In about one third of the countries included (7 out of 22) math OTL showed a statistically significant relation with mathematics achievement. The average effect across the 22 countries (0.025) is even closer to zero than it is in TIMSS grade 4. Seven countries showed negative OTL effects, which is the same amount as those showing significantly positive effects. The strongest negative effect that was found (−0.324; Qatar) is even further away from zero than the strongest positive effect (0.230; New Zealand).

The findings for PISA 2012 showed much stronger effects of OTL on mathematics achievement. In each and every country the OTL effect was significant. The standardized regression coefficients range from 0.119 in Romania to 0.813 in Qatar. The average effect across the 22 countries in PISA was 0.369.

When regression analyses at aggregated levels were carried out, the same dependent and explanatory variables were used as in the analyses at the school level, only this time aggregated at country level. These analyses showed to what extent countries with a high average OTL across all schools also show high average achievement scores as well.

For mathematics the results were fairly similar in PISA and TIMSS (both grades). For TIMSS grade 4 it was found again that math OTL is more strongly related to science achievement than science OTL. In grade 8 (TIMSS results) no significant effects of either math or science OTL on science achievement were found. The standardized regression coefficients of math OTL on mathematics achievement in TIMSS (both grades) and PISA range from 0.464 to 0.533. The *math* OTL coefficient on *science* achievement in grade 4 is 0.430. None of the science OTL coefficients was statistically significant.

All in all the secondary analyses of these international data sets showed a modest effect of OTL for mathematics, next to the unexpected finding that math OTL was

more strongly related to science achievement, than science OTL. Another finding that stood out was the much stronger OTL effects on formal mathematics achievement found in the analysis of the PISA 2012 data set, as compared to the analyses based on TIMMSS. The first hypothetical explanation for this difference that comes to mind is the fact that TIMSS OTL measures were based on teacher responses, and the PISA OTL measures on student responses. The findings leave many questions that will be taken up further on, when discussing implications for further research.

Implications for Educational Policy

The idea of systemic alignment in education could be tackled in various ways. Seen from the center there are two roads of entry: starting at the front with the specification of educational goals as national standards, or starting at the outcome side of policy formation, in the form of putting in place high stakes summative tests or examinations. In earlier chapters these two approaches were indicated as proactive (standards up front) and retroactive, evaluation based. Two additional options would be to simultaneously develop standards and examination programs or do neither, while depending on alternative mechanisms to guarantee connectivity. A schematic description of these four options is rendered in Fig. 6.1.

In the United States the development of common core national standards is a major current policy operation. National Assessments are already in place in the form of NAEP; although States may also use State specific high stakes assessments. The Netherlands has high school autonomy and a strong aversion against "state pedagogy". Educational goals are stated in most general terms as "end terms" and reference levels for mathematics and language at secondary school level. At the same time there are central examinations in secondary education and a high stakes "closure" test at primary education. Countries where neither national standards nor high stakes examinations exist, but which still have high performance on international assessment test are Finland and Belgium. It is assumed that in these countries the quality of education results from alternative measures like: high quality teacher training and formative assessment. The situation indicated in the second row of Fig. 6.1 is more likely in traditional centralistic educational systems, although the accountability movement stimulates implementing summative testing in such countries as well. The development of educational testing in Italy may be seen as an example of this development.

Fig. 6.1 Proactive (standards) and retroactive planning (examinations) in educational policy

National standards	Examinations
X	X
X	0
0	X
0	0

The empirical evidence on the effectiveness of these system level levers of educational improvement is partial, inconclusive and sometimes contradictory (Scheerens 2016, Chap. 9). There is relative consistency in positive support for having central, standard based examinations in place (Bishop 1997; Woessmann et al. 2009), yet when controlling for the socio economic background of studies, some analyses show that the examination effect disappears (Scheerens et al. 2014). The model that liberates control over inputs (such as national curriculum frameworks) while strengthening outcome control by means of examinations and high stakes tests, has much credence in countries which are involved in decentralization and devolution of authority to lower levels in the system.

As far as the proactive approach, featuring central standards and standardized curriculum policies are concerned the results from PISA 2012 (OECD 2014) provide an interesting outlook. A relevant finding is that in countries that have a standardized policy for mathematics, "such as a school curriculum with shared instructional materials, accompanied by staff development and training" (ibid., p. 53) student performance is higher under conditions of autonomy than for countries lacking such a standardized policy. At first sight this conclusion looks contradictory because it seems to refer to the interaction of centralistic, and (standardized policy) and decentral facets of curriculum policy. But school autonomy in the curriculum domains is operationalized in terms of the discretion teachers have over choice of textbooks and curriculum material. The results seem to imply that standardized curriculum frameworks interact positively with teacher autonomy in decision-making about instructional methods. There is also miscellaneous, more casuistic support for the effectiveness of centralized curriculum arrangements. In a comparative study on Latin American countries, Willms and Somers (2000) showed the superiority of educational performance of Cuba. Sahlgren (2015) provides a very interesting analysis of the high educational performance of Finland, which he attributes to the Finish educational system being centralized with little autonomy until the 1990s. He sees the most recent (slight) decline in test scores of Finland as a result of the abandoning of traditional teaching methods. Finally, several upcoming high performing educational systems, such as Singapore and Honk Kong, match detailed proactive approaches in the form of standards and curriculum guidelines with sophisticated assessments. As a matter of fact this would seem to be the more logical approach, since high stakes test and examination development implies the use of standards.

Perhaps the safest conclusion that can be drawn at present is that different strategies might be effective depending on national contexts and traditions in education. Within the context of this study on OTL either "proactive" standards or high stakes assessments are pre-supposed in order to address the alignment issue straightforwardly. A final note of caution with respect to Fig. 6.1 is that the development of examinations requires some idea of national priorities in education, therefore a pure Zero situation on national standards is less probable.

Next to proactive, retroactive or "combined" strategies with respect to national standards and national assessments, this study has highlighted the relatively long chain of intermediary components, when alignment is at stake. Basic intermediary

components are textbooks, school curricula and actual teaching, and depending on the built-up of countries, also state or regional interpretations of national standards. It was noted that the units that offer services in developing these intermediary components may tend to be independent, and it was concluded that the ideal of alignment involves creating connectivity in a context characterized by loose coupling. If such fragmentary organization is the reality, alignment happens more or less by chance, and the challenge is to coordinate and manage connectivity. What this involves is illustrated in a case study of the functioning of the Dutch educational system.

The case study on OTL in Dutch primary education by Appelhof (2016), (not included in this book, and only available in Dutch), shows that during the last fifteen years important developments took place that could be seen as potentially advancing alignment between national standards, teaching methods, actual teaching and testing. The main ingredients were the formulation of "reference levels" initiated by the Committee Meijering, in 2008, the policy initiative concerning "achievement oriented work" as part of the Quality Agendas of the Ministry of Education in 2007, followed up by initiatives from educational publishers, support institutes (CITO and SLO, specifically) and the schools themselves. The case study provides documentation on how educational publishers invested in aligning teaching methods and textbooks to the reference levels, how the test institute (CITO) has done the same for its summative and formative tests, and the SLO (the institute for curriculum development) has supported the development of longitudinal content strategies (Dutch: *doorlopende leerlijnen*). The methods for arithmetic that were described in the case study, show the importance of formative tests; one of the methods (*Rekentuin*) can even be described as being totally centered around adaptive tests. The *Rekentuin* approach comes close to the design of instructional alignment as test preparation, which was offered as a theoretical option in earlier chapters. In addition the RTTI program by Docentplus (Drost and Verra 2015) offers a structured approach, in which teachers are guided in improving existing formative assessments, according to a taxonomy of cognitive operations, ranging from reproduction to insightful application. Alignment of the formative tests to examinations and content standards is an explicit part of the approach.

The government policy to stimulate achievement oriented work is a very relevant context for the furthering of OTL, at school and classroom level, in the Dutch context. Visscher (2015) provides an overview of the results of an ongoing research and development program on "achievement oriented work". The achievement oriented work approach, further abbreviated as AOW, proposes a cyclic approach, in which diagnostic analysis of test results is seen as the first step. Teachers are trained to interpret and use the results of tests, particularly the results of the LVS pupil monitoring system in primary schools, to assess the achievement of their students, and are subsequently trained to use a planning approach to design measures to adapt teaching to the needs of subgroups of students. First outcomes of evaluation studies show positive results. The AOW approach is further refined by means of systematic instructional design methods. Apart from these positive results, the experiences with AOW also indicate that it takes time and effort to teach

teachers to work with test information and apply systematic instructional design methods. Recent work by Vanlommel et al. (2016), in the context of Belgium primary education, points at fundamental problems with implementing rational techniques, like formative assessment and data use in schools. These authors found that a majority of teachers prefer "intuitive" reasoning over data-use in taking important decisions, like pass-fail decisions in progressing to the next grade.

An issue that came up in the Dutch case study by Appelhof (ibid) is the fear that externally developed, refined and well-aligned teaching and assessment methods may harm the professional space and autonomy of teachers. Such sentiments are very important as far as the implementation of rational strategies of alignment is concerned. Although one might argue that these new tools leave enough challenges to the professional expertise of teachers, acceptance may have the nature of an important change in the working culture at school. In the Dutch context, government policy provides mixed signals to teachers and schools, by constantly emphasizing more freedom and autonomy, and apparently not acknowledging that achievement oriented work, partially constrains and externally standardizes work at school.

The results of this study show that the effect of OTL can be considered of "educational significance", when the taught content is compared to content that is actually tested to determine student achievement. Looking more broadly at alignment between various curricular components (like national standards, textbooks, and assessments), the impression from the literature is that alignment at different stages is quite sub-optimal, which was tentatively attributed to independence and loose coupling of the organizational units concerned (government, educational publisher, intermediary levels of government, test developers, and what is actually delivered in teaching).

When the question is raised what government educational policy can do to optimize alignment and OTL, the real options will depend on the overall degree of centralization and decentralization of the system, existing structures and cultural considerations. Still, the general line of thinking is that certain measures at system level can facilitate alignment, and ultimately help in optimizing opportunity to learn at micro level. The following issues should be considered:

(a) Standard based examinations and high stakes tests are to be considered as the basic prerequisite for a rational treatment of the alignment issue. Presupposed is an adequate coverage of state educational standards in particular subject areas in the high stakes tests or examinations. The "instructional sensitivity" of tests (Popham 2001), depends on the transparency of the content structure of tests, sufficient test items per content domain, and a review of the teacheablity of content standards.
(b) The first issue in monitoring alignment is to check the presupposed coverage of national standards in national assessment programs, examinations and high stakes tests. The most probable perspective here would be to operationalize standards into educational objectives. This is the traditional proactive, "deductive" approach. In some cases, when there is strong aversion against

centralistic "state" pedagogy, but high quality examinations are in place, the latter could be used as the starting point for making items, learning tasks and task domains more explicit, also in the service of developing training material and textbooks.

(c) In order to facilitate OTL at micro level, depending on how the educational system is organized, the connectivity of formative tests to summative tests and examinations could be stimulated, and enforced from the center.

(d) Some of the developments in the realm of educational assessment and evaluation go in the direction of enlarging the role that "products of test development" can play in designing teaching methods and the shaping of actual teaching. The experiences in the Netherlands (Appelhof 2016), provide examples of using test results actively in designing teaching. Methods are developed in which formative tests are used adaptively in the service of better differentiation in teaching. A wide practice has come into existence of tests that are part of teaching methods, teachers developing their own tests, on the basis of clear technical guidelines and external support, and test preparation by students, on the basis of items drawn from item banks. In the Netherlands these activities are dependent on choices by autonomous schools, while supported by national policies to stimulate "achievement oriented work". In more general terms, central policies could stimulate test developers to develop item banks, and formative "off springs" of summative tests and examinations.

(e) Finally, it should be mentioned that in actual practice "OTL policies" should be seen as embedded in a context of simultaneously occurring alternative measures to enhance educational quality. The way alignment and OTL have been treated in this report can be seen as an integration of curriculum policies and use of assessments and examinations. Teacher training is an alternative strategy of quality maintenance and improvement, which might to some extent compensate for less developed testing, or seen as a factor that facilitates appropriate use of tests and OTL optimization.

Implications for Teachers

Examining the content that is actually covered in teaching is closest to the actual creation of OTL at school and classroom level. Once again optimizing OTL, and the larger issue of alignment, could be tackled in two ways, indicated in this report as the proactive approach and the retroactive approach. The traditional curriculum development approach would prescribe a continued process of operationalization of educational goals into teacheable learning tasks. This "deductive" approach has been used in the development of school working plans, or school development plans, which were likely to die a quiet death in office cupboards. The alternative "retroactive" approach, described in this report, takes the content of high stakes tests and examinations as point of departure. This is a controversial perspective,

because it could be captured under the heading of "teaching to the test", which is associated with reduced teaching, tunnel vision and cheating. Throughout this report we have been hinting at a legitimate form of test and examination preparation, and in this final section this perspective will be analyzed in more detail, leading up to a series of suggestions to optimize OTL by means of legitimate test preparation.

The theoretical background is the distinction of the two parallel processes of didactic and evaluative specification in Groot's (1986) model, described in Chap. 2. Particularly in settings where state standards are described in general terms, while examinations and high stakes tests are well established, teaching might obtain focus by targeting tested content. This orientation is strongly enforced by accountability policies, not only when these are "high stakes" but also in case of more moderate forms, such as rankings of schools published in the media. Again "teaching to the test" is usually condemned, exactly as one of the disadvantages of accountability policies. The question is whether it is possible to indicate under which conditions "teaching to the test" could be considered as a legitimate and efficient way of enhancing OTL. The ideal type mechanism would be that teachers, on the basis of the information about high stakes tests, would become better informed about which content areas and targeted psychological operations, should be prioritized in teaching and which textbooks should be chosen. Additional benefits could arise when formative assessment would be aligned to the content dimensions of high stakes tests. Such formative assessments could be used to diagnose student progress, provide input for adaptive teaching and evaluate instruction.

When considering how close to reality this ideal type situation is, pitfalls and essential pre-conditions should be examined in more detail. Some of these have to do with characteristics of tests, others with appropriate use by teachers and schools.

In order to provide a good basis for instructional alignment tests should be standard based, "criterion referenced" rather than norm referenced. The structure of the test, i.e. the hierarchy of sub-domains, topics and sub-topics, as well as required performance levels, should be made transparent. Ideally large sets of items (item banks) should be available, at least part of them public and available to schools. Popham (2003) concludes that the like of these conditions were only sub-optimally met in the USA, as he noted that high stakes tests issued by separate states, were often not well aligned with national standards. He also observed that state tests developed by content experts tended to be "overloaded" and insufficiently informative about core knowledge and skills. According to Popham "the curricular intensions handed down by states and districts are often less clear than teachers need them to be for purposes of day-to-day instructional planning". Popham (2001) stresses the importance of the transparency of high stakes tests in the following way: "policymakers … should be educated …to support only high-stakes tests that are accompanied by accurate, sufficiently detailed descriptions of the knowledge or skills measured. A high-stakes test unaccompanied by a clear description of the curricular content is a test destined to make teachers losers. Moreover, because of the item-teaching that's apt to occur, tests with inadequate content descriptors also will render invalid most test-based interpretations about students".

When it comes to the way teachers would ideally make use of test information they should aim for "teaching towards test represented targets, not towards tests" Popham 2003, 17). In other words teachers should capture the core content areas and performance levels embedded in the tests, which stresses the importance of transparency of the test framework; the hierarchy of sub-domains, topics and sub-topics. Ehren et al. (2016) provide empirical evidence from the UK, which shows that teachers' interpretation of core-domains in high stakes tests differed from the interpretation of the test-developers. Perhaps this result should be seen as a further underlining of the call for test transparency. In addition to content alignment, test preparation may also include providing exercise for students in applying different kind of item formats.

The issue of separating legitimate and illegitimate test preparation is addressed most directly by Popham (1991), and his reasoning is cited in some detail below. Popham proposes two kinds of criteria:

"Professional Ethics: No test-preparation practice should violate the ethical standards of the education profession.

Educational Defensibility: No test preparation practice should increase students' test scores without simultaneously increasing student mastery of the content domain tested".

He then describes 5 ways of aligning teaching to tests:

1. *Previous-form preparation* provides *special* instruction and practice based directly on students' use of a previous form of the actual test. For example, the teacher gives students guided or independent practice with earlier, no longer published, versions of the same test.
2. *Current-form preparation* provides *special* instruction and practice based directly on students' use of the form of the test currently being employed. For example, the teacher gives students guided or independent practice with actual items copied from a currently used state-developed high school graduation test.
3. *Generalized test-taking preparation* provides *special* instruction that covers test-taking skills for dealing with a variety of achievement test formats.
4. *Same-format preparation* provides *regular* classroom instruction dealing directly with the content covered on the test, but employs only practice items that embody the same format as items actually used on the test.
5. *Varied-format preparation* provides *regular* classroom instruction dealing directly with the content covered on the test, but employs practice items that represent a variety of test item formats. For example, "if the achievement test uses subtraction problems formatted only in vertical columns, the teacher provides practice with problems presented in vertical columns, horizontal rows, and story form." (Popham 1991, 13–14)

Popham concludes that three of these strategies are not-acceptable. "Previous form preparation is considered educationally unethical because it is aimed at increasing test scores, without furthering student content mastery in a more general sense. Current-form preparation would mostly be considered as professionally and

educationally unethical, and be considered outright as cheating. Same-format preparation is considered educationally inappropriate because it may raise test scores at the cost of students' capacity to generalize what they have learned". Generalized test taking preparation, and varied-format preparation are considered as legitimate strategies, as these strategies train for more generalized skills than the specific test in question.

When Popham empirically investigated whether teachers agreed on his identification of acceptable and non-acceptable test preparation he found that teacher were more lenient, particularly with respect to same format preparation and to special instruction to students "with actual items copied from a currently used" test. Given these results it would appear that deterring teachers from inappropriate forms of test preparation remains a point of concern, although one that could be effectively countered by test quality, more specifically the application of item banks. Together with the empirical findings from the study by Ehren et al. (2016), which pointed out that teachers may have difficulty in inferring the core content from high stakes tests correctly, Popham's results show that appropriate test preparation is not a "run race" and deserves special attention, in contexts like teacher training and applied research.

Finally, an additional strategy for enhancing OTL and aligning teaching to high stakes tests should be mentioned. This strategy consists of considering formative assessments, based on either externally developed or teacher constructed tests, as an effective linking mechanism. In the case study on Dutch education such approaches are illustrated, particularly in the "achievement oriented work" approach (Visscher 2015). A pre-condition is that the formative tests are well-aligned with the relevant high stakes tests and examinations. Another example from the Netherlands, developed primarily for secondary education, but also applicable in other school sectors, is the RTTI approach (Drost and Verra 2015).

Implications for Further Research

While we started out with the statement that the core idea of OTL is almost provocatively simple, in referring to the correspondence between taught and tested content, the conceptual analysis showed that, when seen as part of the larger issue of systemic alignment in education, matters appear to be more complex. When systemic alignment is the issue there are many components that need to be aligned: national standards, standards at intermediary level (state, district, schools), textbooks, assessment programs and actual teaching. Particularly in less centralized educational structures, these components tend to be autonomous and loosely coupled. This makes the alignment issue relatively complex. A key issue is what one might indicate as the curricular validity of high stakes tests and examinations, i.e. a valid representation of state standards by the test. Next, when the potential of high stakes test to effectively and legitimately help schools and teachers to focus their instruction is considered, it was noted that transparence of the test design and

hierarchically ordered content of the tests is a key condition, which may be insufficiently realized in practice. Apart from seeing test preparation as a legitimate way to enhance OTL, it is also a common practice in which less efficient and less legitimate forms cannot be ruled out. Optimizing test preparation is not just a way to improve education, but also a way to avoid and deter from bad practice.

The part of this study that was dedicated to research review indicated that OTL should be considered as having a small, but relative to other levers for improving educational performance, still educationally significant effect on student achievement. Comparable to some other effectiveness enhancing mechanisms, but perhaps smaller than leading authors on OTL effects usually suggest. As was the case with other reviews and meta-analyses on school effectiveness enhancing conditions (Scheerens 2016), there existed large heterogeneity among studies, as far as effect sizes were concerned, but also in the way OTL was operationalized, and studies were conducted. The relative strength of keeping OTL on the agenda in educational policy and practice, but also in educational research, is that the "theory in practice" of how OTL operates and can be enhanced is relatively transparent. There are key-roles for test developers and teachers. Ideas for further research are the following:

1. Given the small scope of this study the emphasis was on studies that had used OTL as the core identifier. We had to keep the analyses of studies that were concentrated on test preparation limited. Even though we identified some relevant studies a logical next step to the current study would be a review (of similar scope as the current one) fully dedicated to test preparation.
2. In this study legitimate test preparation came out as an interesting option for optimizing OTL. The quality of the tests or examinations is quite central for such a perspective on optimizing OTL. As a follow-up study it would be very interesting to analyze the specific criteria examinations or high stakes tests in general would have to meet, in order to be fit to play this leading role. Criteria that were discussed in this report are "curriculum validity", criterion rather than norm-referenced testing, transparency of the test structure, and large sets of items, possibly item banks. Next examinations and high stakes tests used in the Netherlands and one or two other countries, could be analyzed on the basis of these criteria, and empirical data could be collected to explore to what extent teachers in these countries actually use the high stakes tests and examinations to focus their curricular choices.
3. The surprisingly modest effect size of OTL on student achievement that was found in this study suggests a need for more fundamental research on the way OTL is measured. One way to address this is to collect data on OTL from various perspectives: teacher reports (like in TIMSS), student perception (like in PISA, but preferably more detailed), classroom observations and logbooks. The degree of correspondence between various perspectives would provide useful information. Most valid information would probably be obtained through classroom observations and (perhaps) logbooks. On the other hand, teacher and student questionnaires on OTL are much easier to administer. Only if more

demanding methods (observations and logs) are much more valid than information obtained through questionnaires, would it make sense to disregard questionnaire data. As far as student perceptions on OTL are concerned, it seems possible that they are confounded with cognitive ability, prior knowledge and effort. Fast learners, students with more prior knowledge are the ones that work hard and may be more likely to report that a topic was covered than other students. With regard to teacher data, social desirable answers may be a source of bias. An advantage of students' perceptions is that the degree of agreement in answers within classes can be assessed.

4. In the Dutch context it would be very interesting to empirically investigate alignment through content analysis of sources covering components like: reference levels, textbook coverage, formative tests, and formal high stakes tests and examinations. A specific focus on the quality of examinations could be a study in itself. In such a study quality criteria for examining examinations, existing forms of quality control, by the educational Inspectorate and accreditation agencies could be reviewed and strong and weak aspects identified.
5. Perhaps as a replication of the study conducted in England, by Ehren and others about "The Nature, Prevalence and Effectiveness of Strategies Used to Prepare Pupils for Key Stage 2 Maths Tests", an empirical investigation could be made on the way Dutch teachers apply cues from high stakes tests in the Netherlands, in their teaching and classroom assessment practices.
6. As the case study on curricular alignment in the Netherlands showed, there are quite a few examples of advanced test application to enhance student learning. One of these projects could be described in depth, starting out from the conceptual framework developed in this report. An interesting case study might be the RTTI approach by Docentplus (Drost and Verra 2015) in secondary education. A strong focus could be given to the way teachers go about test development and application, and how this affects their teaching.

References

Appelhof, P. (2016). OTL in de Nederlandse onderwijspraktijk: Bevordering van de gelegenheid tot leren in het basisonderwijs, in het bijzonder bij het rekenonderwijs. OTL in Dutch education. In J. Scheerens (Ed.), *Opportunity to learn, instructional alignment and test preparation: A research review*. Utrecht: Oberon.

Bishop, J. (1997). *The effect of national standards and curriculum-based exams on achievement*. Cornell University. Center for advanced Human Relations Studies.

Drost, M., & Verra, P. (2015). *Handboek RTTI*. Bodegraven: Docenplus.

Ehren, M., Wollaston, N., Goodwin, J., & Newton, P. (2016). *Teachers' backward-mapping of patterns in high stakes math tests*. London: London Institute of Education.

Groot, A. D. (1986). *Begrip van evalueren*. 's-Gravenhage: Vuga.

OECD (2014). *PISA 2012 results: Vol. IV. What makes schools successful? Resources, policies and practices.* Paris: OECD Publishing.

Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis, 36*, 399–416.

Popham, W. J. (1991). Appropriateness of teachers' test-preparation practices. *Educational Measurement: Issues and Practice*, Winter.

Popham, W. J. (2001). Teaching to the test. *Educational Leadership, 58*(6), 16–20.

Popham, W. J. (2003). *Test better, teach better: The instructional role of assessment*. Alexandria, Virginia: ACSD.

Sahlgren, G. H. (2015). *Real finish lessons: The true story of an education superpower*. Surrey: Center for Policy Studies.

Scheerens, J. (2016). *Educational effectiveness and ineffectiveness. A critical review of the knowledge base*. Dordrecht, Heidelberg, New-York, London: Springer.

Scheerens, J., Luyten, H., Glas, C. A., Jehangir, K., & Van den Bergh, M. (2014). *System level indicators. Analyses based on PISA 2009 data*. Internal Report. Enschede: University of Twente.

Vanlommel, K., Vanhoof, J., & Van Petegem, P. (2016). Data use by teachers: The impact of motivation, decision-making style, supportive relationships and reflective capacity. *Educational Studies.*

Visscher, A. J. (2015). *Over de zin van opbrengstgericht(er) werken in het onderwijs*. Groningen: RU, Faculteit der gedrags-en maatschappijwetenschappen.

Willms, J. D., & Somers, M.-A. (2000). *Schooling outcomes in Latin America*. Report prepared for UNESCO-OREALC and the Laboratorio Latinoamericano de la Calidad de la Educación [The Latin American Laboratory for the Quality of Education].

Woessmann, L., Luedemann, E., Schuetz, G., & West, M. R. (2009). *School accountability, autonomy and choice around the world*. Cheltenham, UK/Northampton, MA, USA: Edward Elgar.