The Effects of Two Intervention Programs on Teaching Quality and Student Achievement

S.N. Azkiyah
UIN Syarif Hidayatullah Jakarta, Indonesia

Simone Doolaard
Bert P. M. Creemers
M.P.C. (Greetje) Van Der Werf
University of Groningen, The Netherlands

ABSTRACT

This paper compares the effectiveness of two interventions aimed to improve teaching quality and student achievement in Indonesia. The first intervention was the use of education standards, while the second one was the combination of education standards with a teacher improvement program. The study involved 50 schools, 52 teachers, and 1660 students. Teaching quality was measured through classroom observations and student achievement was assessed with a performance test. The multilevel analysis showed student achievement improved only in the second intervention group, whereas the analyses of difference and effect size showed teaching quality in the same group improved only in the first half of the intervention program.

INTRODUCTION

Student achievement in Indonesia is low, as indicated by both national and international measurements (e.g. Jazadi, 2003; Mohandas, 2004). A number of strategies have been adopted in Indonesia to improve education; the most recent is the establishment of education standards, following the wide use of education standards across countries in the world (Choi, de Vries, & Kim, 2009; Neumann, Fischer, & Kauertz, 2010). Next to the accountability purpose, education standards have been argued to serve as guidance for improvement purposes.

Through the provision of how students are expected to perform, regardless of their background, education standards are argued to guide teachers to change their instruction to a more focused and improved form of teaching and learning, and eventually to address the needs of different students to achieve the same minimum goals (e.g. Chambers & Dean, 2000; Dowson, McInerney, & Van Etten, 2007). During the pre-standards era, different teachers for the same subjects and grades may raise different expectation while the same individual teachers may raise different expectation for different students.

However, in both Indonesia and in other countries, not much research, in particular experimental research, has been carried out to examine the effectiveness of education standards. Although limited, available research, especially in the US, has demonstrated that these standards only work for certain groups in which minority and disadvantaged remained left behind (e.g. Lauer et al. 2005; US Department of Education, 2008). Hence, the formulation of education standards does not necessarily tell us how to improve education. There have been studies conducted that could potentially explain the ineffectiveness of standards-based education. The first problem refers to

the language of the standards documents, which is broad and global (e.g. Choi et al., 2009; Dowson et al., 2007). The standards are vague and ambiguous, while their content lacks a sufficient level of grounding (Finn, Petrili, & Vanourek, 1998; Gandal, 1996). The second problem that needs to be addressed is the minimum attention paid to teachers as one of the main actors in education: the absence of adequate guidance to implement the standards in the classroom. Teachers merely receive the copies of the documents and are asked to discuss with their colleagues whether the content corresponds with what they have been teaching (Chambers & Dean, 2000) and to formulate the subject's curriculum across the different grades (Mathison & Freeman, 2008).

In this context, it is highly relevant to look at the results of Educational Effectiveness Research (EER), which offers theory driven and evidence-based information on what works in education, and which could serve as an additional knowledge base for improvement initiatives. Within EER, teachers have been widely found to play a major role (e.g. Luyten & Snijders, 1996; Van Der Werf, Creemers, De Jong, & Klaver, 2000). As argued by Doolaard (1999), the superiority of teacher above the higher levels such as schools could be the fact that classroom teaching and learning process are primary sources of learning.

After controlling for student background characteristics, teacher instruction has been largely recognized as the most important aspect, both in terms of quantity and quality (e.g. Brophy & Good, 1986; Muijs & Reynolds, 2011; Wang & Walberg, 1991). Teacher instruction in this case refers to teacher efforts in maximizing student opportunity to learn (Creemers, 1994). Several activities have been found to provide learning students with opportunities, such as an emphasis on academic goals and achievement (e.g. Cotton, 1995; Doyle, 1986), clear and step-wise presentation of materials as well as effective questioning and feedback (e.g. Scheerens & Bosker, 1997), and clear classroom structures and routines (Brophy & Good, ibid).

In order to contribute to the development of EER especially with respect to teacher effectiveness, Creemers and Kyriakides (2008) developed the dynamic model of educational effectiveness. Several studies have tested the validity of the dynamic model, especially in Cyprus (e.g. Kyriakides, Creemers, Antoniou, & Demetriou, 2010). Moreover, experimental studies in which a classroom-intervention was based on this model showed an increase in teaching quality and student performance (Antoniou, 2009; Antoniou & Kyriakides, 2013).

The above discussion leads to two conclusions. Firstly, education standards could facilitate the development of a shared vision on teaching and learning, which has to advocate higher expectations for all students. However, in regards to the broad and global language used in standards documents, it is essential to study if the standards would be more effective when they were written in more specific and concrete language. Secondly, we know from EER the superiority of teacher or classroom level in improving student outcomes, which implies the importance of assisting teachers in improving their teaching quality, especially with respect to the factors found to be associated with student outcomes.

THE CURRENT STUDY

We hypothesized the combination of both education standards and a teacher improvement program would provide better results than the standards-based education on its own, as has been the case in Indonesia. For this reason, two interventions with two supporting documents were developed. The first one was an elaborated standards document, while the second document included the characteristics of effective teaching, as defined by EER, especially the classroom factors in the dynamic model (Creemers & Kyriakides, 2008).

The first intervention group was provided with the elaborated standards document and the second one was offered both the elaborated standards and a document containing the characteristics of effective teaching. The elaborated standards document here referred to the standards of content in Indonesia as the context of the study and was an attempt to make the general and broad competencies in the government standards clearer, more specific and thereby easier to measure by defining the listed competencies in the standards. In addition, the second group also participated in a teacher improvement program, which was based on the classroom factors of the dynamic model (Creemers & Kyriakides, 2008). More details of the two interventions are explained in the method section.

To investigate the impact of these two intervention conditions, we compared the changes in teaching quality and student achievement to a control group. Following EER, the background characteristics at the student, teacher, and school levels were collected to examine their influence and to obtain more precise estimates of the intervention effects. We hypothesized that both interventions would improve teaching quality and student achievement, but intervention 2 would be more effective than intervention 1. This study was conducted in Indonesia and the focus was on English (reading) because the student attainment rates on the national exams for this subject have been the lowest throughout the years.

Table 1

The Groups and the Interventions

Group	Intervention
1. Experimental 1	The elaborated standards document; the teachers were free to develop their own stragies in implementing and achieving the standards.
2. Experimental 2	The elaborated standards document and a teacher improvement program (teacher training).
3. Control Group	No intervention, teachers used the standard document available for the government.

METHOD

Research Design

The study was a randomized experimental design, in which two interventions were developed to observe their effects, in comparison with a control group (Shadish, Cook, & Campbell, 2002). The participants voluntarily participated and were randomly assigned into the three groups as described in Table 1. The study was conducted within one school year (July 2010 – May 2011), during which a pre- and post- test of students' achievement was conducted, and teaching quality was measured three times.

Sample

The focus of this study was Madrasah Tsanawiyah (MTs), a type of junior secondary education (3 years after 6 years of primary schooling; students' ages range from 12/13 to 14/15 years) under the Indonesian Ministry of Religious Affairs (MORA). Most madrasahs are private, small, attended by students from low-income families, and generally provide education of a lower quality compared to general schools under the Indonesian Ministry of Education (MONE) (ADB, 2006), and therefore improvement is urgently needed.

The research was limited to two neighbouring provinces: DKI Jakarta and Banten. 57 schools, including a total of 59 teachers (M = 44%, F = 56) and 2,431 students (M = 48,5%, F = 51,5%) voluntarily participated in this study. Throughout the year, the number of participants decreased. In this paper, 52 (experimental 1 = 17, experimental 2 = 18, and control group = 17) teachers and 1660 students from 50 schools were retained in the analysis. Information regarding the schools, teachers, and student levels (the details are explained in the sub-section of research instruments and procedures) were gathered and the randomization in this study was generally successful. Only small differences were observed with respect to school accreditation and teacher age, which were neither reflected by the school scores on the English national exam nor by the extent of teaching experience, respectively. At the student level, differences occurred only in the paired comparison between experimental group 2 and the control group with respect to fathers graduated from university and fathers and mothers working as professionals. However, the percentages represented by these items were only small (N $\leq$ 11%).

The Content of Interventions.

The two interventions were aimed at improving student achievement and teachers' teaching quality. The first intervention group was provided with the elaborated standards document, which explained the reading competencies defined in the government standards in more detail. The elaborated standards specifically referred to how the standards' content relates to performance in the Indonesian context. It was expected to provide teachers with a clearer understanding of what to plan, teach, and assess within their classroom. Teachers in both experimental conditions (experimental group 1 and experimental group 2) attended a one-day introductory workshop to discuss the elaborated standards document. Teachers in these groups were advised to find various strategies that enabled them to help different students to achieve the targeted goals.

The second intervention group incorporated a combination of the elaborated standards and a teacher improvement program. In this intervention condition, an additional booklet titled "Becoming an effective teacher of English" was used. The document discussed the characteristics of effective teaching as defined by the classroom factors in the dynamic model, linking these factors with the specific reading skills as mentioned in the elaborated standards. In addition to attending one day workshop on the elaborated standards document, teachers in this group attended one day workshop on effective teaching and six half-day monthly meetings to further discuss the two documents and to advise teachers

on how to implement them in their classroom. The first three meetings focused on orientation, structuring, teaching modelling, the development of application tasks, questioning, and assessment in relation to the competencies defined in the elaborated standards document. The last three meetings concentrated in lesson plan development and peer teaching.

Research Instruments and Procedures

As already indicated, the variables in this study were teachers' teaching quality and student performance on English reading comprehension. Data on teaching quality were collected through classroom observations. The observation instrument consisted of 52 items and was provided on 1 – 5 Likert scale to indicate the frequency (not at all – a great deal) and the quality (minimum – maximum) of the activities observed. It was a modification of the observation instrument used in Cyprus (Creemers & Kyriakides, 2008) to adjust the specific subject in this study and the context of Indonesia. The observation was conducted by a team (five observers in total), with one observer per class observation. The team had been previously trained in using the instrument and the inter-rater reliability was good (generalized Kappa = .72). In line with the scores of most teachers, the scales were re-coded using a rating scale of one to three. A Confirmatory Factor Analysis showed a poor model fit, which also applied to the Exploratory Factor Analysis. Therefore, it was decided to base the main analysis on one scale, which reliability throughout the measurement was excellent ($\alpha = > .9$). The reliability of the subscales appeared to be wide-ranged ($\alpha = .49$ to .93).

The students' performance was measured by a pre-test and a post-test. The reading comprehension test level A2 of the Central Institute for the Development of Tests (CITO) was modified and shortened to match the reading skills described in the elaborated standards document into two rather similar versions. Both versions (pre- and post- test), each including 20 items that were significantly correlated ($r = .275$, $p = .01$), whereas their reliability was modest ($\alpha = .52$) and moderate ($\alpha = .62$), respectively. In the final pilot study, however, the 20 items had a good reliability ($\alpha = .74$).

The data on background characteristics were gathered at the student, teacher, and school levels through multiple choice questionnaires. At the student level, the variables included gender and student socio-economic status (SES), the latter of which included the father's and/ or mother's most recently achieved education and job obtained. At the teacher level, the variables were gender, age, academic degree and the scope of the teaching experience. The variables at the school level contained the status of accreditation, school size, and the mean score on the English national exam, year 2010/2011.

Concerning the procedures, the activities consisted of introductory workshop, 3 measures, and monthly meetings. The introductory workshop occurred of two days. The first day (July 16, 2010) discussed the elaborated standards for both experimental groups. The second day of the workshop (July 17, 2010) was only available to teachers assigned to the second intervention group and introduced the effective teaching program. The first was assessment took place in August 2010, the second in January 2011, and the last in May 2011. The monthly workshop specifically held for the experimental group 2 took place from September 2010 to April 2011.

Data Analysis

Descriptive statistics on the raw mean score of teaching quality over measurements were computed. This analysis was completed with significance tests among the groups (two-tailed, $p < .01$ or $p < .05$, depending on whether student or teacher data were involved) and effect size analysis. In addition, the relationship (r) between the background characteristics and the variables were investigated to determine characteristics to be retained in the multilevel analysis (criterion $r = $ minimally .20). The significant tests and effect size (Cohen's d) analysis were used to test the effects on the interventions on teaching quality. Multilevel analysis (with MLwiN: Rasbash, Charlton, Browne, Healy, & Cameron, 2005) was applied to address the effects of the interventions on student achievement, in which two dummy variables for both interventions were created, with the control group serving as the contrast group. In testing the effects of the interventions the following significance levels (one-tailed) were accepted: $p < .01$ for the data at the student level, and $p < .05$ for the data at the teacher level (because of the small number of teachers).

RESULTS

The Effects of the Interventions on Teaching Quality

First of all, the scales of the observation instrument were reduced from one to five, to one to three because of the small number of teachers who scored four and five. Table 2 reports the mean score of teaching quality in each group from measurement to the next measurement and the statistical analyses employed, i.e. ANOVA test, Bonferroni test, paired t-test and effect size analysis.

At this point, it is important to note that for measurement one, the significance test showed no

Table 2

The Mean Score, the Standard Deviation, and the Effect Size of Teaching Quality Based on the Observer Data collected during Three Measurements

	M1		M2		M3		Effect Size		
Group	Mean	SD	Mean	SD	Mean	SD	M1 to M2	M2 to M3	M1 to M3
Teaching Quality									
Exp. 1	1.80	.26	1.74	.36	1.76	.22	-.19	.07	-.17
Exp. 2	1.94	.37	2.10	.34	2.13	.35	.44*	.09	.53**
Cont.	1.84	.34	1.82	.30	1.91	.23	-.06	.33	.24
Total	1.87	.31	1.90	.37	1.94	.31	.09	.12	.22
ANOVA	$F(2, 98) = .98$		$F(2, .611)= .5.34**$		$F(2, .84) = 3.74**$				
Bonferroni test (mean difference (MD) / (SD)									
Exp 1 and Exp 2		-.14(.11)		-.35(.11)*		-.37(.09)*			
Exp 1 and Cont.		-.04(.11)		-.08(12)		-.15(.10)			
Exp 2 and Cont.		.10(.10)		.27(11)*		-.22(.09)*			

Notes. M = measurement, Exp 1 = experimental 1, Exp 2 - experimental 2, Cont = control group, * *p <.05* (1 tailed), ** *p <.01* (1 tailed)

differences in teaching quality among the three groups and within the paired comparisons (Bonferroni test) of the groups. As expected, significant differences occurred in measurement two and three. As regards these two measurements, the Bonferroni test marked significant differences in the paired comparisons between experimental group 2 and experimental group 1, which also applied to experimental group 2 and the control group. Table 2 also presents the effect size and paired t-test of each group from measurement to the next measurement, from which it is known that while both experimental group 1 and the control group remained stable throughout the three measures experimental group 2 improved significantly from measurement one to two but remained stable from measurement two to three.

Furthermore, Pearson correlation analysis (r) was also performed to test whether differences in teaching quality were related to the background characteristics at both teacher and school levels. The results showed that teaching quality did not sufficiently correlate to the characteristics at both school and teacher levels.

In summary, while no difference was found in the first measurement, significant differences were found in measurement two and three in which experimental group 2 was superior. With respect to the trend, while stability was observed in both experimental group 1 and the control group throughout the three measurements, an improvement was noticed in experimental group 2 from measurement one to two. Thus, the hypothesis that teaching quality in experimental group 1 was better than the control group was rejected while that of the experimental group 2 was accepted. However, teaching improvement in experimental group 2 occurred only in the first half of the program.

The Effects of the Intervention on Student Achievment

In analyzing student achievement, an analysis of covariance (post-test) while controlling for prior achievement (pre-test) was used (see Table 3). Different explanatory variables were introduced for estimating the effects of the interventions resulted from Pearson correlation analysis (r). They included province, school score on English, father's education, and student gender.

After the empty model, the first two models tested the effects of the covariates at the student and school levels separately to estimate the unique variance on each of the levels. The third model combined the effects of both levels whereas model 4 tested the effects of the interventions while controlling for the covariates tested in model 3. The first three models were compared with the empty model and model 4 was compared to model 3 to see how much more variance the two interventions could explain.

The variance of the post-test score at the student level (67%) was twice as much as the variance at the school level (33%). Student prior achievement and gender had significant effects: students who achieved a high score in the pre-test also achieved a high score in the post-test and female students achieved higher than the male students. Surprisingly, SES as represented by father's education, was not significant at p < .01. At the school level, province and a high school prior achievement (8.1 – 9) were found to have significant effects. Students in DKI Jakarta, the capital city of Indonesia, scored better than those in Banten and students in schools with a high score on the English national exam achieved better than students in schools with a low score on the English national exam. Referring to model 3, these covariates explained 16% of the variance.

Next, the two interventions were added. The results showed that the second intervention group had better student achievement rates, which were significant at p < .01 (1-tailed)[1]. Although the students in the first intervention had also improved their achievements, this result was only significant at p < .05 (1-tailed). The model fit as displayed in the decrease in deviance (9.009, df=2)

[1] Learning gain analysis was also performed, which results also showed that experimental group 2 was associated with better performance gain.

Table 3

The Results of the Multilevel Covariance Analysis

	Empty		Model 1 (Student)		Model 2 (Student)		Model 3 (Student + School)		Model 4 (Intervention)	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Fixed Part										
Intercept	9.468**	.270	7.872***	.401	9.584***	.491		.563	7.389***	.609
Student level										
Pretest			.149***	.028			.146***	.028	.146***	.028
Gender (female)			.442***	.153			.395***	.155	.397***	.155
Father education (JHS)			-.186	.216			-.211	.216	-.223	.216
Father education (SHS)			-.055	.209			-.114	.212	-.136	.212
Father education (univ)			.684	.333			.560	.339	.510	.339
School										
Province (Banten)					-1.976***	.484	-1.955***	.484	-1.648***	.444
School Nat. Exam (med.)					.907	.551	.691	.539	.515	.503
School Nat. Exam (high)					2.791***	.924	2.462***	.892	1.877***	.862
Intervention										
Intervention one									.904	.545
Intervention two									1.679**	.537
Random Part										
School level variance										
Intercept	3.875	.784	3.457	.715	2.703	.574	2.423	.531	2.003	.450
School level variance										
Intercept	7.773	.275	7.476	.291	7.639	.275	7.317	.292	7.316	.292
Deviance (-2*loglikelihood)	8268.439		6815.688 1452.751**		7884.852 383.587***		6455.230 1813.209***		6446.221	
Decrease in deviance									9.009***	
Variance explained			0.06		0.11		0.16		0.04	

** *p <.01* (1-tailed), *** *p <.01* (2-tailed)

showed that model 4 was a better model than model 3. The two interventions explained 4% more variance, which made model 4 explain 20% of the total variance. In short, similar with that of teaching quality, the hypothesis that student achievement was better in the second intervention compared to both the first intervention 1 and the control group was met whereas that of the first intervention was rejected.

CONCLUSIONS

The above research findings indicate that only the second intervention (i.e. the combination of education standards and a teacher improvement program) was found to have significant effects on both teaching quality and student achievement. The teachers in experimental group 2 were found to have better student achievement outcomes and to be more effective teachers as their teaching quality improved during the first half of the intervention period. Thus, only the combination of the elaborated standards and the teacher improvement program had produced significant effects.

This study has proved that education standards on its own did not lead to the improvement of either teaching quality or student achievement. On the other hand, the combination of education standards and a teacher improvement program has shown to be more effective. These findings could imply that when teachers participate in effective teacher development program, their teaching quality will improve which eventually is expected to affect student achievement. Avalos (2011) for instance reviewed publications in *Teaching and Teacher Education* over eleven years (2000 – 2011) on teacher development programs and concluded that most studies reviewed show some form of impact of professional development on teachers' knowledge and practice as well as student achievement. Nevertheless, an improvement of teaching quality in this study was found only in the first half of experimental group 2's program. The materials presented in the second intervention might explain why an improvement happened in the first period while stability took place in the second half of the program. The first half of the program (introductory workshops and the first three monthly meetings) conducted before measurement two focused on the eight factors of the Dynamic Model.

The content during the first half of the program likely facilitated teaching improvement in experimental group 2 from measurement one to two. In the second half of the program, the other three monthly meetings attempted to provide further chance for teachers to deepen their understanding of the elaborated standards document and the eight classroom factors of the dynamic model. This idea was accommodated through the development of lesson plans (2 meetings) and peer teaching (1 meeting), which were developed together with teachers in experimental group 2 due to their request. It has to be admitted that the last three meetings did not offer something new with regards to the classroom factors of the dynamic model, which could be the reason why no further improvement was found from measurement two to three.

Furthermore, this study has also demonstrated both the possibility and the complexity of the integration of EER and improvement initiatives. The findings of EER are helpful in determining the direction of improvement. In this study, the classroom factors of the dynamic model were used in developing a teacher improvement program in the second intervention. However, we also faced problems showing the complexity of such integration. For instance we do not know whether the student achievement in experimental group 2 was due to the combination of the elaborated standards and the teacher improvement program, or if it was solely the result of the teacher improvement program. Thus, developing improvement measures is not as easy as identifying factors that work for education and applying them.

DISCUSSION

Suggestions for Future Research

Firstly, it is worth emphasizing the relevance of including the developmental stages of teaching quality. The inclusion of this developmental stage may also address the problem of stability of the second intervention group after measurement two, which was likely to be related with the materials presented during the second half of the intervention program. Specific teaching skills in the next level can help focusing the content or the materials of the intervention. The study in Cyprus carried out by Antoniou (2009), which applied stages of teaching quality suggested by the dynamic model, has resulted in improvement of both teaching quality and student achievement.

Secondly, in regards the question whether the better student achievement in experimental group 2 was the result of the combination of the elaborated standards and the teacher improvement program or only of the teacher improvement program, future research may add another group, which is only focused on a teacher improvement program. Next, future research may opt to include a longer period of intervention, possibly containing more than three points of measurement.

REFERENCES

Antoniou, P. (2009). Using the dynamic model of educational effectiveness to improve teaching practice: building an evaluation model to test the impact of teacher professional development programs. Unpublished PhD thesis. University of Cyprus, Department of Education, Nicosia, Cyprus.

Antoniou, P., & Kyriakides, L. (2013). A Dynamic Integrated Approach to Teacher Professional Development: Impact and Sustainability of the Effects on Improving Teacher Behavior and Student Outcomes. Teaching and Teacher Education, 29(1), 1-12.

Asian Development Bank (ADB) (2006). Indonesia: madrasah education development project. Project Number: 37475-01. Jakarta: SMEC International Pty Ltd.

Avalos, B. (2011). Teacher professional development in teaching and teacher education over ten years. Teaching and Teacher Education, 27(1), 10-20.

Brophy, J.E., & Good, T.L. (1986). Teacher behavior and student achievement. In Wittrock , M.C. (ed.), Handbook of research on teaching (3rd edn, pp. 328 -375). New York: Macmillan.

Chambers, L., & Dean, C. (2000). Noteworthy perspectives on implementing standards-based education. Washington D.C.: Mid-Continent Research for Education, and Learning.

Choi, D., de Vries, H.J., & Kim, D. (2009). Standards education policy development: Observations based on APEC research. Rotterdam; Rotterdam: Erasmus Research Institute of Management, Erasmus University; Erasmus University Rotterdam Host.

Cotton, K. (1995). Effective student practices: A research synthesis. Available at http://www.nwrel.org/scpd/esp/esp95.htm1#1.

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

Creemers, B., & Kyriakides, L. (2008). The Dynamics of Educational Effectiveness. A contribution to Policy, Practice and Theory in Contemporary schools. New York and London: Routledge Taylor & Francis Group.

Doolaard, S. (1999). Schools in change or schools in chain?: the development of educational effectiveness in a longitudinal perspective. Enschede: Twente University Press.

Dowson, M., McInerney, D.M., & Van Etten, S. (2007). The state of play in standards and standards reform. In McInerney, D.M., Van Etten, S., & Dowson, M. (Eds.), Standards in education. Charlotte, NC: Information Age Publishing.

Doyle, W. (1986). Classroom organization and management. In Wittrock, M.C. (ed.), Handbook of research on teaching (3rd edn, pp. 392-431). New York: Macmillan.

Finn, C.E., Jr., Petrilli, M.J., & Vanourek, G. (1998). The State of State Standards. Washington, D.C: Thomas B. Fordham Foundation.

Gandal, M. (1996). Making standards matter, 1996: an annual fifty state report on efforts to raise academic standards. Washington, DC: American Federation of Teachers.

Jazadi, I. (2003). An investigation of current constraints and potential resources for developing learner-centred curriculum frameworks for English language at high schools in Lombok, Indonesia. Unpublished PhD thesis. Adelaide, Australia: Division of Education, Arts and Social Sciences, University of South Australia.

Kyriakides, L., Creemers, B., Antoniou, P., & Demetriou, D. (2010). A synthesis of studies searching for school factors: Implications for theory and research. British Educational Research Journal, 36(5), 807-830.

Lauer, P.A., Snow, D., Martin-Glenn, M., Van Buhler, R.J., Soutemyer, K., & Snow-Renner R., (2005). The influence of standards on K-12 teaching and student learning: A research synthesis. Regional Educational Laboratory.

Luyten, J.W., & Snijders, T.A.B. (1996). School effects and teacher effects in Dutch elementary education. Educational Research and Evaluation, 2(1), 1-24.

Mathison, S., & Freeman, M. (2008). Teachers working with standards and state testing. In Mathison, S., & Ross, E.W. (Eds.) The nature and limits of standards-based reform and assessment: Defending public schools. New York etc.: Teachers College, Columbia University.

Mohandas, R. (2004). Overview of Education for Sustainable Development (ESD) in Indonesia. Jakarta: National Office for Educational Research and development, The Ministry of National Education.

Muijs, D., & Reynolds, D. (3rd ed) (2011). Effective teaching: Evidence and practice. London: SAGE.

Neumann, K., Fischer, H. E., Kauertz, A. (2010). From PISA to Educational Standards: The Impact of Large-Scale Assessments in Germany. International Journal of Science and Mathematics Education, 8(3), 545-563.

Rasbash, J., Charlton, C., Browne, W. J., Healy, M., & Cameron, B. (2005). MLwiN Version 2.0. Centre for Multilevel Modeling, University of Bristol.

Scheerens, J., & Bosker, R. J. (1997). The foundations of educational effectiveness. Oxford: Pergamon.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). Experimental and quasi-experimental sedigns for generalized causal inference. Boston, New York: Houghton Mifflin Company.

US Department of Education (2008). A nation accountable: Twenty-five years after "A nation at risk". Washington D.C.: US Department of Education.

Van der Werf, M.P.C., Creemers, B.P.M., De Jong, R., & Klaver, L. (2000). Evaluation of school improvement through an educational effectiveness model: The case of Indonesia's PEQIP project. Comparative Education Review, 44(3), 329–356.

Wang, M., & Walberg, H. (1991). Teaching and educational effectiveness: Research synthesis and consensus from the field. In K. J. Rehage (Series Ed.) & H. C. Waxman & H. J. Walberg (Vol. Ed.), Effective teaching: Current research (pp. 63-80). Berkeley, CA: McCutchan Publishing.

Correspondence regarding this article should be addressed to S.N. Azkiyah from UIN Syarif Hidayatullah Jakarta, Indonesia. Email may be sent to azkiyah@uinjkt.ac.id