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

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.