Complex interplay effects of classroom instructional and context factors on student growth in English as a foreign language in Vietnam: The case of nonlinear relationship, regularized regression and the relevance of the scaling model

Giang Pham

vom Promotionsausschuss des Fachbereichs Psychologie der Universität Koblenz-Landau zur Verleihung des akademischen Grades Doktor der Philosophie (Dr. phil.) genehmigten Dissertation

Eingereicht am: 04.10.2017

Datum der Disputation: 13.03.2018

Vorsitzende des Promotionsausschusses: Prof. Dr. Melanie Steffens

Erstgutachter: Prof. Dr. Manfred Schmitt

Zweitgutachter: Prof. Dr. Ingmar Hosenfeld

Drittgutachter: Dr. Alexander Robitzsch

Zusammenfassung

Die zentrale Fragestellung dieser Arbeit ist die nach den Effekten unterrichtlicher Faktoren und deren Zusammenspiel auf den Schülerleistungszuwach im Fach Englisch in Vietnam unter Berücksichtigung von Kontextfaktoren. Daneben geht es um methodische Fragestellungen, insbesondere um die Bedeutung der Auswahl eines Skalierungsmodells.

Die verwendeten Daten wurden im Rahmen eines Forschungsprojekts in Vietnam im Schuljahr 2006/2007 erhoben. Neben einem Messwiederholungsdesign mit zwei Messzeitpunkten am Anfang und am Ende des Schuljahres wurde eine Videostudie in der Mitte des Schuljahres durchgeführt. Bei jedem Messzeitpunkt wurden die adaptierten Englisch Tests (C-Test und Hörverstehenstest) und die Fragebögen aus der Studie Deutsch Englisch Schülerleistungen International (DESI) in Deutschland eingesetzt. Zur Gewinnung verhaltensnaher Indikatoren des Unterrichts wurden die Videoaufzeichnungen transkribiert und von trainierten Experten niedrig-inferent kodiert und hoch-inferent geratet. Zur Skalierung der Schülerleistungen wurden mehrere Skalierungsmodelle ausgewählt. Zur Skalierung der Schülerleistung anhand des C-Tests mit einer Testlet-Struktur wurden zwei unidimensionale und zwei Testletmodelle eingesetzt. Um Schülerleistungen anhand des Hörverstehenstests kamen das Raschmodell, das 2PL und das 3PL Modell zum Einsatz. Die Schülerleistungsschätzungen von beiden Messzeitpunkten wurden mithilfe eines gemeinsamen Skalierungsmodells skaliert und miteinander verlinkt. Anschließend wurden die Plausible Values gezogen. Zur Modellierung des Zusammenhangs zwischen den Unterrichtsfaktoren und dem Schülerleistungszuwachs wurden sowohl lineare als auch komplexere Unterrichtseffekte (nicht-linear, additiv, multiplikativ) berücksichtigt. Die anfängliche Leistung und der sozioökonomische Status der Schülerinnen und Schüler werden als Kontextfaktoren betrachtet. Die Analyseverfahren der Wahl waren OLS-Regressionen sowie regularisierten Regressionsmodelle mit lasso (least absolute shrinkage and selection operators).

Die Ergebnisse zeigen hinsichtlich wichtiger fächerübergreifender Merkmale einerseits ein positives Bild der Qualität des Englischunterrichts, aus Sicht der englischen Fachdidaktik jedoch eine mangelde Unterrichtsqualität. Die bedeutsamsten Unterrichtsfaktoren des Schülerleistungszuwachses im C-Test sind Aspekte der Motivierungsqualität sowie der Unterrichtssprache. Für den Zuwachs beim Hörverstehenstest spielten Aspekte der Unterrichtssprache sowie die relative Häufigkeit von Wiederholungsfragen eine wichtige Rolle. Die Hypothesen zu den Unterrichtseffekten wurden durchweg bestätigt. Trotz der Ähnlichkeiten zwischen Schülerleistungsschätzungen anhand verschiedener Skalierungsmodelle hingen die Ergebnisse hinsichtlich der Effekte von Unterrichtsmerkmalen auf den Leistungszuwachs erheblich vom Skalierungsmodell ab.

Abstract

The primary aims of the study are (1) to identify classroom instructional factors which have a crucial effect on the academic growth of ninth-graders in EFL in Vietnam, and (2) to gain insight into their interplay with each other and with context factors. Besides, this study has a strong focus on methodological approaches: (a) using multiple methods in order to deal with the "large *p*, small *n*" problem, (b) to understand the relevance of the scaling model used for the results.

Data from a research project carried out in Vietnam during the school year 2006–2007 were used in this study. Besides a longitudinal design with two measurement points (MPs) using adapted English tests and questionnaires from the DESI-study in Germany, a video study was conducted in the middle of the school year between two MPs. The recorded video data were transcribed, micro-analytically coded, and lessons were rated to gain indicators of classroom instruction. Different IRT scaling models were chosen to estimate student ability in the pretest and posttest. For the C-test, the unidimensional 1PL and 2PL models, the Rasch testlet model, and testlet 2PL model were selected to model student ability. To estimate student ability via the listening comprehension test (LC-test), the Rasch model, the unidimensional 2PL, and 3PL models were applied. The student ability estimates at the two MPs were linked to one common scale using the concurrent calibration approach with different a priori ability distributions. The plausible values (PVs) were generated and treated as student ability estimates for all analyses. To understand the relationship between the instructional variables and student growth, we explored the hypothesized linear and nonlinear, additive and interactive effects of classroom instructional factors. To examine these hypothetical effects, OLS and regularized regression models using lasso (least absolute shrinkage and selection operators) were applied, including main effects as well as quadratic and interaction terms of instructional variables. Initial student ability and the socioeconomic status of students were treated as context variables.

The results show, on the one hand, a positive view of important general instructional quality dimensions of teaching effectiveness and, on the other hand, a strongly teacher-centered and textbook-driven instruction and poor instructional quality from the point of view of EFL didactics. The most important instructional factors of student growth in the C-test were quality aspects of motivation in instruction as well as aspects related to the teaching language. Regarding the LC-test results, language-related aspects together with the relative frequency of repeated questions were the most important predictors of student growth. While the findings confirmed all the hypothesized instructional effects on student growth, aptitude treatment interaction effects of instruction were only confirmed with regard to student growth in the C-test. The different scaling models produced significant differences in the results regarding instructional effects on student growth.

Danksagung

Von ganzem Herzen möchte ich mich bei allen bedanken, die durch ihre fachliche und persönliche Unterstützung zum Gelingen dieser Dissertation beigetragen haben.

An erster Stelle gilt mein Dank Prof. Dr. Manfred Schmitt, Prof. Dr. Ingmar Hosenfeld und Dr. Alexander Robitzsch für die langjährige Unterstützung und für die Bereitschaft, diese Arbeit zu begutachten.

Den ehemaligen Kolleginnen und Kollegen der Arbeitseinheit Entwicklungspsychologie und Bildungsforschung an der Universität Koblenz-Landau – Andreas Helmke, Friedrich-Wilhelm Schrader, Ingmar Hosenfeld, Tuyet Helmke, Susanne Röder, Richard Göllner und Wolfgang Wagner - gilt mein besonderer Dank für den Einstieg in die empirische Bildungsforschung, für die tatkräftige Unterstützung und vielfältige Anregungen sowie für die Zurverfügungstellung von Daten der DESI-Projekte. Ich habe die lernförderliche und harmonische Arbeitsatmosphäre in Landau sehr geschätzt.

In methodischer Hinsicht habe ich sehr stark von den Methodenworkshops des Landauer Graduiertenkollegs «Upgrade» sowie von zahlreichen intensiven Diskussionen mit Alexander Robitzsch profitiert.

Die ehemaligen Kolleginnen und Kollegen am Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens (BIFIE) Salzburg (neben Alexander Robitzsch insbesondere Claudia Luger-Bazinger, Thomas Kiefer, Matthias Trendtel, Robert Fellinger, Roman Freunberger, Ursula Itzlinger-Bruneforth, Takuya Yanagida, Michael Ober) haben mich nachhaltig unterstützt. Ihnen allen danke ich herzlich für die konstruktiven Gespräche sowie für die tolle Zusammenarbeit.

Beim BIFIE Salzburg und dem Institut für Bildungsmanagement und Bildungsökonomie (IBB) an der Pädagogischen Hochschule Zug - ganz besonders Stephan Huber und Guri Skedsmo - bedanke ich mich für die Rahmenbedingungen, die es mir ermöglicht haben, diese Dissertation abzuschließen.

Schließlich möchte ich einen herzlichen Dank an meine Eltern Tuyet und Andreas richten, die mich in allen Belangen immer mit Liebe und Taten unterstützt haben.

Table of Content

List of tables

List of figures

I. Introduction

I.1 General background

Understanding the mechanisms behind the successes and failures of teaching and learning is one of the main interests of interdisciplinary empirical educational research. In this field, researchers from different disciplines, including psychology, pedagogy, didactics, educational sciences, and psychometry, collaborate. After more than half a century of development, various theories, models, research, and methodological approaches have been proposed and numerous empirical studies and meta-analyses have been conducted, which have resulted in a vast number of research findings.

Still, despite or due to the large amount of empirical evidence, throughout the years, one of the most enduring and challenging problems in this field is to find the guiding factors for effective teaching because "*everything seems to work*", as Hattie (2009, p. 1) highlighted at the beginning of his book "*Visible learning – A synthesis of over 800 meta-analyses relating to achievement*". To deal with a list of many relevant factors according to their effect sizes and domains, Hattie (2012) made an effort to synthesize numerous findings. He suggested that higher-level principles of effective teaching and learning make the difference, rather than that any specific teaching method (Hattie, Beywl, & Zierer, 2013, see Chapter IV.1). However, the majority of the findings of the meta-analyses included in Hattie's work were derived from studies conducted in English-speaking and highly developed countries, in particular in the United States, and covered predominantly the effects of student academic outcomes in mathematics, reading, sciences, and social studies. Furthermore, English as a second language was not included in Hatties meta-analyses (Hattie, 2009).

Helmke and Weinert (1997) have pointed out the existence of complex but rather diffuse concepts and theories of effective teaching and good instruction in general. At the core, the most notable reason for this is the complex nature of teaching and learning processes, which influence and are influenced by each other as well as by external and context factors in a highly complex way (*multiple determinants*, see also Berliner, 2006; Creemers, Kyriakides, & Sammons, 2010b, 2010c; Bell et al., 2012). Furthermore, there is a complex interplay of various classroom instructional factors which jointly influence student outcomes: Their effects can be additive or interactive, and they can be either linear or nonlinear (Hasebrook, 2006; A. Helmke, 2014a; Holzberger, Kunter, Praetorius, & Seidel, 2016; Seidel & Prenzel, 2004; Seidel, 2014). Rather than concentrating exclusively on or maximizing a specific factor, the focus should therefore be on an appropriate balance of different factors and various forms of interaction. Oser and Baeriswyl (2002) developed the metaphors "*orchestration*" and "*choreography*"

in order to illustrate the complex interplay of various factors. Even more complex, an optimal dosage of different factors is not seen as being somewhat static and valid for all, but as depending on multiple actors (for whom or which student) and multiple occasions (for which purpose, when) (A. Helmke, 2014a). Due to the multiple determinants that characterize teaching and learning processes as well as their complex interplay, large-scale studies have increasingly been conducted to investigate the quality of classroom instruction (Seidel & Shavelson, 2007).

To capture the complex relationships between multiple factors at multiple levels, advanced methodological approaches (e.g. multilevel analysis, structural equation modelling, multitrait-multimethod analysis, cross-classified and multiple membership multilevel models) have been recommended and have increasingly been applied in this research field (Creemers et al., 2010b; Fondel, Lischetzke, Weis, & Gollwitzer, 2015; Teddlie, Reynolds, & Sammons, 2000). However, there have been some concerns about them being used inappropriately in educational studies (Moss & Haertel, 2016; Nagengast & Trautwein, 2015) as well as in general (Sijtsma, 2016). Together with this, the inconsistency or poor replicability of results in (but not limited to) psychological science (Open Science Collaboration, 2015) and medicine (Begley, 2013) has recently attracted a lot of attention worldwide (Darling-Hammond, 2015; Morganstein & Wasserstein, 2014; Ng & Koretz, 2015; Wasserstein & Lazar, 2016) – to such an extent that it is known as the "*replication crisis*". Even in elaborated, international large-scale assessments (LSAs), discrepancies in the result patterns of studies conducted by the International Association for the Evaluation of Educational Achievement (IEA) and by the Organisation for Economic Co-operation and Development (OECD), such as TIMSS and PISA, on the same subjects at the same time have raised many questions and discussions regarding the methods, reliability, validity, and interpretability of these results, for instance, regarding the modeling and measurement of change at country level (Hutchison & Schagen, 2007; Klieme, 2016; M. Wu, 2010). Such inconsistencies might be caused by various factors at different stages: study operationalization, sampling, data processing, selection of statistical models and methods used to measure constructs, and how relationships between variables are modeled and analyzed. Part of the problem has been identified as the misuse of the *p*-value in reporting and interpreting study results (Wasserstein & Lazar, 2016).

In brief, there are still many challenges and problems in modeling and investigating classroom instructional effects on student outcomes, theoretically and methodologically.

I.2 This study

The present study is situated against this background, but focuses on an area that has been explored comparatively little: an investigation into the complex effects and interplay of different classroom

instructional and context factors of teaching and learning English as a foreign language (EFL) on student outcomes in Vietnam – a developing country in Southeast Asia. With a focus on real classroom processes and instructional factors in a non-experimental design, this study is located in a subdomain of the broad field of empirical educational research: empirical classroom and teaching effectiveness research.

For this purpose, data are used from a research project carried out in Vietnam during the school year 2006–2007 which was led by Dr. Tuyet Helmke and colleagues at the University of Koblenz-Landau in Germany and cooperation partners at the National Institute for Education Strategy and Curriculum (NIESC) in Vietnam. The data collection took place within the framework of a replication study of a large-scale research project initiated in Germany in 2003–2004 – the DESI study. Besides a longitudinal design with two measurement points using standardized tests and questionnaires, this research project is characterized by a video study to gain indicators of classroom instructional quality. Based on the video data, both quantitative and qualitative indicators of classroom instruction can be attained. Moreover, it is possible to extract a large amount of information about classroom instruction, which is difficult or impossible to obtain via different data collection methods.

Within the framework of a doctoral thesis with its constraints concerning length and bandwidth, the dependent variables in this study are limited to academic student outcomes based on test results, and the indicators of classroom instruction are limited to video-based indicators, although further data (such as motivational student outcomes and student perception of instructional quality) are also available via student questionnaires.

I.3 Research aims and special focus on methodological approaches

The primary aims of this study are (1) to identify classroom instructional factors which have crucial effects on the academic growth of ninth graders in EFL in Vietnam and (2) to gain an insight into their interplay with each other and with context factors. In order to achieve these aims, this study has a strong focus on methodological approaches for three main reasons.

First, as a response to "The ASA's Statement on *p*-Values: Context, Process, and Purpose" (Wasserstein & Lazar, 2016) and to the call to "Do things right" (Sijtsma, 2016), an effort will be made in this study to describe in detail the data, sample, and in particular the measurement constructs with regard to their reliability and validity as well as the methodological approaches along with their assumptions. Moreover, multiple methodological approaches are applied in this study, including approaches which

do not rely on *p*-values. Based on them, the size and replicability of the results will be interpreted and discussed.

Second, an investigation into the complex joint effect of many factors requires, on the one hand, modeling of a large number of predictors simultaneously. On the other hand, the number of classes in this study is relatively small (n = 50 classes) in comparison to the number of variables – although it is generally hard to find a significantly larger number of classes in a low-stakes study due to organizational and financial limitations. Hence, the "large *p*, small *n*" or the "*high dimension, low sample size*" problem arises. Using traditional methods such as the ordinary least square method (OLS), this problem can lead to interpretation problems of the results or is even not solvable when the number of predictors exceeds the number of cases (Bühlmann & van de Geer, 2011). Actually, this problem has often arisen in other research fields such as in genetic studies (Mei & Wang, 2016; Tutz, 2012), and several methodological approaches have been introduced in the last few decades to deal with it. Among them, the regularized regression method *lasso* (least absolute shrinkage and selection operator, Tibshirani, 1996) has gained considerable attention and popularity in many research fields (Hastie, Tibshirani, & Wainwright, 2015) with more than 21,200 citations according to Google Scholar at the time of writing this dissertation. However, in the field of behavioral sciences and educational research, this method has rarely been applied until recently (McNeish, 2015; G. Pham, Robitzsch, George, & Freunberger, 2016). In this study, lasso regression and some extensions are applied to identify the most important instructional predictors among many variables and to examine their complex joint effect on student growth.

Third, in studies using standardized test results, the use of different scaling models often leads to different estimates regarding student outcomes (Eckes, 2015; Ng & Koretz, 2015; Robitzsch et al., 2017; Robitzsch, Dörfler, Pfost, & Artelt, 2011; Schroeders, Robitzsch, & Schipolowski, 2014; Trendtel, Pham, & Yanagida, 2016). Consequently, model selection can make a difference for estimating the instructional effects of student outcomes. However, the decision about how test scores are estimated given a wide range of estimation methods is usually left to a technical advisory committee, and the uncertainty in results caused by the selection of a scaling model is often ignored or overlooked by researchers or policy-makers when conducting secondary analyses based on the reported test scores (Ng & Koretz, 2015). Thus, different scaling models are applied in this study to estimate the test scores of students, and differences regarding the measures of classroom instructional effects caused by different scaling models will be discussed.

I.4 Preview

Chapter II gives an introduction to the general teaching and learning situation and issues in EFL in Vietnam. The information provided is valid for the school year of the data collection, 2006–2007. In Chapter III, the original research project in Germany and the research project in Vietnam are briefly introduced. The theoretical background of the study can be found in Chapter IV. The research questions are formulated in Chapter V. Chapter VI addresses the methodological challenges and approaches. Chapters VII–X set out the main results of this study, followed by a discussion of the results, the limitations of the study, and future prospects. Data processing details can be found in the Appendices at the end of this dissertation.

II. Teaching and learning EFL in Vietnam

II.1 Vietnam – an overview

Vietnam is located in Southeast Asia; it borders China in the North, the Laos People's Democratic Republic and Cambodia in the West, and the Pacific Ocean in the East. With a total area of around 331,000km^2, the country is approximately as large as Germany, and has a rapidly growing and young population (The World Factbook, 2015). In 2014, the population size was about 94 million people, with 24% between 0-14, 17% between 15-24, 45% between 25-54, 8% between 55-64, and 8% over 64 years old (The World Factbook, 2015).

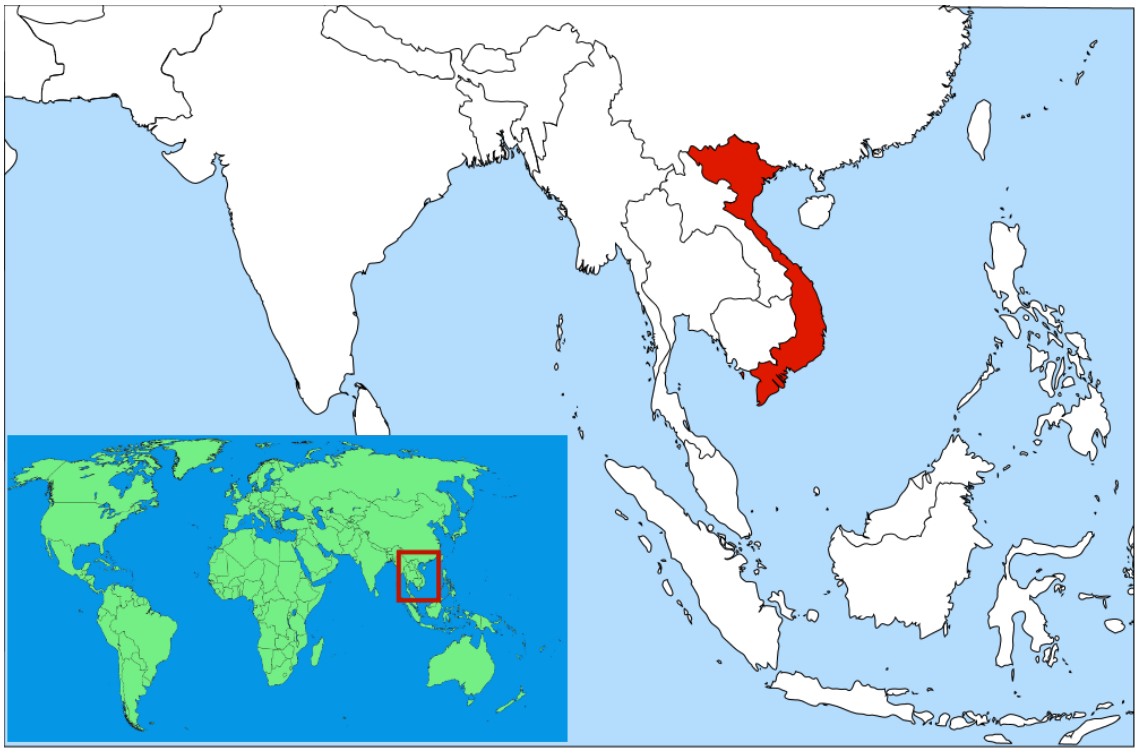

Figure 1: Location of Vietnam

Source: https://commons.wikimedia.org/wiki/Atlas_of_Vietnam (created by Luu Ly, retrieved on 11th March 2017)

Economically, Vietnam has continuously experienced high economic growth since the beginning of its transition from a centrally planned and controlled economy to a market economy (Doi Moi) in 1986. In the period from 1990 to 2009, Vietnam had one of the fastest improvements in living standards and one of the most dramatic reductions in poverty in the world (Asian Development Bank, 2010). From one of the poorest countries with a per capita income below $100, Vietnam became a lower middle income country with an annual per capita income of over $2,000 by the end of 2014 (The World Bank, 2015). On the other hand, serious development challenges are posed by income and non-income inequality, despite the remarkable achievements in reducing poverty (Asian Development Bank, 2015). The period of transition and opening up of the economy saw a widening of the gap between the rich and poor as well as the gap between urban and rural areas (H. T. Le & Booth, 2014). Since the 1986 reforms, the urban population has grown parallel to the increasing share of industry, while agriculture's share of economic output has continued to shrink from about 25% in 2000 to 18% in 2014 (The World Factbook, 2015). In 2014, the urban population was about 33.6% of the total population, with an annual growth rate of 2.95%.

For a long time there have been two independent and dominant core-periphery urban systems (special cities): Ho Chi Minh City (in the South) and Hanoi (the capital, in the North). Vietnam's rapid economic

growth is driven by these two urban systems, which are characterized by high growth and industrial concentration together with their surrounding areas. These two cities together accounted for a total of 15.9% of the country's population, 33.9% of the urban population, and contributed 30.5% of the gross domestic product in Vietnam in 2009 (The World Bank, 2011).

II.2 Education in Vietnam: general indices

Statistically, Vietnam has achieved better results in the education sector in comparison with countries with similar economic development; over 90% of the working-age population is literate (The World Bank, 2010).

Primary education (grades 1–5) is compulsory and has a nominal duration of five years with an official entry age of six years. It is followed by lower secondary education with a duration of four years (grades 6–9) and then upper secondary education (grades 10–12) (UNESCO-IBE, 2011).

Primary education is free of a tuition fee. According to the Primary Universalization Law in 1991, every child must complete primary school by the age of 14 at the latest. On the other hand, there are no sanctions when parents do not follow this. In the school year 2004–2005, more than 98% of children of primary school age attended school, and the enrolment rates for boys and girls were similar (The World Bank, 2010).

Vietnam aims for the *universalization of primary education with a correct age* (99% children enroll in primary school; most 14-year-old children complete primary school at the right age, minimizing repetition and drop-out rates). It also aims for *universal lower secondary education* (The World Bank, 2010). The criteria for the certification of universal lower secondary education are set by the National Assembly (Resolution No. 41/2000/QH10); the government (Decree No. 88/2001/ND-CP) is responsible for checking and certifying the universal lower secondary education results for each administrative unit. Unlike other countries, universal lower secondary education in Vietnam has been implemented without following the model of cumulatively increasing the number of compulsory years.

According to The World Bank (2010), by March 2006, 35 provinces and centrally managed cities were certified by the Ministry of Education and Training as meeting the national standards regarding universal correct-age primary education, and 32 met the standards regarding universal lower secondary education.

Before the school year 2005/2006, there was a graduation examination at the end of each education level (primary, lower, and upper secondary school). From the school year 2005/2006 (until 2014), the final examinations that had been administered at the end of primary and lower secondary education were

removed, leaving one final graduation examination at the end of upper secondary education conducted nationwide.

The lower secondary level is universal: Every primary student who successfully completes primary school is allowed to enter the sixth grade; it is not however compulsory. As for upper secondary, all lower secondary students have to pass an entrance examination in order to continue to upper secondary school. The selection can be made in three ways: either through an examination, through consideration based on ninth-grade learning achievements, or through a combination of both. The decision on which selection method is used in a particular province or centrally managed city is made by the provincial/city People's Committee (The World Bank, 2010). For the upper secondary graduation examination, the Ministry of Education and Training in Vietnam (MOET) introduced a multi-choice testing method for foreign language as part of the reform of the 12th grade assessment practices in 2006 (UNESCO, 2007). The percentage of children of the official lower secondary school age who were enrolled in lower secondary school (the net enrollment ratio) was 68.8% in the school year 1999/2000, 78.9% in the school year 2005/2006, and 87.10% in the school year 2011/2012 (UNESCO Institute for Statistics, 2015).

Despite expansion of schooling in the late 1990s, with primary education becoming nearly universal in 2000, only one out of every three Vietnamese teens aged 15–17 years had access to upper secondary education. In addition, it is worth mentioning the disparities between educational chances in the rural and urban areas, between the poor and nonpoor, and between boys and girls (Asian Development Bank, 2008). This problem is especially pronounced in secondary education and above, as reported in the "*Vietnam Development Report 2014*" by the World Bank (The World Bank, 2013, p. 92): "*primary education is for all, while upper secondary and above is mainly for the wealthy*", because secondary education and above is not free of charge. Moreover, the results of the 2012 Young Lives School Survey revealed that students from wealthier families were on average receiving more periods of instruction per week, their teachers were more qualified, and their school facilities were of better quality (Rolleston, James, Pasquier-Doumer, & Tran, 2013).

The majority of school children in Vietnam attend school on a half-day basis. The half-day tuition time is due to the shortage of teachers and classrooms, which is "*short relative to that in other countries and seen as too short to cover the curriculum adequately*" (The World Bank, 2013, p. 92). On the other hand, many children in urban areas attend informal *extra classes* during the other half-day, which are usually taught by the same teacher and financed by the parents, because wealthier parents demand more schooling than is formally provided (The World Bank, 2013, p. 97). Attendance of extra classes is high in Vietnam: Parents of nearly 50% of lower secondary school students reported having paid for extra classes according to The 2010 Vietnam Household Living Standards Survey; 70 percent of 14- and 15-

year-old students attended extra classes, which amounted to an average of 10 additional hours of instruction per week, equivalent to 27 percent of total instruction time according to the 2009 Young Lives Survey (The World Bank, 2013, p. 97).

Vietnam participated in the Programme for International Student Assessment (PISA) for the first time in the year 2012. In PISA 2012, the results of Vietnamese students aged 15 years were on par with their German peers, and better than those of 15-year-old students in two-thirds of the participating countries (OECD, 2014) in all three domains (mathematics, reading, and sciences). Of course, these results are generalizable only for the 15-year-old population who had not left school at this age. Nevertheless, similar findings were found for younger children by A. Helmke, Schrader, Vo, Le, & Tran (2003) and Rindermann, Hoang, & Baumeister (2013); they found that fifth graders in Vietnam outperformed peers in Germany in mathematics. Results from the Young Lives Survey also confirmed that Vietnamese students performed consistently better in mathematics than their peers in Ethiopia, Peru, and India aged 5, 8, 12, and 15 years (Rolleston, James, & Aurino, 2013). Within the country, according to Rolleston, James, Pasquier-Doumer, et al. (2013), the performance in mathematics and Vietnamese reading of students in disadvantaged areas was notably lower than that of students in urban areas. On the other hand, learning progress did not vary much between advantaged and disadvantaged students, or between girls and boys. In general, girls performed significantly better than boys in Vietnamese reading, and slightly better in mathematics (although not significantly better); children born earlier in a calendar year performed better in both subjects.

Actually, the level of performance of Vietnamese students in international comparison was not totally unexpected. Culturally, Vietnam belongs to the Confucian heritage cultural sphere; this cultural heritage places high value on educational achievement, and students are therefore highly motivated to succeed in school (A. Helmke & Hesse, 2010; D. Y. F. Ho, Peng, & Chan, 2002). The East Asian nations/territories (China, Hongkong, Korea, Japan, Taiwan) and some Southeast-Asian countries (Singapore, Vietnam) are members of this cultural sphere. In international large-scale assessments in mathematics, reading comprehension, and sciences, such as PISA, TIMSS, these countries have constantly achieved outstanding results (Martin, Mullis, Foy, & Stanco, 2012; Mullis, Martin, Foy, & Arora, 2012; Mullis, Martin, Foy, & Drucker, 2012; OECD, 2014a, 2016).

Contrary to the level of student achievement, the educational conditions and classroom instruction in these countries have sometimes been regarded as "*ill*", because they are characterized by large class sizes, poor equipment, extremely teacher-oriented instruction with an overemphasis on examinations and a lack of cultivation of creativity, autonomy, and critical thinking, and excessive pressure to succeed even at the expense of student mental health; the instructional methods are regarded as outmoded D. Y.

F. Ho et al. (2001; 2002). This contrast is well known as the "*paradox of Confucian heritage education*" (D. Y. F. Ho et al., 2002; D. Y. F. Ho, 1994).

In 2004, the class size was $N = 39$ in Vietnam (UNESCO, 2007). In the less advantaged regions, the class size was typically small; it was larger in more advantaged urban areas (Rolleston, James, Pasquier-Doumer, et al., 2013).

In short, there is a consensus among educational researchers that the high level of academic achievement in mathematics, reading, and sciences of students in East Asia (compared with many other countries, and not only developing countries) is not generally attributed to the quality of instruction and the general educational systems in these countries, but to other factors, especially the Confucius learning culture and the total amount of learning time of students (A. Helmke & Hesse, 2010; Hesse, 2007; D. Y. F. Ho et al., 2001; D. Y. F. Ho, 1994). This is also supported by national study results, according to which children of East Asian descent who were born and raised in a Western country (e. g., "Asian Americans") also outperformed native children in this country academically, e. g., Beuchling (2003) in Germany, Jerrim (2014) in Australia, Jonsson & Rudolphi (2011) in Sweden, Byun & Park (2012) and Kao & Thompson (2003) in the United States, Francis & Archer (2005) in the United Kingdom.

In this regard, J. Thompson (2009, p. 23) described and explained instruction quality in Vietnam as follows: "*The style of teaching across educational levels in Vietnam has virtually always been dominated by rote learning, that is memorization and reproduction of information provided in lecture format. For centuries Vietnamese education has been rooted in the Confucian tradition, similar to a number of other East Asian and Southeast Asian societies. Confucian ethics emphasize the supreme knowledge of the teacher and the duty of the student to be passive, obedient and to learn by heart the information provided by the professor. In addition to Confucian influences, which continue to exert underlying albeit strong tides in Vietnamese society and education, the periods of French and Soviet influence also impacted teaching styles. Both systems similarly depend primarily on teaching by lecture and rote memorization.*"

II.3 EFL in Vietnam: level of proficiency

Unlike in mathematics or sciences, East Asian students in general or Vietnamese students in particular cannot be expected to score as well as in EFL students in Western non-English speaking countries. This is due to the fact that Asian native languages including Vietnamese – the official language in Vietnam – are very different from English (for a comparison between English and Vietnamese, see Tang, 2007). This big difference between the two languages means that learners take much longer to reach a high

level of fluency in English than people from regions whose first language is more similar to English (Lightbown & Spada, 2013; Padilla, 2006). Supporting evidence can be found in the first EF English Proficiency Index (EF EPI, 2011), which was based on *online* standardized English test results (including grammar, vocabulary, reading, and listening skills; the test is accessible at www.efset.org) of hundreds of thousands of *voluntary adults* in many countries from 2007 to 2009. Only countries with a minimum of 400 test takers were included in the index.

Rank	Country	EF EPI Score	Level	Rank	Country	EF EPI Score	Level
1	Norway	69.09	Very High Proficiency	23	Italy	49.05	Low Proficiency
2	Netherlands	67.93	Very High Proficiency	24	Spain	49.01	Low Proficiency
3	Denmark	66.58	Very High Proficiency	25	Taiwan	48.93	Low Proficiency
4	Sweden	66.26	Very High Proficiency	26	Saudi Arabia	48.05	Low Proficiency
5	Finland	61.25	Very High Proficiency	27	Guatemala	47.80	Low Proficiency
6	Austria	58.58	High Proficiency	28	El Salvador	47.65	Low Proficiency
7	Belgium	57.23	High Proficiency	29	China	47.62	Low Proficiency
8	Germany	56.64	High Proficiency	30	India	47.35	Low Proficiency
9	Malaysia	55.54	High Proficiency	31	Brazil	47.27	Low Proficiency
10	Poland	54.62	Moderate Proficiency	32	Russia	45.79	Low Proficiency
11	Switzerland	54.60	Moderate Proficiency	33	Dominican Republic	44.91	Very Low Proficiency
12	Hong Kong	54.44	Moderate Proficiency	34	Indonesia	44.78	Very Low Proficiency
13	South Korea	54.19	Moderate Proficiency	35	Peru	44.71	Very Low Proficiency
14	Japan	54.17	Moderate Proficiency	36	Chile	44.63	Very Low Proficiency
15	Portugal	53.62	Moderate Proficiency	37	Ecuador	44.54	Very Low Proficiency
16	Argentina	53.49	Moderate Proficiency	38	Venezuela	44.43	Very Low Proficiency
17	France	53.16	Moderate Proficiency	39	Vietnam	44.32	Very Low Proficiency
18	Mexico	51.48	Moderate Proficiency	40	Panama	43.62	Very Low Proficiency
19	Czech Republic	51.31	Moderate Proficiency	41	Colombia	42.77	Very Low Proficiency
20	Hungary	50.80	Moderate Proficiency	42	Thailand	39.41	Very Low Proficiency
21	Slovakia	50.64	Moderate Proficiency	43	Turkey	37.66	Very Low Proficiency
22	Costa Rica	49.15	Low Proficiency	44	Kazakhstan	31.74	Very Low Proficiency

Figure 2: EF English Proficiency Index 2011 (EF EPI, 2011)

It should be noted that the results cannot be generalized for each country as a whole, and not for school students. The test takers were "*adults who either wanted to learn English or were curious about their English skills*" (EF EPI, 2011, 2014a) and had internet access; thus the sample of each country was self-selected and not representative for the whole population. It can be assumed that the Vietnamese sample in this online test had a higher socio-economic background than average based on the fact that they had access to the internet and that they were interested and motivated in learning English. Thus, it can be assumed that their achievement was above average for the Vietnamese. On the other hand, their performance was among the lowest of all participants from all countries in the 2011 report ("very low proficiency"). Among participants from Asian countries/territories, the Vietnamese sample scored similarly low as the participants of two other Southeast Asian nations (Thailand and Indonesia); samples from India, China, and Taiwan scored a little better ("low proficiency"); participants from Japan, South Korea, Hong Kong, and Malaysia scored the highest ("moderate proficiency").

Since then, the annual EF EPI reports have shown a steady improvement of Asia's participants: "*Since 2007, the regional average EF EPI score has risen 3.52 points, a gain comparable to that of Europe. English, rather than an Asian language, is the lingua franca of the continent*" (EF EPI, 2014a). The EF EPI results of Vietnam have been improving steadily (EF EPI, 2014a, 2014b). In the EF EPI reports in the years 2012 and 2014, the Vietnamese samples reached "low proficiency"; in the EF EPI reports in the years 2013, 2015, and 2016, the Vietnamese samples achieved "moderate proficiency" on average (EF EPI, 2013, 2015, 2016). German samples consistently achieved "high proficiency" from 2007 to 2016 (EF EPI, 2011, 2012, 2013, 2014, 2015, 2016).

An in-depth analysis for Vietnam (EF EPI, 2014b) has revealed that the improvement over time holds true for all age cohorts of participants, and younger adults did not perform better than elder ones (based on self-reported age). On the other hand, the data showed significant differences between regions and between men and women in Vietnam: Participants in Ho Chi Minh city had higher English proficiency than participants from other regions of Vietnam; women scored significantly better than men regarding their speaking skills.

Regardless of whether the EF EPI results hold true for the Vietnamese population and for school students, Vietnam has been aiming to achieve progress in English language education. Nowadays, English is increasingly favored as a foreign language in Vietnam (The World Factbook, 2015) and is arguably the most important foreign language which is taught and learned at education institutions that belong to the national education system (Government of Vietnam, Article 1.1., Decision 1400, 2008, p. 2). In 2008, the Vietnamese government passed Decision 1400 approving the project entitled "*Teaching and learning a foreign language in the national education system, period 2008-2020*", which declared that by 2020 "*foreign languages [will be] a comparative advantage of the development of Vietnamese people*" (Government of Vietnam, Article 1.1., Decision 1400, 2008, p. 1).

In the following sections, we will take a look at several important issues regarding English language education in Vietnam. This aims to allow a better understanding of the general context in which everyday teaching and learning practice in EFL in lower secondary school takes place in Vietnam.

II.4 A brief history of English language teaching in Vietnam

The development of English language teaching in Vietnam has gone hand in hand with the political development of the country; politics has influenced the foreign languages studied in Vietnam's schools and tertiary institutions.

Before Vietnam became part of French colonial Indochina in the late 19th century, Vietnamese was officially written with Nom, which were based on Chinese characters. The French then banned Nom and replaced them with a Romanized alphabet to create the nowadays Vietnamese (Denham, 1992). During the French invasion, English first became present in Vietnam (Hoang, 2010), while French became the medium of instruction in schools and universities for a small minority with access to education (Dang, 1986).

On 2nd September 1945, Ho Chi Minh proclaimed the independence of Vietnam, and the Vietnamese language (with Romanized alphabet) was made the medium of instruction in Vietnamese schools (Ministry of Education, 1990). From 1945 to 1975, after the defeat of the French at Dien Bien Phu, Vietnam was divided into two parts: North Vietnam and South Vietnam. North Vietnam was allied with the former Soviet Union, and South Vietnam with the USA. In this period, English was the most commonly studied foreign language in schools and universities in South Vietnam (Dang, 1986). In North Vietnam, Russian was the most important foreign language in the formal educational system, while English was taught only in some selected classes in some upper secondary schools, and only in towns and big cities as a pilot subject from 1958 (N. Q. Nguyen, 1993). Until 1971, English did not become a school subject in North Vietnam (Denham, 1992).

After the reunification of the country in 1975 until 1986, before Doi Moi, Russian was the most dominant foreign language in schools in the whole country (Hoang, 2010). Nevertheless, there was an "*unplanned*" spread of English in society due to popular demand, because English was seen by the Vietnamese as "*the key which opens many doors*" (Denham, 1992). Since 1980, EFL has been a compulsory subject in lower and upper secondary education, as well as for undergraduates and graduates at tertiary level. However, until 1989, the opportunities to speak English were limited to the classroom in Vietnam (Denham, 1992). During this period, a small number of Vietnamese teachers and interpreters of English were sent to English-speaking countries to do graduate studies in English language teaching. Unfortunately, these training programs were terminated in 1979 as Vietnam was involved in Cambodia. In 1985, Australia resumed English training for Vietnam, which was provided under a bilateral aid program between Australia and Vietnam from 1992. Within this framework, 40 Vietnamese teachers and interpreters were sent to Australia annually to undertake graduate studies in English language teaching until early the 2000s, when the programme was terminated (Thinh, 2006).

The period after Doi Moi in 1986 up to present is characterized by the rapid expansion of English in Vietnam. With a new market economy, the growth of international business and trade, and the increasing number of foreign tourists, the importance of English has been increasing in the Vietnamese society. English proficiency has become a key competence for acquiring a job in tourism, the hospitality

industries, and many other enterprises (Hoang, 2010). English has become the first foreign language and nearly the only foreign language which is taught in schools (97.9% pupils who learned a foreign language in schools learned English, Baker & Giacchino-Baker, 2003). Recently, a foreign language has been made a compulsory subject in the national examination for graduation of upper secondary school; in several large cities, a foreign language is an optional subject in some primary schools from grade 3 (The World Bank, 2010).

II.5 Problems regarding quantity and quality of EFL teachers in Vietnam

Due to the tremendous growth in the number of English classes in Vietnam after 1986, there was a shortage of foreign language teachers, especially experienced ones. To deal with the great demand for English teachers, a large number of unqualified teachers for EFL were recruited; these teachers were not adequately prepared with regard to English skills and teaching methodology; furthermore, the more experienced teachers were not much better, because they completed training in the past, at a time when modern teaching approaches and methods had not reached Vietnam (H. H. Pham, 2001). As a result, the organization of foreign language instruction in Vietnam has faced a number of difficulties, in particular the limited instructional time and low quality of instruction (The World Bank, 2010).

Besides, there are other common difficulties linked to English education in public schools, such as over-crowded classes, poor equipment, and controlled teaching materials by the ministry of education (T. S. Le, 2011). In Vietnam, the textbooks for primary and secondary schools are commissioned and mandated by the Ministry of Education and Training (MOET), which "*prescribes what is taught, what is to be learned, what is assessed, and how much time teachers should spend on the delivery of instruction. Put simply, the textbook becomes the curriculum, and it is understandable that instruction is largely, if not completely, textbook-driven*" (V. Le, 2011, p. 19). Textbook-driven instruction is reinforced by a strong tradition of centralization, within which the teachers themselves learn "*to follow rules established by the ministry and organise their behaviour accordingly*" (Saito & Tsukui, 2008, p. 98).

Therefore, although communicative competence and communicative language teaching (CLT, see Chapter IV.3.3) have been set as the prevailing goal in English education in Vietnam, there is a large gap between what is actually implemented by classroom teachers and the rhetoric (V. C. Le & Barnard, 2009; Nunan, 2003). EFL education in Vietnam is characterized by a central curriculum which is exam-driven and geared to a written examination of grammar, reading, and translation. There is a wide use of teacher-centered, book-centered methods with exclusive focus on reading skills, grammar, and translation methods. In addition, EFL education in Vietnam does not integrate learner feedback policies,

and students have almost never expressed in case they have difficulties understanding something (Denham, 1992; V. C. Le & Barnard, 2009; V. C. Le, 2002; H. H. Pham, 2005; Tomlinson & Bao, 2004).

Most importantly, Ellis (1994) observed a resistance to adopting the communicative approach to teaching English in Vietnam, which he explained with the cultural inappropriateness based on the results of an interview study: "…*it appeared that to adopt the approach, Vietnamese teachers would have to change radically some basic cultural beliefs. It is concluded that for the communicative approach to be made suitable for Vietnamese conditions, it must be both culturally attuned and culturally accepted.*" The cultural issue, which is believed to play an important role as an obstacle for the successful adoption of CLT in Vietnam, is described by V. C. Le, p. (1999, pp. 3–4) as follows: "*Central to pedagogical practices in Vietnam is the traditional view of the teacher-student relationship. This view supports teacher-centred methods and a structured curriculum. The teacher is supposed to be the only provider of knowledge and therefore she/he is highly respected by the students, student parents, and the society as a whole. What the teacher or the textbook says is unquestionably standard norm,*" and "*most teachers believed that reading was the way students could best learn the language.*" This was confirmed by the results of a questionnaire survey conducted by P.-M. Nguyen (2008) with a sample of 647 participants including lecturers, student teachers, university students, and lower and upper secondary school students: The teacher's role in Vietnam is generally perceived as being the ultimate source of knowledge. Teachers in Vietnam believe that students expect them to be a "*guru of knowledge*" rather than a "*facilitator of knowledge*" (P.-M. Nguyen, 2008, p. 39). This lack of readiness to adopt CLT is believed to be one of the main obstacles to developing instructional quality in English teaching in Vietnam (G. Bock, 2000; Canh, 2002; Ellis, 1996; V. C. Le & Barnard, 2009), rather than structural aspects such as class size. Buhn-Wiggers (2014) found no support for the argument that smaller classes might increase graduation rates at the end of secondary school in Vietnam.

Regarding the EFL teaching quality in schools in Vietnam, studies based on large sample sizes and observable data are still rare, and "*large-scale studies, using multiple methods of data collection, are needed*" (V. C. Le & Barnard, 2009, p. 30). Possibly, observable data from a video study based on a relatively large class sample size can meaningfully contribute toward obtaining real life indicators of the teaching quality in EFL in Vietnam.

II.6 EFL competencies of English learners in Vietnam

While many EFL teachers consider the writing skills of Vietnamese English learners as not problematic, this is not the case concerning their speaking skills (Cunningham, 2013; Ha, 2005). This is often

attributed to the big differences between the sound system of Vietnamese and that of English as well as the limited opportunities for hearing and speaking English in Vietnam in the past (Cunningham, 2009). For a long time, improper pronunciation has attracted considerable attention from researchers (especially in the field of linguistics research) in an effort to professionalize the education of English teachers in Vietnam (Cunningham, 2013; Dao, 2007; Ha, 2005; Honey, 1987; D. L. Nguyen, 1970; Tang, 2007; L. C. Thompson, 1965). Based on a cross-linguistic analysis, Tang (2007) pointed out some potential problems that Vietnamese people have in speaking English which are caused by interactions between the Vietnamese and English languages. These are shown in Table 1 below and include problems such as simplifying initial and final consonant clusters, deleting final consonant clusters, substituting English with Vietnamese consonants and vowels, wrong stress of words and sentences, difficulty with word endings that indicate a change in word class, and omitting word endings. According to Tang, these difficulties could arise from the differences between Vietnamese and English, one a monosyllabic language and the other a polysyllabic language (T. A. T. Nguyen, Ingram, & Pensalfini, 2008), regarding phonology (sound level), lexical-semantic (word level), and syntax (grammar).

Table 1: Common problems of Vietnameses in speaking English (Tang, 2007)

Potential Interactions of Vietnamese (L1) with English (L2)

Lang. level	Pattern	Example
Phonology (Sound level)	Simplify initial consonant clusters	*sring* for "string"
	Delete or simplify final consonant clusters	*bok* for "box"
	Substitute with Vietnamese consonants	Dental aspirated "t" for "soft th": [t^hɔt˥] for "thought"
	Substitute with Vietnamese vowels	*cheek* for "chick"
	Intonation pattern influenced by tones	Rising and falling on individual words
Lexical-Semantic (Word level)	Difficulty using words that do not have direct Vietnamese translations	"To do," "to work," and "to make" are all one word in Vietnamese, *làm*
	Difficulty with endings that indicate a change in word class	*so bore* for "so boring"
Syntax (Grammar)	Omit word endings for tense	*walk* for "walked"
	Omit word endings for plurality	*two dollar* for "two dollars"
	Omit word endings for verb agreement	*she walk* for "she walks"
	Omit auxiliary verbs	*You hungry?* for "Are you hungry?"
	Place adjectives after nouns	*car big* for "big car"
	Difficulty with word order in questions	*You want eat what?* for "What do you want to eat?"

Tang also stated that "*quantitative information regarding the frequency and distribution of speech sounds and words of Vietnamese based on both oral language samples and written texts are needed to establish how rare or common linguistic features are and therefore how often they occur in daily language use.*" In fact, the results of several studies with a quantitative approach (albeit with very small

sample sizes – fewer than ten respondents) have recently supported Tang's thesis, such as the study on the quality, quantity, and intelligibility of vowels in Vietnamese-accented English by Cunningham (2010), or the study on the mispronunciation of some English consonants by a Vietnamese sample by Dao (2007) and T. T. T. Nguyen (2007). Ha (2005) pointed out that even students in an English department, who had finished four years of English instruction and taken part in the final exam, also made pronunciation mistakes. It is noted that the quality of language input, especially spoken language received through listening, is a topic that has attracted a lot of attention in research on the acquisition of a second language in the USA, because some research findings have shown its strong influence on student achievement in English as L2 (Padilla, 2006).

Cunningham (2013) assumed that, although pronunciation can be successfully taught and learned, not all pronunciation features are teachable and learnable by Vietnamese students, such as initial and final consonant clusters. Her conclusion was based on the results of a self-developed short course (three weeks, per week two or three hours) on English pronunciation with a focus on intelligibility, which she gave to 110 students of tourism in four classes at a college of business and tourism in Hanoi, a sample which was very aware of the importance of oral proficiency. On a broader level, Jenkins (2000) remarked that "*where the difficulty with an L2 English pronunciation feature is universal…, we are looking at an item that may well be unteachable*" (p. 119).

Regardless of whether this is true or not, Phan and Vo (2012) found out that the mispronunciation of vowels and consonants does not significantly contribute to the (lower) comprehensibility of Vietnamese-accented speech, but only to a (higher) judgement of their accentedness. They conducted an experimental study on the effects of *segmental* and *suprasegmental* errors on the perceptual judgements of the comprehensibility and accentedness of Vietnamese-accented speech by native listeners. Segmentals and suprasegmentals are two facets of pronunciation: Segmentals are "*minimal units of sound defined in phonetic terms*" (Pennington & Richards, 1986) such as consonants and vowels; suprasegmentals are "*referred to as prosody, which includes stress, length, tone, intonation, rhythm and timing*" (Major, 2001) such as sentence stress. Their results showed that suprasemental errors correlated more highly with global judgements of the comprehensibility and accentedness of speech than segmental errors; only the incorrect allocation of stress in a sentence (wrong sentence stress) was a significant predictor of reduced comprehensibility, while incorrect placement of word stress, and vowel and consonant errors were not.

Thus, in this study, I want to investigate the types and frequency of pronunciation errors made by both Vietnamese students and teachers in school classrooms as indicators of product quality. For this purpose, an extension of the video coding guide based on the "*The phonetics and phonology of English and*

pronunciation" written by Eckert and Barry (2005) will be developed. The different types of pronunciation error suggested by Eckert and Barry can be seen as systematic and universal, and cover most of the pronunciation phenomena in the lists of Tang and Phan and Vo.

III. Project introduction

III.1 Genesis – the DESI study in Germany in the school year 2003–2004

Nowadays, English is the global language used for communication between people all over the world. The language skills of the workforce of a country in general and English skills specifically are among the indicators of the global competitiveness of a nation (IMD, 2014). Due to its vocational relevance, English is in increasing demand as a school subject in many countries, even in countries where English is not the first language or lingua franca (The World Bank, 2005, p. 85). Knowledge on how to enhance the quality of teaching and learning English as a foreign language (EFL) in schools is therefore increasingly important.

However, student achievement in EFL and its influence factors have no longer been included in the regular international large-scale assessments of the International Association for the Evaluation of Educational Achievement (IEA) and the Organization for Economic Co-operation and Development (OECD) (TIMSS, PIRLS, PISA…)[1] since the Six Subject Survey in 1970–1971 (IEA, 1971).

Against this background, the first nationwide large-scale assessment of student achievement in EFL in ninth grade was conducted in Germany in the school year 2003–2004 (Beck, Bundt, & Gomolka, 2008; DESI-Konsortium, 2008) – the *DESI* study – to fill this gap. The name "DESI" stands for "*Deutsch-Englisch-Schülerleistungen-International*" *[German English Student Assessment International]*. This was a comprehensive and cooperative interdisciplinary project conducted by English educationalists, linguists, educational researchers, psychologists, and psychometrists, which resulted in numerous empirical findings which served as a basis for the revision of English curricula, textbooks, teaching and learning materials, and in-service and pre-service teacher training programs in Germany. As part of DESI, a video study was conducted in order to acquire authentic data on everyday instruction and investigate its effects on student achievement (T. Helmke et al., 2008). Video study has become increasingly popular in educational research thanks to recent video technological developments and its great potential (cf. Brückmann et al., 2007; Janík, Seidel, & Najvar, 2009; A. Helmke, 2014a). Using

[1] Actually, IEA conducted the Language Education Study in 1995 (http://www.iea.nl/language-education-study), but only phase 1 of the policies of language education in participating countries was completed.

video data, researchers have access to authentic behavioral and communicative indicators of teaching and learning reality, which are not accessible by means of questionnaires or other research instruments. Moreover, video data can be coded, rated, and (re)analyzed with respect to different research questions at any later time point.

Detailed project descriptions and results can be found in the two-volume book about the DESI study (Beck & Klieme, 2007; DESI-Konsortium, 2008). A short summary of the study findings can be found in DESI-Konsortium (2006). In this chapter, the aims, sample, questionnaires, selected methodological issues, and results regarding student progress in EFL and the video study will be briefly outlined.

III.1.1 Aims and study design

DESI is a large-scale longitudinal study funded by the Standing Conference of the Ministers of Education and Cultural Affairs (KMK) in the Federal Republic of Germany. It is an investigation into the language competencies and progress of students in German and English as a first foreign language as well as into the classroom and teaching practices for these school subjects. The study searched for factors and determinants of student achievement and growth in order to supply teaching practice, pre-service education and in-service programs, and the education policy with reliable empirical findings (Klieme, 2008).

To this end, DESI had a complex longitudinal design with two measurement points (MPs) using students, parents, teachers, school principal questionnaires, language tests, and video data (Beck et al., 2008; DESI-Konsortium, 2006). The underlying framework model for the analysis of the effects of student achievement and progress was the "*opportunities-take-up model*" [*Angebots-Nutzungs-Modell*] (A. Helmke & Klieme, 2008, p. 302).

Regarding the school subject EFL, at the beginning (T1) and at the end (T2) of the school year 2003/2004, student questionnaires and English tests were conducted in different domains, including listening comprehensive (LC) and general language proficiency (C-test), to evaluate different language skills. In addition, a teacher questionnaire was applied at the end of the school year. In the middle of the school year, between the two MPs, a video study was carried out in English classes to complete the study with observational data.

The DESI video study attempted to (1) describe and analyze the actual teaching practice based on objectively measurable criteria both qualitatively and quantitatively, (2) acquire behavioral indicators that are relevant to student development of English performance, and (3) take stock of verbal communication in English lessons (T. Helmke et al., 2008).

III.1.2 Sample

The student sample in the DESI study was a nationwide representative sample of all ninth-grade students and classes in Germany, which included around 11,000 ninth-grade students of all school tracks in all states. Teachers and students in 105 English classes voluntarily participated in the video study.

III.1.3 Student and teacher questionnaires

In the DESI study, a student questionnaire was applied at each MP. These questionnaires covered a variety of items and scales including individual background (e.g., gender, age, socioeconomic status of the family), motivational aspects (e.g., academic self-esteem, learning interest, motivation to achieve), school marks, learning strategies, student perception of instructional quality, out-of-school learning-fostering activities, and so on. Some items and scales were administered at both MPs to enable estimates of changes, such as academic self-esteem, learning interest, and student perception of instructional quality. The teacher questionnaire was applied at the second MP; it comprised questions on lesson planning, academic background, teaching experiences, years of teaching the test class, self-perception of own teaching quality, and so on. All documentation on the scale of the DESI questionnaires (Wagner, Helmke, & Rösner, 2009) is published at http://www.pedocs.de/frontdoor.php?source_opus=3252.

III.1.4 Student achievement and growth based on C-test and LC-test

The English achievement tests in the DESI study comprised various subtests to measure different English competencies of students. The test instruments were constructed based on the Common European Framework of Reference for languages (CEFR). The curricular appropriateness of the tests was then examined. They were aligned to the content of the English curriculum for the ninth grade in different school tracks in Germany (for information on the curricular validity of the tests, see Dubberke & Harks, 2008). Among them, only the text reconstruction test (C-test) and the listening comprehension test (LC-test) were conducted at both MPs. Data obtained from these two tests served as a basis for estimates of student academic progress over one school year.

The item pool for each test was large with a wide range of item difficulties to cover all CEFR-based proficiency levels in different English competencies. Because the testing time was limited, the DESI study had a matrix test design with different booklets; each student was given a booklet with a sub-sample of the items. There were anchor items (items which were employed in two or more booklets) to enable linking of the results based on different booklets. Moreover, the presence of multiple booklets made it possible to model student achievement growth without having each student work on the same exercises at both MPs (Beck et al., 2008).

Because students worked on different test booklets, it was not possible to use sum scores to compare student performance within and between two MPs. Instead, as in other large-scale assessments such as PISA and TIMSS, student performance in each test was scaled on the basis of *item response theory* (IRT; see Chapter VI.2) and linked to a common scale using the generalized *Rasch* model, which was implemented with the software ConQuest (Hartig, 2007). Missing data were multiply imputed, and five *plausible values* (PVs, see chapter VI.2.5.2) of the test scores were drawn taking into account all context and questionnaire variables (Hartig, Jude, & Wagner, 2008; Hartig, 2007) in order to achieve optimal population estimates of student outcomes and of the relationships between student outcomes and other factors.

At MP1, the population estimates of student achievement in each test had a normal distribution with a mean $M = 500$ and standard deviation $SD = 100$ (Hartig et al., 2008). Student progress (growth) over two MPs amounted to 27 LC-test points (effect size $d = 0.27$) and 23 C-test points ($d = 0.23$). Students of the highest school track (Gymnasium) achieved higher growth (38 points in the LC-test, 27 points in the C-test) in comparison to students of the lower school tracks (20–25 points in the LC-test, 16–28 points in the C-test) (DESI-Konsortium, 2006).

III.1.5 The video study and its main results

The video study was conducted and implemented, data were analyzed, and results were reported by the research team at the University of Koblenz-Landau. For a full description of the study aims, implementation, and technical report see A. Helmke et al. (2007); for all results, see T. Helmke et al. (2008). The study was designed and implemented analogously to the TIMSS video study 1999 (Stigler, Gonzales, Kawanaka, Knoll, & Serrano, 1996). Two video cameras were implemented in each class; one focused on the teacher and teaching activities, the other on the students and the learning activities. Verbal and visible nonverbal expressions of teachers and students were transcribed one by one. Based on the transcripts, video data were then coded according to a coding manual (low-inference, on a descriptive level) and rated according to a rating manual (high-inference, on a more in-depth level) by multiple trained coders and raters with repeated checks of rater reliability (see Chapter VI.1).

The coding manual was developed primarily based on the *communicative orientation of language teaching* observation scheme (COLT, Spada & Fröhlich, 1995, see Chapter III.2.6.1). Furthermore, both coding and rating manuals were developed based on interdisciplinary theories and empirical findings: the empirical classroom and teaching effectiveness research field in Germany (A. Helmke, 2004; Wellenreuther, 2004), the EFL didactics (Bausch, Christ, & Krumm, 2002; Heuer & Klippel, 1993; Timm, 1998), teaching foreign languages and classroom observation research from English-speaking

countries, in particular the United States (Judd, Tan, & Walberg, 2001; Spada & Fröhlich, 1995; Waxman, Hilberg, & Tharp, 2004), research findings from German and international large-scale assessments, and video studies in classroom instruction research (Baumert et al., 1997; Hiebert et al., 2003; Klieme & Reusser, 2003; Prenzel et al., 2002; Reusser & Pauli, 2004; Stigler, Gallimore, & Hiebert, 2000). Both manuals covered a wide range of classroom instruction factors which had been theoretically and/or empirically found to be significant effects for motivational and academic student outcomes in general as well as in EFL (c.f. A. Helmke, 2014a; see Chapter IV).

The recorded English lessons were described in a multifaceted manner using the video data. On average, English teachers spoke more than half of the lesson time (68% of the entire class speaking time) and twice as much as all students together. Not only the amount of time but also the sequence of teacher-student dialogues were coded. The standard case was a dialogue with two stations, in which the teacher asked a question (station 1), a student replied (station 2); the dialogue was then finished. Dialogues with more than three stations seldom happened in a lesson. Teachers used 13.2% of their entire speaking time to ask questions, but they often did not give students enough time (more than three seconds) to consider an answer. Instead, they immediately gave further guidance, instructions, or asked another question. When students were given time to think, they were able to give an answer after approximately 6.7 seconds on average.

Students independently expressed themselves in English in nearly half of their entire speaking time (47.9%). During these free speaking periods, 32.5% of the student verbal expressions were full sentences, 20.2% sentence fragments, 33.4% one-word sentences, and 13.8% unfinished sentences which were interrupted by the others. During their entire speaking time, students read texts out loud (own texts or other texts) in 26.8% of their speaking time, spoke freely (6.2% of speaking time), or produced other kinds of statements (e.g., repetition). In total, one-fifth of student statements were erroneous (21.6%). Among all of the erroneous student statements, only 52.4% were corrected, mostly by the teacher (85.9%). Students were seldom given the chance to correct their own errors or those made by other students (14.1%).

In general, the video data showed that English lessons were rather teacher-centered and that teachers tended to greatly underestimate their own speaking time. Taking the complexity of the teaching work into account (Doyle, 2006), it is not surprising that, with simultaneous multiple teaching activities, teachers barely had the capacity for self-reflection.

The descriptive information on the English lesson gained through basic coding served in the next steps as a basis for independent variables for correlative analyses with student achievement and growth at

class level. The observable instructional characteristics correlate much higher with growth regarding student performance measured by the LC-test than by the C-test (T. Helmke et al., 2008). An achievement-enhancing English lesson on listening comprehension is characterized by a higher percentage/frequency of student speaking time in general, teacher and student speaking time in English, self-correction by students, teacher patience (waiting after a question), teacher-student dialogues with three or more stations, and a lower frequency of one-word sentences.

Among the more in-depth instructional quality dimensions obtained via ratings, the adequacy of educational aspirations, effective classroom management (task orientation, effective time management, and prevention of disturbances), student motivation (student commitment), and adequately dealing with student mistakes positively correlate with student growth based on the results of the LC-test. On the other hand, a negative effect of the structuredness of English lessons was found. Unlike findings regarding the school subject mathematics, no relationship was found between the difficulty and authenticity of teachers' questions and student outcomes. While narrow-focused questions have been found to have a negative effect on student achievement in mathematics, they were found to have a positive effect on student achievement growth in English based on the LC-test. Moreover, the results revealed a positive effect of reading self-produced texts out loud on LC achievement growth.

Regarding the C-test results as dependent variables, the more teacher speaking time in English and the higher the level of educational aspirations, the stronger student achievement growth. Clarity of instruction, a positive learning climate, support-orientated teaching, positive feedback, teacher commitment, and admission of not knowing about something have all been found to have a positive influence on student growth, as measured by performance in the C-test.

Right after the videotaped lessons, students and teachers answered a short questionnaire about the recorded lessons and their judgment of instructional quality. Based on them, the representativity of the lesson and the diagnostic competence of the teachers were investigated. The student short questionnaire included the following scales: comprehensibility of the learning materials, attention, perceived difficulty, interestingness, active participation in the lesson, and representativity of the experienced lesson. The topics covered in the teacher questionnaire were: nervousness of the teacher during the recording, the central topic of the lesson, satisfaction with the recorded lesson, satisfaction with student behavior, lesson plan and deviations, diagnosis of student learning environment (attitude, comprehensibility, too challenging, not challenging, interestingness), self-estimation of own behavior (speaking time, number of questions, number of answered and unanswered questions, number of closed-ended questions, number of mistakes, number of corrected mistakes), and representativity of the recorded lesson.

One interesting finding based on the results of the short teacher questionnaire was that teachers greatly underestimate their own speaking time. More than 50% of the teachers thought they spoke less than 50% of the class speaking time, while in reality only 3% teachers spoke less than 50% of the class speaking time. Furthermore, 10% teachers spoke for more than 80% of the entire class speaking time, but only 2% were aware that this was the case.

III.2 Research project in Vietnam in the school year 2006–2007

The "I" in "DESI" stands for "International". That means, replication studies in other countries were desired and anticipated, although not commissioned by the Standing Conference of the Ministers of Education and Cultural Affairs in Germany. In the field of language teaching and learning, replication studies are highly encouraged (Language Teaching Review Panel, 2008). In fact, the DESI-instruments were applied within a nationwide study to estimate student performance in the German language in all German-speaking schools in South Tyrol (Beck & Dahl, 2006); a pilot study took place in Austria (DESI-Konsortium, 2006). Outside Europe, the DESI-instruments regarding the school subject EFL, including a video study, were applied in Vietnam in the school year 2006–2007 by the Landau-based research group of Prof. Dr. Andreas Helmke, Dr. Tuyet Helmke, and Dr. Wolfgang Wagner in cooperation with Dr. Tran Thi Bich Tra and colleagues at the National Institute for Education Strategy and Curriculum (NIESC) in Hanoi, Vietnam.

Analogous to in the DESI study, the students in Vietnam were tested twice, namely at the beginning (October 2006) and at the end (April 2007) of the school year. At each measurement point, two English tests (the C-test and the LC-test) and a student questionnaire were applied. The teacher questionnaire was applied once at the second MP. The video study was conducted over two months in the middle of the school year, in January and February 2007. With these data, the observable instructional indicators and their predictive values for student achievement and growth were estimated. In this chapter, the operationalization of this study will be briefly outlined.

III.2.1 Sample

The sample of the study consisted of the English teachers and 2,096 students of 50 ninth-grade classes in the two special cities in Vietnam (Hanoi, Ho Chi Minh City) and a rural province (Bac Ninh). A total of 41 teachers and their class students voluntarily took part in the video study in the middle of the school year. The school and class sample were selected and contacted by the NIESC, the project partner in Vietnam in 2005.

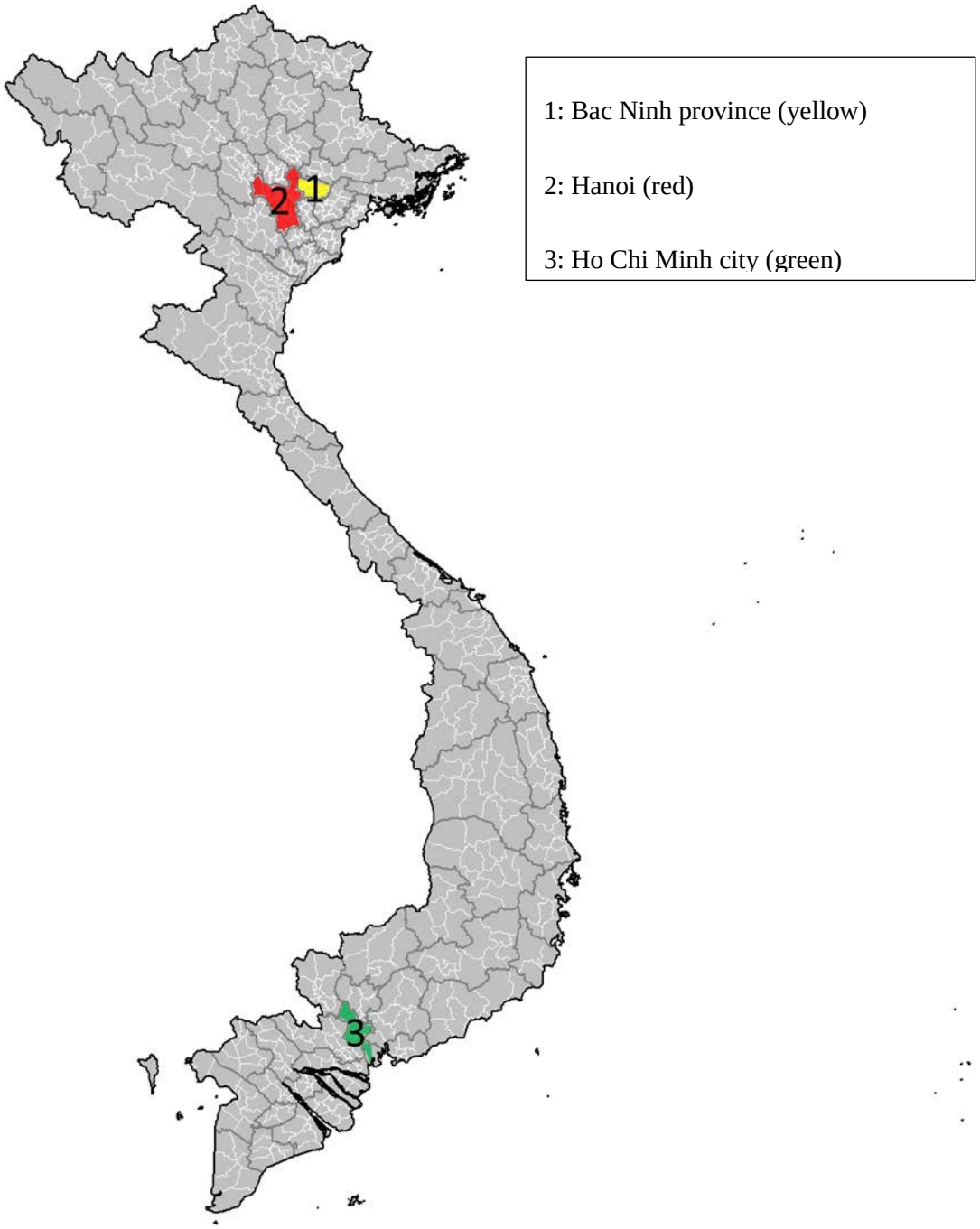

Figure 3: Data collection locations in Vietnam

Source: https://commons.wikimedia.org/wiki/Atlas_of_Vietnam (retrieved on 11.03.2017)

In Bac Ninh province, there were ten lower secondary schools in total at the time, and all of them were selected to take part in the study. In Hanoi and Ho Chi Minh cities, 20 schools were randomly sampled in each city. The number of sample schools in each administrative district in each city was proportional to the total number of lower secondary schools in each district. From each selected school, one ninth-grade English class was randomly chosen, so that the distribution of student achievement in English at

the end of the previous school year (eighth grade) was representative for Hanoi, Ho Chi Minh City, and Bac Ninh. About one-third of the student sample in each city/province performed above-average, average, and below-average according to the official school grade statistics in the previous school year.

III.2.2 The adapted English tests

In order to adapt the DESI test instruments to the Vietnamese sample, a pre-pilot study and a pilot study were conducted by Dr. Tuyet Helmke, Dr. Wolfgang Wagner, and colleagues. One school in Hanoi with average student achievement in the school year before (2005–2006) was chosen for this purpose. All 365 ninth-grade students of this school participated in the pre-pilot and pilot studies.

The pre-pilot study was conducted from October to December 2005. The aim was to test the difficulty and length of the original C-test and LC-test in the DESI study for the Vietnamese sample.

After the pre-pilot study, the test instruments were adapted and tested once again at the end of the school year (the pilot study). The pilot study was conducted from April to June 2006 on the same student sample. Next, the test instruments were adapted once again to finalize them for use in the main study in the following school year.

A comparison of student performance between countries was not anticipated (DESI-Konsortium, 2006). In Vietnam, in particular, due to the different length and difficulties of the test booklets and differences in test administration, the test results of students in two countries could not be reliably linked to a common scale (Holland, 2007).

The main results of the pre-pilot and the pilot study will be provided in the following sections.

III.2.2.1 The C-test

In the DESI study, the C-test battery consisted of twelve short texts (testlets); each testlet had 25 gaps to fill in, with each gap consisting of one item. One test booklet contained six testlets with a total of 150 items. In total there were six C-test booklets with anchor testlets in the different booklets. Students of each school track in Germany were assigned test booklets with appropriate difficulty levels. The anchor testlets which were given to all school tracks to ensure overall test scaling. (Harsch & Schröder, 2007).

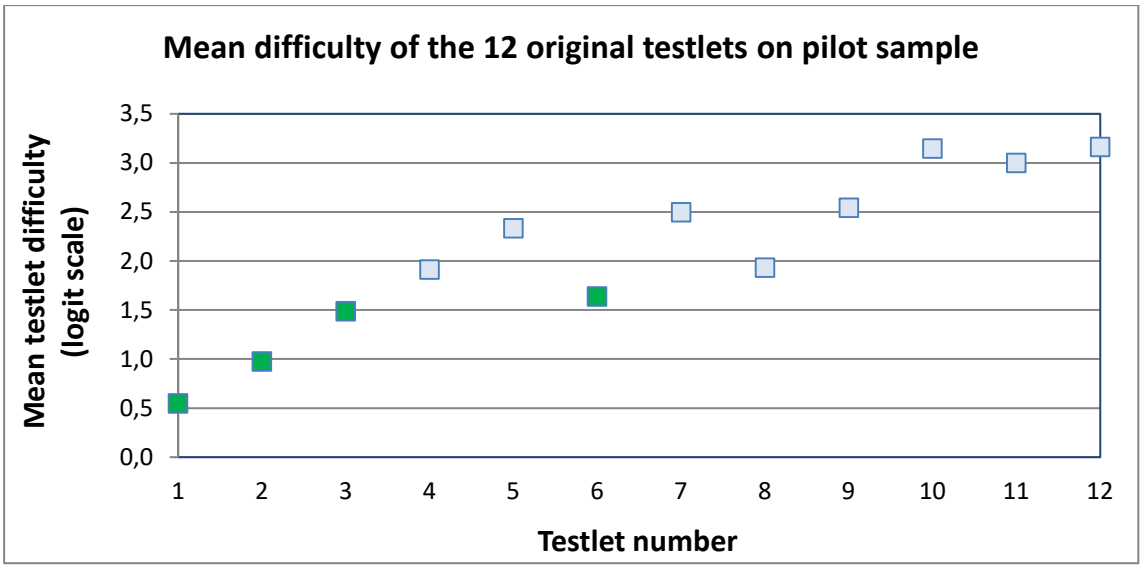

Figure 4: Difficulty of the C-Test testlets in the pilot study

In the pilot study in Vietnam, all six original C-test booklets were evenly and randomly distributed to the student sample at both MPs. Each student was assigned one booklet at one MP and another booklet at the other MP. Because the number of students per item at each MP was rather small, the results of both MPs were analyzed together.

The student responses were coded according to the DESI coding guide: 1 if the answer was correct; 0 if the answer was incorrect; 8 if the answer was not assignable; 9 for missing answers.

The data of the pilot study were analyzed together like in the DESI study using the Rasch model. Figure 4 shows the mean difficulty of all testlets on a logit scale. Figure 5 shows the difficulties of all items in comparison to student ability estimates (with mean $M = 0$ and $SD = 1$). In general, the test items were found to be too difficult for the pilot sample, and the C-test booklets were identified as too long with a very high percentages of not-reached items (all consecutive missing items clustered at the end of a booklet except the first value of the missing series, definition used by PISA, see OECD, 2012, p. 329).

As a consequence, the project researchers decided to shorten the C-test booklets to only four instead of six testlets, and chose the easiest four testlets (which are marked in green in Figure 4) to create the C-test booklet for the main study in the 2006–2007 school year.

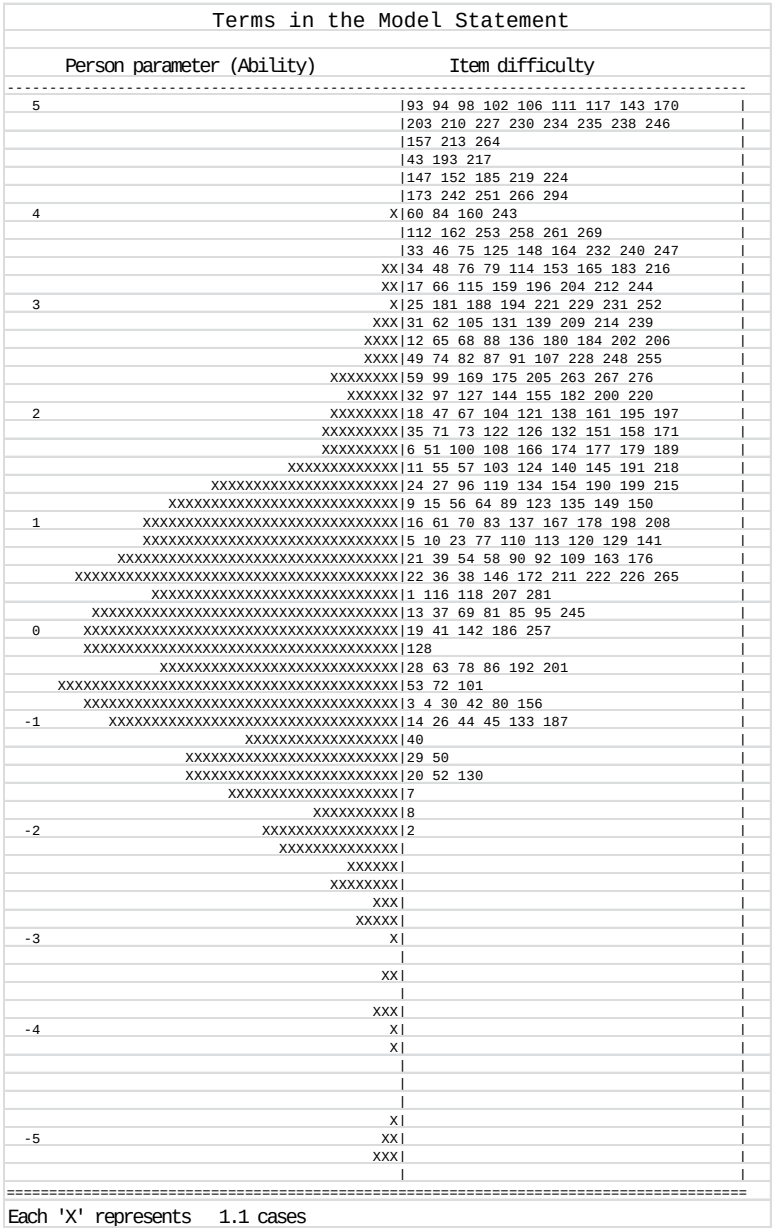

Figure 5: C-test item difficulties and student ability (pilot study)

Based on these chosen four testlets, two C-test booklets were designed for the main study. Both had the same testlets but in different sequences: 1, 3, 2, 6 in version V1 and 2, 6, 1, 3 in version V2 (see Figure 6). In the main study later on, at each measurement point, students who sat next to each other received different booklet versions to avoid a copy effect. Each student received one booklet version at the first MP and the other booklet version at the second MP.

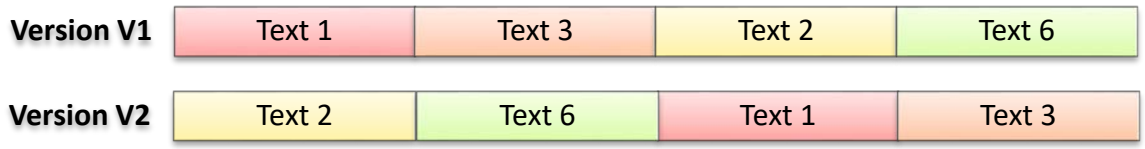

Figure 6: Text order of the two C-Test booklet versions

III.2.2.2 The listening comprehension test

In the DESI study, the listening comprehension test (LC-test) battery consisted of two parts. In the first part, there were 16 multiple-choice items, each related to one of 16 short dialogues (10 to 20 seconds). This part was called the dialogue module. The second part, the conversation part, was based on four longer conversations of about a two-minute length; there were 7 to 10 multiple-choice items on each conversation. The dialogues and conversation recordings were of English native speakers, who spoke in a natural way and speed like in everyday situations. One LC-test CD contained eight short dialogues and one long conversation with pauses in between so that students had enough time to answer the question(s) on it by choosing one of the four given options. All students listened to the same CD and worked on the same items. (Nold & Rossa, 2007)

In the pre-pilot study, the researchers recognized that the original CDs were too difficult for the student sample, even for students with very good school marks in English. Most of them could hardly understand any content of the dialogues and conversations, because they were not familiar with listening to native English speakers speaking at a normal tempo. In English lessons in Vietnam, they were acquainted with listening to the dialogues and conversations at a reduced speech tempo and with particularly clear pronunciation. As a result, the CDs for the LC-test had to be reproduced with the help of two English lecturers at the university of Koblenz-Landau. The same contents were spoken at a much slower tempo with clear articulation. The test with the new CDs was repeated in the pilot study later on.

The results showed that some items were noticeably easier for the Vietnamese sample than for German students due to cultural relevance (differential item functioning, see Appendix B5). Due to this, the project researchers decided to eliminate dialogues and conversations with a different item difficulty rank order to that in the DESI-study from the test, which were supposed to measure both listening comprehensive skills and constructs which were irrelevant to the test construct such as cultural knowledge.

On average, the LC-test with reproduced CDs was still rather difficult for the student sample. Analogous to the C-test, only the easiest modules were chosen to create the LC-test booklet for the main study.

Finally, one dialogue module (comprising eight questions) and one conversation module (with ten questions) were selected to make up the only LC-test booklet for the main study.

III.2.3 The adapted questionnaires

The DESI student and teacher questionnaires were adapted by Dr. Tuyet Helmke and her research team at the University of Koblenz-Landau and at the NIESC in Vietnam. Items and scales that were not relevant to the teaching and learning situation in Vietnam were eliminated.

All items were first translated from German to Vietnamese by one translator, then back from Vietnamese to German by another translator. The back translation was reviewed and validated by members of the DESI research group at the University of Koblenz-Landau. With this procedure, the correctness and cultural appropriateness of the translated questionnaires were ensured.

III.2.4 Data collection

The main data collection was carried out by a group of Vietnamese researchers headed by Dr. Tuyet Helmke (University of Koblenz-Landau) and Dr. Tran Thi Bich Tra (NIESC, Hanoi, Vietnam). The first MP took place at the beginning of the school year from October to November 2006, the second MP at the end of the school year from April to May 2007. At each MP, first the C-test and the LC-test were applied, then the questionnaire. The students had 30 minutes for each test, and 15 minutes for the questionnaire. The total testing time including the introduction and short pauses between the tests/questionnaire was 90 minutes. The test conditions were the same for all classes and at both MPs. In each testing class there were two test supervisors, both were members of the research group. First, the students in Hanoi were tested, then in Bac Ninh province, and finally the students in Ho Chi Minh city.

As described in the previous section, there were two C-test booklet versions which contained the same testlets and items but in a different order. Students who sat next to each other were assigned different booklets. Each student was assigned a different C-test booklet version at a different MP. The LC-test booklet and CD were identical for all class students and at both MPs. Although the same test booklets were reused at MP2, little memory effect was assumed due to the long time interval (nearly one school year) between the two MPs. All test booklets were taken away after the first MP, and the correct answers were not given to the students.

III.2.5 The video study

Analogous to the DESI video study, the video study in Vietnam took place in the middle of the school year between two MPs, from January to February 2007. It added observational data to the replication study, which allowed investigation of the teaching and learning situation on a much more fine-grained level. The video study attempted to (1) describe and analyze the actual practice based on objectively measurable criteria both qualitatively and quantitatively, (2) acquire behavioral indicators that are relevant to student development of English performance, and (3) take stock of oral communication in English lessons. A total of 41 English classes voluntarily participated in the video study; two English lessons were videotaped of each of the classes. The design required the teachers to implement two different approaches per lesson: (a) an authentic lesson based on the central curriculum for all ninth-grade English classes in Vietnam, and (b) an extracurricular speaking lesson with the prescribed topic "Tet" – the New Year festival in Vietnam, which occurred close to the recording time. In the second lesson, teachers were requested to organize the lesson so that they and their students spoke as much as possible. After the videotaped lessons, short questionnaires (see chapter III.1.5) were applied for students and teachers.

The video data of the first lesson can serve as material to investigate regular classroom instruction in EFL school classes in Vietnam. Video data of the second lesson are of particular importance for investigating aspects of language quality, especially the common speaking mistakes made by teachers and students. Short questionnaire data are helpful to examine the representativity of the lesson as well as to investigate the appropriateness of teachers' self-judgment of their own teaching.

III.2.6 The adapted basic coding guides and rating sheet

Analogous to the DESI video study, recorded video data were documented and analyzed using the software Videograph (Rimmele, 2007). First, all verbal products of the lessons were transcribed with each sentence as the basic unit (*turn*). Then, two analysis methods were applied: micro-analytical basic coding and hour-based rating.

III.2.6.1 Transcription and basic coding guides

The DESI coding guide was developed based on the Communicative Orientation of Language Teaching (COLT) observation scheme (Spada & Fröhlich, 1995). The most important coding categories are:

- Dialogue participants (Who spoke to whom?)
- Percentage of teachers' and student speaking time, nonverbal spells
- The language used (English, Vietnamese, mixed)

- Subject (related to tge syllabus, related to discipline, procedural, social)
- Length of student statement (one-word statement, sentence fragment, full sentence, sentence interruption by teacher or by another student)
- Type of the student statement (repetition, reading out text of others, reading out own text, speaking based on instructions, speaking freely)
- Correctness of student utterances
- Syllabus-related teacher's turns (presentation and instruction regarding structure, structuring aids, questions/types of questions, student-oriented teacher responses such as assisstance, feedback/sanction, error handling, characteristics of the questions, self-answering)
- Characteristics of the question (linguistic complexity of the answer required from the students)
- Content of teacher feedback (affectively stressed/neutral positive feedback, mixed feedback, affectively neutral/stressed negative feedback)
- Dealing with mistakes, type of correction (how), type of mistake dealt with by teacher (which)

Moreover, the *contextual* background in which a specific learning episode (see …) took place was taken into account and also coded. The *combinational* (e.g., percentage of sentence fragments in Vietnamese vs. in English) and *sequential* (analysis of occurrences in time sequence, e.g., frequency of teacher questions that follow by waiting time, or frequency of the teacher-student-dialogue pertaining to a specific length and a specific student) approaches were also applied. By means of basic coding, *low-inference* (observable and objectively quantifiable) data were acquired on the English lessons.

In one class lesson, there were phases which were clearly separated from each other in terms of their function (didactic) and form (social, task-based). They were called *episodes*. The meaning and importance of teacher behavior or classroom activities varied depending on the episode within which they occurred. Therefore, the episodes of each English lesson were necessarily classified. The following episodes were defined and coded in the DESI video study: teacher presentation, teacher-oriented conversation, student presentation, student-oriented conversation, educational game, use of audiovisual materials, individual work, pair work, group work, transition.

Several coding categories in the DESI coding guide that were never observed in the recorded lessons in Vietnam, such as learning games other than role-play, were skipped. In addition, the coding guides were adapted to enhance the validity and inter-rater reliability of the basic codings given the new teaching and learning situations in Vietnam (see chapter VI.1).

Furthermore, the coding guide was extended to include a domain-specific topic: the common types of speaking mistake because this is a significant problem in Vietnam (see chapter II.5). For this purpose, the data of the extra-curricular speaking-oriented lesson were coded. Each speaking turn of this lesson was coded as to whether it was correct or incorrect with regard to content, situation/context, phonology, lexicon, and grammar. If a phonological mistake occurred, the type of mistake was coded in a second step. The extended coding scheme broke down the phonological mistakes into the following types, based

on the suggestions of Eckert & Barry (2005), who explained different phonetic phenomena and places where English learners often make mistakes in general, and the suggestions of Tang (2007, see Chapter II.5) regarding the specific problems of Vietnamese learners:

- Using incorrect vowels (e.g. /ɛ, a/ instead of /æ/)
- Lengthening pronunciation of short vowels
- Shortening pronunciation of long vowels
- Skipping or softening hard consonants
- Skipping final consonant clusters
- Rolling 'r'
- Adding 't' to word ending
- Adding 's' to word ending
- Using wrong word stress
- Using wrong sentence stress
- Speaking in a clipped manner
- Other mistakes

The adapted coding guides can be found in Appendix A1 and Appendix A2. Based on these basic codings, the relative frequencies and time percentages of each occurrence were calculated. On this basis, basic coding variables were developed, such as student speaking time and percentage of time during which teacher used mixed languages.

III.2.6.2 Rating sheet

Beside the basic codings, ratings – high-inference measures which require raters to make inferences from a series of classroom events using specific constructs (Rosenshine & Furst, 1971) – were also implemented. The ratings referred to the whole class lesson as a basic unit. As a basis, the DESI rating sheet was adapted to cover a wide range of general and domain-specific instructional quality dimensions (see Chapter IV) such as student orientation, classroom management, dealing with mistakes, quality of motivation, and quality of teacher language. Four rating categories were used to rate each lesson: 1–totally do not agree, 2–rather not agree, 3–rather agree, and 4–totally agree. The entire adapted rating sheet, including all rating variables and categories, can be found in Appendix A3.

IV. Theoretical background

The relationships between student learning outcomes and schooling factors as well as context effects have aroused great interest among international educational researchers since around the late 19th, early 20th century (Lüders & Rauin, 2004). Hattie drew on more than 800 meta-analyses published in the English language until 2009 for his monumental meta-meta-analysis, which encompassed more than

50,000 studies about different factors of student achievement outcomes. Besides prior knowledge, socioeconomic background was found to have one of the most significant individual and context effects on student achievement. More than 35,000 studies have been conducted on teacher and teaching factors, revealing more than 76,000 effects.

Empirical educational research covers a wide range of research areas; each area has its own main research topic and preferred research methods, which can be categorized according to different criteria: focus on the micro level (e.g., cognitive processes in learning, classroom processes, etc.) vs. the macro level (e.g., school factors), focus on general aspects of teaching and learning (e.g., general didactics, empirical teaching-and-learning research, classroom and teaching effectiveness research, psychology of instruction) vs. subject-specific aspects (e.g., subject-related didactics), focus on qualitative vs. quantitative research methods, experimental research vs. field research, focus on elementary vs. secondary school age, and so on (A. Helmke, 2014b; Helsper & Böhme, 2004). Because many topics and aspects do not stand alone, but are connected and interact strongly with each other, there has been a call and tendency toward merging different areas and disciplines and toward using multi-methods in the field of educational research (Krüger & Pfaff, 2004; Moss & Haertel, 2016; Teddlie, Reynolds, & Pol, 2000; Teddlie & Sammons, 2010), so that there actually are no strict boundaries between many research areas (Creemers et al., 2010a; A. Helmke, 2014b; Lüders & Rauin, 2004).

The DESI study is eclectic because the research team that developed it adopted fundamental theories and empirical findings from different research traditions (c.f. chapter III.1), namely research on teaching effectiveness and teaching and learning as well as research on EFL. In this chapter, the main characteristics and findings of these two research areas will be outlined.

IV.1 Academic student achievement and growth

In this study, the term “student achievement” refers to the test results of students at one measurement point; growth is measured by a change in scores between the two MPs (see chapter VI.2). The results of national and international studies on student achievement based on standardized tests in mathematics and reading have commonly suggested $d = 0.4$ to be the benchmark for expected student growth per year at secondary school age (Hattie, 2012).

Prior student achievement has been confirmed as a significant predictor of student achievement at all grade levels, with an average effect size of Cohen’s $d = 0.67$. About 48 percent of variance in student achievement can be explained by student differences in prior achievement (Hattie, 2009). This was

confirmed in the school subjects mathematics and Vietnamese, although the explained variance proportion was smaller, around 20 percent (Rolleston, James, Pasquier-Doumer, et al., 2013).

Furthermore, aggregated prior achievement at class level was found to significantly contribute toward student achievement later on, over and above the contribution of individual prior ability (for a review, see Dumont, Neumann, Maaz, & Trautwein, 2013). This distinct contextual effect is called the *academic class composition effect*. In Germany, research findings have shown that academic school and class composition is the most influential composition variable of schools and classes on student outcomes (Baumert, Stanat, & Watermann, 2006; Gröhlich, Guill, Scharenberg, & Bos, 2010; Stanat, Schwippert, & Gröhlich, 2010). In Asia, this was confirmed in Korea (Kang, 2007). In China, an academic class composition effect was found for mathematics, but not for English (Carman & Zhang, 2012); however, it should be taken into account that the variance between classes regarding initial test scores in English was small in the Chinese study.

Knowing the individual and class composition effects of initial student achievement on student outcomes has led researchers to develop longitudinal designs for research on school and classroom instruction effects (including the DESI study). Also, this serves as a basis for the development of value-added models, which consider the initial achievement of students and classes as important covariates that have to be controlled for in order to appropriately estimate the effectiveness of teacher and classroom instruction for student outcomes (Harris & Herrington, 2015; National Audit Office, 2003; Ray, McCormack, & Helen, 2009; Saunders, 1998, 1999).

IV.2 Effect of socioeconomic background on academic student outcomes

At the latest since the Coleman report (Coleman et al., 1966), the influence of student socioeconomic status (SES) on their learning outcomes has attracted the attention of international educational researchers. On average, the effect of SES on academic student outcomes was $d = .57$, which did not vary much between various types of achievement and different sub-components of SES, such as parental education, parental occupation, or parental income (Hattie, 2009).

For the school subject EFL and for low-income countries, however, there have been differential findings regarding the effect of SES on academic student outcomes. While a positive correlation between student SES and initial student achievement was confirmed, the findings of the DESI study in Germany showed no effects of SES on student outcomes in the C-test after controlling for the effect of initial student achievement (Rolff, Leucht, & Rösner, 2008). In low-income countries, the effect of SES on student achievement has been found to be larger than in high-income countries (Heyneman & Loxley, 1983).

For East-Asian countries including Vietnam (also a low-income country), on the other hand, A. Helmke & Hesse (2010) predicted a small effect of SES on student outcomes due to the culture of parents placing a high value on education, regardless of their social and educational background. Results of PISA 2012 and PISA 2015 supported this prediction, that Vietnam was among countries with below-OECD-average strength regarding the relationship between student performance (in mathematics and sciences) and socio-economic status (OECD, 2014a, 2016).

In Vietnam, Rolleston, James, Pasquier-Doumer, et al. (2013) found that more socially advantaged students and schools often obtained better test scores in mathematics and Vietnamese; and student background factors explained 5% to 7% variance within classes based on test scores in mathematics and Vietnamese.

Similar to initial student ability, many findings have confirmed that the social class composition effect regarding student SES is stronger than the individual effect of SES on academic student outcomes (for a review, see Dumont et al., 2013). Especially students with lower initial achievement suffer from having a socially disadvantaged class composition (De Fraine, Van Damme, Van Landeghem, & Opdenakker, 2003; Zimmer & Toma, 2000).

IV.3 Empirical classroom and teaching effectiveness research

The empirical research field of classroom and teaching effectiveness has its root in scientifically and empirically elaborate underlying theories about the mechanisms of classroom teaching-and-learning processes, and often uses field research with advanced quantitative research methods.

After a long development, there has been a shift from a *person-centered* approach to a *process-product paradigm* (*variable-centered* approach). Rather than searching for stable personal characteristics, which were assumed to facilitate teacher effectiveness according to the early approach (Mayr, 2014, p. 189), the *process-product paradigm* focuses on the relationships and interactions between subject-non-specific aspects of classroom instruction and student outcomes while taking into account mediating processes – other personal, contextual, institutional, and cultural factors (A. Helmke, 2014b, pp. 807–809). During its progress, this research field has profited from and been backed by major advances in other research areas, most of all in the psychology of learning and instruction as well as in psychometrics (Lipowsky, 2006). Recently joining the trend toward interdisciplinary research, studies have merged on the subject-specific aspects of instruction, and there has been renewed interest in the personal characteristics of teachers – especially their professional knowledge and expertise (A. Helmke, 2014b). The development track of this research field, its key elements as well as the relationships and interactions

between them are successfully summarized and illustrated by the *models of instructional provision and uptake* (A. Helmke, 2014a; Lipowsky, 2006; Reusser & Pauli, 1999).

In the following sections, the underlying theoretical framework of the DESI study, the *model of instructional provision and uptake* (A. Helmke & Klieme, 2008) together with the main findings regarding subject-non-specific aspects of effective teaching in this field will be described.

IV.3.1 Model of instructional provision and uptake

Following the conception of Fend (1998) and Helmke and Weinert (1997), Helmke (2014a; 2002) introduced a *model of instructional provision and uptake* (see Figure 7). This framework model takes into account the differences in student outcomes (output) with regard to multilevel multi-determinants: background variables such as student characteristics and initial learning potential (e.g., previous knowledge, intelligence, motivation), family background (e.g., socioeconomic status), context variables including school, regional, and cultural context, class composition, and the main stakeholders and their activities: the teachers and their instruction, diagnosis of student learning, the perception of instruction as well as teaching and learning activities.

This model supplies a framework for studying the teacher and teaching effectiveness, while clearly indicating the difficulties in order to measure them, because an accurate estimate should take into account all other confounding factors.

The core-elements of classroom processes are:

- teaching processes and instructional quality (both subject-specific and general, quality of teaching materials, quantity and quality of opportunities to learn) which are partly conditioned by professional and personal teacher characteristics (e.g., professional competency and expertise, pedagogical orientation, subject-specific competence, expectations, engagement, humor) on the *provision side*; and
- learning activities and processes (inside and outside of the classroom) which are partly conditioned by the initial learning potential on the *uptake side*, and the perception of classroom processes and own teaching by teachers (teacher diagnosis of student learning and awareness of own instruction) and students (student perception of classroom instruction, and own ability and learning activities) as mediation processes, which supply them with orientation for their activities.

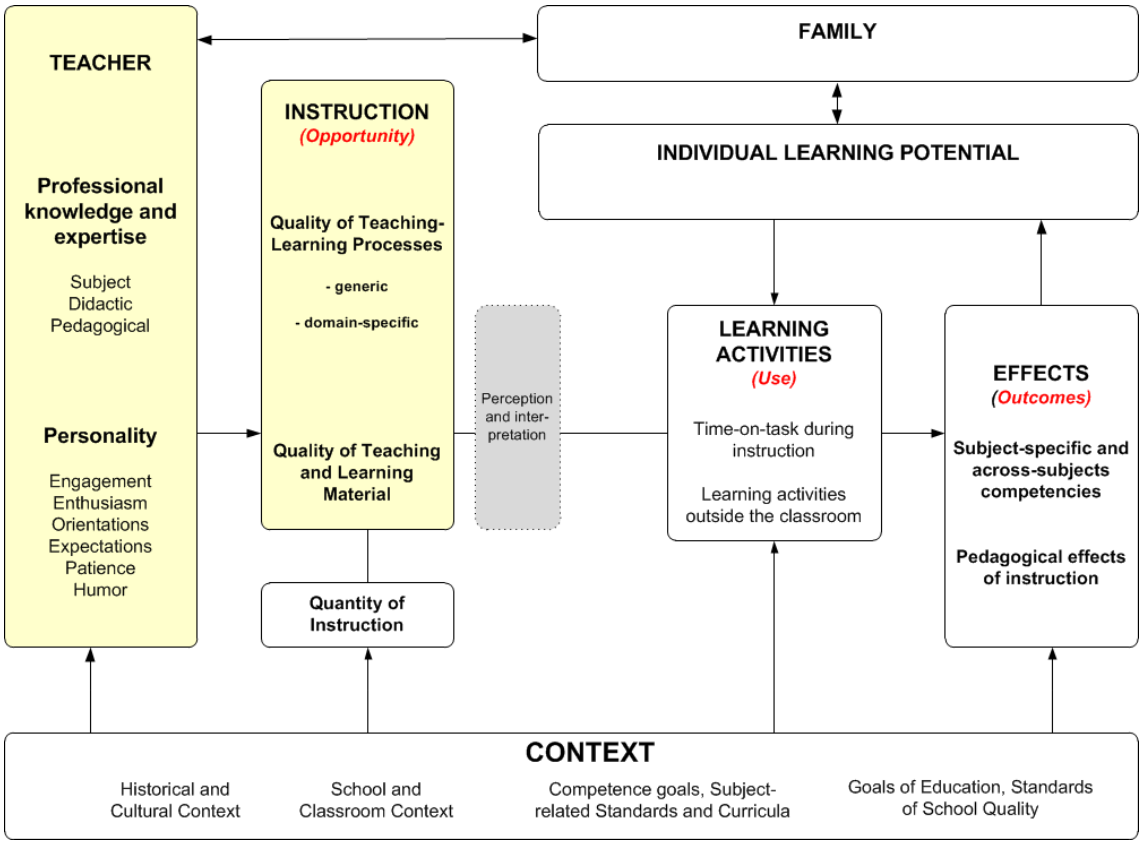

Figure 7: Model of instructional provision and uptake (adapted from A. Helmke, 2014a)

According to this model, in interaction with their students, teachers are responsible for creating learning opportunities that make insightful learning processes possible. Furthermore, this model features essential characteristics of the process-product paradigm:

- *Aptitude treatment interaction* – which refers to the interindividual differences in taking up the available opportunities on the student side.
- *Multidirectionality* of the effects between teaching and learning: Not only students and their teaching are influenced by teachers, but teachers and instruction are also influenced by student initial learning potential and the learning activities.
- *Dynamics* of the network: The present student output becomes future learning potential, and thus influences all future processes.
- Predetermination of the *context effects* at the macro level (cultural, regional, institutional, etc.) which both teachers and students cannot influence.
- *Indirectness* of instructional effects on student outcomes: There is no direct path from teaching variables to student outcomes; the effects of teaching on outcomes are mediated by a student's perception and interpretation of instruction and the learning activities (A. Helmke, 2014a).

IV.3.2 Instructional quality dimensions and teaching effectiveness

Although teacher and teaching effects are indirect, the role of teachers and their teaching is crucial for student outcomes and learning progress, as findings from several longitudinal studies have shown (Gruehn, 2000; Lipowsky, 2006; Rivkin, Hanushek, & Kain, 2005). Especially under-achievers profit largely from good instruction and suffer from bad instruction (Babu & Mendro, 2003; Lipowsky, 2006; Nye, Konstantopoulos, & Hedges, 2004; Rivkin et al., 2005). A considerable proportion of variance in student learning progress is associated with variations in teacher effectiveness (21%, corresponds to an average effect of $d = .32$, Nye et al., 2004). Other findings have shown larger teacher effects in mathematics than in reading (Hattie, 2009, p. 108).

According to the results of thousands of empirical studies in this field, the strongest empirical evidence has revealed that the following general instructional quality dimensions accelerate student academic progress: effective classroom management (Evertson & Weinstein, 2006; Marzano, Gaddy, Foseid, Foseid, & Marzano, 2005; Marzano, Marzano, & Pickering, 2003), clarity, intelligibility, and structuredness of instruction (Hattie, 2009; A. Helmke, 2014a; Meyer, 2004; Slavin, 1994), supportive classroom climate, motivation, individual learning support (Brophy, 2000; Hattie, 2012; A. Helmke, 2014a; T. Helmke et al., 2008; Meyer, 2004), and cognitively activating instruction (Baumert & Kunter, 2006; Baumert et al., 2010; Hattie, 2009; Klieme & Baumert, 2001; Klieme, Pauli, & Reusser, 2009; Marzano et al., 2005).

Regarding professional competencies, a teacher should possess all required knowledge, methods, and skills in their field. According to Baumert & Kunter (2013), the professional competencies of a teacher can be split into the following four domains: specific declarative and procedural knowledge (competence in the narrow sense: knowledge and skills); professional values, beliefs, and goals; motivational orientations; professional self-regulation skills.

Based on the results of nearly one thousand meta-analyses about the effects of student academic achievement, Hattie (2012) reached a synthesis which largely supports the main thesis of the model of instructional provision and uptake, highlighting the effects of the above-depicted instructional quality dimensions and professional competencies. In addition, he emphasized another important aspect of instruction which had previously gained little attention (A. Helmke, 2014b) in this research field: feedback – "*It soon became clear that feedback was among the most powerful influences on achievement*" (Hattie, 2009, p. 173). With this he meant different types of feedback: feedback via frequent testing, teaching test-taking skills, providing formative evaluation to teachers, questioning to provide teachers and students with feedback, and the immediacy of feedback. In particular, feedback

from students to teacher is "*most powerful*", because it can help them to see learning *through the eyes of students*, which is the core of his synthesis.

The synergy of numerous findings pointed toward high-level principles of effective teaching and learning, rather than toward any teaching method by itself proving to be more or less effective, according to Hattie (Hattie et al., 2013, p. 288). The high-level principles of effective teaching can be simply formulated in order to make teaching and learning "*visible"* for the teacher: "*The major argument presented in this book is that when teaching and learning are visible, there is a greater likelihood of students reaching higher levels of achievement*" (Hattie, 2012, p. 14). This lens enables them to activate the learning process of the students, managed by the students themselves ("*I help students to become their own teachers*").

In his summary of the research results in the psychology of second language learning, McDonough (2000) shared Hattie's conclusion: A successful language teacher is not only an instructor, but also a mediator, who provides samples of the language, manages interaction, gives advice on learning, helps students to become independent, provides and manages learning resources, and offers feedback on their performance. Moreover, the crucial aspects of a teacher's behavior, according to studies of teachers and learners in classrooms, are modes of structuring the lessons, provision of feedback and correction of speaking mistakes, coaching in writing, and provision of writing feedback to students.

IV.3.3 Effect mechanism of classroom instruction

According to A. Helmke (2014a), the optimal level of an instructional factor for student progress (and teachers' health) is often, from a theoretical standpoint, not the maximum, but lies *somewhere* in the middle of the entire continuum of the factor scales. Thus, the key for success is appropriate dosage and *balance.*

Due to this, the effects of instructional factors on student progress often turn out to be *nonlinear* rather than linear. Exceeding the optimal point would either not further increase the effect of the factor or even worsen it. In such cases, the assumption of a linear relationship between instructional factors and student outcomes or the use of a measure with this assumption (e.g., correlation, linear regression) could lead to effects being underestimated, overlooked, or misinterpreted.

The relationship between classroom instruction and outcome variables can be both *multiplicative* and *compensatory*. An additive combination of the usual type $Y = \beta_1 x_1 + \beta_2 x_2$ suggests the compensatory assumption that a special strength can compensate for other smallnesses. On the other hand, the multiplicative assumption requires a more complex formula for defining the joint effect of many factors,

since the presence and quality of *one* factor could influence the effect of *another* factor, for example, by including interaction terms in the equation. One example for the multiplicative assumption of the relationship is that an optimal classroom climate is of little use when classroom management is totally inefficient (A. Helmke & Schrader, 1998). These two assumptions suggest that instructional effects should be jointly examined to identify both main and interaction effects of the predictors in order to obtain optimal results and more valuable hints for teachers and their teaching.

Another important interaction is *aptitude treatment interaction* (ATI), which takes into account the interindividual differences between students in taking up the available opportunities (see Chapter IV.3.1). This constitutes one of the most challenging aspects when defining and investigating instructional success, because a specific instruction might be optimal for the learning progress of some students, but not for other students.

IV.4 EFL didactics and the communicative language teaching approach

In the field of EFL didactics, researchers often search for, discuss, and investigate teaching methods which are assumed to be superior and should be promoted (*methods-focused*) based on theories of language acquisition and learning theories.

Since the 1970s, the communicative language teaching (CLT) approach has become one of the most prevalent approaches worldwide in teaching EFL/ESL. In particular, an established classroom observation instrument (the COLT) has been developed to help study this approach. As it was the theoretical background for developing the research instruments in the DESI study, the essence of this approach and the COLT will be outlined in this chapter.

Since its introduction at the beginning of the 1970s by British and American researchers, there has been a world-wide shift toward second/foreign language (L2) teaching methods that focus on the Communicative Approach in Language Teaching (Communicative Language Teaching – CLT) (Cook, 2001; Hunter & Smith, 2012). Asian-Pacific language education policies have shown a strong tendency to follow this trend (Butler, 2011; W. K. Ho & Wong, 2004; Nunan, 2003), as reported for Vietnam; see Ha (2005), Hiep (2007), for example. CLT refers to both processes and goals in classroom learning. The CLT methodology is characterized by less structured and more creative language tasks, putting focus on students, fostering student participation in meaningful L2 interaction in communicative situations to allow them to develop their communicative competencies (Savignon, 2000).

The traditional communicative approach stems from the observation that children acquire their mother-tongue language (L1) through engaging in natural and meaningful communication with others. However, the majority of research evidence does not support the hypothesis that mere exposure to L2 input in communicative situations without a conscious attempt to learn (*implicit learning*, cf. Dörnyei, 2013) is sufficient for L2 acquisition (Lightbown & Spada, 2013). At least since the end of the 1990s, a revised definition of CLT, which is termed the *"principled communicative approach"* (PCA), has been introduced and since then has attracted much attention of international L2 researchers (Celce-Murcia, Dörnyei, & Thurrell, 1997; Dörnyei, 2013). The PCA, while keeping the main principles of the communicative approach (focus on students and their communicative competence through engaging in communication), also pays particular attention to the significance of *explicit learning* for students. Explicit learning refers to the conscious and deliberate attempt to solve problems under teachers' instruction (Dörnyei, 2013). A review of empirical studies which demonstrate a significant advantage of explicit types of L2 instruction over implicit types was given by Norris and Ortega (2000). According to Dörnyei (2013), the PCA is characterized by a simultaneous and equal focus on accuracy (accurate use of L2 structural system), fluency and *formulaic* language (which means knowing language rules such as lexical phrases, idioms, etc.). Lightbown and Spada (2013) particularly emphasize the role of feedback regarding mistakes, especially of feedback in L2 instruction, because the L2 classroom is the "*only place where feedback on error is typically present with high frequency*".

Here, assumptions regarding effective EFL teaching turn out to be quite similar to the assumptions regarding effective teaching in general (cf. Chapter IV.3.3 above): *balance, equilibrium,* and *integration* of more direct instruction of language (including grammatical, lexical, and socio-pragmatic features) with communicative skills within CLT should be strived for, rather than exclusive focus on only one specific technique, skill, or purpose (Spada, 2007). In particular, CLT does not mean avoiding using L1 in classroom instruction. Based on the comprehension of students, teachers should find the right balance between the use of L1 and L2 to make sure that students understand and, at the same time, maximize the use of the target language (W. Wu, 2008).

In order to investigate differences in the *communicative orientation of language teaching* in L2 classrooms and to examine the relationship between L2 instruction and student outcomes, a classroom observation instrument (the COLT observation scheme) was introduced by Allen, Cummins, Mougeon, & Swain (1983). The development of the COLT was preceded by a review of various instruments designed to capture relevant features of the L2 classroom based on theoretical, empirical, and intuitive grounds. This instrument has been continuously further developed (Spada & Fröhlich, 1995) and is considered one of the most established observation instruments in L2 research (T. Helmke et al., 2008).

The COLT scheme contains categories derived from pedagogical issues in CLT literature which describes classroom instruction in terms of the types of teacher and student activities that take place as well as from issues revealed by L1 and L2 acquisition research related to verbal interactions in classroom activities. They served as the foundation for the basic coding guides in the DESI video study (see Chapter III.2.6.1).

V. Research goals and questions

In order to identify crucial instructional factors of student growth in EFL, their complex and interplay effects with each other and with context factors, several goals were set in this study which aim at answering the research questions below.

A. Description of classroom instruction in EFL in Vietnam

The first goal was to characterize real-life teaching and learning practice in English lessons in Vietnam, based on observable indicators of classroom instruction from the videos. To this end, video data were processed and analyzed analogously to the DESI video study by using the adapted DESI coding guides and rating sheet. With reference to them, we attempted to draw up an evidence-based descriptive analysis of how the lessons took place, their similarities, to what respect and extent they differed, and the degree to which teachers and students made speaking mistakes. Based on the findings of the DESI study and on the educational situation and research findings in Asia and in Vietnam so far, it was assumed that:

- EFL lessons in Vietnam at the time of data collection would be teacher-centered and textbook-driven,
- there would be a lack of self-reflection by teachers while teaching,
- teachers would partly be not well prepared in terms of their English speaking skills; not only students but also teachers would make a number of speaking mistakes,
- several instructional quality aspects which can be enhanced by the central curriculum and textbook, such as clarity and structuredness, would be positive, while other dimensions such as individual student orientation or the adaptivity of the lesson would be rated negatively for the same reason.

Accordingly, the following research questions were specified:

1. Does the study confirm the assumption that EFL lessons in this study are teacher-centered and textbook-driven?
2. Do Vietnamese EFL teachers have good self-regulation competencies?
3. Does the study confirm the assumption that many EFL teachers are not adequately prepared in terms of their English skills?
4. What are the most frequent pronunciation errors of teachers and students in this study?

5. In which of the important general quality dimensions of classroom instruction are EFL teachers in this study rated positively, including classroom management, clarity, structuredness, supportive learning climate, motivation, cognitive activation, and feedback?
6. Are teachers in this study rated positively regarding important quality dimensions of effective EFL teaching, including engaging students in communication, equal focus on accuracy and fluency, and dealing with mistakes?

B. Effects of context factors on student achievement and growth at individual and class level

The second goal was to capture the magnitude of student growth in EFL over one school year as well as to investigate both the individual and class composition effects of preknowledge and student socioeconomic background on student achievement and growth in EFL in Vietnam. Based on previous findings, the individual and class composition effects regarding initial student achievement were expected to be larger than the effects of student SES. The corresponding research questions for this goal were:

7. Does prior student achievement have a significant effect on student achievement and growth?
8. Does student socioeconomic background have a significant effect on student achievement and growth? Is the effect of SES smaller than the effect of prior student achievement?
9. Do academic and social class composition have an effect on student achievement and growth in EFL?

C. Instructional effects on academic student growth at class level

Findings in empirical educational and English didactic research have revealed numerous general and domain-specific instructional factors of student outcomes; these were operationalized as the instructional basic coding and rating variables in this study. The third goal was thus to identify the most influential instructional factors of student growth in EFL and to investigate their theoretical complex effects, including their interplay with each other and with context variables. The relationships between classroom instructional factors and student outcomes are assumed to not always be linear but can be nonlinear, not additive, yet rather compensable and interactive, and not invariable but differential (see Chapter IV.1). On this basis, the following research questions were addressed in this study:

10. What are the most important instructional factors for student growth according to the C-test and LC-test?
11. To what extent do the findings confirm the assumption of nonlinear relationships between instructional factors and student progress?
12. Is there empirical evidence supporting the assumption of an interaction effect between instructional factors on student growth?

13. Is there empirical evidence supporting the assumption of an aptitude treatment interaction effect on classroom instruction?
14. Do the findings confirm the assumption of a compensatory effect of instructional factors on student progress?

D. Influence of scaling model on study results

As explained in Chapter I, applying different scaling models to estimate test results can lead to different estimates of student achievement and growth and different estimates of instructional effects on student growth (see Chapter VI.2 for more details). Therefore, the final goal of this study was to examine the consistency of the estimated instructional effects given different scaling models. The corresponding research question was as follows:

15. To what extent is the estimation of instructional effects on academic student progress independent on the selection of a specific scaling model?

VI. Methodological challenges and approaches

In order to answer the research questions, a number of methodological challenges and issues need to be adressed which are relevant to the reliability, validity and replicability of the results. Reliability and validity are two fundamental criteria of educational measurement (Brennan, 2006). Reliability refers to the reproducibility of findings (Brennan, 2001a; Haertel, 2006), while validity refers to the plausibility of the intended uses and interpretations of measurement constructs including inferences and assumptions involved in the interpretations (American Educational Research Association, American Psychological Association, & National Council of Measurement in Education, 2014; Kane, 2006). Actually, there have still been ongoing discussions about the validity concept. Markus and Borsboom (2013) criticized the Kane's argument-based concept of validity that it can be justified independently of the truth, and suggested a concept which involves in addition the justification for accepting a validity argument – which essentially links the arguments to the theoretical attributes of the target construct. Since different constructs have different conceptual models (true score, true relationship between estimates and external variables, true relationship between estimates and universe scores, latent common cause, etc.), they suggested different conceptualizations of validity according to different conceptualization of the constructs.

By no means all theoretical approaches as well as all issues with relevance to the reliability and validity of the results can be addressed within the framework of this dissertation. In this chapter, some challenges

in measuring video based instructional quality, in estimating student achievement and growth as well as in estimating classroom instruction effects on student growth are presented.

Inter-rater reliability has long been a topic of interest in research fields using observation data. Of classroom instructional quality, variability between measures of different occasions has recently attracted a lot a attention. Regarding the measurement of student oucomes, difficulty in choosing appropriate scaling models among many available ones is a hot topic in the fields of psychometrics and educational measurement. Challenges in identifying important instructional factors on student progress, in particular in this study, stem from the typical problem of having small number of classes with large number of video based predictors. In addition, effect of each single instructional factor on student growth is expected small, confounding effects can not be fully controlled due to the non-experiment characters of the study, and the joint effect and interplay between many factors is assumed complex. Furthermore, as in most studies, missing data are present in this study, which might not be assumed completely at random and have thus be dealt with to enhance the reliability of the results.

These problems along with methodological approaches are going to be introduced, discussed and selected with regard to the feasibility and applicability in this study. Problems which are not addressed in this study will be discussed further in Chapter XI.

VI.1 Reliability and validity of video based measures of instructional quality

Instructional quality in this study given the measurement framework is intended to be interpreted as observed indicators of everyday classroom teaching and learning practice in EFL in 9th grade classes in a subpopulation Vietnam in the school year 2006–2007. This subpopulation entails representative 9th class samples of two biggest cities as well as of a small rural province Bac Ninh in Vietnam (see Chapter III.2.1), and thus has urban-dominated characters. This is thus not a representative sample for all 9th grade classes in Vietnam, the results can not be generalized for the whole country.

VI.1.1 Inter-rater reliability of codings and ratings of video data

Researchers have been concerning with reliability problems associated with rating or coding data since a long time ago (Guilford, 1954; Hoyt, 2000; Saal, Downey, & Lahey, 1980; Wirtz & Caspar, 2002). Coding, or low-inference coding by observers (coders), does not require coders to make any high inferences or judgments about the behaviours or expressions of teachers and students. On the other hand, rating requires raters to make own judgments of instructional quality dimensions based on the rating manual. Reliability of codings and ratings originally refered to the inter-rater reliability – the consensus

between multiple coders/raters who coded/rated the same object (Gwet, 2010; Wirtz & Caspar, 2002). With an underlying assumption that there is a true value of coding/rating object, differences between raters are perceived measurement errors caused by coder/rater bias (rater effects). On the other hand, "*if inter-rater reliability is high then […] raters can be used interchangeably without the researcher having to worry about the categorization being affected by a significant rater factor*" (Gwet, 2010, p. 4).

Several measures of inter-rater reliability have often been applied, such as Scott's *Pi* π (Scott, 1955), Cohen's *kappa* κ (Cohen, 1960, 1968), Krippendorff's *alpha* α (Krippendorff, 1980, 2004), Gwet's AC_1 (Gwet, 2008), intraclass correlation ICC2 (Raudenbush & Bryk, 2002). Model-based measures (on the basis of latent variable approaches) have also been suggested and become increasingly popular, for instance measures based on the latent trait or latent class models and their extentions by Uebersax (1993), Uebersax and Grove (1993), Agresti and Lang (1993), Albert and Dodd (2008), Nussbeck and Eid (2015), or based on the mixed membership model suggested by Erosheva & Joutard (2014).

In order to reduce rater bias and thus enhance the reliability of coding and rating results, a training program for all coders/raters with a rating design with multiple coders/raters per variable per recorded lesson had been conducted and done by Dr. Wolfgang Wagner, Dr. Tuyet Helmke and the research group at university of Koblenz-Landau. All coders/raters were trained similarly to the training program for coders/raters of the DESI video study. The procedures are going to be described below.

In training program for coders, five recorded lessons were randomly chosen. All coding variables of each lesson were coded independenly by two to seven coders according to the DESI coding manual (see Chapter VI.1). Table 2 shows a small part of coding data of one variable of one recorded lesson by six coders in the training program.

Table 2: Coding data of one variable of one recorded lesson by six coders in the training program

Time	**Transcription**	**Coder1**	**Coder2**	**Coder3**	**Coder4**	**Coder5**	**Coder6**
00:00:04	s:[Stand up].	0	2	0	2	2	2
00:00:06	s:Good afternoon.	0	2	2	2	2	2
00:00:07	e:Good afternoon teacher.	0	2	2	2	2	2
00:00:10	t:Good afternoon class.	2	2	2	2	2	2
00:00:12	t:Sit down.	2	0	3	2	2	2
00:00:13	t:All of you look cheerful and well today.	2	2	2	2	2	2
00:00:19	c:(The teacher hangs two pictures on the board).	3	3	3	2	4	0
00:00:40	t:Here, look at the two pictures on the board.	0	3	0	2	2	0
00:00:43	t:Are they nice?	2	0	2	2	2	2
00:00:45	e:Yes.	0	0	2	2	4	2
00:00:46	t:They are painted by students at our school.	2	0	2	2	2	2

Time	Transcription	Coder1	Coder2	Coder3	Coder4	Coder5	Coder6
00:00:52	t:Very nice.	2	2	2	2	2	2
00:00:53	t:Now, eh, what do you think about when you look at these pictures?	0	0	0	0	4	0
00:01:00	t:What do you think about when you look at these pictures?	0	0	0	0	4	0

Note: Column "Time": recorded time from the beginning of the video. Column "Transcription": one line represent data of one turn (one expression, or one act) corresponding to the recorded time, s = student speaks, e = whole class speak, t = teacher speaks, c = comment (in this turn no one speaks). Columns "Coder1" to "Coder6": coding values of each turn by each coder.

In the training program, Krippendorff's α (for calculation details see Krippendorff, 2011) was calculated for each coding categorie of each variable as well as for each variable (over all categories) of each lesson. In case an inter-rater reliability was low, intensive discussions were held and the coding manual was accordingly optimized, and all coders rated the corresponding variable of the corresponding lesson once again according to the revised coding manual. The training program finished when Krippendorff's α of all variables of all seven lessons exceeded the cut-off value of 0.667 according to Krippendorff (2004, 2011). In addition, a coder fit for each coder was calculated. This measure was developed by Dr. Wolfgang Wagner for the training program of the DESI video study (A. Helmke et al., 2007). It based on the Krippendorff's coincidence and distance matrix, and represents the relation between proportion of mismatch estimates related to each rater and the total mismatch caused by all raters. Raters with high proportion of mismatch were eliminated from the coding process later on. For economic reasons, it was not possible to have multiple coder design for the remaining recorded lessons in the main coding process. They were distributed equally to the coders. To ensure the coding quality, random checks were done. After coding five to seven recorded lesson, each coder had to rated one randomly chosen variable of one lesson once again, which he or she already rated previously. In case the intra-rater reliability of the check variable was lower than 0.8, a short refresher training was offered.

In the coding process, each coder coded all turns of one lesson. In the rating process, raters had to make high inferences or judgments about the whole recorded lesson based on the DESI rating manual, and gave only one rating per variable per recorded lesson. The entire rating sheet in this study is found in Appendix A3.

In the training program for raters, all raters rated all variables of eight selected lessons. During the training program, raters and trainers discussed thoroughly. Accordingly, the rating manual was extended if necessary to reduce rater disagreement. In the main rating process, each variable of each recorded lesson was rated by two raters. As a measure of inter-rater reliability for rating data, intraclass correlation ICC2 (for calculation details see section VI.3.1) of each variable over all recorded lesson and over all

raters (N = 2 per lesson) was calculated. In case the ICC2 of one variable was smaller than 0.7, trainers led a discussions with all raters to find out reason for diverse ratings and to extend the rating manual if necessary. The rating process finished when ICC2 of all rating variables exceeded 0.7.

Final coding and rating data for each lesson were randomly drawed from all available codings and ratings, and are regarded reliable referred to the inter-rater reliability. In the training phases, all statistical calculations were done using the statistics software SAS by Dr. Wolfgang Wagner.

VI.1.2 Generalizability of ratings of classroom instruction

In the last five to ten years, the reliability and validity issues in measuring instructional quality via classroom observation has gained a renewed attention in the international educational research field. The issue of reliability in measuring instructional quality via observations has been being extended to cover the consensus between ratings/codings over multiple occasions in which the instruction were rated. Regarding this, Kennedy (2010) asked "*To what extent is the quality of teachers' everyday practice – actual classroom behavior – really a function of enduring personal qualities that they bring with them, and to what extent is it a function of schedules, materials, students, insitutional incursions into the classroom, and the persistent clutter of reforms that teachers must accommodate?*" (p. 597). Researchers in the USA and in Germany have repeatedly reported the inconsistency of teachers' observation ratings across multiple lessons, multiple classrooms, and multiple years (Bell et al., 2012; Hill et al., 2012; Morgan, Hodge, Trepinski, & Anderson, 2014; Polikoff, 2015; Praetorius, Lenske, & Helmke, 2012; Praetorius, Pauli, Reusser, Rakoczy, & Klieme, 2014).

As the video study was conducted in January – February 2007, two lessons were recorded in each class. The first recorded lesson in each class was a curriculum-based lesson, the other was an extracurriculum speaking lesson in which students should speak in English as much as possible for the purpose of investigating teachers' and student language quality (see chapter III.2.5). The second recorded lesson was not an typical EFL lesson, and thus was not appropriate to use for the purpose of investigating the everyday classroom instruction in EFL in Vietnam. Hence, for each class only data of one recorded lesson were available for analyses.

Under the assumption that quality of classroom instruction with regard to one teacher in combination with one specific class in one school year can only be measured with high reliability based on multiple measures throughout the school year, results regarding instructional quality in this study based on measures of only one lesson can be short of reliability. In this study, no correlation between measurement errors and student growth is assumed. As a consequence, relationships between classroom instruction and student growth are assumed attenuated on average. Of course, this assumption might not

hold in the reality given small sample size as Loken and Gelman (2017, p. 585) pointed out according to their simulation study with added random noises to "ideal" data: "*For large-N studies, added error always reduces the effect. For small N, the reverse can be true*." (with large *N* they meant a sample size of 3000, and with small *N* they meant a sample size of 50 – equal the class size in this study). However, their results also suggested that exaggeration of observed estimates is less likely than attenuation under presence of measurement errors.

Seeing from a different angle, judgement (rating) of instructional quality is not only dependent on measurement occasion, but also on the perspective from which the instruction is observed such as student perspective, teacher perspective, observer perspective (Clausen, 2002; A. Helmke et al., 2014, 2017; G. Pham et al., 2012). Ratings of classroom instruction quality in this study reflects exclusively the perspective of educational researchers in Germany (who developed the rating criteria) and the trained raters. A comparison between rating data of one recorded lesson on the one hand and student perception of instructional quality over one school year on the other hand (data from the student questionnaire) can give an insight into the extent the video-based indicators of classroom instruction quality in this study can be generalized for instructional quality of the class over one year from student perspective. The results of this comparison will be presented and discussed in chapter XI.

VI.1.3 Representativity of the recorded lesson

A common problem which possibly occurs in all observation studies is related to the representativity of the lessons in the presence and observation of third parties (researchers, observers). It can be assumed, that teachers would have given their best in these lessons. Thus, the quality of the recorded lessons might be systematically better (according to teachers' own concept of good teaching) than of an ordinary lesson. This would result in additional measurement errors of classroom instruction and thus lower reliability of instructional quality estimates. On the other hand, it is not plausible to assume that measurement errors in this case would correlate with student progress. A possible consequence is thus once again the underestimation between instructional quality and student growth.

In order to gain insight into the extent the recorded lessons woud have differed from an ordinary lesson without the presence of research team, short questionnaires for both students and teachers right after the recorded lesson were applied (cf. section III.1.5). Results regarding the representativity of the lessons are going to be presented in chapter 0.

VI.2 Estimating student achievement and growth via IRT approaches

Since late 1970s and early 1980s, item response theory (IRT) has been becoming the mainstream framework in the field of international educational assessment (Yen & Fitzpatrick, 2006). IRT is "*a family of statistical models used to analyze test item data*", which "*provides a unified statistical process for estimating stable characteristics of items and examinees and defining how these characteristics interact in describing item and test performance*" (Yen & Fitzpatrick, 2006, p. 111). In comparison to classical test theory (CTT) which focusses on properties of intact tests (test scores), IRT focuses on item responses and makes stronger assumptions (Brennan, 2011).

CTT assumes that observed raw score (X) of a test is the sum of unobserved true score (T) and error score (E), with expected value of E is zero, E and T do not correlate within and across items for identifiability reason, and test items are assumed interchangable (be it classically parallel, tau-equivalent, essentially tau-equivalent, or congeneric; see Eid, Gollwitzer, & Schmitt, 2010; Brennan, 2011; Robitzsch, 2016a). IRT assumes that examinees with a particular ability level (θ) have a certain probability $P(X_i = x|\theta)$ to give a specific response (x) to an item i with a specific set of item parameters, and test items are not interchangable. CTT relates test scores to examinee ability (T) via linear function, while IRT relate item scores to examinee ability (θ) via nonlinear functions. Within the IRT framework, a replication requires using a set of items with identically the same parameters (strictly parallel form), while requirement for a replication is smaller within the CTT framework (e. g. using a tau-equivalent form; for a more detailed comparison between CTT and IRT see Brennan, 2011).

There is a variety of IRT approaches and models, each one rests upon different set of underlying assumptions with regard to the (dimensionality of) measurement constructs as well as to item response and item characteristic functions, accordingly yields different person ability estimates or *test scores* (Kolen & Brennan, 2004; Robitzsch, 2016; Trendtel, Pham, et al., 2016; Yen & Fitzpatrick, 2006). In order to interpret the test results appropriately, it is important to make clear which assumptions are presumed, and to what extent the results can be generalized.

IRT models are favoured in LSAs (over CTT models) due to many reasons (see Berezner & Adams, 2017). But first and foremost, IRT models have been developed specifically to support the process of test development and construct validation, while CTT models are not because of the interchangeability assumption (Berezner & Adams, 2017; Robitzsch, 2016).

Within the IRT framework, test scores of different MPs can be linked to each other via linking process. There have been many linking approaches introduced until now (Holland & Dorans, 2006; Kolen & Brennan, 2004; Trendtel, Pham, et al., 2016).

In the following sections, characteristics of the C-test and LC-test as well as approaches to estimate test scores and growth based on these tests are going to be described.

VI.2.1 The C-test and the testlet structure

Since its introduction in the early 1980s (Raatz & Klein-Braley, 1981, 1985), C-test has become one of the popular test instruments in the field of language assessment (Grotjahn, 1995, 2010). In Germany, the C-Test has been applied in the important language tests and studies such as in the large-scale-study DESI (Harsch & Schröder, 2007), in the Test Deutsch als Fremdsprache (TestDaF; test of *German as a foreign language*, Eckes & Grotjahn, 2006), in onDaF (the online placement test of German as a foreign language; Eckes, 2010). A C-test consists of four to eight texts, in each text every k^{th} word is partly obliterated to be an empty gap. The examinees should fill in the gaps to complete the words. The C-test is widely used as a test of *general language proficiency* (Grotjahn, 2006; Harsch & Schröder, 2007; Schroeders et al., 2014), in particular of general written language ability (Hastings, 2002; Vockrodt-Scholz & Zydatiß, 2010).

The applied C-test in our study included four original texts C1, C2, C3 and C6 from the item pool of the DESI study C-test (cf. chapter III.2.2.1), each text contains 25 gaps. After taken the semantic and grammar phenomena into account, the gaps in the texts C1, C2 and C6 were created by erasing the second half of every third word in a text, in the text C3 by deleting the second half of every two words (Harsch & Schröder, 2007). A gap is considered an item. All items of one text share a common theme (*stimuli*). An aggregation of items on a single stimulus (in our case one text is considered a stimulus) can be called a *testlet* as suggested by Wainer & Kiely (1987). The C-test with its testlet structure can be called a *testlet-based test*.

Sharing a common stimulus, *text specific dependencies* (or *local item dependencies* - LID) between items of one testlet – which is not shared with items of other testlets – can be expected. The LID between DESI C-Test items were investigated by Harsch & Hartig (2010). Weak to moderate dependencies were found, strongest by items of testlet C3 due to the shortest distance between two adjacent gaps.

VI.2.2 Unidimensional IRT models

In the DESI study, test data were coded binary, and the Rasch or one parameter logistic model (1PL, Rasch, 1960) was applied to estimate item parameter and test scores based on C-test and based on LC-test (Hartig, 2007). The Rasch model assumes that an item *i* of test *I* has an item difficulty b_i for a group of examinees with ability θ_p, and the probability that they give a correct answer to item i ($P[X_{pi} = 1]$) is defined as:

$$P[X_{pi} = 1] = \frac{\exp(\theta_p - b_i)}{1 + \exp(\theta_p - b_i)} \qquad (1)$$

In formula (1), X_{pi} is the observed response of person *p* to item *i* which has two possible values 1 (correct) and 0 (incorrect). Formula (1) can be written shortly as logit $P(X_{pi} = 1) = \theta_p - b_i$. When the person ability θ_p is equal to the difficulty b_i, the probability that person *p* gives the correct answer to item *i* is 0.5. If $\theta_p > b_i$, the probability of having the correct answer is accordingly higher and vice versa.

Item and person parameters can be estimated independently of each other, and only sum scores are necessary for estimation: "*we may estimate the item parameters independently of the personal parameters, the latter having been replaced by something observable, namely, by the individual total number of correct answers. […] we may estimate the personal parameters without knowing the item parameters which have been replaced by the total number of correct answers per item.*" (Rasch, 1960, p. 325). By using the Rasch model, it is assumed that the probability of a correct response of person *p* to the item *i* is caused by only one common trait which is interpreted as the interested ability (unidimensional assumption). Given this ability construct, all responses are stochastically independent (local independence assumption). Note that, the ability construct itself "*is not necessarily one-dimensional*" conceptually (Rasch, 1960, p. 326).

Later on, other unidimensional IRT models were introduced for dichotomous items, most notably the unidimensional two parameter logistic (2PL) and three parameter logistic (3PL) models by Birnbaum (1968). The Rasch model was applied by all PISA studies until 2012, in 2015 the unidimensional 2PL model was applied for the first time (OECD, in press). In the TIMSS study, the unidimentional 3PL model has been in use (Foy, Brossman, & Galia, 2013).

The unidimensional 2PL model includes a second item parameter – item discrimination a_i – in the probability equation: logit $P(X_{pi} = 1) = a_i(\theta_p - b_i)$. Item discrimination a_i varies between items, and shows how well item differentiates between more and less capable examinees on the ability construct. For 2PL model, the raw sum scores are not sufficient any more for estimating item and person

parameters, but weighted sum scores – weights correspond to the item discrimination estimates. This way, some item groups will be overrepresented, and some other underrepresented in the ability construct depending on the covariances between them (Robitzsch, Freunberger, Itzlinger-Bruneforth, Breit, & Schreiner, 2015). Hence, Robitzsch (2016) interpreted ability construct via Rasch model as an equally weighted representation based on all test items, and ability construct via unidimensional 2PL model as an data-based weighted representation of test items (see also Goldstein, 1980; Brennan, 2001b).

The unidimensional 3PL model includes another item parameter – a lower asymptote or guessing parameter g_i – in the probability equation:

$$P(X_{pi} = 1) = g_i + (1 - g_i) \cdot \frac{\exp(a_i(\theta_p - b_i))}{1 + \exp(a_i(\theta_p - b_i))}$$

3PL model allows that examinees with very low ability may nevertheless answer the item correctly (e. g. by guessing), and that different items can have different lower asymtotes (pseudoguessing levels). The guessing parameter g_i is often applied for items of closed-ended format. Although this model has becoming one of the most popular scaling models in the field of educational measurement, researchers have still been having concerns regarding the ambiguity of interpretation, stability and accuracy of the estimates (Hambleton, Swaminathan, & Rogers, 1991; Han, 2012; Holland, 1990; Kolen, 1981; Lord, 1980; San Martin, del Pino, & De Boeck, 2006).

Beyond 3PL model, several models for dichotomous have been suggested such as the 4PL and 5PL models (M. A. Barton & Lord, 1981; Fox, 2010; Loken & Rullison, 2010; Magis, 2013; Robitzsch, 2016), the nonparametric IRT model (Rossi, Wang, & Ramsay, 2002). However, these models have not been widely used due to conceptual drawbacks and difficulties in estimating additional parameters.

In case test data are coded polytomous, such as if sum scores of subgroups of items are used as response data, unidimensional IRT models for polytomous items can also be applied such as the graded response model by Samejima (1969), the partial credit model by Masters (1982), the generalized partial credit model by Muraki (1997), the nominal item response model by R. D. Bock (1972) and so on (cf. Trendtel, Pham, et al., 2016).

For an overview of unidimensional IRT models see van der Linden and Hambleton (1997), for several more recent unidimensional IRT models see Robitzsch (2016).

VI.2.3 IRT models for testlet-based tests

Test scores of a testlet-based test via a unidimensional scaling model with LID assumption reflect both the common factor represented by all test items and testlet specific factors. If the construct of interest is the common factor excluding testlet specific commonality, the measured reliability of ability estimates via a unidimensional model might be overestimated (Cao, Lu, & Tao, 2014; DeMars, 2006; Eckes, 2014, 2015; Ip, 2000; Li, Bolt, & Fu, 2006; Sireci, Thissen, & Wainer, 1991; Wainer & Wang, 2000).

To model the testlet structure of the testlet based test items, different IRT approaches have been introduced. Beside unidimensional models for polytomous items using testlet sum scores as response data, the most well-known IRT models for this purpose are the testlet models (Bradlow, Wainer, & Wang, 1999; Li et al., 2006; Wainer, Bradlow, & Wang, 2007; Wang & Wilson, 2005) based on the ideas of bi-factor models (Brunner, Nagy, & Wilhelm, 2012; Gibbons & Hedeker, 1992; Gignac, 2014; Reise, 2012). For an overview see Rauch & Moosbrugger (2011); a graphical comparison between unidimensional model with LID assumption (left) and testlet models (right) is depicted in Figure 8.

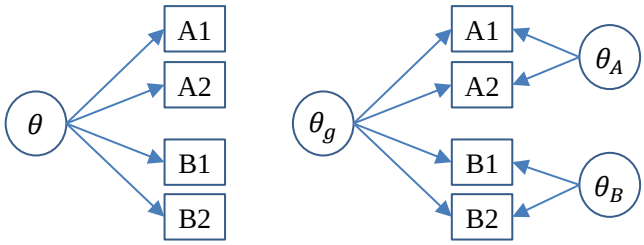

Figure 8: Unidimensional model with LID assumption (left) and testlet model (right)

Testlet models incorporate an aditional random effect $\gamma_{pd(i)}$ for person *p* on testlet *d(i)* to which item *i* belongs to the probability equation, which describes the interaction between persons and items (local item dependence) within the testlets. Testlet 1PL, 2PL, 3PL models (Wang & Wilson, 2005) are formulated as follows:

$$P(X_{pi} = 1) = g_i + (1 - g_i) \cdot \frac{\exp(a_i(\theta_p - b_i + \gamma_{pd(i)}))}{1 + \exp(a_i(\theta_p - b_i + \gamma_{pd(i)}))}$$

with $P(X_{pi} = 1)$ the probability that person *p* gives correct response to item *i* of testlet *d*, θ_p the ability of person *p*, b_i the item difficulty parameter of item *i*, g_i the guessing parameter of item *i*. If $g_i = 0$ and $a_i = 1$ we have the testlet 1PL or Rasch testlet model. If $g_i = 0$ and $a_i \neq 1$ we have the testlet 2PL model (cf. Wainer, Bradlow, & Du, 2000). If $g_i \neq 0$ and $a_i \neq 1$ we have the testlet 3PL model. The following assumptions were set to facilitate parameter estimation using a hierarchical Bayesian

framework (Wainer & Wang, 2000): $\theta_p \sim N(0,1), \gamma_{d(i)} \sim N(0, \sigma^2_{\gamma_{d(i)}}), a_i \sim N(\mu_a, \sigma^2_a), b_i \sim N(\mu_b, \sigma^2_b)$, $\log[g_i(1-g_i)] \sim N(\mu_g, \sigma^2_g)$. $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. The larger the testlet specific variance $\sigma^2_{\gamma_{d(i)}}$ is, the greater the proportion of total variance in test score that is attributable to the testlet and not to the general ability construct. For the means μ_a, μ_b, μ_c, the prior distributions with mean zero $\mu_a \sim N(0, V_a), \mu_b \sim N(0, V_b), \mu_c \sim N(0, V_c)$ are typically chosen.

Having recently been proposed and gained increasing attention, another model family for handling LID is one which treats local dependencies as a source of disturbance and allows and explicitly models correlated residuals. For instance the copula models (Braeken, Tuerlinckx, & De Boeck, 2007; Braeken, 2011), which are similar to the locally dependent unidimensional models (Ip, 2010). The item response function of copula models is identical to the 1PL logistic model with the distinction that the residuals are allowed to be correlated within a testlet. Thus, the advantage of the copula model over the testlet model is that the meaning of the theta ability is similar to that of the widely accepted unidimensional model, though the copula models and the testlet models are statistically equivalent (Ip, 2010). Schroeders et al. (2014) showed furthermore, that in case residual correlations were above zero and for tests with many testlets or with large testlets, the parameter estimates by copula model were less biased than results of other models (Rasch model, PCM, Rasch testlet model), although differences in parameter estimates between different models were generally small. Practically however, it is challenging to apply the copula model to estimate results of a test/testlet with many items (Robitzsch, 2016). For the C-test in this study with 4 testlets and 25 items per testlet, it was not able to apply this model for parameter estimates.

VI.2.4 Model selection for estimating student ability

VI.2.4.1 Arguments based on reliability and model fits

Several researchers prefered testlet models to estimate person ability based on testlet-based tests due to the reliability argument (see above), and suggested to select models based on model fits. A comparison of model fits suggests that 2PL and 3PL scaling models fitted the data better than the corresponding 1PL models, and the testlet scaling models fitted the data better than the unidimensional models in case of C-test (see Appendix D1). Nevertheless, many scaling models have sufficiently good fits: RMSEA < .06 (with the exception of Rasch model), and SRMSR < .08 (Hu & Bentler, 1999). Hence, it is not decisive to reject any model based on model fits, especially when the parsimony of the model and interpretability of person parameters are also taken into account.

VI.2.4.2 Validity arguments

Robitzsch (2016) argued that rather than any empirical values, model selection should base primarily on validity arguments (see also Goldstein, 1980). A testlet model should be favoured if the testlet specific factors are regarded irrelevant to the interested ability construct; otherwise a unidimensional model would be a better choice. The choice between an 1PL model or a more-parameter model should be based on the theoretical consideration, if test items should have different weights in defining the ability construct according to their empirical discriminations (Robitzsch et al., 2015; for item discrimination see Appendix B4). In this study, validity arguments serve as main criteria in model selection.

VI.2.4.3 Validity of the tests and test scores

Since one goal of this study is to investigate the relationship between classroom instruction and student growth, it is desirable to obtain the student ability measures corresponding to the outcomes of the EFL classroom instructions in Vietnam. Nevertheless, the match between content of the English tests in this study and of lower secondary EFL curriculum and textbooks in Vietnam can not be quite assured, though the original tests were piloted and adapted (see chapter III.2.2).

In order to enhance the test validity beforehand, item difficulties and discriminations as well as test length were analyzed (see Appendix B). Too difficult items or items which do not discriminate well between low and high ability students (based on their sum scores) were eliminated. The missing analyses confirmed that the tests were not too long given the available testing time, thus most non-responses were supposed intendedly omitted rather than due to time out in both MPs. This reinforced the way non-responses were treated as incorrect but not ignored (see Appendix B1).

To gain indicators to judge the validity of the tests and test scores, correlations between student ability estimates based on two tests and different scaling models and midterm school marks in English as well as other school subjects mathematics and Vietnamese were calculated. They are presented and discussed in chapter X. Based on them it can be examined if the test scores reflect EFL ability better than mathematic competence (regarded as an indicator of intelligence) as well as Vietnamese competence (regarded as an indicator of general language competence), and which test and scaling model yielded test scores which correspond better to a curriculum based indicator of student achievement in EFL.

VI.2.4.4 Model selection for estimating student ability via the C-test

Actually, testlet specific factors might or might not be regarded relevant to the interested ability construct, and are often partly of interest. If two or more items belong to a lexical phrase within a testlet, responses on them are expected to correlate more strongly to each other than with other test items. In this case, item group specific variance is expected high, which is however not necessarily irrelevant to the interested general EFL proficiency construct. On the other hand, testlet specific factors can also represent position effect of the testlet items (Debeer & Janssen, 2013; Hartig & Buchholz, 2012; Robitzsch, 2009), which is not regarded relevant to the ability construct.

If the contents of the C-test and of the EFL curriculum in Vietnam do not match well, testlet models might be the better choice, since their test scores are assumed less test specific than test scores of unidimensional models. If however part of one C-test testlet by chance matches the EFL curriculum and content of EFL textbook better, we might not want to eliminate this specific variance from the variance of the interested ability construct.

Overall, it is not clear-cut which model family (unidimensional vs. testlet models) is better with respect to the interested construct (general written language proficiency which is relevant to the EFL curriculum in Vietnam). In fact, the "optimal" model could lie somewhere in between, that a part of testlet specific variance should be kept while the rest should be eliminated from the general factor variance. Therefore, both unidimensional and testlet models are considered equally (in)appropriate to model student ability via C-test.

On these grounds, the unidimensional 1PL model (Rasch model, M1), unidimensional 2PL model (M2), Rasch testlet model (M1T) and testlet 2PL model (M2T) were selected to model student general proficiency based on the C-test in this study. The 3PL scaling models with guessing parameters are not applied in this case, since it is supposed extremely unlikely that a correct answer can be given just by random choice.

VI.2.4.5 Model selection for estimating student ability via the LC-test

The LC-test in this study consists of 18 multiple choice items, eight items belong to the diaglogue part, ten items belong to the conversation part (see chapter III.2.2.2). Theoretically, items of the conversation part which share the same stimulus could have high local item dependency. However, empirical results show that $Q3$-statistics (Yen, 1984) of conversation items were low in absolute and not higher than $Q3$-statistics of dialogue items (see Appendix C3b). Hence, testlet models were not applied to estimate item and person parameters based on the LC-test. To obtain test scores based on the LC-test, three established

unidimensional 1PL – Rasch (M1), 2PL (M2), and 3PL (M3) models for dichotomous items were selected.

VI.2.5 Estimates of student achievement and growth

VI.2.5.1 Linking item and person parameters over two MPs

Difference scores or change scores between two MPs are treated as measures of student growth in this study. The necessary condition to use the change score is that the ability estimates of two MPs must be comparable and on the same metric. For that purpose, a link between ability estimates of two MPs must be made via linking process. For an overview of different linking approaches using common items see Kolen & Brennan (2004), Dorans, Pommerich, & Holland (2007), or Trendtel, Pham, et al. (2016). One of the most widely used linking approaches using common items in LSAs is the concurrent calibration approach to estimate item parameters of both MPs simultaneously. In the process, different a priori ability distributions of the sample at two MPs are assumed and estimated (multiple group analysis). Since the metric of person parameters in an IRT model is the same as of item parameters, the metrics of student ability estimates at two MPs via this process are identical. This approach is easily applicable for all IRT models, and is in particular favoured in linking parameters of multidimensional models such as the testlet models (Simon, 2008). Thus, this approach was chosen for linking student ability estimates in this study. Using this, it is assumed that the items function invariably over different MPs (time-invariant), and the differences in responses overtime are caused by development in student ability. To examine this assumption, differential item functioning analyses are executed (see chapter Appendix B5).

The calibration and scaling processes were done in the programming environment R (R Core Team, 2015). To estimate item and person parameters based on unidimensional models, R package **TAM** (Kiefer, Robitzsch, & Wu, 2016) was applied using the marginal maximum likehood estimation method (see Fischer, 2007; Robitzsch, 2016). To estimate item and person parameters based on testlet models, R packages **mirt** (Chalmers, 2012) was used which applies the confirmatory maximum likelihood method with a dimension reduction EM algorithm for modeling testlets (Chalmers, 2012; Gibbons & Hedeker, 1992).

In the first step, item parameters as well as means and variances of student ability distributions at each measurement points were estimated. Afterwards, student ability parameters were estimated. Ability estimates in this study are not used to assign individual test scores but to obtain group estimates of achievement and growth and of their relationship with classroom instructional and context factors. For these purposes, plausible values (PVs, see M. Wu, 2005 and von Davier, Gonzalez, & Mislevy, 2009)

were drawed and used as student ability estimates as in LSAs. The PVs and the process to draw them are explained and described in the next section.

VI.2.5.2 Drawing plausible values of student ability and missing data imputation

The *plausible values* (PVs) were first developed for the analyses of the 1983–1984 NAEP (National Assessment of Educational Progress) data by Mislevy, Sheehan, Beaton and Johnson (Mislevy, Beaton, Kaplan, & Sheehan, 1992; Mislevy, Johnson, & Muraki, 1992; Mislevy, 1991) based on Rubin's work on multiple imputations (1987). Since then, PVs have been being used as student ability estimates and as a tool for secondary analyses in LSAs such as NEAP, PISA, TIMSS, DESI (Hartig, 2007; NAEP, 2016; OECD, 2014b, n.d.; Olson, Martin, & Mullis, 2009; Robitzsch, Pham, & Yanagida, 2016).

PVs are multiple imputations of the latent ability of each student, and represent the ranges of ability estimates that a student might reasonably have according to a scaling model given the student's item responses X. Instead of calculating a point estimate (most possible value based on an estimation method) for θ, a range of possible values of θ with an associated probability for each of these values is estimated (posterior distribution of θ for a student). PVs are random draws from this posterior distribution. Using PVs, measurement errors associated with ability estimates are taken into account, and less biased (or unbiased) population parameters can be obtained (von Davier et al., 2009; M. Wu, 2005). This technique can also be applied for other (non-IRT) latent variable models (Asparouhov & Muthén, 2010).

In such a process, when the relationships between student ability θ and other covariates $\mathbf{Y}$ are *not* considered in estimating the distribution of possible values of θ, the PVs are called unconditional PVs. Otherwise, they are called *conditional* PVs. In order to estimate the conditional posterior distribution of θ, both the measurement model (scaling model) and the analysis model are taken into account. The scaling model formulates the probability $P(\mathbf{X}|\theta)$ that students with ability θ have item response pattern $\mathbf{X}$. The analysis model formulates the probability $P(\theta|\mathbf{Y})$ that students with background characteristics $\mathbf{Y}$ have ability estimate θ. The conditional posterior distribution of θ is formulated according to Bayes rule as $P(\theta|\mathbf{X},\mathbf{Y}) \propto P(\mathbf{X}|\theta)P(\theta|\mathbf{Y})$ (for a more detailed explanation see Robitzsch, Pham, et al., 2016). Simulation studies have shown that unbiased parameter estimates (e. g. correlations, regression coefficients) in secondary analyses can be obtained using conditional PVs (Mislevy, 1991; M. Wu, 2005).

Since understanding the relationships between student ability θ and other covariates (such as student SES, classroom instructional variables) is one main goal in this study, conditional PVs were drawn as estimates of student abilities. For this purpose, all available variables from the questionnaires (407

variables from the student questionnaires at both MPs together with their class mean values and 323 variables from the teacher questionnaire at T2), four additional context variables (two dummy-coded variables regarding the school location, the class size, and the participation status of the classes in the video study), and all coding and rating variables from the video study (see chapter III.2.6) were included in the analysis models.

Due to the presence of missing data in covariates, an integrated treatment process of missing data on covariates and latent variables was applied (Robitzsch, Pham, et al., 2016). First, the missing data were imputed by chained equations (MICE approach, van Buuren, 2012) under the *missing at random* assumption (Lüdtke & Robitzsch, 2010; Rubin, 1976) using R package **mice** (van Buuren & Groothuis-Oudshoorn, 2011) with supplement functions from R package **miceadds** (Robitzsch, Grund, & Henke, 2016). The multilevel data structure (students are nested within classes) were taken into account using random intercept model for the imputation of level 1 (individual level) variables.

In the next step, after all missing data were replaced by imputed values, the analysis model was specified. Not only linear relationships but also nonlinear relationships and interaction effects between covariates and student ability were modelled. Since the number of predictors including quadratic and interaction terms turned out to be very large (more than 5000), partial least squares technique (Abdi, 2010) was applied for the latent regression model $\theta = \mathbf{Y\beta} + \varepsilon$. That means, a smaller number of uncorrelated factors was stepwise extracted under the criterion of retaining as much as possible of the variation present in the **Y** matrix. Since the dependent variable θ in the latent regression is still unknown, point estimates of the student ability are used instead to extract the PLS-factors for the next step: Warm's weighted likelihood estimates WLEs (Warm, 1989) for unidimensional scaling models, maximum a posteriori (MAP) estimates (Embretson & Reise, 2000) for testlet models. For this purpose, the R package **pls** (Mevik & Wehrens, 2007) was used.

To draw conditional PVs of student ability based on unidimensional scaling models, the individual unconditional posterior distribution of θ (individual likelihood) was extracted in one previous step. Then, the conditional posterior distribution of θ was calculated via fitting the latent regression $\theta = \mathbf{Y\beta} + \varepsilon$ based on the extracted individual likelihood and PLS-factors using the function tam.latreg() in the R package **TAM** (Kiefer et al., 2016). Finally, the plausible values was drawn using function tam.pv() of the same R package.

To draw conditional PVs of student ability based on testlet models, the unidimensional plausible value imputation method (Mislevy, 1991; Asparouhov & Muthén, 2010; Blackwell, Honaker, & King, 2017a, 2017b) was applied using MAP point estimates $\hat{\theta}$ of student ability together with their measurement

errors $SE(\hat{\theta})$ instead of individual likelihood for practical reasons. Assuming $\hat{\theta} = \theta + e$, and $\theta = \mathbf{Y\beta} + \varepsilon$, PVs of θ are drawn from the posterior distribution $P(\theta|\hat{\theta}, \mathbf{Y})$ using the function mice.impute.plausible.values() in the R package **miceadds** (Robitzsch, Grund, et al., 2016).

PVs of student ability of two MPs were drawn separately. To ensure that the relationship between ability estimates of two MPs were dealt with appropriately, PVs of student ability at one MP together with its quadratic term were treated as covariates to draw PVs for student ability at the other MP.

The process of data imputation and drawing PVs of student ability was done simultaneously, iteratively, and multiple times. Within each iteration, first the class level variables, then the individual background variables were imputed; afterwards the PVs of student ability at the first (T1) and the second measurement point (T2) were drawn. Imputed values and PVs of one iteration served as starting values for the next iteration. Imputed values and PVs after 40 iterations were saved and treated as one imputation dataset. A total of 10 imputed datasets were generated for further analyses in this study. A transformation was done so that the pooled mean and standard deviation of the student achievement at T1 was $N(0,1)$. For a more detailed explanation and a technical description of the whole data imputation and PVs drawing process using R see Robitzsch, Pham, et al. (2016). Student growth was then calculated as the difference between student achievement at T2 and T1 based on the transformed PVs. An alternative integrated multivariate normal distribution approach for imputing missing covariates and handling measurement error prone variables has been suggested by Blackwell et al. (2017b) and implemented in R package **Amelia** (Honaker, King, & Blackwell, 2011).

VI.3 Modelling the relationship between instructional and context factors and student growth

VI.3.1 Validity and reliability of class mean values of student achievement, student growth and SES

To investigate the relationship between student growth and classroom instructional and context factors, the level of analysis (class level vs. individual level) is the next topic worth mentioning. In this study, analyses at class level (level-2) are of main interest. While video-based classroom instructional variables are direct measures at class level, student achievement, student growth, and student SES at class level were obtained by aggregating individual student values.

Speaking of level-2 construct based on aggregations of within-group individual (level-1) values, Lüdtke et al. (2008) differentiate between two types of level-2 construct: reflective and formative constructs.

When all level-1 indicators within each level-2 group are designed to measure the same level-2 construct, and scores associated with different individuals within the same level-2 group are interchangeable, aggregation of level-1 values builds up a *reflective* level-2 construct – which reflects the "latent" level-2 construct which is assumed to "cause" the level-1 indicators within each group. In this case, variation within each level-2 group is regarded as measurement error. The reliability of a reflective level-2 construct depends on two factors: the number of individuals of each group (N_j), and the homogeneity/heterogeneity within groups (intraclass correlation ICC, Raudenbush & Bryk, 2002). The intraclass correlation ICC is given by $\mathrm{ICC} = \tau^2/(\tau^2 + \sigma^2)$, σ^2 represents the within-group variablity, and τ^2 captures the between-group variability. This way, ICC measures the proportion of the variance of the interested variable that is between groups (level-2 units). The reliability ICC2 (Snijders & Bosker, 2012) of a reflective level-2 construct is expressed by $\mathrm{ICC2} = \frac{\tau^2}{\tau^2+\sigma^2/N_j}$. The unreliability of a manifest group mean can lead to biased estimation in analyses using it. For this reason, multilevel latent variable approaches were developed which correct for unreliability of group means (Goldstein & McDonald, 1988; Longford & Muthén, 1992; Muthén & Satorra, 1989; Muthén, 1989), which have been implemented in the latent variable modeling software Mplus (Asparouhov & Muthén, 2006; Muthén & Asparouhov, 2011; Muthén, 2002).

In constrast, when the focus of level-1 measures is on a level-1 construct (e. g. individual student test scores), level-1 individual within the same level-2 unit are likely to have different level-1 true scores and their scores are not interchangeable, the aggregation of level-1 values results in a *formative* level-2 construct, which does not reflect the same construct measured at level-1. In this case, the reliability of aggregated values depends on the sampling ratio (the percentage of level-1 individual within a level-2 group) and the reliability of the individual level-1 measures. Within group variation can be regarded as a group characteristic but not measurement error, hence it is inappropriate to use ICC to estimate the (un)reliability of the group mean, especially when the sampling ratio (percentage of participants within groups) approachs 100% (Lüdtke et al., 2008).

Theoretically, class mean achievement and growth as well as class mean SES (for calculation details see Appendix E) could be seen as rather formative than reflective constructs given the definitions mentioned above. Given this consideration, reliability correction is not required in further analyses using these variables.

Practically, if no or very little variation between classes is captured and within-class variation is large, it would make little sense to do analyses at class level using class mean values. Thus, ICC of these variables were examined. The proportion of variation between classes regarding student achievement

and growth is substantially larger than zero (ICC of student achievement ≥ .48, and ICC of student growth ≥ .15, see chapter VIII.1.2). Similarly, proportion of variance between classes amounts to 45% total variance of student SES (*ICC* = .45), that means there are substantial differences between classes regarding student SES. Together with large class sizes (range between 27 and 59), the reliability ICC2 of class mean values are very high (see chapter VIII). Even if these constructs should be regarded as reflective ones, analysis results with and without reliability correction would be very similar.

Altogether (and given 100% sampling ratio), no corrections for unreliability of these class mean values were done in further analyses in this study.

VI.3.2 Examining contextual effects using multilevel covariate model

In order to examine contextual effecs regarding student prior achievement and SES (individual and class composition effects, see chapter IV.1 and IV.2), multilevel modelling (MLM) technique is required to model data at both individual and class level simultaneously. This have traditionally been done using the multilevel random intercept model (Raudenbush & Bryk, 2002; see Lüdtke et al., 2008). This model is formulated in two equations as follows:

- Within classes (within level): The individual dependent variable Y_{ij} (student outcome of student *i* in class *j*) is predicted by differences between students within classes regarding the predictor via the class specific linear regression: $Y_{ij=}\beta_{0j} + \beta_{1j}\left(X_{ij} - \bar{X}_{..}\right) + r_{ij}$, β_{0j} is the class specific intercept, β_{1j} is the class specific regression slope, r_{ij} is the regression residual at individual level, the predictor variable X_{ij} is centered at the grandmean $\bar{X}_{..}$.
- Between classes (between level): the respecting class specific slope and intercept of class *j* are defined by $\beta_{0j} = \gamma_{00} + \gamma_{01}\bar{X}_{.j} + u_{0j}$, and $\beta_{1j} = \gamma_{10}$, with γ_{00} is the grand mean of dependent variables over all classes, γ_{01} is the regression coefficient associated with class mean value of the predictor, γ_{10} is the slope relating $\bar{X}_{.j}$ to the intercepts from the equation at within level (equal for all classes), and u_{0j} is the level-2 residual of regression for class intercepts β_{0j}.

The total equation representing the relation between *Y* and *X* is accordingly:

$$Y_{ij} = \gamma_{00} + \gamma_{10}\left(X_{ij} - \bar{X}_{..}\right) + \gamma_{01}\bar{X}_{.j} + u_{0j} + r_{ij}$$

Since $(X_{ij} - \bar{X}_{..})$ and $\bar{X}_{.j}$ are not independent, γ_{01} is the specific effect of the class mean after controlling for individual differences on *X*, called *contextual effect*. Consequently, a class composition effect is regarded present if γ_{01} is statistically sigficant ($p < .05$, Lüdtke et al., 2008).

In this study, since no reliability correction for class mean value $\bar{X}_{.j}$ was applied (see previous section), the analysis model is called the *multilevel manifest covariate* (MMC) model (Lüdtke et al., 2008). This was done using the function BIFIE.twolevelreg() in the R package **BIFIEsurvey** (BIFIE, 2015).

VI.3.3 Estimating linear effects of instructional factors on student growth

Traditionally, effects of instructional factors on student outcomes have been investigated using ordinary least square (OLS) regression models for data at one level, using multilevel models for data at different levels (see above), and using path models or structural equation models to examine different types (direct, indirect) of effects (Creemers et al., 2010b; Teddlie, Reynolds, & Sammons, 2000). All of them require a sufficient sample size at highest level of analysis with regard to the number of predictors in model and the expected effect sizes for model convergence and accurate estimation (Cohen, 1992; Eid et al., 2010; Green, 1991; Kelley & Maxwell, 2003; Knofczynski & Mundfrom, 2008; Maas & Hox, 2005; Tanaka, 1987). This is not fulfilled in this study given a small sample size at class level (N = 50), large number of model predictors especially when interaction and compensatoric effects are included, and the instructional effects are expected small since this is not an experimental study and measurement errors of instructional variables are assumed high (see chapter VI.1).

With focus on analyses using data of one level (class level), there actually are established analysis methods which are applicable for dataset with a large number of predictors and small number of cases, such as random forests (Breiman, 2001), or regularized regressions (Fahrmeir, Kaufmann, & Kredler, 1996; Tutz, 2012). Among them, the regularized regressions using *lasso* (*least absolute shrinkage and selection operator*, Tibshirani, 1996, Hastie et al., 2015) are ones of the most popular methods used in cases with a large number of model predictors due to the sparsity of the model, consequently the interpretability of the results as well as the ability to avoid overfitting problem and thus better prediction. With lasso, small coefficients are set to zero, and large coefficients are nonzero but shrunk in absolute value.

In statistics, machine learning, engineering, finance, medical and genetic and other research fields, lasso regressions have been widely used (Hastie et al., 2015; Tutz, 2012). Lately, applications of the lasso in several established statistical modelling approaches have been proposed, such as in structural equation models (Epskamp, Rhemtulla, & Borsboom, 2017; Huang, Chen, & Weng, 2017; Jacobucci, Grimm, & McArdle, 2016), IRT approaches and differential item functioning analyses (Sun, Chen, Liu, Ying, & Xin, 2016; Tutz & Schauberger, 2015), latent class models (Y. Chen, Li, Liu, & Ying, 2017).

In the behavioural sciences and educational research fields however, these method are still less wellknown despite its great relevance and potential (McNeish, 2015; G. Pham et al., 2016). Thus, in this study, regularized regressions with lasso are chosen to be the main analysis methods to answer the research questions 10–14 regarding instructional effects on student growth. Since these methods are less

known in this research field, OLS regression analyses are also applied in models in which the number of predictors are not too large for the purpose of result comparison.

In the following sections, regression models using OLS and lasso applied in this study are introduced.

VI.3.3.1 Estimating linear effects of instructional factors using multiple linear regression with OLS and lasso

With Y_i represents the class mean growth of class i, and $Z = (z_{i1}, \dots, z_{ij}, \dots, z_{ip})$ represents p predictors associating with class i (e. g. context and instructional factors), the multiple linear regression model assumes that

$$Y_i = \beta_0 + \sum_{j=1}^{p} z_{ij}\,\beta_j + e_i$$

where β_0 (intercept) and $\beta = (\beta_1, \dots, \beta_j, \dots, \beta_p)$ (regression coefficients) are $p + 1$ unknown parameters and e_i is an error term. The OLS method provides the estimates of the unknown parameters by minimizing the least-squares objective function:

$$L^{OLS}{}_{\beta} = \sum_{i=1}^{N}\left(Y_i - \beta_0 - \sum_{j=1}^{p} z_{ij}\,\beta_j\right)^2 = (Y - \boldsymbol{Z}\beta)^T(Y - \boldsymbol{Z}\beta), \qquad (2)$$

Y is the outcome vector, $\boldsymbol{Z}$ is the predictor matix, N is the sample size. The solution $\hat{\beta}$ of the least squares minimization is given as $(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T Y$. If p is large, the interpretation of the final model is challenging, since all least-squares estimates β are typically nonzero. If $p > N$, the least-squares estimates are not unique, and there is an infinite set of solutions so that $L^{OLS}{}_{\beta} = 0$, which all fit the data completely (this is known as the *overfitting* problem).

To solve this problem, the lasso applies a *lasso penalty* to *regularize* the estimation process. With lasso, the parameters are estimated by minimizing the problem:

$$L^{Lasso}{}_{\beta} = \frac{1}{2N}\sum_{i=1}^{N}\left(Y_i - \beta_0 - \sum_{j=1}^{p} z_{ij}\,\beta_j\right)^2 + \lambda \cdot \sum_{j=1}^{p}|\beta_j|, \qquad \lambda \geq 0, \qquad (3)$$

the lasso penalty is $\sum_{j=1}^{p}|\beta_j|$, λ is called the *shrinkage* parameter (Friedman, Hastie, & Tibshirani, 2010), N is the sample size. The additional factor $1/2N$ in comparison to (2) makes λ in (3) comparable for different sample sizes, while it would make no difference in (2) (Hastie et al., 2015). lasso solution $\hat{\beta}^{lasso}(\lambda)$ is depending on the chosen λ value by applying the soft-thresholding operator $S(\hat{\beta},\lambda)$ in which:

$$\hat{\beta}^{lasso}(\lambda) = S(\hat{\beta},\lambda) = \begin{cases} \hat{\beta} - \lambda, & if\ \ \hat{\beta} > 0\ and\ \lambda < |\hat{\beta}| \\ \hat{\beta} + \lambda, & if\ \ \hat{\beta} < 0\ and\ \lambda < |\hat{\beta}| \\ 0, & if\ \lambda \geq |\hat{\beta}| \end{cases}$$

For cases with large p or when $p > N$, there are many algorithms to estimate $\hat{\beta}$, for instance an effective and fast algorithm was given by Friedman et al. (2010) based on an iterative coordinatewise gradient method.

One problem arises when $\beta = 0$, the penalty functions become nonseparable, and the coordinate descent is not guaranteed to converge using the standard Newton Raphson algorithm. To solve this, Friedman et al. (2010) implemented a proximal Newton-type method having a closed form expression for the starting solutions, and each subsequent solution is warm-started from the previous close-by solution (for proximal Newton-type methods see Lee, Sun, & Saunders, 2014).

To select an optimal value of λ, a list a possible values is generated, starting with a so large one that $\lambda \geq |\hat{\beta}|$ for all $\hat{\beta}$, and ending with a very small values of λ). The optimal value can be selected using information-criterion based model selection (such as Akaike information criterion – AIC, Bayes information criterion – BIC), or often using *k*-fold cross validation.

In *k*-fold cross validation, the sample is randomly partitioned into *k* equal size subsamples. The process is repeated *k*-times, each time $k - 1$ subsamples are used as training data to obtain a set of estimates, and the remaining subsample is retained as the validation data for testing the model. Then, the *k* results are averaged to produce a single estimation. The λ–value which produces the minimal cross-validated mean square errors is selected. An example of how λ is chosen based on *k*-fold cross validation process is given by G. Pham et al. (2016). Using cross validation process, prediction accuracy for future data can be enhanced, which means the generalizability of the results can be better assured (James, Witten, Hastie, & Tibshirani, 2013; Kohavi, 1995).

VI.3.3.2 Identifying important linear effects based on multiple regressions

Although identifying the importance of regressor variables is not a new topic and several criteria have been suggested and often used for this purpose, variable importance has not been very well defined as a concept (Grömping, 2009). In this study, three approaches are applied to identify most important effects of instructional factors on student growth including the statistical significance of OLS regression coefficients (based on *p*-value), the local effect size Cohen's f^2, and the lasso regression coefficients.

VI.3.3.2.1 Statistical significance of the OLS regression coefficients

The first approach involves the statistical significance of OLS multiple regressions of regressor variables based on their *p*-values. If $p < .05$, it has normally been interpreted that the effect is statistically significantly different than zero, and the predictor is statistically significant or important.

However, a *p*-value is the probability for a given statistical model assuming the null hypothesis is true. Thus, *p*-value can rather be used to reject the null hypothesis given a specififed model than to confirmed if the alternative hypothesis is true or if an effect is truly nonzero (Nuzzo, 2014; Stelzl, 1982; Wasserstein & Lazar, 2016). Even then, "*a p-value near 0.05 taken by itself offers only small evidence against the null hypothesis*" (Johnson, 2013; see also Greenland et al., 2016), since it provides no clues to justify all the model assumptions involved. A *p*-value less than .05 suggests no more than it is worth to "*repeat the experiment*", and if subsequent studies also yield $p < .05$, the conclusion is the observed effect is unlikely to be caused by chance (Goodman, 2008, p. 135).

Should *p*-values are reported, it should furthermore be taken into account that *p*-values can be influenced by sample size: "*An effect that fails to be significant at a specific level in a small sample can be significant in a larger sample*" (Moore, McCabe, & Craig, 2009, p. 465).

As a basis for proper inference of the results in general, Wasserstein & Lazar (2016) and Greenland et al. (2016) encouraged researchers to supplement or even replace *p*-values with other estimates and approaches.

VI.3.3.2.2 Effect size Cohen's f^2

In this research field, it is extremly unlikely to expect that an effect is truly zero where "*everything seems to work*" (Hattie, 2009, p. 1). The relevant research question should therefore be "*what matters most*" rather than "*what matters*". This is actually the question of the relative importance of regressor variables. To answer this question, Cohen (1988) suggested the use of *effect sizes* together with rule of thumb thresholds to categorize them into *small*, *medium* and *large* effect sizes. The effect size Cohen's *d* is

apparently most well known, and widely used to quantify an effect based on experimental study design or to quantify growth between measurement points. For multiple regressions, Cohen (1988) suggested the effect size f^2, which is equally useful but has been relatively less common (Selya, Rose, Dierker, Hedeker, & Mermelstein, 2012).

In a regression model, the *global* effect size Cohen's f^2 is defined as follows:

$$f^2 = \frac{R^2}{1 - R^2},$$

R^2 is the proportion of variance of the outcome accounted for by all model predictors. The *local effect size* Cohen's f^2 of each single predictor in model can also be calculated, which reflects the proportion of variance uniquely accounted for by one variable, over and above that of all other variables. Considering C is the set of all model covariates (e.g. context variables), z is the variable of interest (e.g. an instructional variable), R_C^2 is the proportion of variance of the outcome accounted for by C, and R_{total}^2 is the proportion of variance of the outcome accounted for by C and z together, the local effect size Cohen's f_z^2 associated with z is calculated as:

$$f_z^2 = \frac{R_{total}^2 - R_C^2}{1 - R_{total}^2}$$

(Cohen, 1988, p. 410).

Having p regression models, each model includes the same set of context variables C and one instructional variable z_j as model predictors, the local effect f_z^2 of p instructional variables z_1, z_2, …, z_p can be calculated and compared with each other. This can be regarded as one among the state-of-the-art methods for identifying the relative importance of the instructional predictors in relation to each other with regard to the outcome (after controlling for context effects) (Grömping, 2009).

Cohen also suggested the thresholds to categorize the effect size f^2 into small, medium, and large effects:

- Small effect size: $f^2 \geq .02$
- Medium effect size: $f^2 \geq .15$
- Large effect size: $f^2 \geq .35$

Accordingly, effects with $f^2 < .02$ can be ignored in the presence of all other model covariates.

VI.3.3.2.3 lasso regression coefficients

Applying multiple regression with lasso, only a subset of predictors has nonzero coefficients. They are regarded as important to the prediction accuracy of the regression model. Note that, among predictors which correlate highly with each other, lasso tends to pick one and ignore the rest (their coefficients are set to zero, Friedman et al., 2010).

***

Multiple regressions using OLS and an extension of lasso regression (see section VI.4.2) were applied to investigate the linear instructional effects after controlling for context effects in this study (research question 10). In each model, all context and one instructional effect are predictor, and class mean growth of the C-test and LC-test is the outcome.

VI.3.4 Modelling and identifying nonlinear effects of instructional factors

VI.3.4.1 Modelling nonlinear relationship between instructional factors and outcomes

To examine the theoretical assumption regarding the nonlinear relationship between instructional effects and student outcomes (see chapter IV.3), curvilinear regression models can be applied (Cohen, Cohen, West, & Aiken, 2003). Curvilinear regression uses a linear model to fit a curved line to data points by including polynomial term(s) of the variable in question as further model predictor(s) beside the main linear effect:

$$Y_i = \beta_0 + z_k\beta_{1k} + z_j^2\beta_{2j} + z_j^3\beta_{3j} + \cdots + z_j^k\beta_{kj} + e_i$$

A second-order polynomial regression model (*quadratic regression*) includes the main effect and the quadratic term of the interested variable, and fits a curved line with one bend to the datasets. A higher-order polynomial terms includes the higher-order and all lower-order terms and fits a curved line with more bends to the datasets. For any dataset, a (very high order) polynomial regression model can always be found which fits the data perfectly ($R^2 \approx 1$), but it might be worse than a simpler model in predicting future data (overfitting problem). Overfitting is undesirable, and one should not use a model that is more flexible or more complex than it needs to be (parsimony, Hawkins, 2004). According to Cohen et al. (2003), the choice of model should be based on theory. Therefore, to examine the assumption that the maximum (or miminum) value might not be the optimum value regarding the effect of an instructional factor, the quadratic regression model is chosen. A linear model examines if the main effect is roughly

positiv or negativ; a quadratic regression assumes a parabolic relationship and identifies accordingly the predictor value at which the empirical maximum/minimum of the outcome is reached.

VI.3.4.2 Identifying nonlinear relationship between instructional factors and outcomes

Both linear and quadratic regression models are likely not the "true" model but approximations of this, and they are both misspecified to some extent (Berk et al., 2014). To identify whether the relationship between an instructional factor and student growth is nonlinear means to decide whether the quadratic term contributes to overall prediction above and beyond the linear term. For the purpose of model selection, several approaches have been suggested.

VI.3.4.2.1 The *F*-test for least square quadratic regression

Traditionally, the *F*-test for gain in prediction by the addition of the quadratic term is used for this purpose (Cohen et al., 2003), with $F = f_{z^2}^2 \times \frac{N-k-1}{\Delta k}$ $(df = (\Delta k, N-k-1))$, $f_{z^2}^2$ is the local effect size Cohen's f^2 of the quadratic term z^2 (see section VI.3.3.2.2), N is the class sample size, k is the total number of model predictors, Δk is the number of additional predictor(s) (in this case $\Delta k = 1$). When F is sufficiently large to meet the significance criterion (Appendix Tables D.1 and D.2 in Cohen et al., 2003), the null hypothesis is rejected or the contribution of the quadratic term z^2 is identified statistically significant (Cohen et al., 2003, pp. 171, 205).

VI.3.4.2.2 Regularization method based on the strong hierarchical lasso

Among other approaches, regularization method based on the lasso can be thought of. The interpretation would be easy, since if the coefficient of the quadratic term is nonzero, the nonlinear relationship can be interpreted as important with regard to future prediction of the outcome. However, since a quadratic term can be seen as an interaction term of a variable with itself, the marginality principle should be taken into account for the sake of the interpretability of the model results (Nelder, 1977). This principle requires that an interaction term (or a quadratic term) can be selected into the model only if the main effect is also included (strong rule of hierarchical structure). The original lasso method does not takes this kind of relationship between model predictors into account, and thus does not ensure that the strong rule of hierarchical structure can be hold in the final model. To solve this problem, several regularization methods for variable selection with hierarchical structure using lasso have been proposed recently (Bien, Taylor, & Tibshirani, 2013; Hao, Feng, & Zhang, 2016; Lim & Hastie, 2015; Yuan, Joseph, & Zou, 2009; Zhao, Rocha, & Yu, 2009). In this study, the *strong hierarchical lasso* suggested by Bien et al. (2013) is applied. This method is applicable for the general case of having many quadratic and

interaction terms in model, enables a simple interpretation of the effect of the hierarchy demand on the solution, and is implemented in the R package **hierNet** (Bien & Tibshirani, 2014).

The general regression model is formulated as follows:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j z_{ij} + \frac{1}{2} \sum_{j,k=1}^{p} \gamma_{jk} z_{ij} z_{ik} + e_i, \tag{4}$$

the β_j coefficients are associated with main effects, while the γ_{jk} coefficients are associated with interaction effects including quadratic terms when $j = k, \gamma_{jk} = \gamma_{kj}, \beta_0$ is the intercept, and e_i is an error term. This model can also be written shortly as:

$$Y = \beta_0 1 + \tilde{Z}\phi + \varepsilon,$$

where 1 is the vector of ones, $\tilde{Z}$ is a design matrix for main and interaction effects, ϕ represents all model coefficients, and $\varepsilon \sim N(0, \sigma^2)$. The optimization problem L_β^{Lasso} (see section VI.3.3.1) for this regression model becomes:

$$L_\phi^{Lasso} = \frac{1}{2} \sum_{i=1}^{N} (Y_i - \beta_0 1 - \tilde{Z}_i \phi)^2 + \lambda \sum |\phi| = \frac{1}{2} \left\| Y - \beta_0 1 - \tilde{Z}\phi \right\|^2 + \lambda \|\phi\|_1,$$

which is refered to as the *all-pairs lasso* as called by Bien et al.

To select a subset of model variables and estimates the nonzero coefficients under strong hierarchy ($\hat{\gamma}_{jk} \neq 0 \Rightarrow \hat{\beta}_j \neq 0$ and $\hat{\beta}_k \neq 0$), Bien et al. used the *k*-fold cross validation approach (see section VI.3.3.1). They searched primarily for interactions that have large main effects, and favoured models with *practical sparsity* (number of raw variables one must include in model to make predictions at a future time) over models with *parameter sparsity* (number of nonzero coefficients in model) due to the practical importance.

It is done by a lasso-like procedure which involves adding a set of convex constraints to the lasso. The full formulation of the lasso optimization problem is rewritten as:

$$L_\phi^{Lasso} = q(\beta_0, \beta, \gamma) + \lambda \|\beta\|_1 + \frac{\lambda}{2} \|\Gamma\|_1,$$

subject to $\Gamma = \Gamma^T$ (Γ is the matrix of coefficients with γ_{jk} the element in jth row and kth column). $q(\beta_0, \beta, \gamma)$ is the loss function $\frac{1}{2}\sum(Y - \beta_0 1 - \tilde{Z}\phi)^2$, and the penalty term is rewritten as $\lambda\|\phi\|_1 = \lambda\|\beta\|_1 + \frac{\lambda}{2}\|\Gamma\|_1$.

The proposed *strong hierarchical lasso* is a modification of the lasso, in which β is replaced by two vectors $\beta^+, \beta^- \in \mathbb{R}^p$: $\beta = \beta^+ - \beta^-$ $(\beta^+ \geq 0, \beta^- \geq 0)$ with the constraints $\beta_j^+\beta_j^- = 0$ for $j = 1, \dots, p$ so these are negative and positive parts and $\beta_j^+ + \beta_j^- = |\beta_j|$, and a constraint on $\|\gamma_j\|_1$ is included: $\|\gamma_j\|_1 \leq |\beta_j|$, γ_j denotes the jth row of Γ. The optimization problem with the strong hierarchical lasso is thus:

$$L^{hierNet} = q(\beta_0, \beta^+ - \beta^-, \gamma) + \lambda 1^T(\beta_j^+ + \beta_j^-) + \frac{\lambda}{2}\|\Gamma\|_1,$$

subject to $\Gamma = \Gamma^T$, ${\beta_j}^+ \geq 0, {\beta_j}^- \geq 0$, $\|\gamma_j\|_1 \leq \beta_j^+ + \beta_j^-$ for $j = 1, \dots, p$. Due to the constraint $\|\gamma_j\|_1 \leq |\beta_j|$, if $\hat{\gamma}_{jk} \neq 0$, then $\|\hat{\gamma}_j\|_1 > 0$ and $\|\hat{\gamma}_k\|_1 > 0$, and $\hat{\beta}_j \neq 0$ and $\hat{\beta}_k \neq 0$. If the best fitting model would have large $\|\gamma_j\|_1$ and moderate $|\beta_j|$, this can be accommodated by making β^+ and β^- both large.

Note that, the results of a quadratic regression should be interpreted carefully and with cautious especially at extreme values on *Z* if data at these extreme values are sparse. Extrapolation of a quadratic regression beyond the extreme values of predictors *Z* is particularly dangerous and undesirable (Cohen et al., 2003; Hawkins, 2004). Regarding this caution, the use of regularization quadratic regression with strong hierarchical lasso based on cross validation method might be advantageous over least square quadratic regression, apart from other mentioned advantages.

***

The nonlinear instructional effects (research question 11) after controlling for context effect were investigated on this basis. In each model, all context and one instructional effect including its quadartic term were included as predictor, and class mean growth of the C-test and LC-test is the outcome.

VI.3.5 Investigating interaction and compensatory effects of instructional factors on student growth as well as the aptitude treatment effect

The regression model in equation (4) includes interaction terms $Z_j Z_k$ for all predictors j and k. Thus, this can be applied to examine the interaction effects between instructional factors as well as the aptitude

treatment effect at class level, in which interaction between interactional factors and the prior class achievement are treated as model predictors.

Furthermore, a multiple regression in general is typically used to examine the additive or compensatory effect between two or more variables (Eid et al., 2010). Hence, with this model, the compensatory effect between instructional factors is modelled and can be effectively examined.

Due to the large number of available instructional variables and consequently extremely large number of interaction terms in comparison to the number of classes, the investigation into thousands of interaction effects between all pair of variables were skipped for practical reasons. After all, these effects (if exist) happen simultaneously and jointly in combination with addtive (compensatory) effects (see chapter IV.3). This can be investigated via a regularized regression model with strong hierarchical lasso using R package **hierNet** including all main and interaction effects as well as quadratic terms of the interested variables, which was implemented to answer the research questions 12–14 in this study. For such a model with a large number of predictors given small sample size, regression models based on least squares cannot be applied. To reduce the model complexity as well as due to practical importance, only instructional factors with identified important linear and/or nonlinear effects are selected to be included in this model.

VI.4 Dealing with multiple imputed datasets and sampling error

VI.4.1 Rubin's rules

To achieve estimates based on multiple imputed datasets, Rubin's rules (1987) were applied (see also Enders, 2010; Bruneforth, Oberwimmer, & Robitzsch, 2016). Accordingly, each analysis was performed multiple times, each time based on one imputed dataset. The *multiple imputation point estimate* $\hat{\mu}$ (e.g. mean, regression coefficient) is defined as the arithmetic average over all respective *m* estimates $\hat{\mu}_m$ (l = *1,2,…M*; *M* is the number of imputed datasets, in this study *M* = 10):

$$\hat{\mu} = \frac{1}{M}\sum_{m=1}^{M}\hat{\mu}_m$$

The estimation variance of a point estimate $\hat{\mu}$ according to Rubin (1987) is calculated by combining two components: the variance component within each imputed dataset $V_w(\hat{\mu})$ (within-imputation variance), and the variance component caused by variation between imputed datasets $V_B(\hat{\mu})$ (between-imputation

variance). The between-imputation variance $V_B(\hat{\mu})$ is the product of the sum squares of differences between each $\hat{\mu}_m$ and $\hat{\mu}$ with a constant factor:

$$V_B(\hat{\mu}) = (1 + \frac{1}{M}) \cdot \sum_{m=1}^{M} (\hat{\mu}_m - \hat{\mu})^2$$

As usual in standard statistical inference for a sample, the sampling error variance should be taken into account to calculate the standard error of the estimates associating with each of the imputed dataset. Due to the small sample size at the level of analysis ($N = 50$), the bootstrap resampling method for class sample (Efron & Tibshirani, 1986) was applied to estimate the sampling variance $V_w(\hat{\mu})$ associated with each imputed dataset to achieve higher power level of the results (Diaconis & Efron, 1983). For this purpose, $S = 100$ bootstrap samples were randomly generated. The total number of datasets was thus equal to 10 (number of imputations) times 100 (number of bootstrap samples) is 1000 datasets. Each boostrap sample consists of a subset of the class sample, and is randomly drawn from the class sample with replacement, so that the size of each bootstrap sample is the same of that of the original class sample ($N = 50$). Correspondingly there are *replicate weights* associated with each bootstrap sample. For each of the imputed dataset m, 100 estimates $\hat{\mu}_{s,m}$ were calculated using 100 bootstrap samples with the corresonding replicate weights. The sampling variance or within-imputation variance of each imputed dataset m is given by:

$$V_{w,m}(\hat{\mu}_m) = \frac{1}{S} \cdot \sum_{s=1}^{S} (\hat{\mu}_{s,m} - \hat{\mu}_m)^2$$

The within-imputation variance $V_w(\hat{\mu})$ of the multiple imputation point estimate $\hat{\mu}$ is the arithmetic average over all $V_{w,m}(\hat{\mu}_m)$:

$$V_w(\hat{\mu}) = \frac{1}{M} \cdot \sum_{m=1}^{M} V_{w,m}(\hat{\mu}_m)$$

Finally, the total variance associated with an multiple imputation point estimate $\hat{\mu}$ is:

$$\mathrm{Var}(\hat{\mu}) = V_B(\hat{\mu}) + V_w(\hat{\mu})$$

The standard error of each estimate $\hat{\mu}$ is the square root of $\mathrm{Var}(\hat{\mu})$: $\mathrm{SE}(\hat{\mu}) = \sqrt{\mathrm{Var}(\hat{\mu})}$.

To specify the confidence intervals or *p*-values of parameter estimates which are assumed to have a t-distribution (e. g. means or regression coefficients), the fraction of missing imputation (FMI) is involved, which quantifies the missing data's influence on the total variance of a parameter estimate:

$\text{FMI} = \frac{V_B(\hat{\mu})}{\text{Var}(\hat{\mu})}$. The degrees of freedom df used to specify the confidence interval of estimates is defined as:

$$df = \frac{S-1}{\text{FMI}}$$

The above described procedures are implemented in the R-package **BIFIEsurvey** (BIFIE, 2015), and were applied to estimate all study results except results of lasso regressions. The method to estimate lasso regression coefficients based on multiple imputed datasets is described below.

For pooling *F*-tests of analysis of variance, the adjusted denominator degrees of freedom used for inferences in multiple imputation with small sample sizes suggested by Reiter (2007) was applied (see also Van Ginkel & Kroonenberg, 2014).

VI.4.2 lasso regression and multiple imputations

With one dataset, the results of a lasso regression (see chapter VI.3.3.1) can be easily interpreted, since some regression coefficients are set to zero, and the others are nonzero (variables are selected). However, using multiply imputed datasets, the variable selection can be inconsistent across the multiple datasets: a coefficient can be nonzero using one imputed dataset, but is zero using another imputed dataset. While both multiple imputation (MI) and variable selection based on lasso regression have become increasingly popular, variable selection on multiply imputed dataset has remained a longstanding statistical problem. Only recently, a couple of statistical approaches have been proposed for this purpose (Q. Chen & Wang, 2013; Guo et al., 2015; Musoro, Zwinderman, Puhan, Riet, & Geskus, 2014). Among them, the MI-lasso method (Q. Chen & Wang, 2013) – which is an extention of the lasso method – has gained most resonance (30 citations until the time of this writing according to Google Scholar). This method based on the group lasso penalty, and is implemented to yield a consistent variable selection across multiple datasets by fitting regression models on all imputed datasets jointly.

Denote $\hat{\beta}_{m,j} = \hat{\beta}_{1,j}, \dots, \hat{\beta}_{M,j}$ $(m = 1, \dots, M)$ be the M estimated coefficients for variable z_j in M imputed datasets. If z_j is unimportant, all $\hat{\beta}_{m,j}$ should be zero, otherwise they should all be nonzero. The joint optimization function over all imputed data is:

$$L_{\beta_{m,j}}^{MI-lasso} = \sum_{m=1}^{M}\sum_{i=1}^{N}\left(Y_i - \beta_{m,0} - \sum_{j=1}^{p}\beta_{m,j}z_{m,ij}\right)^2 + \lambda\sum_{j=1}^{p}\sqrt{\beta_{1,j}^2 + \cdots + \beta_{M,j}^2},$$

Where $\sum_{j=1}^{p}\sqrt{\beta_{1,j}^2+\cdots+\beta_{M,j}^2}$ is called the group lasso penalty. This penalty guarantees the desired consistency of variable selection across all imputed datasets (all zero or all nonzero, see Yuan & Lin, 2006). The difficulty here is also (cf. also section VI.3.3.1), that the penalty function is not differential at the origin point. To solve this problem, Chen and Wang used the quadratic approximation as proposed by Fan and Li (2001) and solved the problem iteratively (this approach can also be applied for the normal lasso regression, albeit it is less efficient than the approached used by Friedman et al., 2010, and has been implemented in R package **LAM**, Robitzsch, 2017).

Given $\hat{\beta}_{m,j}^{(t)}$ the set of estimates at the t^{th} iteration, as long as $\sqrt{(\hat{\beta}_{1,j}^{(t)})^2+\cdots+(\hat{\beta}_{M,j}^{(t)})^2}>0$ the approximation is:

$$\sqrt{\beta_{1,j}^2+\cdots+\beta_{M,j}^2}\approx\frac{\beta_{1,j}^2+\cdots+\beta_{M,j}^2}{\sqrt{(\hat{\beta}_{1,j}^{(t)})^2+\cdots+(\hat{\beta}_{M,j}^{(t)})^2}}$$

Correspondingly, the group penalty can be approximated by $\sum_{j=1}^{p}c_j\beta_{m,j}^2$, with $c_j=1/\sqrt{(\hat{\beta}_{1,j}^{(t)})^2+\cdots+(\hat{\beta}_{M,j}^{(t)})^2}$. The regression coefficients of the next iteration $\hat{\beta}_{m,j}^{(t+1)}$ can be obtained by solving M separate ridge regressions (Tutz, 2012). The iterations continued until the convergence is reached. To avoid the possible problem that a group of coefficients stay at zero once they are shrunken to zero, Chen and Wang fixed $\hat{\beta}_{1,j}^{(t)}=\cdots=\hat{\beta}_{M,j}^{(t)}$ when $\sqrt{(\hat{\beta}_{1,j}^{(t)})^2+\cdots+(\hat{\beta}_{M,j}^{(t)})^2}\leq\sqrt{M}\delta$ and chosed a very small value of $\delta=10^{-10}$.

Finally, Rubin's rules are applied to achieve ultimate regression coefficients.

This method was applied to identify important linear instructional effects in this study.

For regularized regression model with strong hierarchical lasso, no statistical approaches for multiply imputed dataset have been found until the time of this writing. Hence, pooled regression coefficient estimates according to Rubin's rules together with the number of nonzero estimates (over 10 imputed datasets) are used to interpret the importance of model predictors.

VI.5 Dealing with different results associated with different scaling models

Study results based on different scaling models will be compared and discussed with respect to the last research question (question 15) about the impact and relevance of selected scaling model on the estimates of instructional effects on student growth (see chapter X).

In addition, due to the fact that the "true" scaling model might be somewhere in between, averaging results via model averaging might be considered better estimates of the interested measures (see Burnham & Anderson, 2002, Levin & Williams, 2003, Brock, Durlauf, & West, 2007, Weigel, Knutti, Liniger, & Appenzeller, 2010, Robitzsch, 2016a). Assuming that all selected models have the same possibility of being the "true" model, estimates via model averaging $\bar{\mu}$ are calculated as the arithmetic average of estimates based on all scaling models μ_r (estimates based on scaling model *r*). The variance of the average estimate over different models is:

$$\mathrm{Var}(\mu) = \frac{\sum_{r=1}^{R} \mathrm{Var}(\mu_r) + \sum_{r=1}^{R} (\mu_r - \bar{\mu})^2}{R},$$

with *R* is the number of applied scaling models. With this procedure, the uncertainty of model selection can be taken into account in estimating the standard error of the estimates (Berk, Brown, & Zhao, 2010; Robitzsch, 2016). To average parameters, the R-package **MuMIn** (K. Barton, 2016) was applied.

VII. Video-based descriptive results of classroom instruction in EFL in Vietnam

In this chapter, the hypotheses regarding classroom EFL instruction in this study (Research Questions 1–6) will be examined. The analysis will be based on the basic coding and rating indicators together with students' and teachers' responses to the short questionnaires after the recorded lessons. The results of the curriculum-oriented lessons serve as a basis for understanding the quality of regular instruction. The results of the extra-curricular communication-oriented lessons will be used to gain an insight into the language quality (speaking mistakes) of students and teachers.

As mentioned in Chapter III.2.5, 41 out of 50 classes voluntarily took part in the video study. Table 3 shows the participation rates in the video study in the three data collection places.

Table 3: Participation rate in the video study

Place	Number of classes in the survey	Number of classes in the video study	Participation rate in the video study
Hanoi	20	17	85%
Ho Chi Minh city	20	14	70%
Bac Ninh province	10	10	100%

As shown in Table 3, 100% of the classes in Bac Ninh province in the survey study took part in the video study (N = 10). Of the attended classes in the test and survey study, 70% of the classes in Ho Chi Minh city (N = 14) and 85% of the classes in Hanoi (N = 17) participated in the video study. In the following sections, the results will be reported based on video data and short questionnaire data from these 41 classes.

VII.1 Representativity of the recorded lessons

The representativity of the videotaped lessons was judged by the teachers and students through short questionnaires right at the end of the recorded lessons. According to the teachers, the curriculum-oriented lessons proceeded as usual, and they were satisfied with themselves as well as with the students. The relevant descriptive statistics are shown in detail in Figure 9.

In fact, the teachers did not feel really nervous during the recorded lesson. None of them felt “very” nervous; five (12%) of them were “rather” nervous. Thus, the instruction was not affected by teachers’ nervousness according to the teachers. Indeed, 77% (30/39, two non-responses) of teachers answered that the instruction did not deviate from their initial lesson plan. Altogether, 90% of teachers (36/40, one non-response) stated that the recorded lessons were typical of a normal English lesson; 83% of teachers (34/41) were rather or very much satisfied with their performance in the recorded lesson, and 93% (38/41) were rather or very much satisfied with the students’ participation and performance during the lesson.

Teachers' judgment of the videotaped lesson	
How typical was the recorded lesson in comparison to a normal lesson?	very much: 18; rather: 18; little: 2; not at all: 2; missing: 1
Were you nervous during the videotaped lesson?	very much: 0; rather: 5; little: 20; not at all: 16; missing: 0
If yes, did the nervousness affect the instruction?	very much: 0; rather: 1; little: 17; not at all: 19; missing: 4
Did the lesson deviate from your plan?	yes: 9; no: 30; missing: 2
How satisfied are you with yourself in the lesson?	very much: 5; rather: 29; little: 7; not at all: 0; missing: 0
How sastisfied are you with the students in the lesson?	very much: 10; rather: 28; little: 3; not at all: 0; missing: 0

Figure 9: Representativity of the videotaped lesson according to teachers' judgement

According to the students, the lessons were better prepared and proceeded better than usual. In Figure 10, the blue bars represent the means of the class student judgments, and the horizontal lines represent the standard deviation of the class sample.

The students found the lesson more interesting, diversified, quieter, and stated that more materials were provided. Furthermore, they participated more actively and frequently in the lesson. The mother tongue language Vietnamese was less often used by both the teachers and the students. The frequency with which each student was called upon by the teachers was more or less the same as usual.

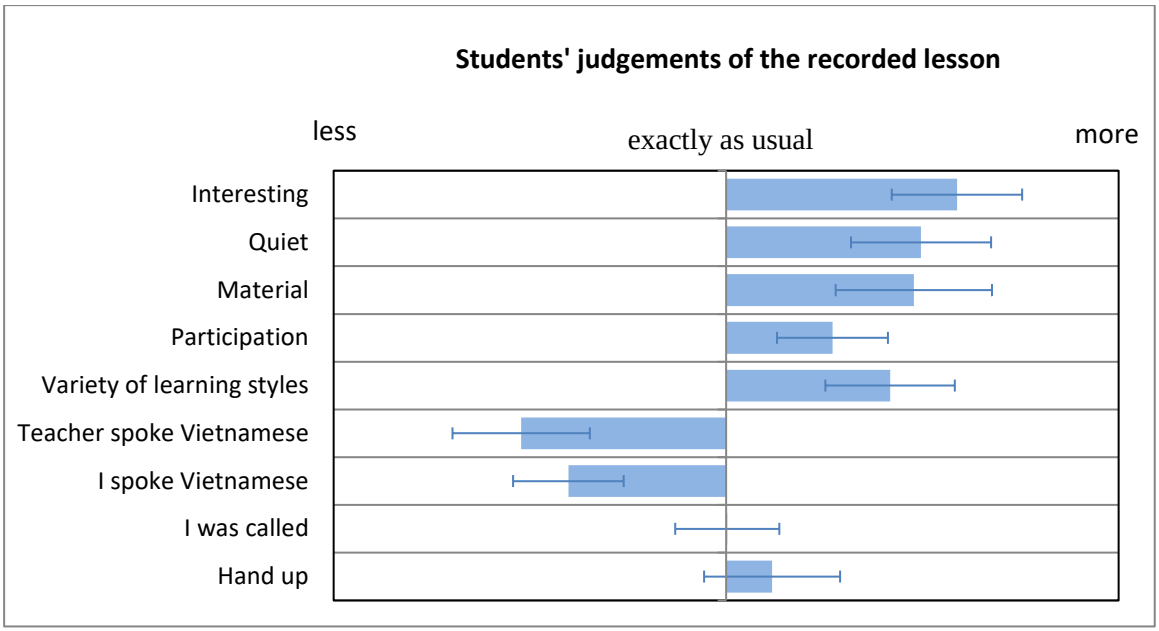

Figure 10: Students' judgment of the representativity of the recorded lessons

VII.2 Are the EFL lessons teacher-centered and textbook-driven?

To answer the first research question ("*Does the study confirm the assumption that EFL lessons in this study are teacher-centered and textbook-driven?*"), the following basic coding indicators of classroom instruction were analyzed: time-on-task, lesson episodes, speaking time and language of the teachers and students, communication patterns in lessons, types of student statements, and syllabus-related teacher activities. In addition, the results based on the following rating variables were also taken into account: task orientation, lesson monitoring, and the variation/adaptivity of the lessons.

VII.2.1 Time-on-task

A school lesson in Vietnam lasts 45 minutes. Figure 11 shows the proportion of time devoted to different components in the curriculum-oriented lessons.

Lesson time for the syllabus-related contents and learning processes (syllabus-related subjects) was an average of 87.5% (*SD* = 4.6%), approximately 39+/- 2 minutes. The other 4 to 6 minutes were used for procedures such as organizational matters (not syllabus-related), preparation, and transition between phases. The time for social activities was about 20 seconds or 0.7% of the lesson time (*SD* = 0.7%), for instance, for greetings, jokes, and laughter.

During the recorded lessons, no serious disciplinary problems occurred. Hence, teachers barely had to spend time on discipline-related activities (*M* = 0.04%, *SD* = 0.09%). These figures confirm previous

homogeneous findings about the high level of discipline in school lessons in East Asia (A. Helmke & Hesse, 2010; D. Y. F. Ho et al., 2002, pp. 41–42; Stigler & Hiebert, 1999).

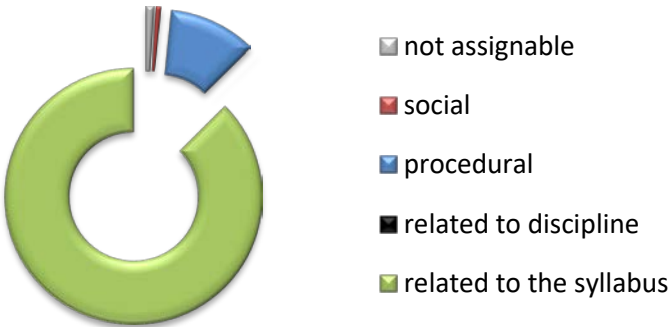

Figure 11: Lesson time and components

VII.2.2 Task orientation

One of the rating variables was the task orientation of instruction. The raters rated how frequently teachers used lesson time for subject-related tasks without idling and wasting time (min = 1, max = 4). In accordance with the results of the previous section, the rating results showed that lesson time in all lessons was mostly used for textbook content but not for other themes or purposes. The mean rating of all lessons approached the maximum category (M = 3.8, SD = 0.4). Ratings did not vary much between lessons, only rating categories 3 and 4 were assigned.

All recorded lessons started right after the ritual choral greeting with an exercise and ended with a learning game or an exercise. The beginning and ending of conversations was not observed in all lessons, possibly because it was not integrated into the textbook.

VII.2.3 Lesson episodes

The time percentages of the lesson episodes (100% is equal to total lesson time) are visualized by the blue bars; one standard deviation of each value is represented by half the length of the thin vertical line in Figure 12 (limited by line 0 of the x-axis). Teacher-centered discussions (discussions that were led by teachers) played a major part in all recorded lessons. They amounted to 53.8% of the entire lesson time on average with a standard deviation SD = 14.2% (approx. 6.4 minutes). In contrast, student-centered discussions (discussions which were led by a student or a group of students) were seldom observed with an average amount of time of only 32.4 seconds (1.2%). In 33 out of the 41 recorded lessons, there were no discussions led by students.

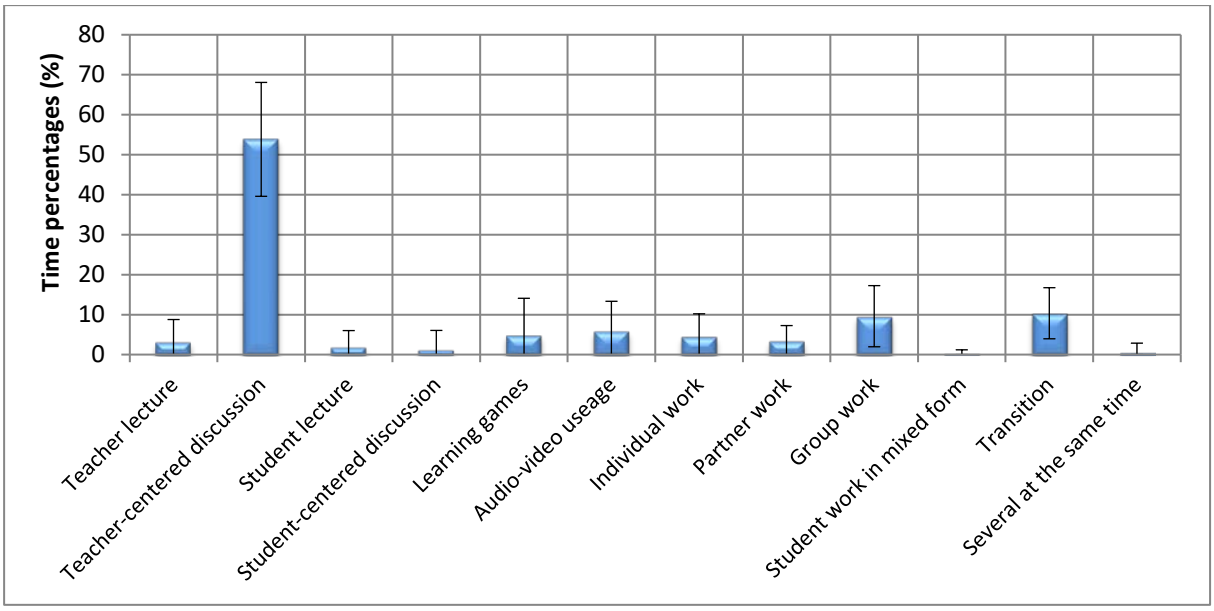

Figure 12: Duration time percentages of lesson episodes

Note: the blue bars represent the mean time percentages of the lesson episodes; the thin black vertical lines represent one standard deviation to either side of the mean values, limited by line 0 of the x axis

Lectures (prepared, structured, long presentations which were generally initiated by the person speaking and were not a response) took place infrequently, with neither students nor teachers giving lectures (the respective means were 2.0% and 3.3% of the total lesson time). Teachers in 25 classes did not give any lectures, and students did not give any lectures in 31 classes.

Unlike lectures, learning games were implemented and video tapes were played in nearly every lesson with a duration of 4.9% and 6% of the total lesson time on average. Both games and the usage of audio-video were an integral part of the textbook with concrete examples and descriptions, including a CD.

Similarly, student work, which took several forms such as individual work, partner work, or group work, was often conducted following a concrete suggestion in the textbook. Together, they took up 17.9% of the total lesson time (*SD* = 7.9%). Among these student work forms, group work was the most popular, with an average time duration of nearly 10% of total lesson time (approx. 4.5 minutes).

Between two subsequent lesson episodes, a transition often took place. Such a transition is characterized by the fact that one teaching method/class arrangement or episode comes to an end, but the next does not begin immediately (e.g., students change places, move around tables, clear up their desks, etc.). These lesson episodes were coded as a *transition* (procedural teaching phase). The time spent on transitions between lesson episodes was *M* = 10.3% (*SD* = 6.4%) on average, comparable to the time

spent on group work. However, a typical lesson might be less varied regarding the social form and lesson episodes, according to the students' judgments (see Chapter VII.1).

On the whole, the instruction proceeded in a very similar manner. The lessons varied negligibly with regard to the time duration of each lesson episode. The main difference related to the amount of time devoted to teacher-centered discussions, group work, and transitions. In lessons with more time for teacher-centered discussions, there were fewer transitions ($r = –.50$, $p < .001$). In addition, lessons with more time for transitions were often characterized by the fact that there was more time for student partner work ($r = .30$, $p < .05$).

VII.2.4 Lesson communication: time, language, and pattern

VII.2.4.1 Speaking time and language

On average, the teachers' speaking time amounted to 53% of total lesson time (*SD* = 9%, see Figure 13). English was used as the teaching language, and teachers spoke in English in 83% of their speaking time.

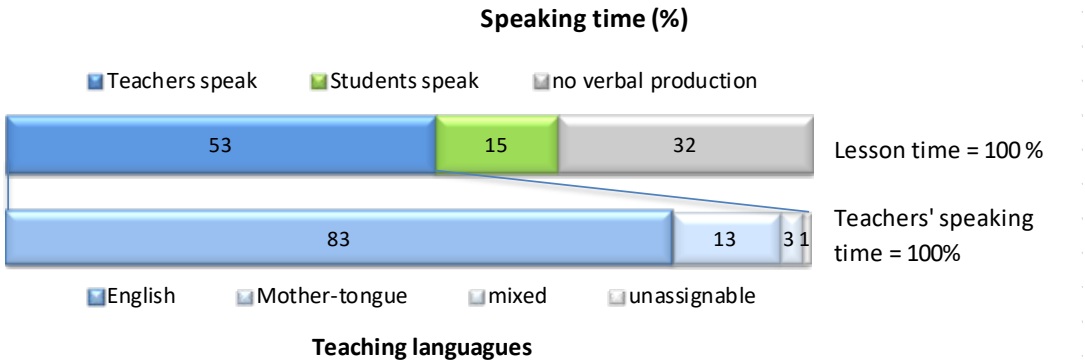

Figure 13: Teachers' speaking time and teaching languagues (%)

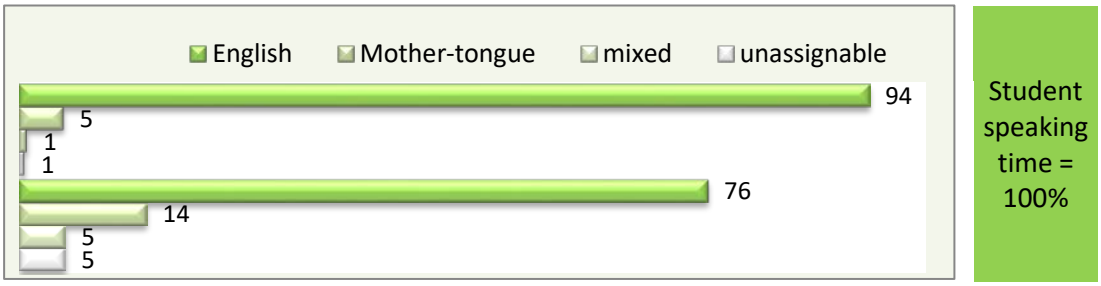

Figure 14: Students' speaking time by languages (%) in English lessons

Together, the students spoke in 15.2% of total lesson time (*SD* = 4.9%), equivalent to 22% of lesson time with verbal production. When they spoke, students mainly used English as the communication language in lessons, with an average speaking time in English *M* = 94% of their total speaking time (*SD* = 7.6%, see Figure 14).

However, according to the results of the short student questionnaire, the amount of time in which teachers and students spoke in Vietnamese in regular (not videotaped) lessons might be higher (c.f. Chapter VII.1)

Among all student speaking turns (100%, relative frequency), 58.2% were individual turns (an individual student spoke, equivalent to 75.4% of total student speaking time); 41.8% were group turns (students spoke in a group) or class turns (all students chorused). Excluding group turns and class turns, the individual student verbal contributions amounted to 11.4% of the total lesson time.

VII.2.4.2 Communication pattern

The communication pattern shows how the class conversations proceeded (see Figure 15). In 76% of teacher speaking time, teachers spoke to a group of students/the whole class but not to an individual student; this corresponded to 59% of total class conversation time. During those phases, teacher-centered discussions or teacher lectures took place rather than dialogues. Dialogues between the class teacher and students (represented by the arrow between the two blue bars in Figure 15) took place during 31% of the total class conversation time. The teacher's part in dialogues added up to 19% of class conversation time or 60% of total dialogue time. Almost all dialogues between the teacher and individual student (T-S dialogues) proceeded in the same specific sequence: *teacher requested – student replied once – teacher gave feedback* (this pattern was true for 82.9% of the T-S dialogues with *SD* = 12.5%). Only 11% of all T-S dialogues encompassed two student speaking turns (by the participating student). Longer T–S dialogues, in which the participating student had more than two speaking turns in one dialogue, or S–T dialogues with a student initiating the conversation occurred rarely. Similarly, dialogues between students (S–S dialogues) during class conversation phases rarely happened, amounting to only 4% of the total class conversation time.

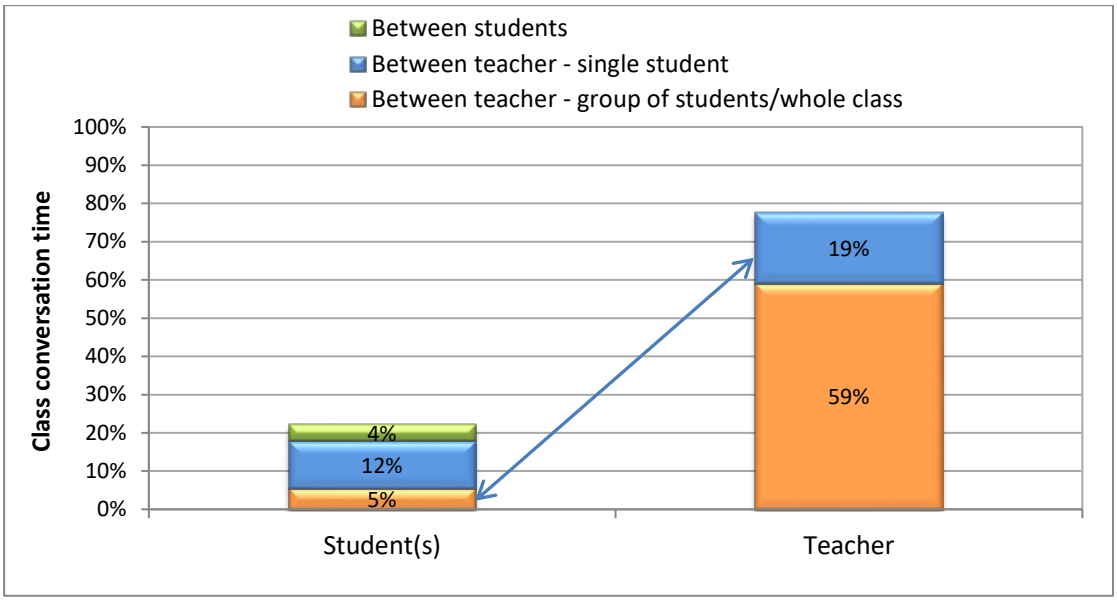

Figure 15: Communication pattern (time percentage) during class conversation time

VII.2.5 Types of student statements

To analyze the different types of student statements, the total speaking time in English of all students together is regarded as 100% (equivalent to 94% of total student speaking time). Analogous to in the DESI-video study, all types of student statements were coded in seven categories: speaking freely (independent statements that were not restricted by any instructions), speaking based on instructions, reading out own text (student reads out loud texts that he/she produced himself/herself beforehand, e.g., during silent work or as homework), reading out text of others (e.g., student reads aloud from the textbook), statement of non-knowledge (e.g., “I don’t know”), repetition (a statement repeated verbatim), and unassignable statements. The results are shown in Figure 16.

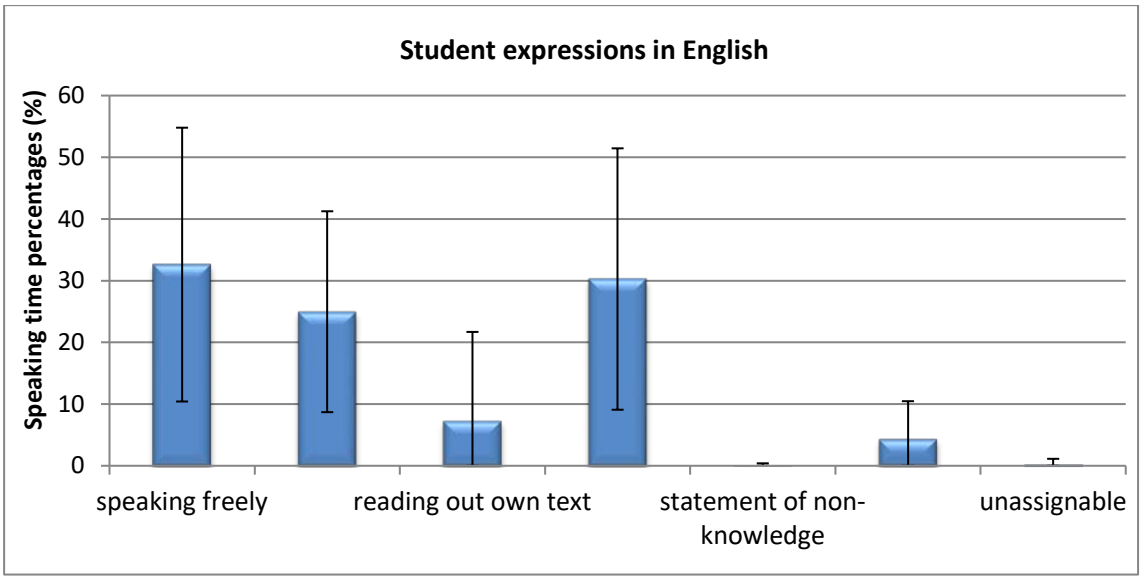

Figure 16: Relative students' speaking time by mode of English expressions

The category "statement of non-knowledge" was rarely observed in all classes (*M* = 0.07%, *SD* = 0.3%).

In contrast, students often read out given texts (*M* = 30.3%, *SD* = 21.2%) and spoke based on instructions (*M* = 25%, *SD* = 16.3%). The category "repetition" of statements was observed less often (*M* = 4.4%, *SD* = 6.0%). Together, these three categories add up to 59.7% (*SD* = 24.7%) of student speaking time in English.

In comparison to these three types of statement together, "reading out own text" was much less frequently observed (*M* = 7.4%, *SD* = 14.3%). In addition, it was totally absent in 23 out of 41 lessons.

In 32.6% of total student speaking time in English (corresponding to 34% of all student utterances in a lesson), students made statements independently. In this respect, the classes also varied strongly (*SD* = 22.1%).

With regard to the student independent statements, the length/completeness of the statements was also coded. The results are shown in Figure 17. In the following paragraph, the total number of student independent statements in English in a lesson is regarded as 100%.

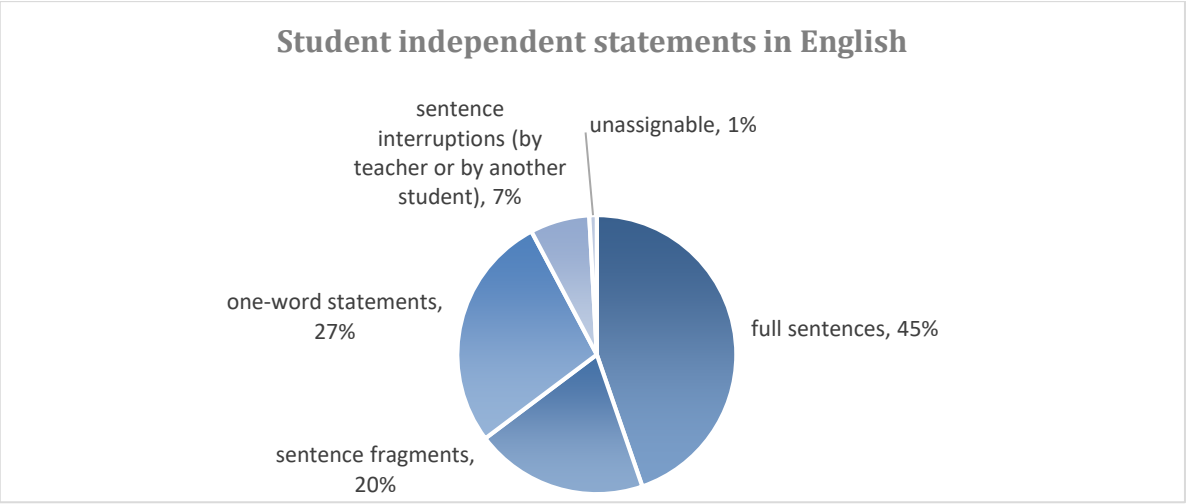

Figure 17: Frequencies of different modes of students' free statements in English

Students often used full sentences (not including one-word statements) to contribute verbally, which constituted 44.7% (*SD* = 23%) of all students' independent statements on average. An additional 27.5% of student independent statements were one-word statements (such as "yes", "no", *SD* = 19.7%) and one-fifth were sentence fragments (*M* = 20%, *SD* = 14.4%). In 6.9% of cases on average (*SD* = 8.1%), students' free statements were interrupted by the teacher or another student. The high value of standard deviations shows that there were big differences between classes concerning the frequency of students' independent statements.

VII.2.6 Syllabus-related teacher activities

This section describes how the syllabus-related teacher activities in the class sample proceeded. An overview of the time percentages of different syllabus-related teacher activities can be found in Figure 18 (100% is equal to the total time of all teacher turns).

The major part of teaching activities was composed of teachers' direct instruction, including verbal instructions, presentations, explanations, requests (both verbal and nonverbal), nonverbal demonstrations, rhetorical questions (to which no answer was expected), and structuring aids (*M* = 73%, *SD* = 6%).

It was followed by teacher questions (of low and high complexity, *M* = 15%, *SD* = 4%), repetition of questions (*M* = 3%, *SD* = 2%), or comprehension questions (*M* = 1%, *SD* = 1%). As response to student answers, teachers mostly gave positive feedback (*M* = 5%, *SD* = 3%) or dealt with student mistakes (*M* = 1%, *SD* = 1%); other types of responses such as giving hints, supporting, giving negative or mixed feedback were seldom observed.

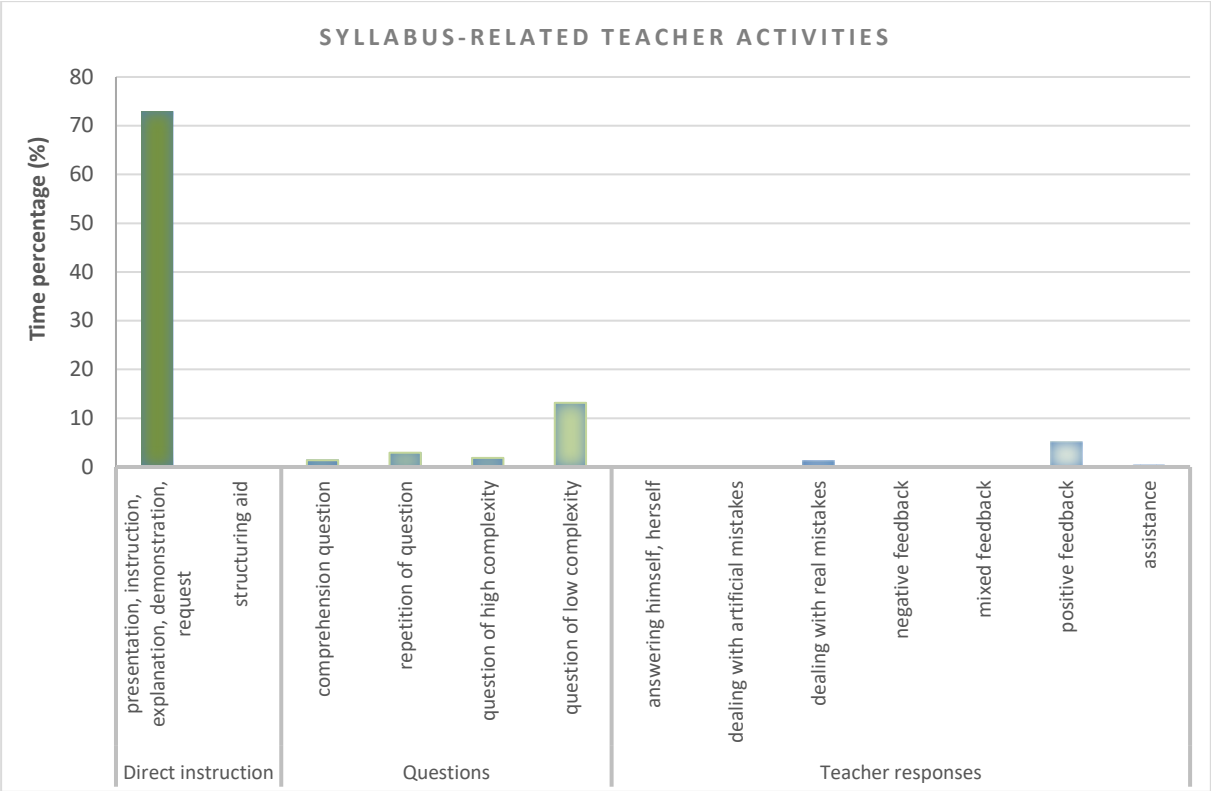

Figure 18: Time percentages of syllabus-related teacher activities

VII.2.7 Lesson monitoring

With regard to lesson monitoring (rating variable, min = 1, max = 4), more than 75% of teachers frequently expected an exact, specific answer or work as a correct answer, managed to obtain it, and tended to ignore or block other student ideas (narrow focused monitoring). Mean rating of this aspect was $M = 2.9$ ($SD = 0.7$).

Another aspect of lesson monitoring is student orientation of instruction – how and how often teachers involve student questions, suggestions, and ideas in lessons. The mean rating of instruction in this regard was $M = 2.2$ ($SD = 0.8$), which indicated that the majority of teachers ($N = 29$) tended to not elaborate on students' concerns.

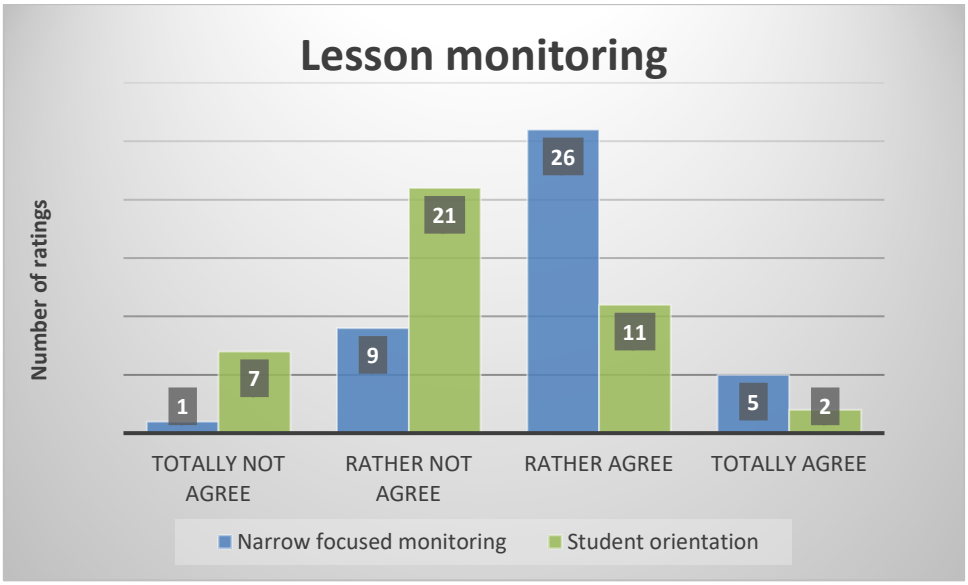

Figure 19: Rating results regarding lesson monitoring

VII.2.8 Variation and adaptivity of the lessons

The variation and adaptivity of lessons was rated low with *M* = 1.4 (*SD* = 0.6, min = 1, max = 4). That means that the raters did not observe the teachers' intention or effort to vary exercises, materials or examples in order to adapt to different students' abilities (see also Appendix F1).

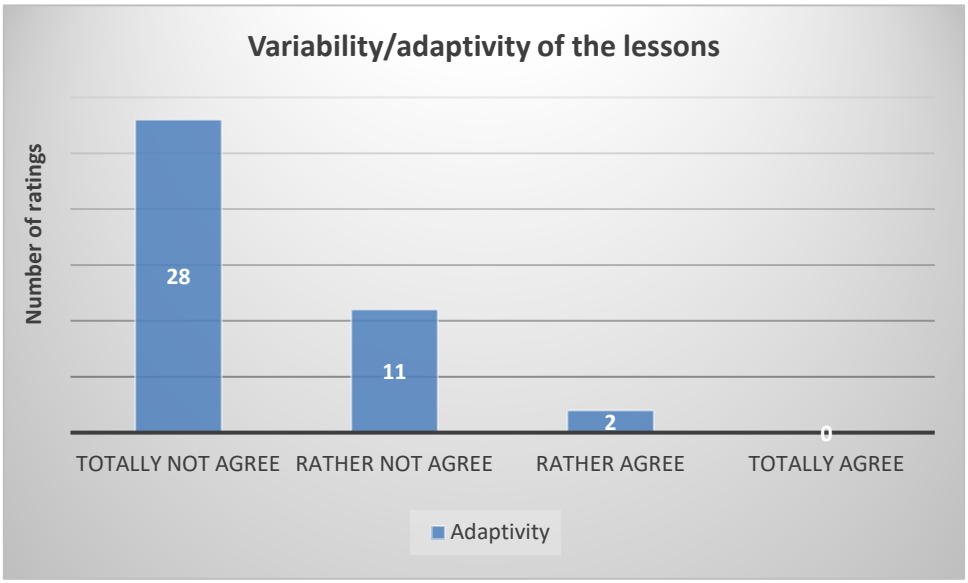

Figure 20: Ratings of the variability, adaptivity of the lessons

VII.2.9 Chapter summary and answer to Research Question 1

The EFL lessons in this study can be regarded as highly task- and teacher-oriented. Lesson time was mostly spent on syllabus-related activities in the form of teacher-centered discussions. Direct instruction dominated in the teacher activities. Teacher speaking time was approximately four times as high as student speaking time. The textbook played a central role in EFL lessons: Topics which were not in the textbook were rarely mentioned or dealt with in lessons. Student suggestions, ideas, and themes which deviated from the prearranged topics or expected answers tended to be ignored or even blocked by the teacher. The amount of time in which students spoke freely was small, around 3.4% of the whole lesson time. Moreover, according to the students (results of the short questionnaire), they normally participated less in a regular lesson than in the recorded lesson.

Hence, the first research question "*Does the study confirm the assumption that EFL lessons in this study are teacher-centered and textbook-driven?*" can be answered with "Yes."

VII.3 Self-regulation competences of the teachers

VII.3.1 Teacher judgment of own speaking time

To answer the second research question ("*Do Vietnamese EFL teachers have good self-regulation competences?*"), a short questionnaire was implemented to analyze the accuracy of teacher judgment of own speaking time in the recorded lessons. Right after the recorded lesson, teachers were asked to estimate in percent their own relative speaking time during the class speaking phases in which either the teacher or the students expressed themselves verbally (100% = whole class speaking time).

While video data suggested that the speaking time of teachers was an average of $M = 77.8\%$ ($SD = 6.7\%$), most teachers judged that their speaking time was between 20% and 40% ($M = 36.8\%$, $SD = 13.5\%$) of the total speaking time (see Figure 21). Ten percent of teachers judged their own speaking time to be 20% and below, while none of them spoke for less than 50% of all class speaking time. This means that teachers strongly underestimated their own speaking time.

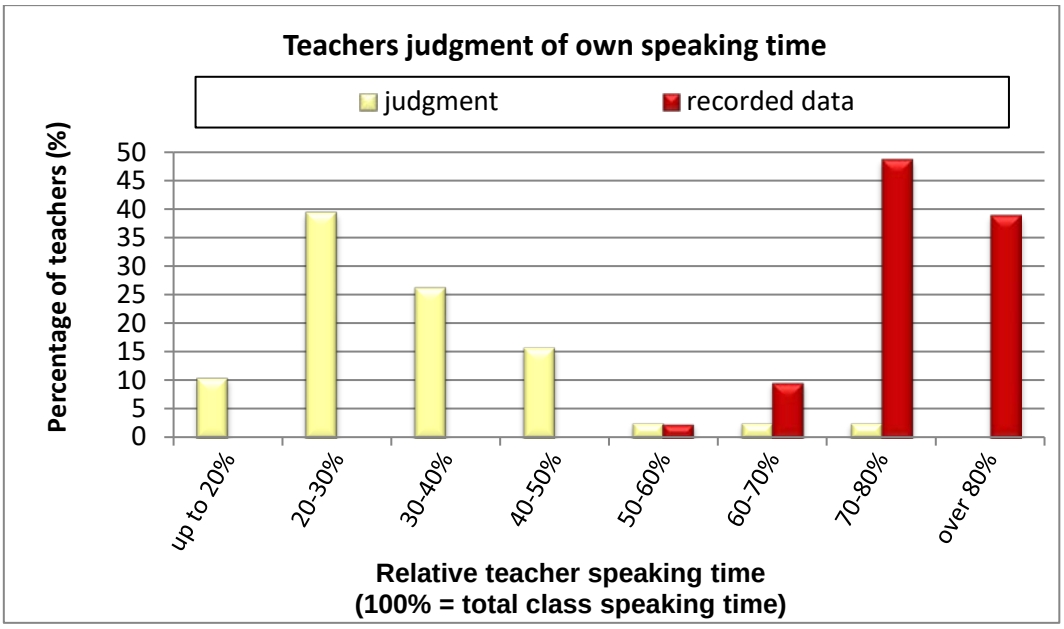

Figure 21: Estimated vs. recorded teacher speaking time

VII.3.2 Chapter summary and answer to Research Question 2

The results regarding the accuracy of teacher judgment of their own speaking time revealed that EFL teachers in this study misjudged their share of speaking time to a large extent. This shows that teachers do not accurately self-reflect while teaching or do not have good self-regulation competences. The second research question "*Do Vietnamese EFL teachers have good self-regulation competencies?*" can thus be answered with "No."

Because this phenomenon was also found among EFL teachers in Germany (c.f. chapter III.1.5), this can be assumed to be a universal characteristic of English teaching and teachers. However, without accurate self-reflection and self-regulation competences, teachers have no reliable basis for developing their own instructional teaching quality and for judging teaching effects on students. A possible conclusion is that professional programs are necessary to help teachers enhance their self-reflection accuracy and self-regulation competencies.

VII.4 Student and teacher speaking mistakes and teacher language

Based on previous findings (see section II.5), speaking mistakes were expected to be a big problem among both students and teachers in this study. In this section, the results of the extra-curricular speaking lesson will be presented with data on the types and frequency of speaking mistakes made by the students and teachers. All of the teacher and student statements in English or mixed languages (in case a few

Vietnamese words were used in an English sentence) were taken into account. All statements in the Vietnamese language, those that were not understandable, or statements made by student groups or the whole class were excluded. For convenience, all selected statements will be referred to as statements *in English*. The following coding categories were used: frequencies with which teachers and student made mistakes (time percentages and relative frequencies), types of mistakes, and types of phonological mistakes.

In addition, the rating results of the curriculum-oriented lesson regarding teacher language have also been taken into account. Based on them, the Research Questions 3 and 4 ("*Does the study confirm the assumption that many EFL teachers are not adequately prepared in terms of their English speaking skills?*", and "*What are the most frequent pronunciation errors of teachers and students in this study?"*) can be answered.

VII.4.1 Frequencies the teachers and students made speaking mistakes

Figure 22 shows the speaking time percentages of teachers and individual students: total speaking time, speaking time in English, and speaking time in English with mistakes.

As expected, the average speaking time of individual students in this lesson (20.8%, see

Figure 22, top bar) was considerably higher than in the curriculum-oriented lesson (11.4%). Most of the time (95.7%), they spoke in English. The teacher spoke in English 82.9% of the speaking time. On average, there were 385 teacher speaking turns in English (per lesson) with a mean duration of 2.8 seconds per expression (*SD* = 0.5s). The average number of individual student speaking turns in English was 112 (*SD* = 51.5, per lesson); the mean duration of each turn was 4.7 seconds (*SD* = 1.4s).

The first coding category was the accuracy of the English statements. Each statement was coded if it was correctly articulated (that means without any mistakes). On average, students made mistakes 52.7% (*SD* = 20.1%) of their speaking time in English, corresponding to 38.4% (*SD* = 18%) of individual student statements in English.

The teachers also made a number of speaking mistakes. On average, the teachers made mistakes 41.6% (*SD* = 12.7%) of their speaking time in English, or in 34.4% (*SD* = 10.6%) of all their English statements. In one lesson, the students made no mistakes, while the teacher made mistakes in 50% of all of her statements in English.

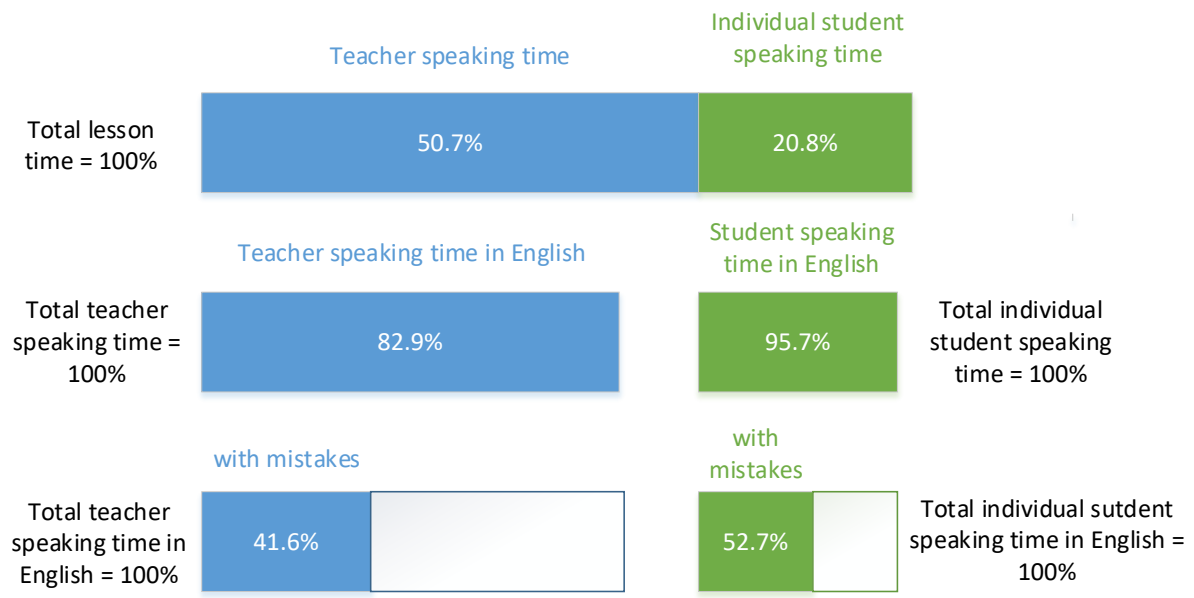

Figure 22: Speaking time in English of teachers and students with mistakes

VII.4.2 Types of mistakes

Next, the types of mistakes made by the teachers and students were analyzed. Five coding categories for mistakes were developed and coded with regard to content, vocabulary, situation/context, grammar, and phonology in cooperation with EFL experts (see Chapter III.2.6). One statement could include more than one mistake.

The coding results (see Figure 23) revealed that the most frequent speaking mistake that both the students and the teachers made was with regard to pronunciation: 83.1% (SD = 16.5%) of all student statements with mistakes included phonological mistakes. The corresponding value for teachers was 85.1% (SD = 8%).

The second most frequent speaking problem of the students was grammar mistakes: 25.6% (SD = 14.2%) of all student statements with mistakes contained grammatical mistakes. Likewise, this was also the second largest speaking problem of the teachers; 19.6% (SD = 8.2%) of all their statements with mistakes included grammatical mistakes.

Furthermore, incorrect vocabulary was found in 5.2% (SD = 6.8%) of all teacher statements with mistakes, and the teachers made mistakes regarding the situation/context of a statement in 4.4%

(SD = 4.3%) of all their statements with mistakes. On the other hand, students did not often make speaking mistakes concerning vocabulary (1.1%) or situation/context (1.7%) on average.

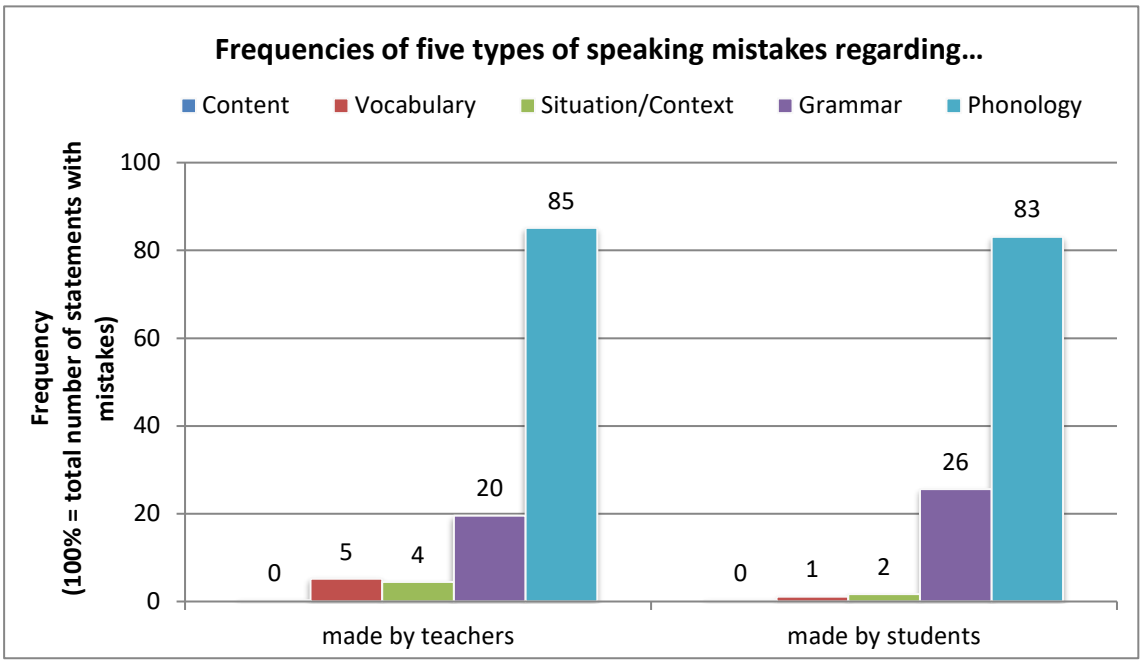

Figure 23: Relative frequency of five types of speaking mistake in speaking-oriented lesson

Both teachers and students least frequently made mistakes in terms of the speaking content (0.3% of all teachers' and students' English statements with mistakes).

VII.4.3 Common types of phonological mistakes

Table 4 shows the relative frequency of each of the twelve coded phonological mistakes made by the teachers and students in this study (see Chapter III.2.6).

The results confirm previous findings about the common mistakes English learners in this study often made (c.f. Chapter II.6).

The most common phonological mistake of both the teachers and students was "skipping final consonant clusters." This problem was found in 50.8% (*SD* = 13.9%) of all teachers' statements with phonological mistakes. The corresponding value for students was 60.3% (*SD* = 16.6%).

The second most frequently made phonological error was relevant to the pronunciation of hard consonants, namely "skipping or softening hard consonants" (teachers: *M* = 46.1%, *SD* = 14.3%, students: *M* = 43.1%, *SD* = 19.1%).

Together, these two errors occurred in one out of two utterances of both teachers and students who made phonological mistakes.

Third, "using incorrect vowels" was found in 14.1% (*SD* = 7.3%) of all teacher and 19.6% (*SD* = 10.2%) of all student statements with phonological mistakes.

Table 4: Relative frequency of phonological error types

	…by teachers		…by students	
Phonological error type	mean (%)	sd (%)	mean (%)	sd (%)
Skipping final consonant clusters	50.8	13.9	60.3	16.6
Skipping or softening hard consonants	46.1	14.3	43.1	19.1
Using incorrect vowels	14.1	7.3	19.6	10.2
Using wrong word stress	7.8	8.9	9.2	9.6
Shortening pronunciation of long vowels	7.3	4.5	10.8	8.2
Adding 's' to word ending	5.2	5.1	4.3	5.9
Rolling 'r'	2.8	4.7	0.3	1.0
Adding 't' to word ending	1.2	2.6	0.1	0.6
Lengthened pronunciation of short vowels	1.0	1.7	0.8	1.8
Speaking in a clipped manner	0.2	0.7	0.9	3.5
Using wrong sentence stress	0.0	0.2	0.1	0.5
Others	3.5	5.2	2.3	5.1

Note: 100% = all faulty expressions with phonological errors

While students and teachers frequently made mistakes at word level (*segmental errors*), they did not often make mistakes at sentence level (*suprasegmental errors*), such as using incorrect sentence stress (teachers: *M* = 0.0%, *SD* = 0.2%, students: *M* = 0.1%, *SD* = 0.5%).

VII.4.4 Teacher language

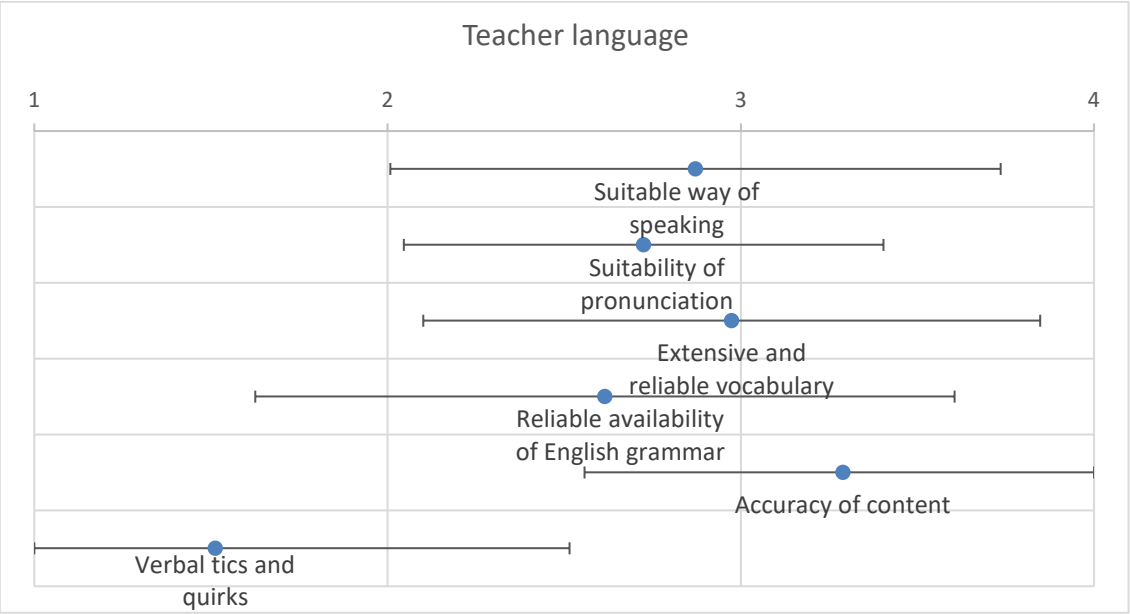

Figure 24: Ratings of teacher language

Note: blue points are mean rating; horizontal lines represent one standard deviations to each side of the means (limited by minimum and maximum rating values)

The rating results of how well teachers expressed themselves in English varied between classes. In general, teachers possessed an extensive and reliable repertoire of English vocabulary (M = 3.3, SD = 0.7) and did not have problems conveying lesson content adequately (Accuracy of content: M = 3.3, SD = 0.7). About one-third of them did not speak suitably (did not use formulations that correspond with the way the English/Americans speak, M = 2.9, SD = 0.9) and had a strong Vietnamese accent when speaking (mean rating of suitability of pronunciation including accent-free pronunciation: M = 2.7, SD = 0.7). Seventeen teachers had observable difficulties in using correct grammar when speaking (M = 2.6, SD = 1.0). Verbal tics and quirks were observed in seven teachers (e.g., frequently saying “OK” without any meaning, frequently repeating words unintentionally); however, this was not a problem for the majority of teachers (M = 1.5, SD = 1.0).

VII.4.5 Chapter summary and answer to Research Questions 3 and 4

The results confirmed the observation of Cunningham (2009) that EFL teachers make speaking mistakes. The teachers in this study made mistakes in more than one-third of their statements and in nearly half of their speaking time in English. They made different types of mistakes, most frequently regarding incorrect grammar and inadequate pronunciation. More than one-third of the teachers had problems with English grammar when speaking, and one-fourth did not have an extensive and reliable

repertoire of English vocabulary. Most of them did not sound like and express themselves like a native speaker, and one-third did not speak suitably. These results underline the fact that a number of EFL teachers in this study were indeed not adequately prepared in terms of their English speaking skills. Thus, the Research Question 3 ("*Does the study confirm the assumption that many EFL teachers are not adequately prepared in terms of their English speaking skills?*") can be answered with "No, many of them are not."

Due to this, they are likely to pass their mistakes onto their students, which is suggested by the positive (and statistically significant) correlations between the relative frequencies teachers and students made grammar mistakes ($r = .32, p = .003$), used incorrect word stress ($r = .31, p = .02$), and skipped or softened hard consonants ($r = .37, p < .001$). A statistically significant correlation between the relative frequencies with which teachers and students made mistakes was not found in general ($r = .11, p = .53$), but only for the sample in Hanoi ($r = .48, p = .006$). The reason might be that, in Hanoi only, an English teacher normally teaches a class from the 6th to the 9th grade (which is not always the case in Ho Chi Minh city and Bac Ninh province). Because the students were taught by one teacher in Hanoi for a longer time, it is more likely that these students adopted the mistakes that the teacher made.

The pronunciation mistakes that teachers and students most often made in this study were at word and sound level, for example, skipping final consonant clusters, skipping or softening hard consonants, using wrong vowels, using incorrect word stress, and shortening the pronunciation of long vowels. Mistakes at sentence level such as using incorrect sentence stress did not tend to be a problem for the teachers and students in this study. According to Phan & Vo (2012), only the incorrect allocation of stress in a sentence contributes significantly to reduced comprehensibility of speaking. Although it is still a controversial issue as to whether it should be a goal of foreign language teaching and learning for learners to sound like a "native speaker", intelligibility is an indisputable standard that learners should strive to achieve (Lightbown & Spada, 2013).

VII.5 General quality dimensions of classroom instruction

In this section, the rating results regarding the important general quality dimensions of classroom instruction will be presented in order to answer Research Question 5: "*In which of the important general quality dimensions of classroom instruction are EFL teachers in this study ranked positively, including classroom management, clarity, structuredness, supportive learning climate, motivation, cognitive activation, and feedback?*" The relevant basic coding indicators such as the use of lesson time (time-on-

task), syllabus-related teacher activities (assistance, teacher feedback) were additionally taken into account.

VII.5.1 Classroom management

Two important indicators of the quality dimension classroom management were described in the Sections VII.2.1 (the use of lesson time) and VII.2.2 (task orientation). The lesson time was mostly used for textbook content but not for other themes or purposes. The rating result was in accordance with this; the average result was *M* = 3.1 (*SD* = 0.8), which means that the teachers were good at managing lesson time and at allocating it to different learning phases, according to the raters. The mean rating of task orientation of all lessons, as mentioned in Section VII.2.2, was nearly as high as the maximum possible rating result (*M* = 3.8, *SD* = 0.4).

Regarding discipline (free of disruptions, disciplinary interventions), the instruction was rated quite high with *M* = 3.4 (*SD* = 0.8).

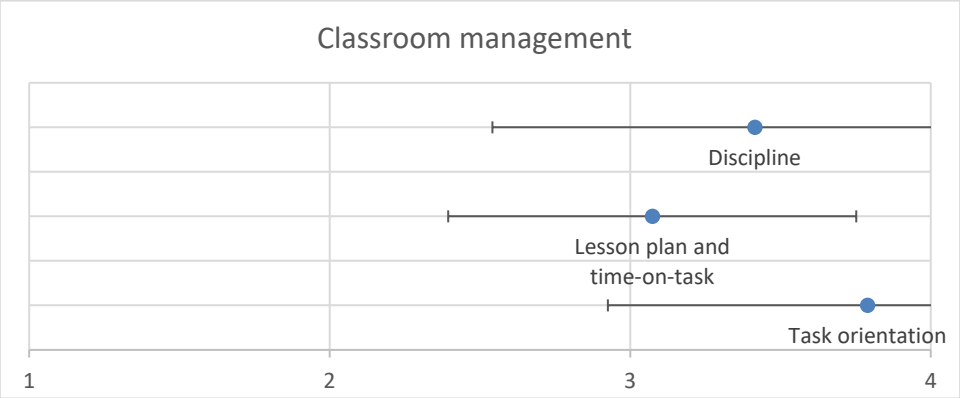

Figure 25: Mean ratings of quality dimension classroom management

VII.5.2 Clarity

Two aspects of clarity as an instructional quality dimension were rated: the clarity/coherence with regard to content (the contents were communicated in a clear and understandable manner; coherent, logical presentation that is easily understandable; presentation focused on key points; suitable examples; clearly understandable questions), and the conciseness of teacher language (clear diction, full, clear, well-planned sentences, cohesive and fluent formulations; clear articulation). On a 4-point rating scale ranging from 1 – totally not agree to 4 – totally agree, the mean rating of clarity/coherence with regard to content was *M* = 3.5 (*SD* = 0.7) and the mean rating of the clarity and conciseness of the teacher language was *M* = 3.0 (*SD* = 0.8).

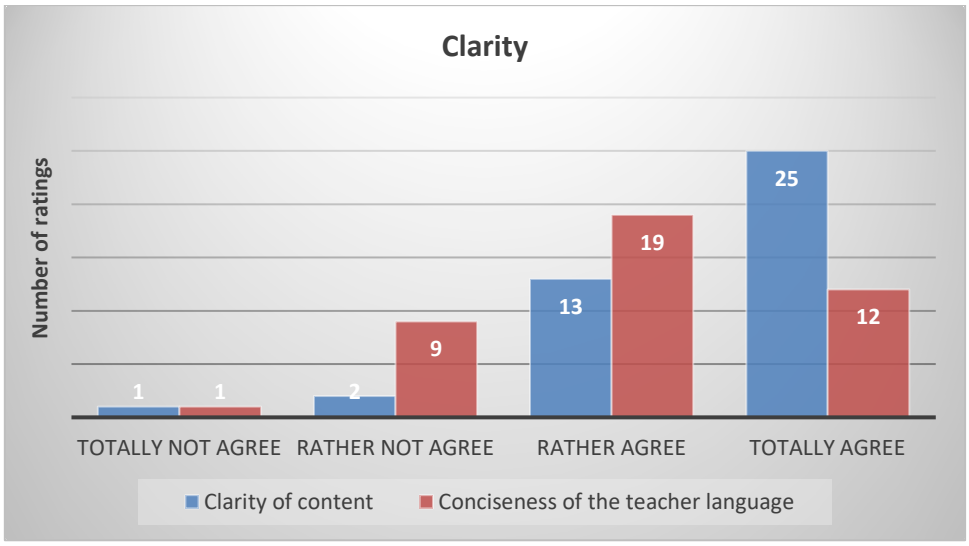

Figure 26: Ratings of clarity of the lessons

VII.5.3 Structuredness

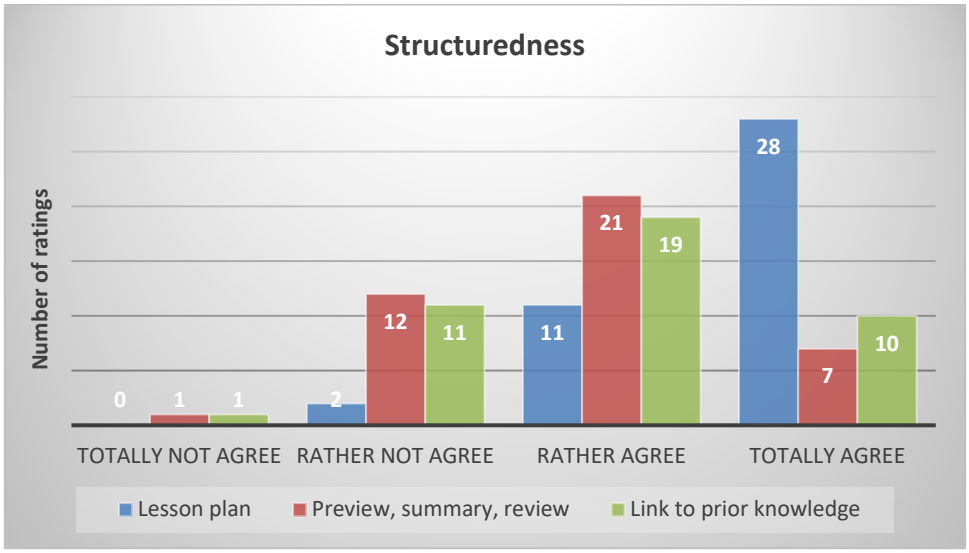

Figure 27: Ratings of the structuredness of the lessons

The structuredness of instruction refers to the degree to which information, content, skills, and goals of instruction are organized, arranged, and presented so that students can easily understand, learn, and achieve the learning goals. The lesson plan of most lessons was rated as highly structured (systematic, logical, with evident main topic, $M = 3.6$, $SD = 0.7$). However, because the lesson plan was provided and described in detail in the textbook, this rating result was actually less relevant to the quality of individual instruction.

Another aspect of the structuredness of the lesson was rated: teacher effort to give structure to the lesson by including previewing (at the beginning), summarizing (at the end), reviewing, or highlighting important contents. In this regard, the average rating of lessons was *M* = 2.8 (*SD* = 0.7).

The third aspect that was rated was related to the effort of teachers to create links between the current topic and prior knowledge or future topics. In this respect, the mean rating of the classes was *M* = 2.9 (*SD* = 0.8).

VII.5.4 Supportive classroom climate

Positive interpersonal relations and a positive mistake-making environment are among the key factors of a supportive classroom climate which were rated in this study.

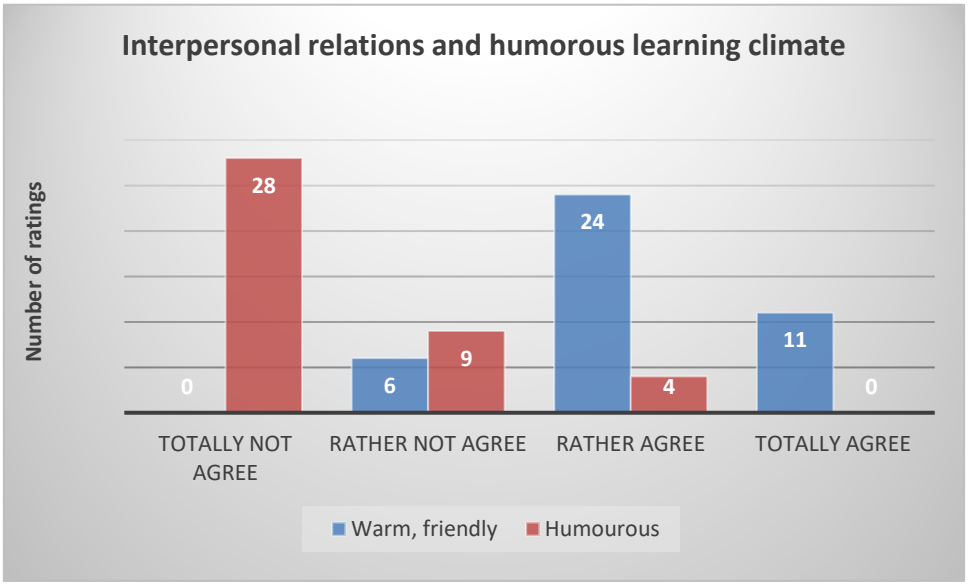

Figure 28: Ratings of interpersonal relations and humorous learning climate

The learning climate and the relationship between the teacher and students were positively judged in most classes. The teachers were rated as mainly friendly and encouraging with regard to their relationship with students and interested in the students' personal and private matters. The students treated the teachers with respect, and the relationships between the teachers and students were mainly rated as warm and trust-based (*M* = 3.1, *SD* = 0.6). However, none of the teachers tended toward having fun or making jokes in lessons; many of them maintained a serious attitude while teaching, and none of the lessons were really humorous (*M* = 1.4, *SD* = 0.7).

The way teachers treated mistakes was rated as constructive (*M* = 3.3, *SD* = 0.5). The students also showed no negative attitudes or did not react negatively toward mistakes. In most classes (*N* = 33), the

students stayed affectively neutral when someone made a mistake (rating category 3). The mean rating of students' attitudes toward mistakes was *M* = 3.1 (*SD* = 0.4).

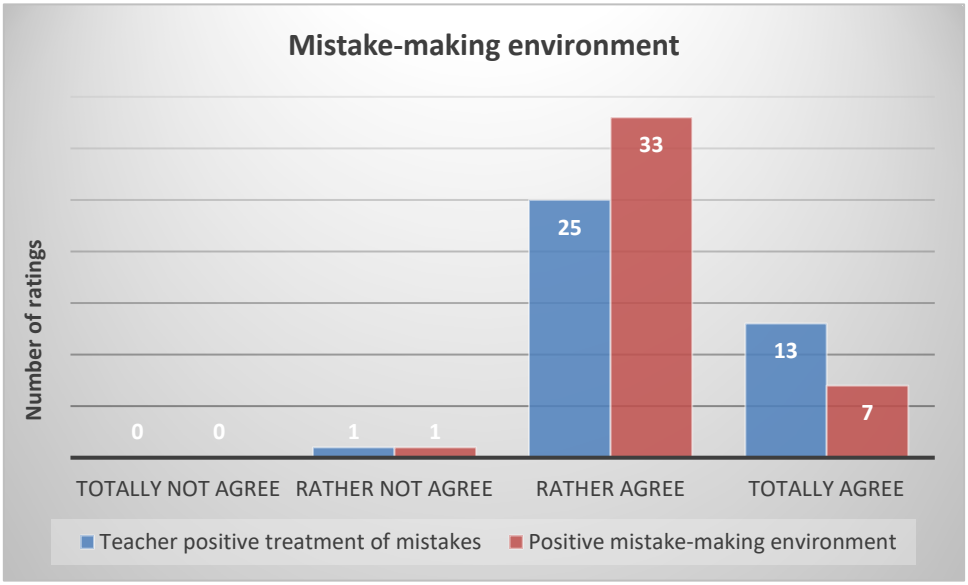

Figure 29: Ratings of the mistake-making environment

VII.5.5 Quality of motivation

In most lessons, both the teachers and students appeared to be stimulating, energetic, and active (mean rating of teacher commitment and enthusiasm: *M* = 3.0, *SD* = 0.6; mean rating of student commitment: *M* = 3.2, *SD* = 0.6). In terms of teachers encouraging and stimulating student statements, the mean rating of the teachers was *M* = 2.8 (*SD* = 0.8). Teachers differed more in terms of the quality of their motivation. No or only a few authentic (relevant to the every-day life of students) examples or exercises were used in the 27 lessons. The mean rating of the authenticity and real-life relevance of the lessons was *M* = 2.3 (*SD* = 0.8).

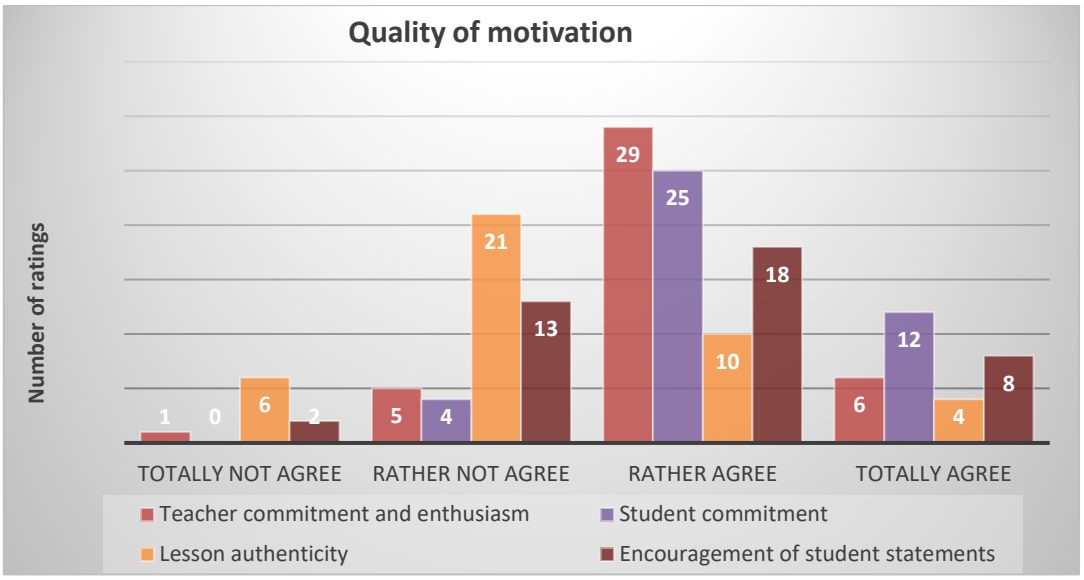

Figure 30: Ratings of dimension quality of motivation

VII.5.6 Cognitively activating instruction

Two rating variables were relevant to the quality dimensions cognitively activating instruction (stimulation of students' insightful learning): fostering learning and thinking strategies as well as proposal of demanding, advanced topics.

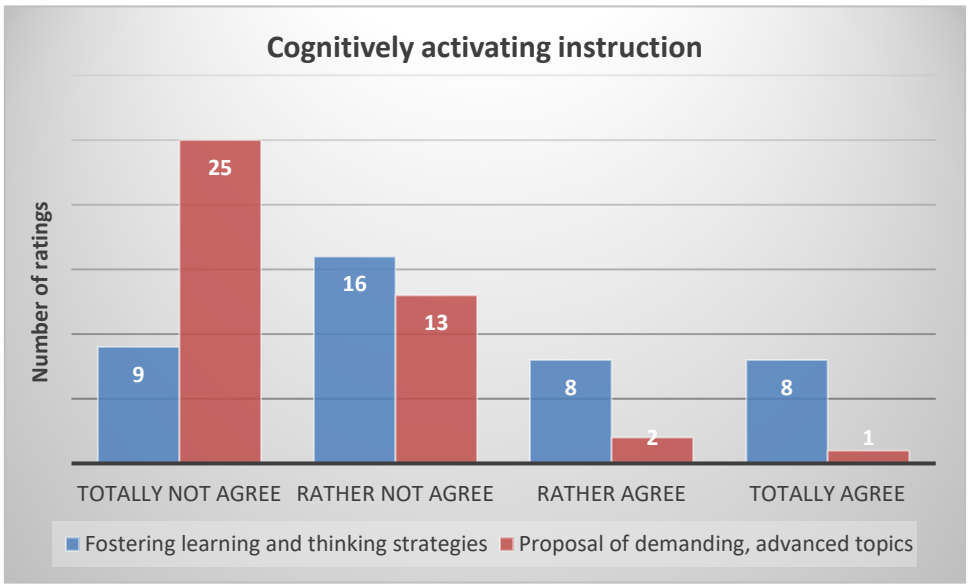

Figure 31: Ratings of the dimension cognitively activating instruction

Teachers varied strongly with regard to the frequency and number of suggestions concerning learning and thinking strategies, such as for learning and practicing vocabulary. Nine teachers never made any suggestion in this regard (rating = 1), while eight others did this quite often, each time making multiple

suggestions (rating = 4). Ratings regarding this aspect were an average of *M* = 2.4 with a standard deviation *SD* = 1.0. On the other hand, advanced topics were never proposed in the 25 recorded lessons, and not often worked on in all recorded lessons, except for in one (*M* = 1.5, *SD* = 0.7).

VII.5.7 Feedback

On average, feedback accounted for 10% (SD = 4%) of all teacher turns based on the basic coding results with regard to syllabus-related teacher activities. The teachers gave feedback to one out of ten student statements in English (M = 10%, SD = 4%); most feedback was positive.

Indeed, 95% of all teacher feedback was positive (*SD* = 6%). Positive feedback was usually given verbally without an emphasized affective affirmation such as "OK" or "correct." Positive feedback with an emphasized affective affirmation (such as "very good") amounted to 9% (SD = 10%) of all positive feedback.

After a student answer or statement, if a teacher was not sastisfied with it, he/she did not often give negative feedback, but rather called upon another student, gave hints, or answered the question himself/herself. Only 3% (*SD* = 5%) of all teacher feedback was negative. The remaining 1% was made up of mixed feedback. Eighteen teachers never gave negative feedback during the recorded lesson. Negative feedback with emphasized affective affirmation was given only once by one teacher.

VII.5.8 Chapter summary and answer to Research Question 5

The results in this study showed that the teachers managed lessons well, implementing effective lesson plans and using time well. The lessons were extremely task-oriented with very little time used for social interactions and procedural matters, with few and small disturbances. Teachers were therefore considered good at classroom management. However, a regular lesson would be noisier according to the results of the student short questionnaire.

The clarity of the content and the conciseness of the teacher language were also rated positively. Actually, because the lessons were highly textbook-driven, the clarity of lesson content can be partially attributed to the textbook quality.

Likewise, the structuredness of the lessons was rated very positively, which, however, cannot be attributed to instructional quality alone but also to the quality of the textbook. The lessons differed to some extend with regard to how the lesson content was structured with a preview, summary, review, and highlighting and whether connections were made to prior knowledge and topics were successfully linked.

Both teachers and students showed a high level of commitment in lessons. Teachers made an effort to encourage students to make statements. However, they often expected a specific response or answer, and ignored other ideas of students. In most lessons, no explicit signs were found that teachers tried to supply students with varying competence levels with different exercises, materials, tasks, or questions. Furthermore, there were no signs that teachers intended to elaborate on learning difficulties or propose advanced topics for high performers. There were no examples with relevance to the everyday-life of the students, and the authenticity of the lessons was rated negatively. All in all, the quality of motivation of the instruction can be seen as rather positive, while individual support, student orientation, and cognitive activation were rated less positively.

A friendly and warm classroom climate was observed in most lessons, although most teachers lacked humor.

Feedback was often given by teachers, and it was mostly positive. If they were not satisfied with students' statements, teachers often gave hints, or called upon other students rather than give negative feedback. Feedback was usually given with a neutral attitude. Feedback with emphasized affective affirmation was quite rare.

In general, the EFL teachers were rated positively regarding important quality dimensions of classroom instruction, especially classroom management, clarity, structuredness, supportive learning climate, and feedback. Dimensions in which teachers had rather low ratings were individual support, student orientation, and cognitive activation.

VII.6 Quality dimensions of effective EFL teaching

Some of the most important instructional quality dimensions specific to the school subject EFL are: engaging students in communication, equal focus on accuracy, fluency, and formulaic language, and dealing with mistakes. Based on both basic coding and rating indicators, the quality of these dimensions will be presented in this section in order to answer Research Question 6 *("Are teachers in this study rated positively regarding important quality dimensions of effective EFL teaching, including engaging students in communication, equal focus on accuracy, fluency, and formulaic language, and dealing with mistakes?"*). In addition, data from the teacher questionnaire regarding dealing with student mistakes were included; without these additional data it would not have been possible to interpret the video results.

VII.6.1 Engaging students in communication

Involvement of as many students as possible (M = 2.9, SD = 0.7) was judged as a salient lesson goal by raters in almost all lessons. At least 25% of students (rating = 2) in all classes, at least 50% of students in 29 classes (70%, rating = 3), and more than 75% (rating = 4) of students in four classes were called upon by the teacher or contributed verbally at least once in the lesson. Rating Category 1 (less than one-fourth of all students were called upon by the teacher in the lesson) was not assigned by any rater in any class. Rating Category 2 was assigned only to classes with at least 39 students. This result was in accordance with the positive rating result for the encouragement of student statements by teachers (see Section VII.5.5).

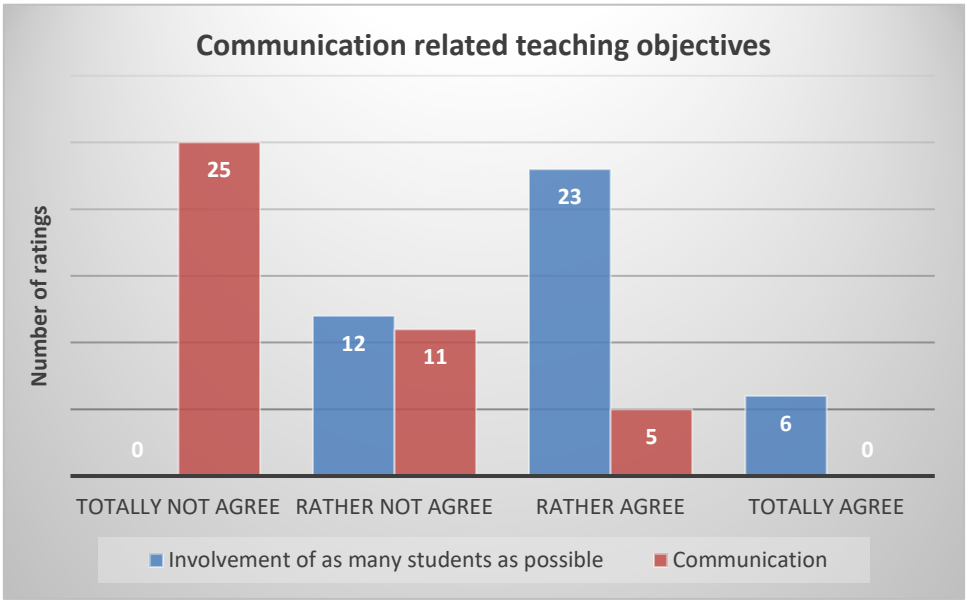

Figure 32: Ratings of communication-related teaching objectives

On the other hand, as presented in Section VII.2.4, the total student speaking time in the curriculum-oriented lesson was 15%, while teachers spoke in 53% of the total lesson time. Furthermore, a major part of the lessons was teacher-centered discussions. Unsurprisingly, communication was not rated as a teaching objective (M = 1.5, SD = 0.7, min = 1, max = 4). There were hardly any discussions in which the teacher and the students communicated with each other in an every-day manner and in which the teacher behaved more like a communication partner rather than playing the role of the teacher.

VII.6.2 Equal focus on accuracy and fluency

Throughout the recorded lesson, teachers focused more on the *fluency* (M = 3.2, SD = 0.7) than the *accuracy* of students' verbal contributions (M = 2.5, SD = 0.6), based on the rating data. In most lessons,

a good flow of communication with only a few, rather short interruptions was observed. The teachers only occasionally addressed mistakes. The results of the next section shed more light on the treatment of mistakes.

VII.6.3 Dealing with mistakes

As reported in Section VII.4, 38% of all individual student statements (SD = 18%) and 34% of all teacher statements in English (*SD* = 11%) in extra-curricular speaking lessons contained mistakes. Sometimes, teachers made the same mistakes as students. Sometimes, the teachers' corrections also contained mistakes. Occasionally, teachers corrected or hinted at a "mistake" in a student statement which was actually correct. In one lesson, the teacher corrected students' expressions 28 times while students made mistakes only nine times (311%). On average, 16% of individual student expressions with mistakes (SD = 16%) were dealt with by teachers. In seven classes, teachers never handled students' mistakes.

Unfortunately, it was not possible to distinguish between when teachers intentionally did not handle a student's mistake and when they did not recognize the mistake, because they also made mistakes.

According to the results of the teacher questionnaire, it cannot be assumed that teachers would have intentionally ignored students' mistakes (see Table 5). Based on teacher self-judgments, they usually corrected students' mistakes of all types (especially regarding grammar, pronunciation, and vocabulary) with explanations.

Table 5: Frequencies of dealing with different student error types

	mean	**sd**
Frequency of dealing with pronunciation mistakes		
Ignoring	1.9	1.0
Correction without explanation	2.1	0.9
Correction with explanation	**3.2**	0.7
Correction by students themselves	1.8	0.9
Correction by another student	2.5	0.7
Using the corrected form as a part of the answer	2.7	0.9
Frequency of dealing with grammatical mistakes		
Ignoring	1.6	0.9
Correction without explanation	1.4	0.6
Correction with explanation	**3.5**	0.7
Correction by students themselves	2.2	0.9
Correction by another student	2.8	0.9
Using the corrected form as a part of the answer	2.9	0.9

	mean	sd
Correction frequency in learning phases with focus on vocabulary/grammar		
grammatical mistake	**3.7**	0.5
pronunciation mistake	**3.6**	0.5
misunderstanding a word	**3.5**	0.6
error regarding the content	**3.3**	0.6
using improper words regarding the context	3.0	0.8
using improper words in oral communication	3.0	0.8
Correction frequency in learning phases with focus on communication		
grammatical mistake	2.9	0.9
pronunciation mistake	**3.4**	0.6
misunderstanding a word	3.2	0.7
error regarding the content	3.1	0.7
using improper words regarding the context	3.0	0.7
using improper words in oral communication	3.0	0.8

Note: the answers range from 1 = "never" to 4 = "usually"

Based on the basic coding results of teacher responses (e.g., giving clues, suggestions, corrections) in the extra-curricular speaking lesson, when it was obvious that a teacher found a student statement to be incorrect, teachers mainly corrected the mistake themselves (M = 86%, SD = 21%) or hinted at the mistakes (M = 10%, SD = 16%). Further explanations regarding the mistakes were given quite infrequently (M = 1%, SD = 3%). Students were rarely given the chance to find (M = 3%, SD = 8%) or to correct (M = 1%, SD = 5%) their own mistakes or the mistakes of other students. It was never observed that the teachers asked students to explain the mistakes.

In the curriculum-oriented lessons, when teachers handled students' mistakes (relative frequency), they most often corrected phonological mistakes (49%, SD = 33%), 24% (SD = 28%), followed by grammar mistakes, 9% (SD = 24%), content-related mistakes, 7% (SD = 13%), vocabulary mistakes, and 6% (SD = 14%) situation/context-related mistakes; 2% (SD = 7%) of corrections were related to different kinds of mistakes simultaneously. For 2% (SD = 8%) of all teacher corrections, it was unclear which kind of mistakes they were handling.

VII.6.4 Chapter summary and answer to Research Question 6

In all classes, at least 25% of students made statements in the lessons. On the other hand, the content of the discussions, dialogues, and conversations was mainly teacher- and task-centered, and did not integrate any elements of every-day language. In the process, the teachers paid more attention to fluency than to the accuracy of student statements. Student mistakes were often not handled. However, taking

into account the fact that teachers frequently made mistakes and had limited English skills themselves, it is assumed that they were not able to recognize the student mistakes rather than intentionally refraining from handling student mistakes.

With regard to the quality dimensions specific for teaching EFL, the results in this study showed a differentiated view of classroom instruction: a rather positive view of structure at first sight (easily observable indicators), but negative when the focus shifted to deep structural features of instruction.

VIII. Academic student outcomes and socioeconomic background

In this chapter, descriptive statistics of student achievement and growth as well as of socioeconomic background will be reported first, based on multiply imputed data and the PVs of all 2,096 students of the 50 participating classes. In the sample, 56.6% of the students were girls; 53.4% were boys. The class size ranged from 27 to 59 students per class with a mean class size of $\bar{N}$ = 41.9. Next, the results regarding the relationships between initial student ability and social background, on the one hand, and student achievement and growth, on the other hand, will be presented. Based on these results, the Research Questions 7–9 related to the effects of contextual factors on student achievement and growth at individual and class level will be answered.

VIII.1 Descriptive statistics of student achievement and growth

For each test and scaling model, the test scores (PVs) of all students at T1 across 10 imputed datasets were rescaled to produce a pooled mean of $M = 0$ and a pooled standard deviation of $SD = 1$. The individual student test scores at T2 were transformed accordingly. Student growth was measured as the difference between T2 and T1. In addition, the growth in effect size Cohen's d (Cohen, 1988) was calculated and will be reported:

$$d = \frac{M_{\mathrm{T2}} - M_{\mathrm{T1}}}{\sqrt{\frac{SD_{\mathrm{T1}}^2 + SD_{\mathrm{T2}}^2}{2}}},$$

M_{T1} and SD_{T1} are the mean and standard deviation of student test achievement at T1; M_{T2} and SD_{T2} are the mean and standard deviation of student achievement at T2. Estimates at class level were the class mean values.

VIII.1.1 Individual level

In each test, all scaling models yielded similar means (*M*) and standard deviations (*SD*) regarding student achievement at T2. Over one school year, growth in individual student test scores was on average $\bar{d}$ = .52 (*SE* = .08) in the C-test (ranging from .47 to .62 depending on the scaling model) and $\bar{d}$ = .40 (*SE* = .06) in the LC-test (ranging from .37 to .44 depending on the scaling model).

In the C-test, the testlet models (Rasch testlet model M1T or 2PL testlet model M2T) yielded slightly higher growth estimates (M1T: *d* = .62, M2T: *d* = .51) than those of the unidimensional scaling models (Rasch model M1: *d* = .47, unidimensional 2PL model M2: *d* =.49). Taking into account the standard error of the growth estimates, we found no statistically significant differences between models in this regard.

Table 6: Descriptive statistics of the test scores of both tests at both MPs

Scaling model		**M1**		**M2**		**M1T**		**M2T**		**M3**		**Average**	
Test results		***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***			***M (SE)***	***SD (SE)***
C	T1	0	1	0	1	0	1	0	1				
	T2	.47	.98	.49	1.00	.63	1.02	.51	.98			.52 (.12)	.99 (.06)
	Growth estimate	***d***	***SE***	***d***	***SE***	***d***	***SE***	***d***	***SE***			***d***	***SE***
		.47	.04	.49	.04	.62	.07	.51	.05			.52	.08
		M	***SD***	***M***	***SD***					***M***	***SD***	***M (SE)***	***SD (SE)***
LC	T1	0	1	0	1					0	1		
	T2	.46	1.08	.43	1.20					.40	1.13	.43 (.12)	1.13 (.09)
	Growth estimate	***d***	***SE***	***d***	***SE***					***d***	***SE***	***d***	***SE***
		.44	.05	.39	.04					.37	.04	.40	.06

Note: C = C-test, LC = listening comprehension test. M1 = Rasch model, M2 = unidimensional 2PL model, M1T = testlet 1PL model, M2T = testlet 2PL model, M3 = unidimensional 3PL model. Average = average results of all scaling models. T1 = first MP, T2 = second MP. *M* = mean, *SD* = standard deviation, *d* = growth estimated in effect size Cohen's *d*, *SE* = standard error.

In the LC-test, the growth estimate was highest based on the Rasch model (M1: *d* = .44) and lowest based on the unidimensional 3PL model (M3: *d* = .37). The results of all three scaling models did not differ significantly from each other. An increase in student variance at T2 in the LC-test over time (student variance at T1 was set to 1; student variance at T2 was *SD* = 1.08–1.20 depending on the scaling model) was observed, but it was not statistically significant.

VIII.1.2 Class level

At class level, large differences between classes were found. The intraclass correlation ICC (Raudenbush & Bryk, 2002, see Chapter VI.3.1) ranged from .48 to .66 in the C-test and from .62 to .68 in the LC-test at each MP (Table 7). That means that the differences between classes accounted for 48% to 68% of total student variance at each MP in each test.

A high ICC and large class size together led to a high reliability of the class mean for student achievement. The reliability of mean achievement of the classes' ICC(2) (Snijders & Bosker, 2012, see Chapter VI.3.1) ranged from .96 to .99 (see Table 7).

Table 7: Reliability of the aggregated test scores at class level

Test	Scaling model	MP	var_w	var_b	ICC	min *N*	max *N*	mean *N*	min ICC(2)	max ICC(2)	mean ICC(2)
C	M1	T1	.52	.47	.48	27	59	41.9	.96	.98	.97
	M2	T1	.51	.48	.48	27	59	41.9	.96	.98	.98
	M1T	T1	.39	.61	.61	27	59	41.9	.98	.99	.99
	M2T	T1	.34	.64	.66	27	59	41.9	.98	.99	.99
	M1	T2	.48	.46	.49	27	59	41.9	.96	.98	.98
	M2	T2	.51	.49	.49	27	59	41.9	.96	.98	.98
	M1T	T2	.37	.66	.64	27	59	41.9	.98	.99	.99
	M2T	T2	.33	.63	.66	27	59	41.9	.98	.99	.99
LC	M1	T1	.35	.60	.63	27	59	41.9	.98	.99	.99
	M2	T1	.35	.63	.64	27	59	41.9	.98	.99	.99
	M3	T1	.35	.64	.65	27	59	41.9	.98	.99	.99
	M1	T2	.38	.80	.68	27	59	41.9	.98	.99	.99
	M2	T2	.55	.90	.62	27	59	41.9	.98	.99	.99
	M3	T2	.45	.83	.65	27	59	41.9	.98	.99	.99

Note: C = C-test, LC = listening comprehension test, M1 = Rasch model, M2 = unidimensional 2PL model, M1T = testlet 1PL model, M2T = testlet 2PL model, M3 = unidimensional 3PL model. var_w = variance within classes, var_b = variance between classes, min *N* = smallest class size, max *N* = largest class size, mean *N* = average class size, ICC(2) = reliability of class mean achievement, min ICC(2) = ICC(2) by smallest class size, max ICC(2) = ICC(2) by largest class size, mean ICC(2) = ICC(2) by average class size.

Estimates of student variance *within* and *between* classes in the C-test varied according to the scaling model (see Table 7, columns 3, 4). The results of the unidimensional scaling models were similar to each other and differed from those of the testlet models, which were also similar to each other. The results of the testlet models (M1T, M2T) showed lower student variance *within* classes and higher variance *between* classes at both MPs in the C-test. In the LC-test, the results of the three unidimensional models M1, M2, and M3 were similar in this regard.

The differences in the growth estimates generated by the scaling models were observed (see Table 8). In the C-test, two unidimensional models (M1, M2) yielded quite similar results. The same holds true for the two testlet models. The unidimensional models and the testlet models differed in terms of the variance estimates of student growth and ICC. The unidimensional models generated higher variance estimates of growth in the C-test (σ^2 = .36) than the testlet models (σ^2 = .20 – .23).

The variance estimates of student growth within classes generated by the unidimensional models (var_w = .31) were about six times as high as the variance between classes (var_b = .05, ICC = .15) and higher than the variance estimates within classes generated on the basis of the testlet models (var_w = .12 – .13). The testlet models resulted in larger variance estimates of student growth between classes (var_b = .08 – .10) than those generated by the unidimensional models and amounted to more than 40% of total variance regarding growth (ICC > .40). This resulted in a higher reliability of class mean growth based on the testlet models (ICC(2) = .97) than that based on the unidimensional models (ICC(2) = .88).

Table 8: Descriptive statistics of the growth estimates of both tests at both MPs

Test	Scaling model	*M*	*SE*	*SD*	ICC	var_w	var_b	min ICC(2)	max ICC(2)	mean ICC(2)
C	M1	.47	.04	.60	.15	.31	.05	.83	.91	.88
	M2	.49	.04	.61	.15	.31	.05	.82	.91	.88
	M1T	.63	.07	.48	.44	.13	.10	.96	.98	.97
	M2T	.51	.05	.45	.41	.12	.08	.95	.98	.97
LC	M1	.46	.05	.68	.36	.30	.17	.94	.97	.96
	M2	.43	.05	.73	.28	.39	.15	.91	.96	.94
	M3	.39	.05	.77	.24	.46	.14	.89	.95	.93

Note: C = C-test, LC = listening comprehension test, M1 = Rasch model, M2 = unidimensional 2PL model, M1T = testlet 1PL model, M2T = testlet 2PL model, M3 = unidimensional 3PL model. *M* = mean growth, *SE* = bootstrapped standard error of the mean, *SD* = standard deviation, ICC = intraclass correlation, var_w = variance within classes, var_b = variance between classes, ICC(2) = reliability of the growth estimates at class level, min ICC(2) = ICC(2) by smallest class size, max ICC(2) = ICC(2) by largest class size, mean ICC(2) = ICC(2) by average class size.

In the LC-test, the variance estimates of growth based on three scaling models were similar to each other. The total variance estimates of student growth were higher than in the C-test ($\sigma^2 = .47 - .60$), and ICC ranged from .24 (M3) to .36 (M1), resulting in ICC(2) values ranging from .93 to .96, respectively.

The growth estimates at class level in Cohen's *d* ranged from .63 to .77 in the C-test (see Table 9) and ranged from .44 to .53 in the LC-test.

Table 9: Growth estimates at class level

Test	Scaling model		*d*	*SE*
C-test	M1	Rasch 1PL	.67	.06
	M2	Unidimensional 2PL	.68	.06
	M1B	Testlet 1PL	.77	.08
	M2B	Testlet 2PL	.63	.06
LC-test	M1	Rasch 1PL	.53	.07
	M2	Unidimensional 2PL	.48	.05
	M3	Unidimensional 3PL	.44	.06

Note: *d* = Cohen's d, *SE* = standard error of *d*.

VIII.1.3 Relationship between test results of the C-test and LC-test

The Rasch model (M1) and the unidimensional 2PL model (M2) were used to scale student scores in both tests. The results based on each of these two scaling models showed a strong positive correlation between student achievement in two tests at each MP at both individual level ($r \geq .76, p < .001$) and class level ($r \geq .90, p < .001$, see Table 10).

Table 10: Correlations between student achievement in two tests based on the same scaling model

Level	Scaling model	MP	*r*	*SE*	*p*
Individual	Rasch	T1	.76	.03	.000
		T2	.78	.02	.000
	Unidimensional 2PL	T1	.78	.03	.000
		T2	.78	.02	.000
Class	Rasch	T1	.90	.03	.000
		T2	.91	.02	.000
	Unidimensional 2PL	T1	.91	.02	.000
		T2	.92	.02	.000

Note: *r* = correlation between student achievement in both tests based on the same scaling model, *SE* = standard error of *r*, *p* = significant level.

Student growth as measured in the two tests also correlated positively, but with smaller correlation coefficients at individual level ($r \leq .18, p < .01$, see Table 11). At class level, the correlation between student growth in the C-test and LC-test based on the Rasch model was not statistically significant ($r = .21, p = .10$); with the unidimensional 2PL model, the corresponding result was $r = .26$ ($p = .05$).

Table 11: Correlations between student growth in two tests

Level	Scaling model	*r*	*SE*	*p*
Individual	Rasch	.15	.05	.002
Class		.21	.13	.101
Individual	Unidimensional 2PL	.18	.05	.000

Class		.26	.13	.048

Note: *r* = correlation between student achievement in both tests based on the same scaling model, *SE* = standard error of *r*, *p* = significant level.

VIII.2 Relationship between initial student achievement and student outcomes

VIII.2.1 Individual level

At individual level, the initial student achievement correlated positively with student achievement at T2 ($r \geq .74$, $p < .001$, see Table 12). In the C-test, it explained 73% of posttest variance on average ($R^2 = .67$ to .80 depending on the scaling model). In the LC-test, 60% of student variance in the posttest was explained by student achievement in the pretest (R^2 ranged from .55 to .63 according to the scaling model). The local effect Cohen's f^2 (see Chapter VI.3.3.2) of initial student achievement on student achievement at T2 was high in both tests ($f^2 \geq .35$).

Table 12: Correlations between student achievement at T1 and student achievement at T2 and growth (individual level)

		Student achievement at T2 as outcome					Student growth estimates as outcome				
Test	Scaling model	*r*	*SE*	*p*	R^2	f^2	*r*	*SE*	*p*	R^2	f^2
C	M1	.82	.02	.00	.67	2.0	–.34	.05	.00	.11	.13
C	M2	.82	.02	.00	.67	2.0	–.30	.05	.00	.09	.10
C	M1T	.89	.02	.00	.79	3.7	–.19	.14	.17	.05	.05
C	M2T	.89	.02	.00	.80	4.0	–.26	.09	.00	.07	.07
LC	M1	.79	.03	.00	.62	1.7	–.20	.07	.00	.04	.04
LC	M2	.80	.03	.00	.63	1.7	–.07	.06	.22	.01	.01
LC	M3	.74	.03	.00	.55	1.2	–.21	.06	.00	.04	.05

Note: C = C-test, LC = listening comprehension test, M1 = Rasch model, M2 = unidimensional 2PL model, M1T = testlet 1PL model, M2T = testlet 2PL model, M3 = unidimensional 3PL model. *r* = correlation between initial student achievement at T1 and outcome, *SE* = standard error of *r*, *p* = significance level of *r*, R^2 = proportion of outcome variance explained by initial student achievement, f^2 = local effect Cohen's f^2.

Considering individual student growth as the outcome variable, we found negative correlations between initial student achievement and student outcome. The correlation coefficients varied between –.34 and –.19 in the C-test and between –.21 and –.07 in the LC-test (see Table 12). The correlations based on ability estimates generated from the testlet 1PL model (M1T) in the C-test and from the unidimensional 2PL model (M2) in the LC-test were not statistically significant ($p > .05$); the results based on the ability estimates generated with other scaling models were statistically significant.

On average, 9% of variance in student growth in the C-test was explained by the pretest scores (R^2 varied from .05 to .13, depending on the scaling model). In the LC-test, this variance was 3% (R^2 varied from .01 to .05, depending on the scaling model). The average correlation coefficients between student achievement at T1 and growth were $\bar{r}$ = –.27 (*SE* = .10, *p* = .01) in the C-test and $\bar{r}$ = –.16 (*SE* = .09, *p* = .07) in the LC-test. The local effect Cohen's f^2 of initial student achievement on growth in both tests was small (average results: $.02 \leq f^2 < .15$).

VIII.2.2 Class level

Similar results were found for the relationships between initial achievement and student outcomes at class level (see Table 13). The test results at both MPs correlated positively and strongly with each other for each scaling model ($r \geq .89$, $p < .001$) in each test. Initial class achievement had large local effects Cohen's f^2 on class achievement at the posttest ($f^2 > .35$).

Table 13: Level 2 correlations between student achievement at the two MPs and between T1 and growth

		Correlation between class achievement at T1 and T2					Correlation between class achievement at T1 and growth				
Test	Scaling model	*r*	*SE*	*p*	R^2	f^2	*r*	*SE*	*p*	R^2	f^2
C	M1	.94	.02	.00	.88	7.1	–.22	.15	.16	.05	.05
C	M2	.94	.02	.00	.88	7.4	–.15	.17	.36	.03	.03
C	M1T	.92	.03	.00	.84	5.7	–.10	.22	.66	.03	.03
C	M2T	.93	.02	.00	.87	6.7	–.21	.15	.18	.05	.05
LC	M1	.88	.04	.00	.78	3.6	.04	.13	.77	.00	.00
LC	M2	.91	.03	.00	.83	4.8	.17	.12	.14	.03	.03
LC	M3	.90	.03	.00	.82	4.5	.06	.15	.70	.01	.01

Note: C = C-test, LC = listening comprehension test, M1 = Rasch model, M2 = unidimensional 2PL model, M1T = testlet 1PL model, M2T = testlet 2PL model, M3 = unidimensional 3PL model. *r* = bivariate correlation, *SE* = standard error of *r*, *p* = significance level, R^2 = proportion of outcome variance explained by class achievement at T1, f^2 = local effect Cohen's f^2.

In the C-test, the results of all scaling models pointed toward a negative correlation between initial class achievement and class mean growth, but it was not statistically significant. The local effect Cohen's f^2 of initial class achievement on student growth was small ($f^2 \geq .02$ & $f^2 < .15$). In the LC-test, the correlation between class achievement at T1 and growth was not negative; it was small in absolute terms and not statistically significant. The value of the local effect Cohen's f^2 suggested that the effect of initial achievement on student growth at class level in the LC-test was negligible ($f^2 < .02$ on average).

VIII.2.3 Academic class composition effect on academic student outcomes

The results of the *multilevel manifest covariate model* (*MMC*, Lüdtke et al., 2008, see Chapter VI.3.2) showed significant effects of academic class composition on academic student achievement at T2 after controlling for individual student effects (see Table 14 and Table 15). The regression coefficient of initial student achievement at class level (level 2) constituted the contextual effect (composition effect, see Chapter VI.3.2).

Table 14: Academic class composition effect on student achievement at T2 in the LC-test

	M1		M2		M3		Average	
Fixed effects	β	*SE*	β	*SE*	β	*SE*	β	*SE*
Intercept	.45	.06	.42	.05	.43	.05		
Level 1	.61***	.05	.73***	.05	.49***	.05	.61***	.11
Level 2	.49***	.11	.35***	.08	.54***	.09	.46***	.12
Variance	R^2	*SE*	R^2	*SE*	R^2	*SE*	R^2	*SE*
Level 1	.35	.07	.34	.06	.19	.06	.29	.10
Level 2	.80	.06	.84	.05	.83	.05	.82	.06
Total	.66	.05	.66	.04	.61	.04	.64	.05

Table 15: Academic class composition effect on student achievement at T2 in the C-test

	M1		M2		M1T		M2T		Average	
Fixed effects	β	*SE*	β	*SE*	β	*SE*	β	*SE*	β	*SE*
Intercept	.46	.04	.49	.04	.61	.06	.49	.05		
Level 1	.67***	.02	.69***	.02	.81***	.05	.81***	.04	.75***	.08
Level 2	.20***	.05	.25***	.07	.22	.13	.17	.11	.21*	.10
Variance	R^2	*SE*	R^2	*SE*	R^2	*SE*	R^2	*SE*	R^2	*SE*
Level 1	.49	.04	.48	.04	.69	.03	.67	.04	.58	.11
Level 2	.89	.04	.89	.04	.85	.05	.88	.05	.88	.05
Total	.68	.03	.69	.03	.79	.04	.81	.03	.74	.07

Note: Model's dependent variable = student achievement at T2, predictors = student achievement at T1 at individual level (level 1, grand-mean centered) and at class level (level 2). M1 = Rasch model, M2 = unidimensional 2PL model, M1T = testlet 1PL model, M2T = testlet 2PL model, M3 = unidimensional 3PL model. Average = averaged results over all scaling models. Intercept = expected outcome when all model predictors are zero, Level 1 = individual level (within classes), Level 2 = class level (between classes), β = regression coefficient, R^2 = proportion of explained variance, *SE* = standard error of estimates. Significance level of regression coefficients: *** $p < .001$, * $p < .05$

In the LC-test, the effect of pretest achievement at individual level (level 1) on posttest achievement ranged between $\beta = .49$ and $\beta = .61$ (average $\beta = .61$, $SE = .11$, $p < .001$). The posttest achievement of students was also influenced by the class mean achievement at T1; the composition effect ranged between $\beta = .35$ and $\beta = .54$ (average $\beta = .46$, $SE = .12$, $p < .001$). The results of all three scaling models in the LC-test were significant, and the corresponding estimates of the composition effect were not

significantly different from each other. Differences in initial class achievement accounted for 80% to 84% of variance in achievement between classes at T2 (average R^2 between = .82).

Individual differences in initial student achievement accounted for 19% to 35% student variance in achievement at T2 (average R^2 level 1 = .29). In total, 61% to 66% of variance in student achievement at T2 was accounted for by differences in initial student achievement at both individual and class level together (average R^2 total = .64).

In the C-test, the composition effects based on student ability estimates calculated with four scaling models were similar in absolute value. However, the estimates of class composition effect calculated with the two unidimensional scaling models were statistically significant, while the results calculated with the two testlet models were not. The average composition effect for all four models was β = .21 (*SE* = .10, p = .04). Differences in initial class achievement explained between 85% and 89% of variance in achievement between classes at T2 (average R^2 between = .88). Individual differences in initial student achievement explained between 48% and 69% of student variance in achievement at T2 (average R^2 within = .58). In total, 68% to 81% of variance in student achievement at T2 was accounted for by differences in initial student achievement at both individual and class level together (average R^2 total = .74).

VIII.3 Differences in student achievement and growth regarding the SES of students

VIII.3.1 Individual level

The socioeconomic status (SES) of students ranged from –3.4 to 3.6 with a mean of *M* = 0 and standard deviation of *SD* = 1 (see Appendix E). Unlike the small effect of SES on student achievement reported in previous studies in mathematics and the sciences (see Chapter IV.2), students with a higher SES scored significantly higher in both English tests and at both MPs in this study.

In the C-test, the correlation coefficients between student achievement and SES were an average of $\bar{r}$ = .55 (*SE* = .05, p < .001; r ranged from .51 to .60 depending on the scaling model) at T1, and $\bar{r}$ = .54 (*SE* = .05, p <.001; r ranged from .51 to .58) at T2. In the LC-test, the average correlation coefficient between student achievement and SES was $\bar{r}$ = .57 (*SE* = .04, p < .001; r ranged from .56 to .59) at T1 and $\bar{r}$ = .55 (*SE* = .04, p < .001; r ranged from .54 to .56) at T2.

With an increase of 1 SES point (1 *SD* of student SES), we expected an increase of between .50 and .58 *SD* of student achievement in the C-test at T2 depending on the scaling model (β = .50–.58, average $\bar{\beta}$ = .54, SE = .06) and an increase of between .53 and .55 *SD* of student achievement in the LC-test at T2 (β = .53–.55, average $\bar{\beta}$ = .55, SE = .04). On average, the individual student SES accounted for 25% to 34% of the variance in student achievement in the C-test at T2 (average R^2 = .29) and for 29% to 31% of the variance in student achievement in the LC-test at T2 (average R^2 = .30).

The 25% of the students with the highest SES scored much higher than the 25% of the students with the lowest SES. Across all scaling models, the differences between these two groups was an average of $\bar{d}$ = 1.69 at T1 and $\bar{d}$ = 1.67 at T2 in the C-test and $\bar{d}$ = 1.74 at T1 and $\bar{d}$ = 1.72 at T2 in the LC-test (see Table 16).

Table 16: Differences in student outcomes between 25% of students with the highest and lowest SES

Test results		Group 1	Group 2	$\bar{d}$	*SE*
C	T1	Top 25% SES	Bottom 25% SES	1.69	.23
	T2	Top 25% SES	Bottom 25% SES	1.67	.21
	Growth	Top 25% SES	Bottom 25% SES	–.05	.23
LC	T1	Top 25% SES	Bottom 25% SES	1.74	.18
	T2	Top 25% SES	Bottom 25% SES	1.72	.17
	Growth	Top 25% SES	Bottom 25% SES	.20	.18

Note: C = C-test, LC = listening comprehension test. Group 1 – Top 25% SES: 25% student with highest SES, Group 2 – Bottom 25% SES: 25% students with lowest SES. $\bar{d}$ = averaged Cohen's *d* between Group 1 and Group 2 across all scaling models, *SE* = standard error of $\bar{d}$.

On the other hand, no statistically significant correlations were found between student SES and growth: average $\bar{r}$ = –.02 (*SE* = .09, p = .79; r ranged from –.07 to .01) in the C-test and average $\bar{r}$ = .07 (*SE* = .07, p = .25; r ranged from .05 to .11) in the LC-test. The effect size Cohen's f^2 of student SES on student growth in both tests based on the results of all scaling models was extremely small ($f^2 <$ .02) and thus negligible.

VIII.3.2 Class level

The class sample in this study varied regarding the mean SES of class students, and students within classes were relatively homogenous regarding their SES. The proportion of variance between classes amounted to 45% of total variance in student SES (ICC = .45). Accordingly, the reliability ICC(2) of the class mean SES was an average of .97 for an average class size (*N* = 42); the minimum value ICC(2) was .95 for classes with the smallest class size (*N* = 27). Thus, the class mean SES was a reliable measure at class level.

Similarly to at individual level, class mean achievement in both tests at both MPs correlated positively with the class mean SES ($r \geq .84$, $p < .001$, see Table 17). The correlation coefficients did not vary significantly between tests, MPs, and scaling models.

The classes with a higher initial ability were also classes with a higher class mean SES. The correlation between class mean SES and the class's initial achievement in the C-test varied between .84 and .86 ($p < .001$), and the correlation between class mean SES and the class's initial achievement in the C-test varied between .86 and .90 ($p < .001$).

Based on the local effect Cohen's f^2, the class mean SES can be regarded to have a large effect ($f^2 > .35$) on class achievement at T2 in both tests.

Table 17: Level 2 correlations between class mean SES and student outcomes

		Correlation between class mean SES and class mean achievement at T2					Correlation between class mean SES and class mean growth				
Test	Scaling model	*r*	*SE*	*p*	R^2	f^2	*r*	*SE*	*p*	R^2	f^2
C	M1	.84	.03	.00	.73	2.7	.02	.20	.94	.00	.00
C	M2	.84	.03	.00	.73	2.7	.07	.21	.75	.01	.01
C	M1T	.84	.04	.00	.73	2.7	.13	.22	.54	.03	.03
C	M2T	.86	.03	.00	.74	2.8	-.03	.20	.88	.00	.00
LC	M1	.85	.03	.00	.67	2.1	.17	.13	.21	.03	.03
LC	M2	.86	.03	.00	.68	2.1	.25	.13	.06	.06	.07
LC	M3	.90	.02	.00	.72	2.5	.13	.15	.37	.02	.02

Note: C = C-test, LC = listening comprehension test, M1 = Rasch model, M2 = unidimensional 2PL model, M1T = testlet 1PL model, M2T = testlet 2PL model, M3 = unidimensional 3PL model. *r* = bivariate correlation, *SE* = standard error of *r*, *p* = significance level, R^2 = proportion of outcome variance explained by class mean SES, f^2 = local effect Cohen's f^2.

The correlation coefficients between class mean SES and class mean growth in the C-test based on the results of all scaling models were centered around zero and not significantly different from *zero*. Based on the local effect Cohen's f^2, the class mean SES had a negligible effect on student growth in the C-test ($f^2 < .02$ on average).

Correlations between class mean SES and class mean growth in the LC-test across all scaling models pointed toward a positive relationship between the two factors, but it was also not statistically significant. The local effect Cohen's f^2 of the class mean SES on student growth was small ($f^2 \geq .02$ & $f^2 < .15$).

VIII.3.3 Social class composition effect

Table 18: Social class composition effect on student achievement at T2 in the C-test

	M1		M2		M1T		M2T		Average	
Fixed effects	*β*	*SE*	*β*	*SE*	*β*	*SE*	*β*	*SE*	*β*	*SE*
Intercept	.46	.05	.48	.05	.62	.07	.50	.06	.51	.09
Level 1	.16***	.02	.16***	.02	.17***	.03	.18***	.03	.17***	.03
Level 2	.71***	.09	.73***	.09	.86***	.10	.83***	.10	.78***	.11
Variance	R^2	*SE*	R^2	*SE*	R^2	*SE*	R^2	*SE*	R^2	*SE*
Level 1	.03	.01	.03	.02	.04	.02	.05	.03	.04	.02
Level 2	.76	.05	.76	.06	.75	.06	.76	.06	.76	.06
Total	.39	.05	.39	.05	.50	.05	.52	.06	.45	.08

Note: Model's dependent variable = student achievement at T2 in the C-test, predictors = student SES at individual level (level 1, grand-mean centered) and at class level (level 2). M1 = Rasch model, M2 = unidimensional 2PL model, M1T = testlet 1PL model, M2T = testlet 2PL model. Average = average results across all scaling models. Intercept = expected outcome when all model predictors are zero, Level 1 = individual level (within classes), Level 2 = class level (between classes), β = regression coefficient, R^2 = proportion of explained variance, *SE* = standard error of estimates. Significance level of regression coefficients: *** $p < .001$

The effect of social composition on student achievement in the C-test was confirmed (see Table 18). This effect ranged between $\beta = .71$ and .86 ($p < .001$) across all scaling models, with an average $\beta = .78$ ($SE = .11$, $p < .001$) after controlling for individual SES effects. Differences in the class mean SES between classes explained between 75% and 76% of variance in achievement between classes at T2 (average R^2 between = .76). The differences in individual SES explained only 3% to 5% of variance between students regarding their achievement at T2 (average R^2 within = .04).

Table 19: Social class composition effect on student achievement at T2 in the LC-test

	M1		M2		M3		Average	
Fixed effects	*β*	*SE*	*β*	*SE*	*β*	*SE*	*β*	*SE*
Intercept	.46	.07	.44	.08	.40	.07	.43	.08
Level 1	.18***	.02	.19***	.03	.18***	.03	.19***	.03
Level 2	.90***	.13	.97***	.13	.96***	.12	.94***	.13
Variance	R^2	*SE*	R^2	*SE*	R^2	*SE*	R^2	*SE*
Level 1	.05	.02	.04	.02	.04	.02	.04	.02
Level 2	.69	.06	.70	.05	.74	.05	.71	.06
Total	.49	.05	.45	.05	.49	.05	.48	.05

Note: Model's dependent variable = student achievement at T2, predictors = student SES at individual level (level 1) and at class level (level 2). M1 = Rasch model, M2 = unidimensional 2PL model, M3 = unidimensional 3PL model, Average = average results of all scaling models. Intercept = expected outcome when all model predictors are zero, Level 1 = individual level (within classes), Level 2 = class level (between classes), β = regression coefficient, R^2 = proportion of explained variance, *SE* = standard error of estimates. Significance level of regression coefficients: *** $p < .001$

Quite similar results were found for the LC-test (see Table 19). The effect of social composition on student achievement at T2 in the LC-test was confirmed (level-2 β ranged between .90 and .97 across scaling models; average $\beta = .94$, $SE = .13$, $p < .001$). Differences between classes in the LC-test were,

to a large extent, explained by differences in the class mean SES (average R^2 between = .71, ranging between .69 and .74 across scaling models). Individual differences in SES accounted for only 4% to 5% of variance in student achievement at T2 (averaged R^2 within = .04).

VIII.4 Comparison of the effects of prior achievement and SES on student achievement and growth

In the C-test, the initial test scores and student SES at both levels together explained between 69% and 81% of variance in student achievement at T2 depending on the scaling model (average R^2 total = .75, *SE* = .07, see Table 20). This was comparable to the result of the MMC model with only the initial test scores (individual and class level) as model predictors (average R^2 = .74, *SE* = .07, see Table 15) and significantly higher than the explained variance proportion R^2 of the MMC model with only student and class mean SES as predictors (average R^2 = .45, *SE* = .08, see Table 18).

To compare the effect sizes of these two variables, the local effect size Cohen's f^2 was calculated. The local effect size of initial class achievement (both levels together) on student achievement at T2 in the C-test was f^2 = 1.22, while the Cohen's f^2 of student SES (both levels together) was f^2 = .04 on average.

Table 20: Joint effect of student SES and initial test scores on student achievement at T2 in the C-test

	M1		M2		M1T		M2T		Average	
Fixed effects	*β*	*SE*	*β*	*SE*	*β*	*SE*	*β*	*SE*	*β*	*SE*
Intercept	.47	.04	.49	.04	.63	.06	.52	.05	.52	.08
Level 1 SES	.02	.02	.03	.02	-.01	.03	.00	.03	.01	.03
Level 1 achievement T1	.67***	.02	.68***	.02	.82***	.05	.81***	.05	.74***	.08
Level 2 SES	.22*	.09	.21*	.09	.36*	.15	.25n.s.	.15	.26	.14
Level 2 achievement T1	.04	.07	.06	.09	-.16	.19	-.09	.16	-.04	.16
Variance	R^2	*SE*	R^2	*SE*	R^2	*SE*	R^2	*SE*	R^2	*SE*
Level 1	.49	.04	.48	.04	.69	.03	.67	.04	.58	.11
Level 2	.91	.04	.91	.04	.88	.05	.89	.04	.90	.04
Total	.69	.03	.70	.03	.81	.03	.81	.03	.75	.07

Note: Model's dependent variable = student achievement at T2 in the C-test, predictors = student SES and student initial achievement in the C-test at individual level (level 1, grand-mean centered) and at class level (level 2). M1 = Rasch model, M2 = unidimensional 2PL model, M1T = testlet 1PL model, M2T = testlet 2PL model. Average = average results of all scaling models. Intercept = expected outcome when all model predictors are zero, Level 1 = individual level (within classes), Level 2 = class level (between classes), β = regression coefficient, R^2 = proportion of explained variance, *SE* = standard error of estimates. Significance level of regression coefficients: *** $p < .001$, * $p < .05$

At class level only, the local effect of initial student achievement was f^2 = 1.32, larger than the local effect of the class mean SES which was f^2 = .17.

At individual level, differences regarding SES did not explain differences in student achievement at T2 after controlling for differences in initial student achievement (local effect size f^2 = .002 and thus negligible). The proportion of variance in student achievement at T2 which was accounted for by differences in initial student achievement was R^2 = .58, corresponding to a large local effect size of f^2 = 1.29.

Similar results were found for the LC-test. The proportion of variance of student achievement at T2 which was accounted for by initial achievement and student SES at both levels was R^2 = .65 (*SE* = .05) on average, similar to the R^2 accounted for by initial student achievement only (R^2 = .64, *SE* = .05) and higher than the R^2 accounted for only by individual student and class mean SES (R^2 = .48, *SE* = .05).

Table 21: Joint effect of student SES and initial test scores on student achievement at T2 in the LC-test

	M1		M2		M3		Average	
Fixed effects	*β*	*SE*	*β*	*SE*	*β*	*SE*	*β*	*SE*
Intercept	.46	.06	.43	.05	.42	.05	.43	.05
Level 1 SES	.08***	.02	.08**	.03	.10**	.04	.09**	.03
Level 1 achievement T1	.59***	.05	.71***	.05	.47***	.05	.59***	.11
Level 2 SES	.23	.17	.15	.15	.15	.17	.17	.17
Level 2 achievement T1	.24	.17	.21	.11	.37**	.13	.27	.16
Variance	R^2	*SE*	R^2	*SE*	R^2	*SE*	R^2	*SE*
Level 1	.36	.07	.35	.06	.20	.05	.30	.10
Level 2	.81	.06	.85	.05	.84	.04	.83	.05
Total	.67	.04	.67	.04	.62	.04	.65	.05

Note: Model's dependent variable = student achievement at T2 in the LC-test, predictors = student SES and student initial achievement in the LC-test at individual level (level 1, grand-mean centered) and at class level (level 2). M1 = Rasch model, M2 = unidimensional 2PL model, M3 = unidimensional 3PL model. Average = average results of all scaling models. Intercept = expected outcome when all model predictors are zero, Level 1 = individual level (within classes), Level 2 = class level (between classes), β = regression coefficient, R^2 = proportion of explained variance, *SE* = standard error of estimates. Significance level of regression coefficients: *** $p < .001$, ** $p < .01$, * $p < .05$

The local effect sizes of initial student achievement and SES on student achievement at T2 are shown in Table 22. Similarly to the results in the C-test, the local effect size of SES on student achievement at T2 was much smaller in comparison to that of initial student achievement. At individual level, the local effect size of student SES on student achievement at T2 was negligible (f^2 = .01). Between classes and overall, the student SES had a small local effect size on student achievement at T2 (f^2 > .02 and

f^2 < .15). At all levels, initial student achievement had a large local effect on student achievement at T2 (f^2 > .35).

Table 22: Local effect sizes of initial student achievement and of SES on student achievement at T2 in the LC-test

Model predictors	R^2 total	f^2 total	R^2 level 2	f^2 level 2	R^2 level 1	f^2 level 1
Initial achievement (level 1 and 2)	.64	.50	.82	.76	.29	.37
SES (level 1 and 2)	.48	.03	.71	.06	.04	.01
SES and initial achievement (level 1 and 2)	.65		.83		.30	

Note: level 1 = individual level (predictor at level 1 was grand-mean centered), level 2 = between level. R^2 total = proportion of total explained variance of student achievement at T2, f^2 total = local effect size of predictors on student achievement at T2, R^2 between = proportion of explained variance between classes regarding student achievement at T2, f^2 between = local effect size of predictors on class achievement at T2, R^2 within = proportion of explained variance of student achievement within classes at T2, f^2 within = local effect size of predictors on student achievement within classes at T2.

VIII.5 Chapter summary and answers to Research Questions 7–9

Based on evidence found in this study, the Research Questions 7 to 9 can be answered as follows:

Research Question 7. *Does prior student achievement have a significant effect on student achievement and growth?*

Student growth in EFL over one school year was similar to the average result at secondary school age found for other school subjects and in other countries ($d = 0.4$, Hattie, 2012, see Chapter IV.1). Classes differed with regard to initial student ability and student achievement in the posttest.

The results in both tests pointed to a large and significantly positive effect of prior student ability on student achievement in the posttest, which confirmed previous findings in international studies and in Vietnam (see Chapter IV.1). Students and classes with a higher ability at T1 scored significantly higher at T2. In addition, students and classes with a higher initial ability in one test also scored higher in the other test at T2. The correlations between initial class achievement and the posttest results were higher at class level than at individual level. Students and classes that had a higher achievement at T1 still had a higher achievement at T2 in both tests.

However, students with higher test scores at T1 tended to make less progress over one school year. Actually, the negative correlation between the initial test score and growth could be due to regression toward the mean effect (Nesselroade, Stigler, & Baltes, 1980; Steyer, 2002) because the variance of student achievement in the posttest was not significantly higher than the variance of initial student achievement (Nachtigall & Suhl, 2002; Rogosa, 1995). This negative effect was statistically significant

in the C-test but not in the LC-test (on average), and the effect sizes were small. At class level, no statistically significant relationship was found between initial student achievement and growth at class level.

There was a small positive correlation between individual student growth in the two tests, but no correlation between class mean growth in the two tests.

In short, as expected, prior student achievement had a significant effect on student achievement, but it only had a small negative effect on student and class mean growth in the C-test, and no effect on student and class mean growth in the LC-test.

Research Question 8. *Does student socioeconomic background have a significant effect on student achievement and growth? Is the effect of SES smaller than the effect of prior student achievement?*

The findings in this study revealed a large effect of student SES on student achievement in the posttest at both individual and class levels. About one-third of the variance in student achievement at T2 in both tests was explained by differences in individual student SES. The achievement gap between the 25% of students with the highest and lowest SES was about $d = 1.7$ in both tests at the beginning of the school year (about three times higher than average student growth in one school year), and that did not change much after one school year. At class level, the classes varied with regard to the class mean SES, and it was a strong predictor of student achievement at T2 in both tests.

These findings contradict previous findings in other school subjects in Vietnam, for example, the PISA results in 2012 and 2015, and do not support the assumption of a small or no relationship between student SES and achievement based on Vietnam's cultural and political background (see Chapter IV.2).

However, the SES effect was smaller than the effect of the initial test score on student achievement at T2. After controlling for the effect of prior student achievement, the local effect of student SES on student achievement in the posttest was considerably lower and was small in both tests. At individual level, the local effect of SES on student achievement after controlling for the pretest effect was negligible in both tests. Between classes, the local effect of student SES on student achievement at T2 was small in the LC-test and medium-sized in the C-test.

To sum up, student SES is confirmed to have a significant and large effect on student achievement at T2. In comparison to the effect of initial student ability on student achievement in the posttest, the effect of student SES is smaller. Student SES was not a significant predictor of student growth at both levels; although classes with a higher average SES tended to progress more in the LC-test, the effect was small and not statistically significant. In the C-test, no relationship was found between class mean SES and class mean growth.

Research Question 9. *Do academic and social class composition have an effect on student achievement and growth in EFL?*

The study results confirmed both academic and social class composition effects on student achievement at T2, but not on student growth in EFL. However, classes with a higher initial ability at T1 in each test also had a higher SES, because these two aspects correlated strongly with each other. Due to this, only one class composition variable will be taken into account as a covariate in further analyses.

At this point, it should be mentioned that other contextual effects on student achievement were found, too. For example, girls, younger students (born in 1992 or later), and students who had not repeated a class scored significantly higher than their counterparts at both tests and at both MPs (see Appendix F2). Also, classes with larger class sizes scored higher in both tests and at both MPs (average $\bar{r}$ = .45, p < .01 in the C-test and average $\bar{r}$ = .32, p < .01 in the LC-test). However, the positive effect of class size was explained completely by the differences in mean SES between classes (see also Rolleston, James, Pasquier-Doumer, et al., 2013, see Chapter II.2). Likewise, the effects of the other context factors on student achievement in the posttest were negligible and non-significant after having controlled for the effect of initial student ability at both levels. None of them had considerable and significant effects on student growth at both levels.

IX. Instructional effects on academic student growth at class level

This chapter sets out the results of regression models carried out to estimate the effects of video-based instructional classroom variables on student growth at class level. In order to control for possible confounding contextual effects, initial class achievement (at T1) was included as a covariate in the regression models with class mean growth in the C-test as the dependent variable; class mean SES as a covariate in the regression models with class mean growth in the LC-test as the dependent variable (see previous chapter). For the analyses, we used 10 imputed datasets with multiply imputed data and the PVs of all 2,096 students of the 50 participating classes. The results will be summarized and interpreted in order to answer the Research Questions 10–14.

IX.1 Preselection of instructional factors

Among the 216 available basic coding variables, many correlated strongly positively or negatively with each other or had similar meanings. Thus, before conducting the analyses, we removed redundant or non-meaningful variables. Basic coding variables with non-interpretable content (such as "unassignable"), never or seldom occurred (third quartile $Q3 < 0.5$ either regarding the relative time percentages or the relative frequencies, such as the time spent on discipline-related activities, student statement of non-knowledge) were eliminated from further analyses. Of each basic coding category, only the relative time percentage or relative frequency was chosen based on the content of the coding category, for example, the speaking time percentages of students and teachers or the relative frequency of transitions. Among the group of variables whose time percentages or relative frequencies added up to 100%, if any two of them correlated strongly negatively with each other ($r \leq -0.8$), only one was selected for further analyses. Among the group of variables that correlated strongly with each other ($r \geq 0.8$) and were similar in terms of content (e.g., total student speaking time and sum of individual student speaking time), only one was kept. In total, 59/216 basic coding variables were preselected for further analysis.

Rating variables with a ceiling or floor effect ($M \leq 1.5$ or $M \geq 3.5$) and variables in which almost all classes were rated positively or negatively with a small SD ($M - SD > 2.5$ or $M + SD < 2.5$) were not included in further analyses. Overall, 12/32 rating variables were eliminated.

Among the teaching materials and multimedia (see Appendix F1), the black board, video, television, internet, and the language lab were not included because of either non-/seldom observation or because almost all teachers used them. Due to the co-occurrence of PCs and projectors, only the PC was kept in further analyses.

Altogether, 81 video-based instructional variables were preselected as instructional factors to be further analyzed in this chapter.

IX.2 Linear relationship between instructional factors and student growth

In order to identify linear relationships between instructional factors and class mean growth, lasso and OLS multiple linear regression analyses were executed with class mean growth as the dependent variable (see Chapter VI.3.3). One instructional variable was included as a model covariate in each regression model: class mean achievement at T1 (for analyses with class mean growth in the C-test as the dependent variable) and class mean SES (for analyses with class mean growth in the LC-test as the dependent variable). The *z*-scores of all model variables were used for interpretation purposes.

IX.2.1 Linear effects of student growth in the C-test

The results of the lasso regression analyses revealed three instructional factors with nonzero lasso regression coefficients. Among all examined instructional factors in this study, they could be regarded as having the most important linear effect on class mean growth in the C-test after controlling for the effect of initial class achievement based on the lasso regression coefficients.

Table 23: Predictors of class mean growth in the C-test with nonzero lasso regression coefficients

	Scaling model of the dependent variable			
	M1	**M2**	**M1T**	**M2T**
Predictor	β_l	β_l	β_l	β_l
Encouragement of student statements (rating)	.11	.18	.15	.
Teacher speaking time in mixed languages (time percentage)	.	–.10	–.09	.
Affectively stressed positive feedback (relative frequency)	.	.11	.	.

Note: M1 = Rasch model, M2 = unidimensional 2PL model, M1T = testlet 1PL model, M2T = testlet 2PL model. β_l = lasso regression coefficient after controlling for effect of initial class achievement, "." means coefficient was set to zero.

Based on the testlet 2PL model (M2T), no instructional factors were identified as important for class mean growth in the C-test. Only the rating variable "Encouragement of student statements" had a nonzero positive lasso regression coefficient with regard to student growth in the C-test based on the three scaling models. The basic coding variable "Teacher speaking time in mixed languages" (time percentage) had a nonzero lasso regression coefficient with regard to growth in the C-test based on two scaling models. Another basic coding variable "Affectively stressed positive feedback" (relative frequency) was identified as a significant predictor of student growth in the C-test, only if the unidimensional 2PL model was used to estimate the test scores.

Actually, these three variables had the highest average local effects Cohen's f_z^2 on student growth based on the results of the OLS regressions (see Table 24).

Based on the results of the OLS multiple regressions, the variable "Encouragement of student statements" had an average local linear effect Cohen's f_z^2 = .13 (ranging between .08 and .16 depending on the scaling model of the test results) and an average regression coefficient β = .31 (SE = .18, p = .08). The variable "Teacher speaking time in mixed languages" had an average local linear effect f_z^2 = .11 (ranging between .08 and .13 depending on the scaling model) and an average regression coefficient β = –.31 (SE = .17, p = .06). And the variable "Affectively stressed positive feedback" had an average local linear effect f_z^2 = .08 (ranging between .05 and .12 depending on the scaling model) and an average regression coefficient β = –.26 (SE = .20, p = .19). According to the categorization of effect sizes by Cohen (1988), all of them had a small effect on student growth after controlling for the effect of initial class achievement.

All other instructional factors also had a small (average $f_z^2 \geq .02$) or negligible local effect size on class mean growth in the C-test. The 27 instructional factors which had an average local effect size Cohen's $f_z^2 \geq .02$ on student growth in the C-test and their regression coefficients are listed in Table 24 in decreasing order of their average local effect sizes.

Table 24: Predictors of class mean growth in the C-test with an averaged local effect size $f_z^2 \geq .02$

	Scaling model of the dependent variable									
	M1		M2		M1T		M2T		Average	
Predictor	β	f_z^2	β	f_z^2	β	f_z^2	β	f_z^2	β	f_z^2
Encouragement of student statements (rating)	.35*	.16	.34*	.15	.31	.14	.24	.08	.31	.13
Teacher speaking time in mixed languages (time percentage)	–.33*	.13	–.34*	.13	–.31	.11	–.26	.08	–.31	.11
Affectively stressed positive feedback (relative frequency)	.29	.10	.32	.12	.24	.07	.21	.05	.26	.08
Time used for social activities (time percentage)	.22	.05	.23	.05	.28	.09	.27	.09	.25	.07
Student grammar mistakes (relative frequency)	.22	.06	.21	.05	.24	.07	.28*	.09	.24	.07
Teacher question of low complexity (relative frequency)	–.26	.07	–.25	.07	–.24	.06	–.12	.02	–.22	.06
Student speaking time in mixed languages (time percentage)	–.23	.05	–.25	.06	–.20	.04	–.21	.04	–.22	.05
Teacher question of high complexity (relative frequency)	.18	.03	.17	.02	.27	.07	.27	.07	.22	.05
Repeated questions (relative frequency)	–.17	.03	–.17	.03	–.17	.04	–.24	.07	–.19	.04
Educational games (time percentage)	–.21	.05	–.22	.05	–.14	.02	–.16	.03	–.18	.04
Dealing with different types of mistakes simultaneously (relative frequency)	.17	.05	.22	.06	.16	.03	.12	.02	.17	.04
Teacher commitment (rating)	.14	.02	.14	.02	.16	.04	.17	.04	.15	.03
Student orientation (rating)	.12	.02	.13	.02	.19	.05	.19	.05	.16	.03

	Scaling model of the dependent variable									
	M1		M2		M1T		M2T		Average	
Predictor	β	f_z^2	β	f_z^2	β	f_z^2	β	f_z^2	β	f_z^2
Using PC	.15	.02	.13	.02	.12	.02	.21	.05	.15	.03
Speaking time of group of students (time percentage)	.12	.02	.11	.01	.11	.02	.22	.06	.14	.03
Teacher language: Suitable way of speaking (rating)	–.16	.03	–.16	.03	–.12	.02	–.13	.03	–.14	.03
Teacher let students correct mistakes (relative frequency)	–.16	.03	–.16	.03	–.12	.02	–.12	.02	–.14	.03
Lesson authenticity (rating)	–.17	.03	–.21	.04	–.06	.01	–.11	.02	–.14	.03
Teacher speaking time in Vietnamese (time percentage)	–.16	.02	–.16	.03	–.17	.04	–.13	.02	–.16	.03
Pointing out student mistakes (relative frequency)	.10	.01	.10	.01	.15	.03	.22	.06	.14	.03
Relative frequencies of student mistakes regarding situation/context	–.14	.03	–.15	.02	–.15	.03	–.11	.01	–.14	.02
Teacher reliable availability of English grammar	–.17	.03	–.16	.03	–.13	.02	–.07	.01	–.13	.02
Affectively neutral positive feedback (relative frequency)	.16	.03	.13	.02	.14	.03	.13	.02	.14	.02
Student speaking time in Vietnamese (time percentage)	–.15	.02	–.17	.03	–.15	.03	–.04	.01	–.13	.02
Student statements interrupted by teacher or another student (relative frequency)	.15	.02	.12	.02	.14	.03	.10	.02	.13	.02
Teacher availability of extensive and reliable English vocabulary (rating)	–.16	.03	–.12	.02	–.07	.02	–.10	.02	–.11	.02
Teaching objective: Accuracy (time percentage)	.14	.02	.13	.02	.10	.02	.11	.02	.12	.02

Note: M1 = Rasch model, M2 = unidimensional 2PL model, M1T = testlet 1PL model, M2T = testlet 2PL model. Average = average results of all scaling models. β = OLS regression coefficient (after controlling for effect of initial class achievement), f_z^2 = local linear effect Cohen's f^2. Statistically significant level: $^* p < .05$

A medium or large effect according to Cohen was not found for any factor on average. "Encouragement of student statements" was the only factor which had a medium local effect on student growth when the unidimensional scaling models were applied to estimate test scores.

Among all variables with a small average effect size on student growth in the C-test, several patterns of results which were similar to the results in the DESI-study were found. For instance, the variables quality of motivation, student orientation, and feedback (encouragement of student statements, teacher commitment, student orientation, positive feedback) were positive predictors of student growth and had comparably higher local linear effects. The variables related to the use of Vietnamese or the use of mixed languages by the students and teachers had a small negative effect on class mean growth in the C-test.

On the other hand, some variables had a relationship with student growth which was difficult to interpret; for instance, classes in which students made a relatively high number of grammar mistakes compared to other kinds of mistakes progressed more; teachers with a more suitable way of speaking had more

extensive vocabulary, or classes with a higher rating of lesson authenticity progressed less. These might be spurious, because the instructional variables correlated to a varying degree with each other. For instance, classes in which students made a comparatively high number of grammar mistakes were also classes in which students made fewer phonological mistakes (r = –.46, p = .005) with higher ratings regarding teacher commitment (r = .34, p = .005), and teachers in these classes gave positive feedback more frequently than in other classes (r = .32, p = .03). This made it clear that the regression results did not necessarily imply causality between the factors and the dependent variable.

Taking the statistical significance of the OLS regression coefficients into account, "Encouragement of student statements" and "Teacher speaking time in mixed languages" were regarded as significant predictors of student growth in the C-test, when applying the unidimensional models. Based on the Rasch testlet model, the relative frequency of "Student grammar mistakes" was also a significantly positive predictor. However, the latter was not easy to interpret. The regression coefficients based on ability estimates calculated with the testlet models were mostly not statistically significant.

IX.2.2 Linear effects of student growth in the LC-test

In the LC-test, three variables with the highest local linear effect based on the OLS regression results also had a nonzero lasso regression coefficient after controlling for the effect of class mean SES (see Table 25 and Table 26).

Table 25: Predictors of class mean growth in the C-test with a nonzero lasso regression coefficient

	Scaling model of the dependent variable		
	M1	**M2**	**M3**
Predictor	β_l	β_l	β_l
Teacher speaking time using Vietnamese in transitions (time percentage)	–.17	–.18	–.17
Student speaking time in mixed languages (time percentage)	–.13	–.17	–.30
Repeated questions (relative frequency)	.	–.06	–.24

Note: M1 = Rasch model, M2 = unidimensional 2PL model, M3 = unidimensional 3PL model. β_l = lasso regression coefficient after controlling for effect of class mean SES, "." means coefficient was set to zero.

"Teacher speaking time using Vietnamese in transitions" and "Student speaking time in mixed languages" were identified as important linear effects of student growth in the LC-test based on the lasso regression results, regardless of which scaling models were used to estimate the test scores. The relative frequency of "repeated questions" had two nonzero lasso regression coefficients, when ability estimates based on the unidimensional 2PL or 3PL models were used. The lasso regression coefficients of each factor were similar in terms of sign and differed in absolute value. The absolute value of the lasso

coefficients of the latter two factors using ability estimates via the unidimensional 3PL model were considerably higher than when using ability estimates via other scaling models.

These three instructional factors had an average local linear effect $f_z^2 > .10$ on student growth in the LC-test and a statistically significant average regression coefficient ($p \leq .05$) after controlling for the effect of class mean SES.

Another factor with a statistically significant average OLS regression coefficient was a positive predictor: "Teacher lecture (time percentage)" (average f_z^2 = .08, β = .27, SE = .13, p = .03). Three other variables, "Affectively neutral positive feedback" (relative frequency, negative predictor), "Dealing with grammar mistakes" (relative frequency, positive predictor), and "Dealing with phonological mistakes" (relative frequency, negative predictor), were also identified as important predictors of student growth in the LC-test using the statistical significance of the OLS regression coefficient, but only when the Rasch model or the unidimensional 2PL model was applied to estimate student ability.

In total, 27 instructional factors had a small average local linear effect on student growth in the LC-test. Medium or large average effects were not found according to the average effect size based on all scaling models, with the exception of "Teacher speaking time using Vietnamese in transitions" based on test scores estimated via the unidimensional 2PL model and "Student speaking time in mixed languages (time percentage)", and "Repeated questions (relative frequency)" based on the test results of the 3PL scaling model. All factors with an average local linear effect on class mean growth in the LC-test $f_z^2 \geq .02$ are listed in Table 26 below in decreasing order of their average local effect sizes.

Table 26: Predictors of class mean growth in the LC-test with average local linear effect $f_z^2 \geq .02$

	Scaling model of the dependent variable							
	M1		M2		M3		Average	
Predictor	β	f_z^2	β	f_z^2	β	f_z^2	β	f_z^2
Teacher speaking time using Vietnamese in transitions	–.34**	.14	–.34**	.15	–.32*	.12	–.34*	.14
Student speaking time in mixed languages (time percentage)	–.29	.08	–.33*	.11	–.42*	.19	–.34*	.13
Repeated questions (relative frequency)	–.28	.09	–.28	.09	–.36*	.15	–.31*	.11
Affectively neutral positive feedback (relative frequency)	–.32*	.11	–.26	.07	–.29	.09	–.29	.09
Teacher lecture (time percentage)	.29*	.09	.28*	.09	.25	.07	.27*	.08
Student phonological mistakes (relative frequency)	–.19	.04	–.23	.06	–.28	.09	–.23	.06
Total student speaking mistakes (relative frequency)	–.23	.06	–.21	.05	–.23	.05	–.22	.05
Dealing with grammar mistakes (relative frequency)	.25*	.06	.24*	.06	.17	.03	.22	.05

	Scaling model of the dependent variable							
	M1		M2		M3		Average	
Predictor	β	f_z^2	β	f_z^2	β	f_z^2	β	f_z^2
Time used for social activities (time percentage)	.23	.06	.21	.05	.15	.03	.20	.04
Teacher speaking time in Vietnamese (time percentage)	–.23	.05	–.21	.05	–.15	.02	–.20	.04
Struturedness: Preview, summary, review, highlight (rating)	–.20	.05	–.19	.05	–.15	.03	–.18	.04
Teacher speaking time to individual student (time percentage)	–.20	.04	–.20	.05	–.15	.03	–.18	.04
Teacher speaking time in mixed languages (time percentage)	–.19	.04	–.20	.05	–.17	.03	–.19	.04
Dealing with phonological mistakes (relative frequency)	–.19	.04	–.20*	.04	–.17	.03	–.19	.04
Student one-word statements (relative frequency)	.21	.05	.15	.03	.17	.03	.18	.04
Teaching objective: Fostering learning and thinking strategies (rating)	–.17	.03	–.16	.03	–.18	.04	–.17	.03
Teacher commitment (rating)	.17	.03	.19	.04	.11	.02	.16	.03
Using PC	.15	.02	.18	.03	.17	.04	.17	.03
Dealing with real student mistakes (relative frequency)	.18	.04	.17	.03	.14	.02	.17	.03
Dealing with vocabulary mistakes (relative frequency)	–.12	.02	–.12	.02	–.20	.05	–.15	.03
Dealing with mistakes regarding situation, context (relative frequency)	.13	.03	.11	.02	.17	.04	.14	.03
Student orientation (rating)	.18	.04	.19	.04	.06	.01	.14	.03
Speaking time of group of students (time percentage)	.14	.02	.15	.03	.16	.03	.15	.03
Structuredness: Link to prior knowledge (rating)	–.15	.02	–.14	.02	–.17	.03	–.15	.03
Lesson authenticity (rating)	–.15	.02	–.11	.01	–.20	.04	–.15	.02
Speaking time: whole class chorused (time percentage)	–.18	.04	–.13	.02	–.11	.01	–.14	.02
Student grammar mistakes (relative frequency)	.11	.01	.11	.02	.18	.03	.13	.02

Note: M1 = Rasch model, M2 = unidimensional 2PL model, M3 = unidimensional 3PL model.
Average = average results over all scaling models. β = OLS regression coefficient (after controlling for effect of class mean SES), f_z^2 = local linear effect Cohen's f^2. Statistically significant level: * $p < .05$, ** $p < .01$

Regarding the LC-test, a similar pattern of results was found to that in the DESI-study. For instance, classes with less progress in the LC-test were characterized by more teacher speaking time in Vietnamese or in mixed languages, in lesson conversations or in transitions, and student speaking time in mixed languages; students with teachers who were rated higher regarding teacher commitment and student orientation progressed more; classes in which students made mistakes more frequently while speaking, in particular phonological mistakes, progressed less. This study found the same negative relationship between the structuredness of the lesson and student growth as in the DESI-video study. Two aspects of the structuredness of lessons (regarding preview, summary, review, highlight, and

regarding linking to prior knowledge) had a negative relationship with student growth, although the effect was not statistically significant and the effect size was small.

Once again, it is difficult to interpret some relationships; for example, classes in which students made a higher number of grammar mistakes than other kinds of mistakes progressed more, or class students who had teachers who gave a comparatively high amount of affectively neutral positive feedback progressed less; in addition, classes with a higher rating regarding lesson authenticity progressed less.

IX.2.3 Chapter summary and answer to Research Question 10

Based on the regression analyses, the regression coefficients and effect sizes of each of the instructional factors were estimated, which generated an average estimate of the relationship between each of them and the dependent variable (positive or negative). Overall, the selection of variables based on the lasso regression coefficient led to interpretable and plausible results regarding the important predictors of student growth in the C-test at class level. Variables with nonzero lasso regression coefficients were always those with the highest local linear effect based on the OLS regression results. On the other hand, variables selected based on the statistical significance of the OLS regression coefficients were sometimes difficult to interpret.

An important remark should, however, be made: The selected variables based on both the lasso (nonzero) and OLS regression coefficients (statistically significant) differed for different scaling models in both tests. Although the sign of the (nonzero) regression coefficients of the factors was identical, their absolute values and/or statistical significances varied across all scaling models.

In terms of the local effect size, all instructional factors had a small or negligible average local linear effect on student growth in both tests based on all scaling models. Medium effects were found only for several variables when implementing a specific scaling model. Actually, a small local effect of each factor was not unexpected due to multiple reasons, amongst others the non-experimental design of the study (Seidel & Shavelson, 2007), and because the cumulative and joint effect of many factors together over one school year was between small (d = 0.44) and medium (d = 0.77) at class level – depending on the test and the scaling model. The benchmarks set by Cohen – largely based on the results of experimental studies – might be too high in this study context with regard to the selection of variables. On the other hand, it is rather arbitrary to select and difficult to justify any cut-off value for the selection of variables based on the effect size.

Against this background, the selection criterion based on the lasso regression coefficient was considered most practical and plausible among these three criteria for selecting the variables. Based on this criterion,

the Research Question 10 (*"What are the most important instructional factors of student growth in the C-test and LC-test?"*) can be answered as follows:

The most important instructional factors of student growth in the C-test after controlling for the pretest effect were "Encouragement of student statements", "Teacher speaking time in mixed languages", and "Affectively stressed positive feedback", depending on the scaling model. If the testlet 2PL model were applied as the scaling model, none of the instructional predictors in this study would be regarded as important predictors of student growth according to the C-test.

For student progress in the LC-test, "Teacher speaking time using Vietnamese in transitions" and "Student speaking time in mixed languages" were the most important predictors of student growth after controlling for the SES effect, regardless of which of the three unidimensional scaling models was used to estimate student ability. "Repeated questions" (relative frequency) was also identified as one of the most important predictors of student growth in the LC-test, when the Rasch model was not used as the scaling model.

IX.3 Nonlinear relationships between instructional factors and student growth

To explore nonlinear relationships between instructional factors and student growth under the theoretical assumption that the optimum level of instructional variables is not always the maximum (see Chapter IV.3.3), second-order polynomial regressions (see Chapter VI.3.4) were executed. That means that a quadratic term of each instructional factor was added as an additional model predictor to the regression models beside the main effect (z-score) and the covariate (initial class achievement with regard to class mean growth in the C-test, and class mean SES with regard to class mean growth in the LC-test). Of course, the analysis of nonlinear relationships was not performed for instructional factors that were dummy-coded, because the quadratic term of one dummy variable is identical to the main variable. Furthermore, no analysis was conducted for some rating variables which were assumed to not have a nonlinear relationship: The highest rating category corresponded to the optimum level of these variables, such as "social environment, warmth, cordiality" (the highest rating category was applied when there was consistently friendly interaction, with signs of personal interest, warmth, and cordiality). Alternatively, for "positive treatment of mistakes", the highest rating category was applied when mistakes were treated in a constructive (stimulating and supporting the learning process) and motivating (boosting willingness to learn) manner. Thus, the nonlinear relationships with student growth were analyzed for 64 selected instructional variables.

IX.3.1 Nonlinear relationship between instructional factors and student growth in the C-test

Similar to the results in Section IX.2.1, the results of the strong hierarchical lasso regression models (see Chapter VI.3.4) including the quadratic terms of the instructional factors revealed no considerable nonlinear relationships between the instructional factors and student growth in the C-test with student ability estimates based on the testlet 2PL model.

Table 27: Instructional factors with a significant nonlinear effect on class mean growth in the C-test based on results of hierarchical lasso regression analysis

	M1			M2			M1T			M2T		
Predictor	β_{lz}	β_{lq}	*N* nonzero	β_{lz}	β_{lq}	*N* nonzero	β_{lz}	β_{lq}	*N* nonzero	β_{lz}	β_{lq}	*N* nonzero
Student reading out own text in English (relative frequency)	–.03	–.03	9	–.06	–.06	10	.05	–.15	10	.00	.00	1
Teacher speaking time using Vietnamese in transitions (time percentage)	.02	–.11	10	0	0	0	–.01	–.21	10	0	0	0

Note: M1 = Rasch model, M2 = unidimensional 2PL model, M1T = testlet 1PL model, M2T = testlet 2PL model. β_{lz} = lasso regression coefficient of the linear terms, β_{lq} = lasso regression coefficient of the quadratic terms. *N* nonzero = number of imputed datasets with nonzero lasso regression coefficients of the quadratic term of the corresponding factor.

When calculating student growth estimates based on student ability estimates via the other two or three scaling models as dependent variables and initial class achievement as the model covariate, the quadratic terms of "Student reading out own text in English (relative frequency)" and "Teacher speaking time using Vietnamese in transitions" were found to have nonzero lasso coefficients over almost all imputed datasets (*N* nonzero = 9 or 10).

The *F*-test results based on the OLS regression models suggested that the relationship between these two variables ("Student reading out own text in English (relative frequency)", "Teacher speaking time using Vietnamese in transitions") and class mean growth in the C-test based on the test results of at least one scaling model was nonlinear ($p < .05$), too. In addition, "Teacher lecture" (time percentage) also had a nonlinear relationship with student growth in the C-test based on the results of the *F*-test. All of them had a negligible linear effect on student growth in the C-test (local linear effect based on the OLS regression results $f_z^2 < 0.2$; lasso regression coefficient of the main effect was set to zero; see Section IX.2.1).

The OLS regression coefficients of these three variables (β_z = regression coefficient of the linear term, β_q = regression coefficient of the quadratic term) and the *F*-test results of the model comparison between the regression models with and without the quadratic term (*F*-ratio with adjusted degrees of freedom (1, 14.6)) are shown in Table 28.

In addition, the total local effect of each instructional factor on class mean growth f_t^2 (for both the main and quadratic terms together) is also set out in Table 28. It was calculated based on R^2 of the regression model with both the linear and quadratic terms of this factor (together with initial class achievement as model covariate) and of the regression model without them (which included only the covariate as a predictor). f_t^2 could be regarded as the local effect size of the corresponding instructional factor on class mean growth in the C-test if the nonlinear relationship was taken into account.

Table 28: Instructional factors with a significant nonlinear effect on class mean growth in the C-test based on *F*-test results

	Scaling model of the dependent variable																			
	M1				M2				M1T				M2T				Average			
Predictor	β_z	β_q	*F*	f_t^2	β_z	β_q	*F*	f_t^2	β_z	β_q	*F*	f_t^2	β_z	β_q	*F*	f_t^2	β_z	β_q	*F*	f_t^2
Student reading out own text in English (relative frequency)	.08	–.36*	**5.0***	.13	.05	–.34*	4.4	.12	.13	–.37*	**5.2***	.13	.09	–.29	3.4	.09	.09	–.34*	4.5	.12
Teacher speaking time using Vietnamese in transitions (time percentage)	.25	–.41*	**5.2***	.11	.23	–.37*	4.1	.09	.18	–.38*	**4.6***	.11	.20	–.33	3.7	.08	.22	–.37*	4.4	.10
Teacher lecture (time percentage)	–.16	.38	**4.9***	.12	–.14	.40	**5.4***	.13	–.14	.30	3.2	.08	–.13	.24	1.9	.05	–.14	.33	3.8	.09

Note: M1 = Rasch model, M2 = unidimensional 2PL model, M1T = testlet 1PL model, M2T = testlet 2PL model. Average = average results of all scaling models. β_z = regression coefficient of the linear term, β_q = regression coefficient of the quadratic term. *F* = *F*-ratio with degrees of freedom (1, 14.6). f_t^2 = local effect of the factor based on R^2 of regression model with both linear and quadratic terms and model without instructional predictor. Statistical significance level of regression coefficients and *F*-ratio: ***bold** $p < .05$.

The *F*-test results differed when using different scaling models. Taking the average results of all three variables into account, the *p*-values of the F-test were all larger than .05. The nonlinear relationship between class mean growth in the C-test and "Student reading out own text in English" (relative frequency) as well as "Teacher speaking time using Vietnamese in transitions" was confirmed using all scaling models except the testlet 2PL model.

To obtain a better perception understanding of the nonlinear relationship between these two instructional factors and class mean growth in the C-test, the expected class mean growth in the C-test corresponding to each of these factors is depicted in the following figures (bold line in each figure). Additionally, the density curve (thin line) of each instructional variable (*z*-score) is also shown in each figure.

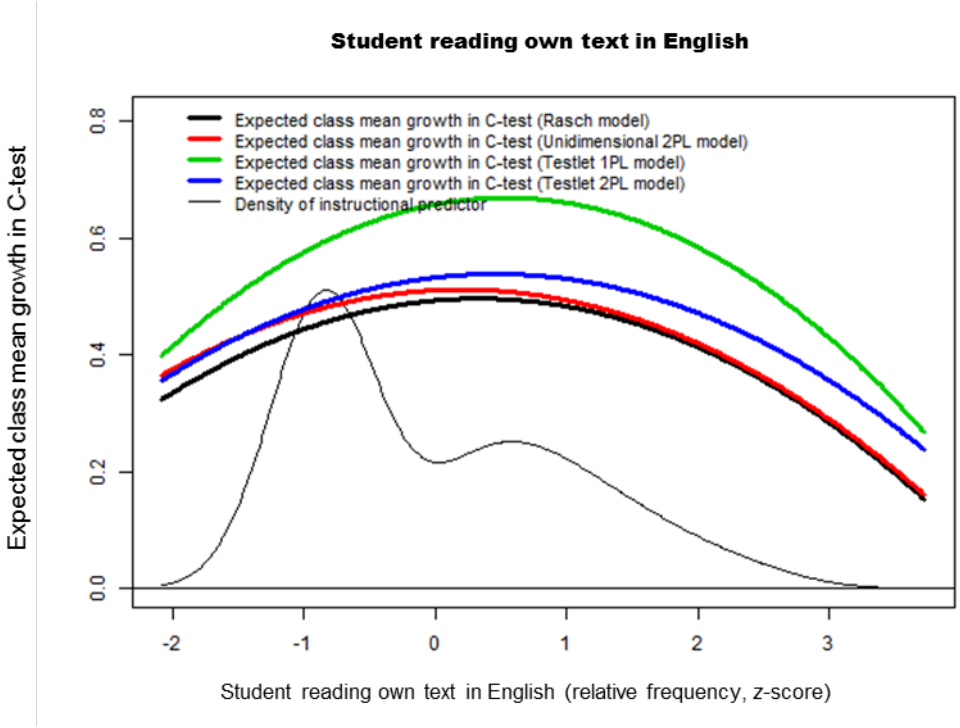

Figure 33: Relationship between "Student reading own text in English" (relative frequency) and class mean growth in the C-test

A positive relationship was found between the relative frequency of "Student reading own text in English" and class mean growth in the C-test, until this frequency approached circa 6% (*z*-score ≤ 0.6, see Figure 33). When this frequency exceeded 6%, a negative relationship was observed.

During transitions between lesson episodes, teachers used Vietnamese 9% of the time on average (*SD* = 12%). When this number exceeded around 10% (*z*-score ≥ 0.6, see Figure 34), a negative trend regarding the relationship between using Vietnamese in transitions and class mean growth in the C-test was found. Under 10%, a positive trend was observed.

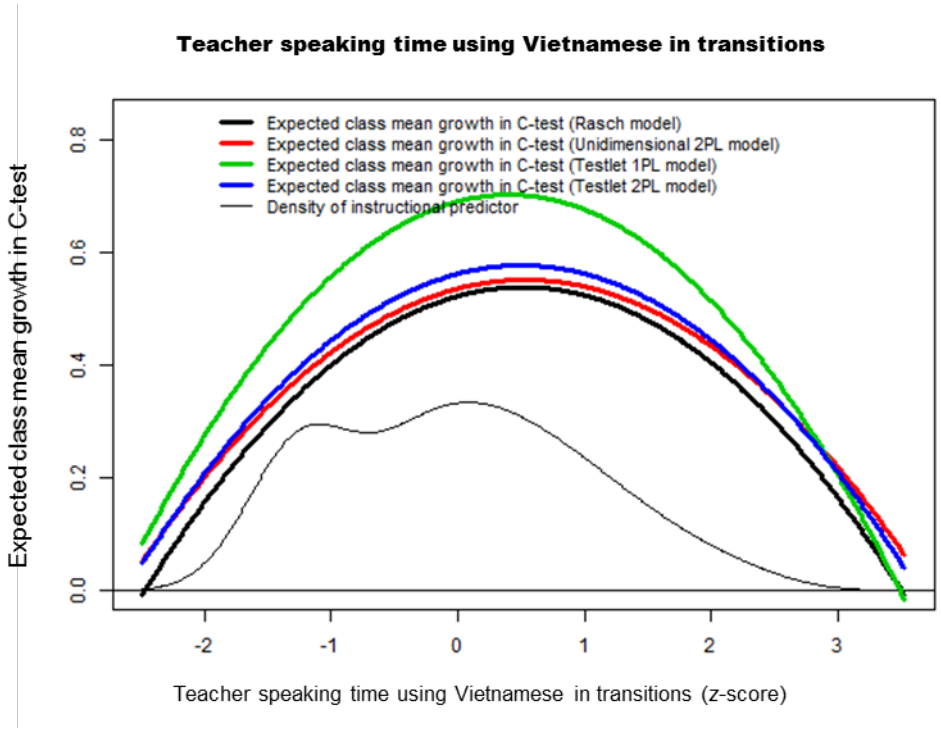

Figure 34: Relationship between "Teacher speaking time using Vietnamese in transitions" (time percentage) and class mean growth in the C-test

IX.3.2 Nonlinear relationship between instructional factors and student growth in the LC-test

Table 29: Instructional factors with a significant nonlinear effect on class mean growth in the LC-test based on results of hierarchical lasso regression analysis

	M1			M2			M3		
Predictor	β_{lz}	β_{lq}	*N* nonzero	β_{lz}	β_{lq}	*N* nonzero	β_{lz}	β_{lq}	*N* nonzero
Lesson authenticity (rating)	–.16	20	10	–.05	.09	10	–.19	.19	10
Teaching objective: Involvement of as many students as possible	–.06	–.06	10	0	0	0	–.09	–.10	10
Total teacher speaking time (time percentage)	0	0	0	0	0	0	.01	.12	10
Narrow focused monitoring (rating)	–.01	.01	5	–.02	.02	6	.00	.02	7

Note: M1 = Rasch model, M2 = unidimensional 2PL model, M1T = testlet 1PL model, M2T = testlet 2PL model. β_{lz} = lasso regression coefficient of the linear terms, β_{lq} = lasso regression coefficient of the quadratic terms. *N* nonzero = number of imputed dataset with nonzero lasso regression coefficients of the quadratic term of the corresponding factor.

Strong hierarchical lasso regression results revealed a considerable nonlinear relationship between "Lesson authenticity" and student growth in the LC-test based on the test results of all scaling models (see Table 29). With class mean growth in the LC-test based on the scaling models M1 (Rasch model)

and M3 (the unidimensional 3PL model), a nonlinear relationship of "Teaching objective: Involvement of as many students as possible" was also found. The predictor "Total teacher speaking time" (time percentage) had a nonlinear relationship with class mean growth when M3 was used as the scaling model for ability estimates in the LC test. The nonlinear relationship found between "Narrow focused monitoring" and class mean growth in the LC-test did not vary much between the different scaling models but it did vary between the 10 imputed datasets. For all 10 imputed datasets, the average lasso regression coefficients of this factor were nonzero but small.

Table 30: Instructional factors with a significant nonlinear effect on class mean growth in the LC-test based on F-test results

	Scaling model of the dependent variable															
	M1				M2				M3				Average			
Predictor	β_z	β_q	F	f_t^2	β_z	β_q	F	f_t^2	β_z	β_q	F	f_t^2	β_z	β_q	F	f_t^2
Lesson authenticity (rating)	–.27	.42*	**10.3***	.26	–.23	.43*	**10.8***	.26	–.31	.37*	**8.0***	.23	–.27	.42*	**9.7***	.25
Time used for social activities (time percentage)	.11	.28	3.8	.14	.06	.36	**6.5***	.20	.04	.27	3.5	.10	.07	.30	**4.6***	.15
Narrow focused monitoring (rating)	–.02	.27	3.8	.09	–.06	.30	**5.4***	.14	.07	.28	4.3	.10	–.01	.28	4.5	.11
Teaching objective: Involvement of as many students as possible	–.05	–.24	3.0	.08	–.04	–.28	4.1	.10	–.07	–.32*	**5.4***	.14	–.05	–.28	4.2	.10

Note: M1 = Rasch model, M2 = unidimensional 2PL model, M3 = unidimensional 3PL model.
Average = average results of all scaling models. β_z = regression coefficient of the linear term, β_q = regression coefficient of the quadratic term. F = F-ratio with adjusted degrees of freedom (1, 14.6). f_t^2 = local effect of the instructional factor (both linear and quadratic terms). Statistical significance level of F-ratio: *__bold__ $p < .05$.

Based on the F-test results, four instructional factors were identified as having a significant nonlinear relationship with class mean growth in the LC-test. The regression coefficients of these variables (β_z = regression coefficient of the linear term, β_z = regression coefficient of the quadratic term), the F-ratio with adjusted degrees of freedom (1, 14.6), and total local effect f_t^2 of each instructional factor on class mean growth are shown in Table 30 above.

Among these factors, two factors had a statistically significant average F-ratio of $\alpha = .05$. They were "Lesson authenticity (rating)" and "Time used for social activities (time percentage)". With the exception of "Time used for social activities (time percentage)", the nonlinear relationship of the other three predictors with student growth in the LC-test was also confirmed by the results of the hierarchical lasso regression model.

The relationship between "Lesson authenticity" and student growth in the LC-test was, however, difficult to interpret. The data showed a negative effect of lesson authenticity on class mean growth in the LC-test on average, but the instruction which was rated as very authentic (rating category 4) was associated with above-average student growth in the LC-test (see Figure 35). Caution is needed, as there were very few units of instruction in rating category 4; hence interpretation should be suspended here (see Cohen et al., 2003). Apart from that, a negative relationship (linear or average effect) between lesson authenticity and student growth in the LC-test was not anticipated. A possible explanation might be the cultural and social differences between everyday life in Vietnam and in western English-speaking countries. Thus, the relevance of tasks, materials, examples to the every-day life of the students in Vietnam might not contribute to the ability of students to understand conversations and dialogues in culturally and socially different contexts.

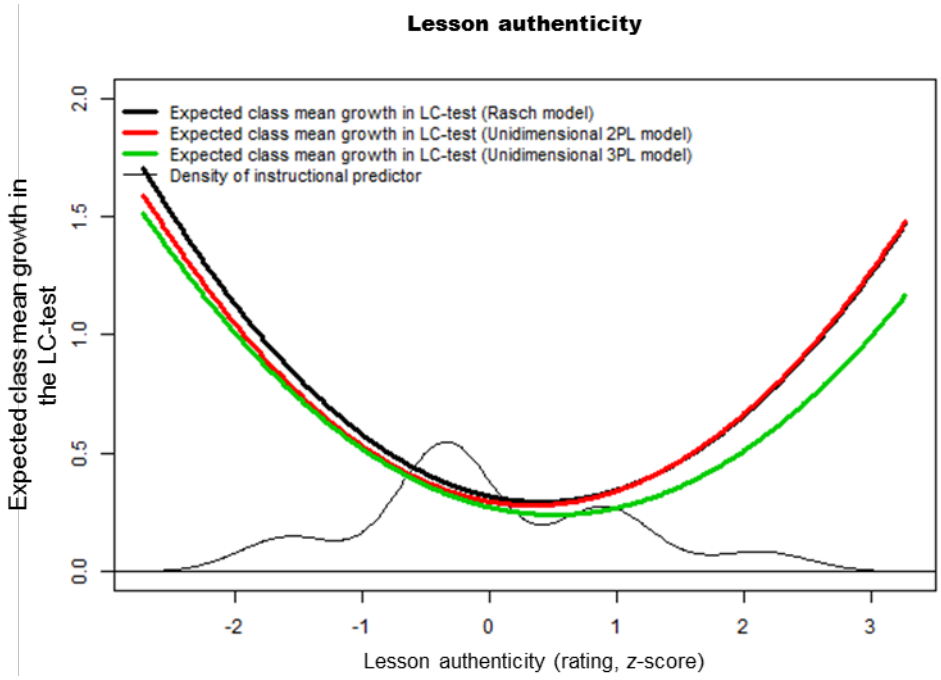

Figure 35: Relationship between "Lesson authenticity" and class mean growth in the LC-test

Similarly, the nonlinear effect of "Narrow focused monitoring" on student growth in the LC-test was also not anticipated. Instructions which were highly narrow-focused and instructions which were not narrow-focused were all associated with higher student growth in the LC-test. That suggested that "Narrow focused monitoring" per se was neither a negative nor a positive instructional predictor in EFL, but that its effect depended on other factors.

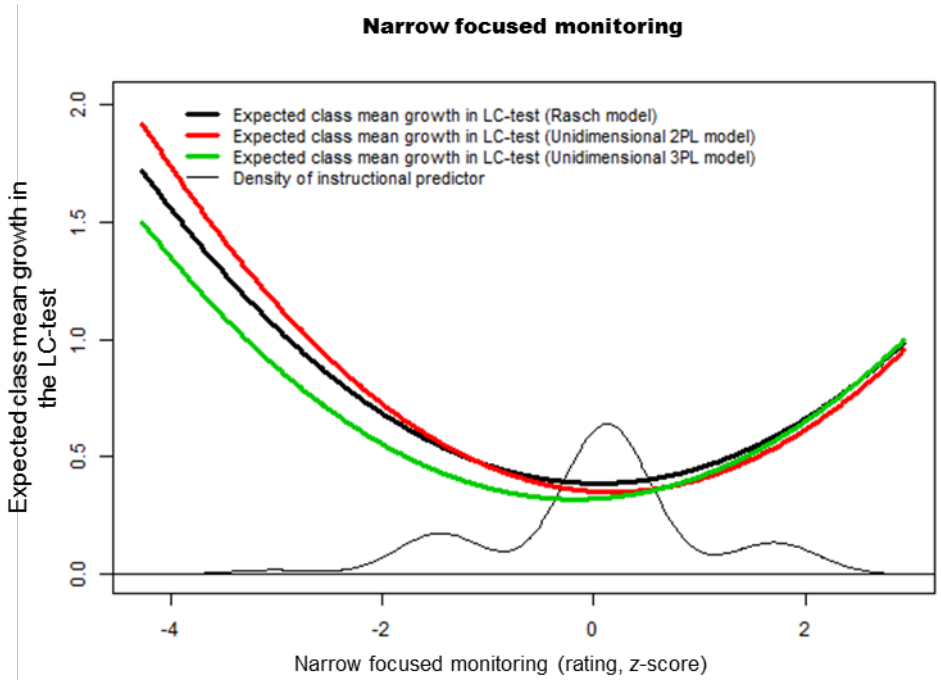

Figure 36: Relationship between “Narrow focused monitoring” (rating) and class mean growth in the LC-test

The relationship between “Teaching objective: Involvement of as many students as possible (rating)” and class mean growth in the LC-test suggested that maximum growth was not expected at the maximum value of the predictor, but that the optimum value was approximately equal to the mean value in this study (see Figure 37).

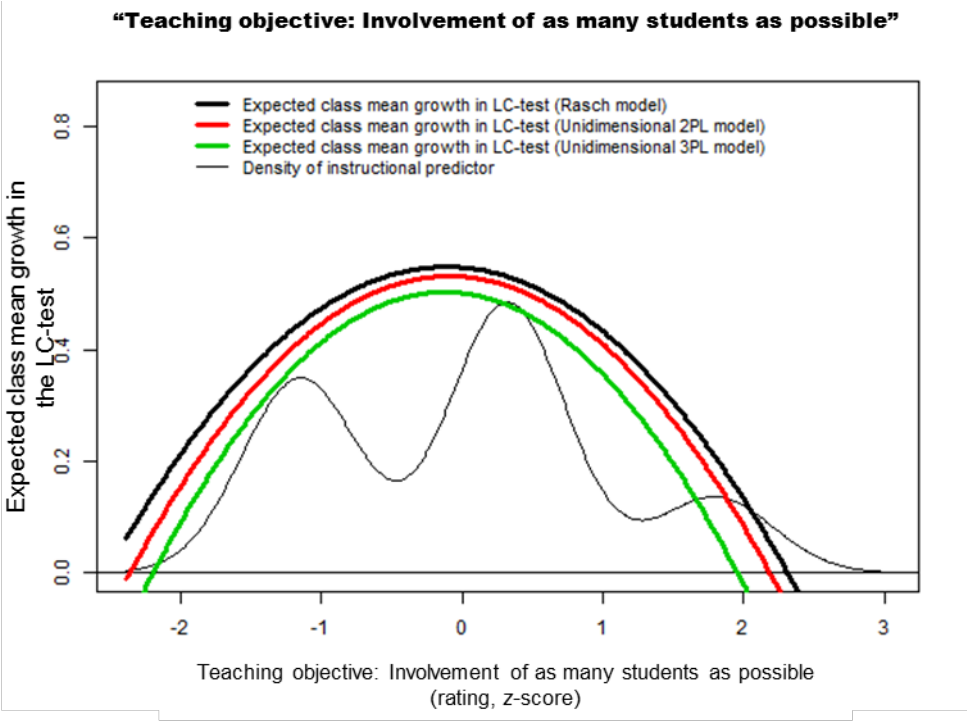

Figure 37: Relationship between "Teaching objective: Involvement of as many students as possible" (rating) and class mean growth in the LC-test

IX.3.3 Chapter summary and answer to Research Question 11

Research Question 11. *To what extent is the assumption of nonlinear relationships between instructional factors and student progress confirmed?*

The relationship between several instructional factors and student growth was found to be nonlinear in both tests, which meant the formulation "the more, the better" or "the less, the worse" did not apply for all variables and all cases. To enhance student growth based on the C-test, it was neither optimal to not use Vietnamese at all in transitions nor to use Vietnamese frequently in transitions. The same applied for the relationship between "Student reading out own text in English" (relative frequency) and student growth in the C-test. The nonlinear relationship between these two factors and student growth in the C-test was confirmed by both the nonzero coefficient of the quadratic term via strong hierarchical lasso regression models and the statistically significant *F*-test when comparing the OLS regression models.

The results regarding the nonlinear relationship between student growth in the LC-test and instructional factors were, to some extent, unexpected and difficult to interpret. The hypothesized nonlinear relationship was confirmed for the rating variable "Teaching objective: Involvement of as many students as possible" by both the lasso regression and the *F*-test results. Two other instructional factors – "Lesson authenticity" (rating) and "Narrow focused monitoring" (rating) – were also found to have a nonlinear

relationship with class mean growth in the LC-test based on both the hierarchical lasso regression and F-test results. However, their relationships with student growth were not really hypothesized and anticipated. The mean value of the factors corresponded approximately to the minimum expected class mean growth in the LC-test. A possible interpretation could be that these results are due to the cultural and social differences between everyday life in Vietnam and in western English-speaking countries (associated with the variable "Lesson authenticity"), due to confounding effects which were absent in the explanation model, or due to sparse data for the low and high values of the predictors and, hence, interpretation should be suspended. Another factor, "Total teacher speaking time" (time percentage), was suggested to have a nonlinear relationship with the class mean growth in the LC-test based on the lasso regression results, but not based on the *F*-test results ($.05 < p < .10$). In contrast, "Time used for social activities" (time percentage) was only suggested to have a nonlinear relationship with the class mean growth in the LC-test based on the *F*-test results.

In brief, the assumption of nonlinear relationships between instructional factors and student progress was confirmed for a few factors in both tests. Once again, differences associated with different scaling models in both tests were found.

IX.4 Joint effect of instructional factors on student growth

To examine the joint effects of instructional factors on student growth at class level, a regression model for each dependent variable was applied (see Chapter VI.3.4). As model predictors, we included all instructional factors which had been identified to have an important linear and nonlinear effect on student growth based on lasso regression results (see Section IX.2, IX.3). In each model, the quadratic term of the model predictors as well as all interactions between them were also included as model predictors. Regarding student growth in the C-test, the selected model covariate was initial class achievement; regarding student growth in the LC-test, it was class mean SES (c.f. Chapter VIII, and Section IX.2). In addition, in order to investigate the aptitude treatment interaction effect with regard to classes with different initial achievement in the LC-test, the initial class achievement in the LC-test was also included as a model covariate in the explanation model for class mean growth in the LC-test.

Due to the large number of predictors in the model (for student growth in the C-test: 27 predictors, for student growth in the LC-test: 44 predictors) together with the small number of classes ($N = 50$), only (strong hierarchical) lasso regression models were executed to investigate the joint effect of instructional factors on student growth. Using an OLS multiple regression model would lead to difficulty in differentiating between true relationships and noise as well as to overfitting problems (see Chapter VI.3.4).

IX.4.1 Joint effect of instructional factors on student growth in the C-test

In previous sections, three instructional factors were identified as having a significant linear effect and two others were identified as having a significant nonlinear effect on class mean growth in the C-test. They were “Encouragement of student statements”, “Teacher speaking time in mixed language”, “Affectively stressed positive feedback”, “Student reading out own text in English”, and “Teacher speaking time using Vietnamese in transitions.” These five instructional factors together with initial class achievement were included as predictors in an explanation model for student growth in the C-test. Their main effect (linear, additive) together with all quadratic terms (nonlinear effect) and interactions between each of them as pairs (interaction) were investigated via the strong hierarchical lasso model. The results are shown in Table 31 below.

Table 31: Joint effect of instructional factors and initial class achievement on class mean growth in the C-test

	Scaling model							
	M1		M2		M1T		M2T	
	β_{lz}	N nonzero	β_{lz}	N nonzero	β_{lz}	N nonzero	β_{lz}	N nonzero
Additive effects								
Initial class achievement	–.32	10	–.22	10	–.14	10	0	0
Encouragement of student statements	.20	10	.19	10	.19	10	0	0
Teacher speaking time in mixed languages	–.28	10	–.25	10	–.18	10	0	0
Affectively stressed positive feedback	.21	10	.24	10	.16	10	0	0
Student reading out own text in English (relative frequency)	–.05	10	–.08	10	–.04	9	0	0
Teacher speaking time using Vietnamese in transitions	.11	10	.08	10	.01	9	0	0
Nonlinear effects	β_{lq}	N nonzero	β_{lq}	N nonzero	β_{lq}	N nonzero	β_{lq}	N nonzero
Initial class achievement	.12	10	.09	10	.07	6	0	0
Encouragement of student statements	.03	7	.03	6	.05	7	0	0
Teacher speaking time in mixed languages	0	0	.00	2	.00	1	0	0
Affectively stressed positive feedback	0	0	.00	1	.00	2	0	0
Student reading out own text in English (relative frequency)	–.07	10	–.06	9	–.05	7	0	0
Teacher speaking time using Vietnamese in transitions	–.11	10	–.05	9	–.06	9	0	0
Interaction effects	β_{li}	N nonzero	β_{li}	N nonzero	β_{li}	N nonzero	β_{li}	N nonzero
With initial class achievement								
Encouragement of student statements	–.04	6	–.02	6	–.04	6	0	0

	Scaling model							
	M1		M2		M1T		M2T	
Teacher speaking time in mixed languages	–.09	10	–.05	10	–.02	3	0	0
Affectively stressed positive feedback	–.05	9	–.06	10	–.01	3	0	0
Student reading out own text in English	.00	1	0	0	.00	1	0	0
With "Encouragement of student statements"								
Teacher speaking time in mixed languages	–.04	7	–.02	6	.00	0	0	0
Affectively stressed positive feedback	.02	4	.02	5	.01	2	0	0
Student reading out own text in English	.00	2	0	0	0	0	0	0
Teacher speaking time using Vietnamese in transitions	.00	2	0	0	–.01	2	0	0
With "Teacher speaking time in mixed languages"								
Affectively stressed positive feedback	–.14	10	–.16	10	–.12	10	0	0
Student reading out own text in English	.00	1	.00	1	.00	1	0	0
With "Affectively stressed positive feedback"								
Student reading out own text in English (relative frequency)	0	0	0	0	.01	1	0	0
With "Student reading out own text in English (relative frequency)"								
Teacher speaking time using Vietnamese in transitions	.05	9	.03	8	.00	4	0	0
R_l^2	.35		.30		.24		0	

Note: M1 = Rasch model, M2 = unidimensional 2PL model, M1T = testlet 1PL model, M2T = testlet 2PL model. β_{lz} = lasso coefficient of the linear terms, β_{lq} = lasso coefficient of the quadratic terms, β_{li} = lasso coefficient of the interactions. *N* nonzero = number of imputed datasets with nonzero lasso regression coefficients of the corresponding variable. Interactions with all zero-coefficients are not shown. R_l^2 = proportion of explained variance of the dependent variable by all model predictors together based on the strong hierarchical lasso regression model.

Similarly to in previous sections, all regression coefficients were zero if student ability estimates in the C-test were estimated via the testlet 2PL model. Based on student growth measured via the unidimensional scaling models and the testlet 1PL model, the lasso regression coefficients of the model predictors were identical in sign, similar in absolute value, and were all nonzero. In general, the absolute values of the regression coefficients were smaller with regard to a dependent variable estimated via the testlet 1PL model than one estimated via the unidimensional models.

The main effects of all selected instructional predictors were nonzero; the highest absolute coefficient was found for "Teacher speaking time in mixed languages", followed by "Encouragement of student

statements" and "Affectively stressed positive feedback." Their effect was as large as the main effect of initial class achievement. Taking the sign of these effects into account, the results revealed that a positive effect of the indicators of the quality of motivation ("Encouragement of student statements" and "Affectively stressed positive feedback") could, to a large extent, compensate for the negative effect caused by "Teacher speaking time in mixed languages."

The results again confirmed a considerable nonlinear effect of "Student reading out own text in English (relative frequency)" and "Teacher speaking time using Vietnamese in transitions" on class mean growth in the C-test. The maximum class mean growth in the C-test was not expected to be the maximum value of these two variables, but it was expected to be an optimum value between 0.5 and 1 (*z*-score). Exceeding this optimum value, no further improvement, even a decline, in student growth was expected. In addition, a nonlinear effect of initial class achievement was suggested by the model results; classes with low or high achievement at T1 progressed more than classes with average initial achievement. The relationship between "Encouragement of student statements" and class mean growth in the C-test was also shown to be a flat curve, where a positive effect was observed when the predictor was higher than or equal to 1 (*z*-score, equivalent to rating category 2).

For simplicity, in order to understand the joint effect between two variables, an OLS regression model was applied.

The highest interaction effect was found between "Teacher speaking time in mixed languages" and "Affectively stressed positive feedback", which did not correlate with each other ($r = –.06, p = .70$). By applying an OLS regression model with these two variables and their interaction as model predictors, 23% of variance in class mean growth in the C-test estimated via the Rasch model, 25% via the unidimensional 2PL model, and 13% via the testlet 1PL model was explained. To illustrate this, the expected class mean growth in the C-test associated with the Rasch model based on the joint effect of these variables via the OLS regression results is depicted in Figure 38.

For teachers who did not give affectively stressed positive feedback at all, the model-based expected class mean growth in the C-test was around 0.4 points of the ability scale (average growth, c.f. Chapter VIII.1), regardless of how frequently they used mixed languages in lessons. For teachers who gave affectively stressed positive feedback most frequently and did not used mixed languages, expected class mean growth in the C-test was highest, around 0.8 scale points or more. On the other hand, for teachers who most frequently gave affectively stressed positive feedback and also used mixed languages most frequently, the lowest growth in the C-test was expected for their classes (approaching zero). The classes with teachers who did not (often) use mixed languages and gave more affectively stressed positive

feedback progressed more. Figure 38 shows that giving affectively stressed positive feedback was only associated with a positive effect on student growth if the teacher did not often used mixed languages in lessons.

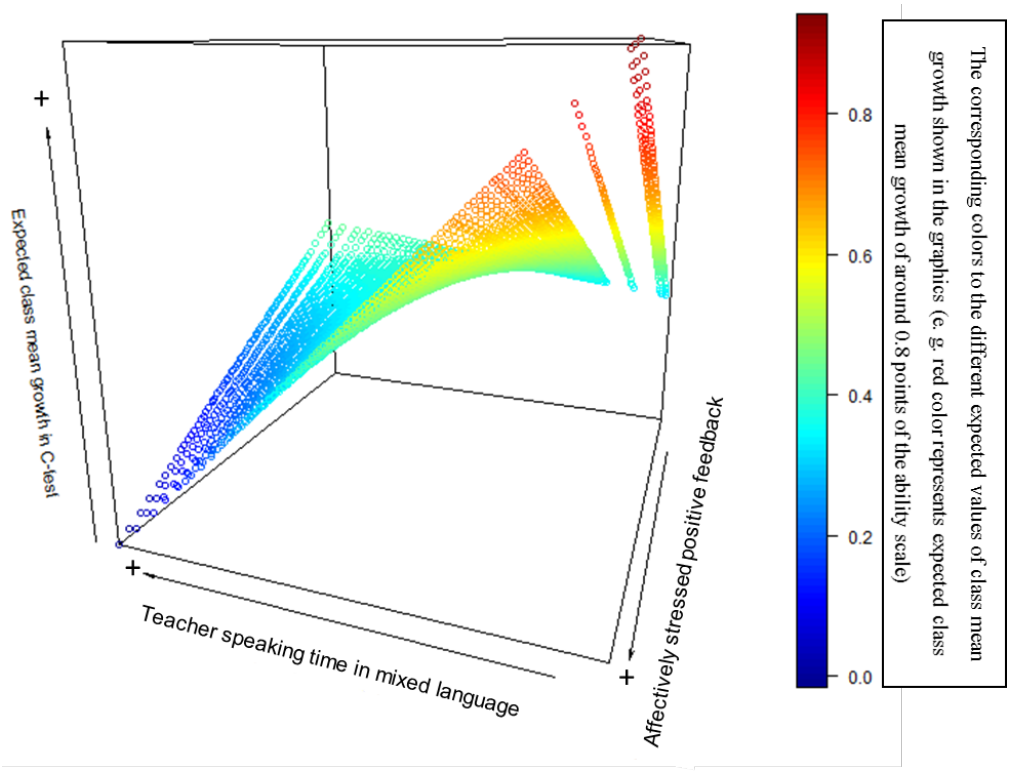

Figure 38: Expected class mean growth in the C-test (Rasch model) based on the joint effect of "Teacher speaking time in mixed languages (time percentage)" and "Affectively stressed positive feedback (relative frequency)" taking their interaction into account.

Note: this figure was created using R package plot3D (Soetaert, 2016)

A similar relationship was also found between "Teacher speaking time in mixed languages (time percentage)" and another motivation aspect of teaching – "Encouragement of student statements", but it was smaller.

Other considerable interaction effects were also revealed, such as between initial class achievement and "Teacher speaking time in mixed languages (time percentage)", "Affectively stressed positive feedback (relative frequency)" (not confirmed by the testlet model), and "Encouragement of student statements" (not confirmed in all imputed datasets); or between "Student reading out own text in English (relative frequency)" and "Teacher speaking time using Vietnamese in transitions" (not confirmed by the testlet model or in all imputed datasets).

There was a small negative (not statistically significant) correlation ($r = -.19, p = .11$) between "Teacher speaking time in mixed languages" and initial class achievement in the C-test. It indicated that teachers in classes with lower initial achievement used mixed languages a little more often. The expected class mean growth in the C-test (Rasch model) based on the joint effect of these two variables (main additive effects together with interaction between them and the quadratic term of initial class achievement) via an OLS regression model (R^2 = 27%) is illustrated in Figure 39.

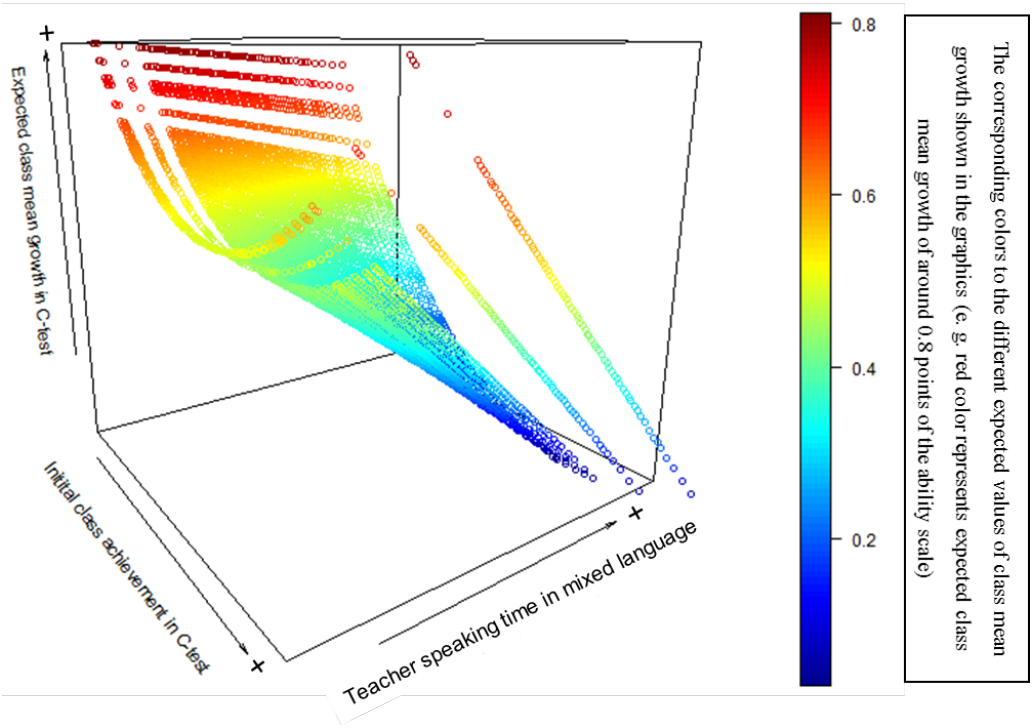

Figure 39: Expected class mean growth in the C-test (Rasch model) based on joint effect of initial class achievement in the C-test (Rasch model, including quadratic term) and "Teacher speaking time in mixed languages (time percentage)" taking their interaction into account.

Note: this figure was created using R package plot3D (Soetaert, 2016)

For all initial achievement levels, classes with teachers who more often used mixed languages progressed less. This negative effect of using mixed languages was larger in classes with higher initial achievement. That means that classes with higher initial achievement suffered more from having an EFL teacher who often used mixed languages than classes with lower initial achievement. However, only a few classes in this study were located in the lower right-hand corner in Figure 39 due to the fact that teachers in classes with high initial achievement used mixed languages in lessons less often than in classes with low initial achievement. Hence, the results should be interpreted with caution.

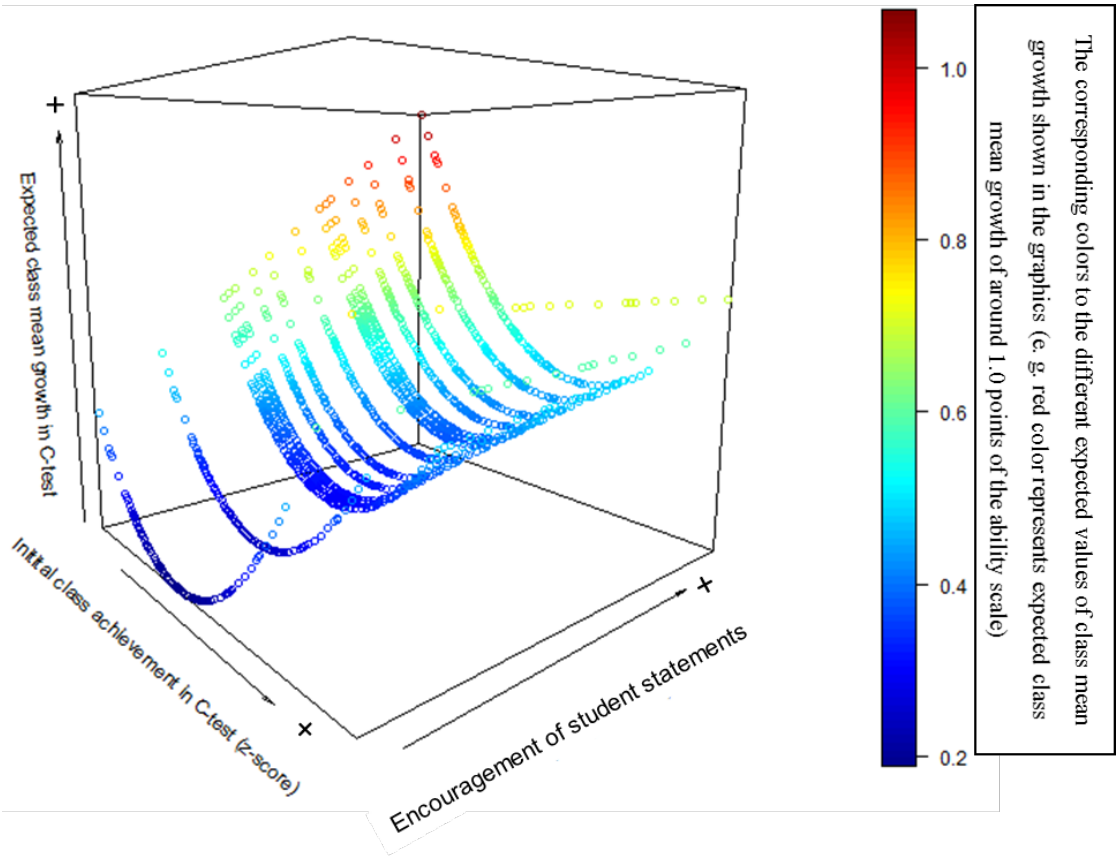

Figure 40: Expected class mean growth in the C-test (Rasch model) based on the joint effect of initial class achievement in the C-test (Rasch model, including the quadratic term) and "Encouragement of student statements (rating)" taking their interaction into account.

Note: this figure was created using R package plot3D (Soetaert, 2016)

Taking the interaction effect with initial class achievement into account, the positive effect of "Affectively stressed positive feedback" and "Encouragement of student statements" was larger for classes with *lower initial achievement* and smaller or even negative for classes with high initial achievement (see Figure 40 and Figure 41). That means that classes with lower initial achievement profited more from being motivated and encouraged by teachers. For classes with high initial achievement, being motivated and encouraged more by teachers did not help them to progress more in EFL. Classes with high initial achievement progressed most when their teacher did not give affectively stressed positive feedback at all and progressed less when they were given a high amount of such feedback in lessons.

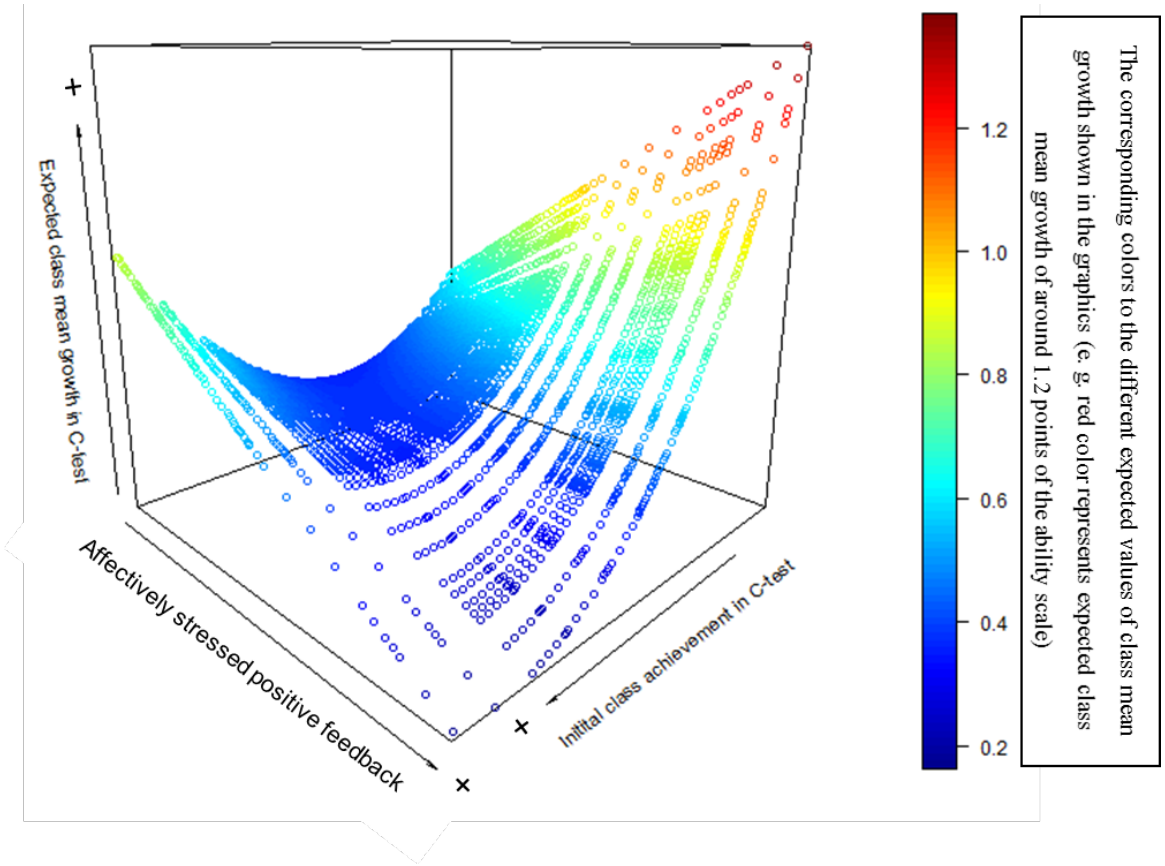

Figure 41: Expected class mean growth in the C-test (Rasch model) based on the joint effect of initial class achievement in the C-test (Rasch model, including the quadratic term) and "Affectively stressed positive feedback (relative frequency)" taking their interaction into account.

Note: this figure was created using R package plot3D (Soetaert, 2016)

The joint effect of "Student reading out own text in English (relative frequency)" and "Teacher speaking time using Vietnamese in transitions (time percentage)" points toward two tips for effective teaching of EFL (see Figure 42; not confirmed when the testlet model was used to calculate test results). First, not using Vietnamese at all in transitions is not helpful for enhancing student growth in the C-test. And, second, highest growth is not expected for classes in which students read out their own texts in English most frequently or when teachers most often use Vietnamese in transitions. Maximum progress was achieved in this study by classes with a circa 6% relative frequency of students reading their own texts and in which the teacher spoke Vietnamese in ca. 10% of his or her speaking time in transitions.

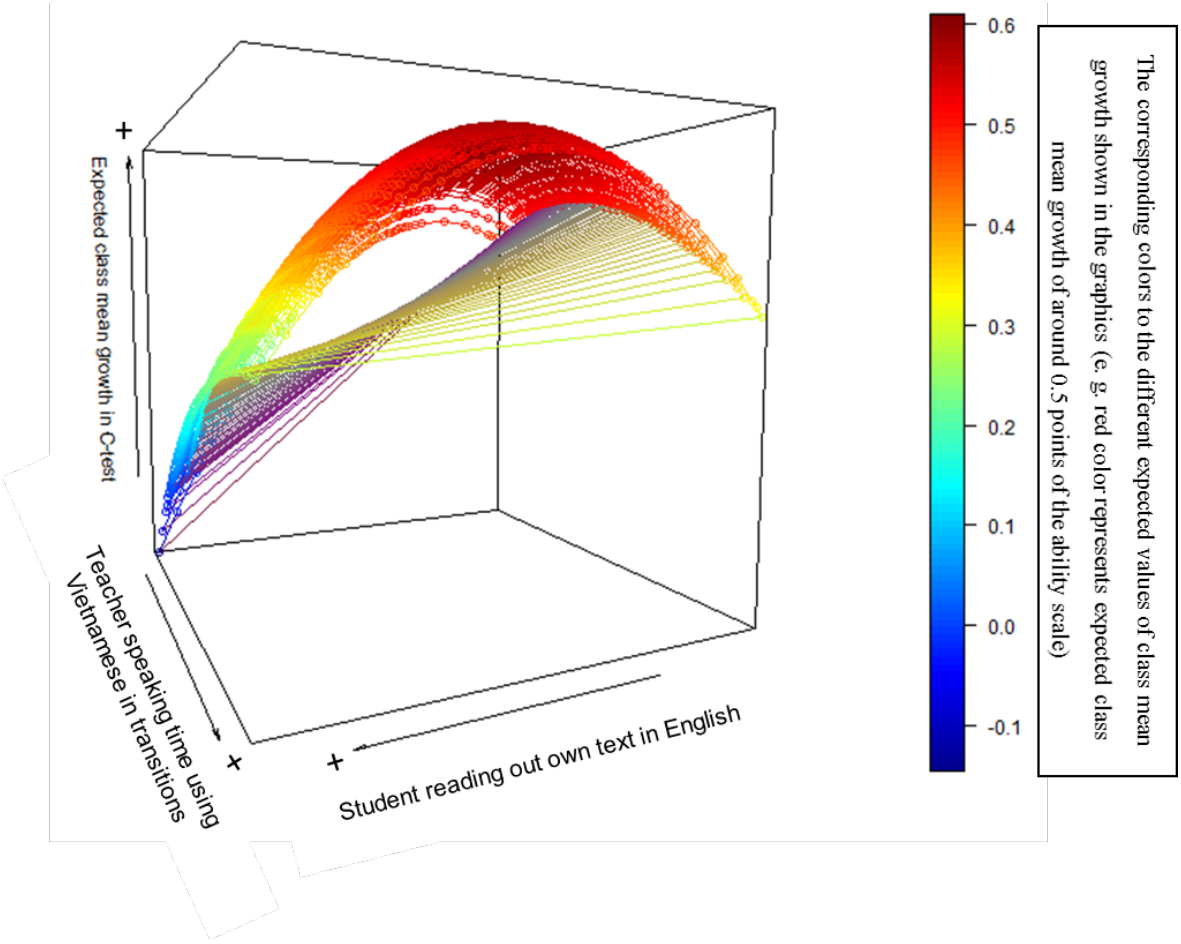

Figure 42: Expected class mean growth in the C-test (Rasch model) based on the joint effect of "Student reading out own text in English (relative frequency)" and "Teacher speaking time using Vietnamese in transitions (time percentage)" including their quadratic terms and interaction.

Note: this figure was created using R package plot3D (Soetaert, 2016). Lines were added just for better readability.

Overall, the joint effect of all predictors together via the hierarchical lasso regression model (see Table 31 above) explained 35% of variance in class mean growth in the C-test calculated using the Rasch model, 30% using the unidimensional 2PL model, 24% using the testlet 1PL model, and it did not explain any differences in class mean growth in the C-test when using the testlet 2PL model.

IX.4.2 Joint effect of instructional factors on student growth in the LC-test

The joint effect of instructional factors on class mean growth in the LC-test was analyzed using the same method as in the previous chapter. Together with the class mean SES and initial class achievement as model covariates, the instructional factors with significant linear and nonlinear effects on class mean growth in the C-test based on the results of lasso regression analyses were included as predictors in one strong hierarchical lasso regression model. They were "Teacher speaking time using Vietnamese in

transitions", "Student speaking time in mixed languages", "Relative frequency of repeated questions", "Lesson authenticity", "Teaching objective: Involvement of as many students as possible", "Total teacher speaking time", and "Narrow focused monitoring." Not only their main effects but also their quadratic terms and interactions were modeled. The results are shown in Table 32.

Table 32: Joint effect of instructional factors, class mean SES and initial class achievement on class mean growth in the LC-test

	M1		M2		M3	
	β_{lz}	N nonzero	β_{lz}	N nonzero	β_{lz}	N nonzero
Additive effect						
Initial class achievement	–.01	7	.01	4	–.01	2
Class mean SES	.04	7	.04	7	.02	6
Teacher speaking time using Vietnamese in transitions	–.23	10	–.19	10	–.17	10
Student speaking time in mixed languages	–.05	10	–.07	10	–.10	10
Relative frequency of repeated questions	–.15	10	–.15	10	–.21	10
Lesson authenticity	–.22	10	–.19	10	–.23	10
Teaching objective: Involvement of as many students as possible	.01	9	–.01	5	.00	9
Total teacher speaking time	.01	8	.00	5	.01	6
Narrow focused monitoring	–.10	10	–.15	10	.00	10
Nonlinear effect	β_{lq}	N non–zero	β_{lq}	N non–zero	β_{lq}	N non–zero
Initial class achievement	–.01	6	–.01	3	.00	1
Class mean SES	–.01	2	–.01	2	.00	1
Teacher speaking time using Vietnamese in transitions	.00	1	0	0	.00	1
Student speaking time in mixed languages	–.04	10	–.05	10	–.09	10
Relative frequency of repeated questions	–.06	9	–.09	10	–.13	10
Lesson authenticity	.14	10	.12	10	.11	10
Teaching objective: Involvement of as many students as possible	.00	2	–.01	3	–.01	3
Total teacher speaking time	.02	7	.02	4	.01	5
Narrow focused monitoring	.04	6	.06	8	.07	9
Interaction	β_{li}	N non–zero	β_{li}	N non–zero	β_{li}	N non–zero
With initial class achievement						
Student speaking time in mixed languages	0	0	0	0	.00	1
Lesson authenticity	0	0	0	0	.01	2
Total teacher speaking time	.00	2	.00	1	0	0
With class mean SES						
Teacher speaking time using Vietnamese in transitions	.00	1	0	0	0	0
Student speaking time in mixed languages	0	0	.01	1	.00	1
Lesson authenticity	.01	4	.01	2	.01	6
Teaching objective: Involvement of as many students as possible	.00	1	0	0	0	0
Narrow focused monitoring	.00	1	0	0	0	0
With "Teacher speaking time using Vietnamese in transitions"						
Student speaking time in mixed languages	.00	2	.00	1	.00	2
Relative frequency of repeated questions	.00	1	.01	3	.00	1

	M1		M2		M3	
Teaching objective: Involvement of as many students as possible	–.03	8	–.01	4	–.03	9
Narrow focused monitoring	.18	10	.11	10	.09	9
With "Student speaking time in mixed languages"						
Relative frequency of repeated questions	.00	1	.00	1	.01	2
Lesson authenticity	.00	1	0	0	0	0
With "Relative frequency of repeated questions"						
Lesson authenticity	.06	8	.04	8	.07	9
Narrow focused monitoring	–.02	5	–.01	2	–.01	3
With "Lesson authenticity"						
Total teacher speaking time	0	0	0	0	.00	1
Narrow focused monitoring	.00	1	–.01	2	–.01	1
R_l^2	.31		.30		.35	

Note: M1 = Rasch model, M2 = unidimensional 2PL model, M3 = unidimensional 3PL model. β_{lz} = lasso coefficient of the linear terms, β_{lq} = lasso coefficient of the quadratic terms, β_{li} = lasso coefficient of the interactions. *N* nonzero = number of imputed datasets with nonzero lasso regression coefficients of the corresponding variable. Interactions with all zero-coefficients are not shown. R_l^2 = proportion of explained variance of dependent variable by all model predictors together based on the strong hierarchical lasso regression model.

The regression results confirmed that initial class achievement did not have a considerable effect on class mean growth in the LC-test. The lasso coefficients of both the linear and nonlinear effects were zero or near to zero depending on the imputed dataset, regardless of which scaling model was used to generate the test results. The linear effect of SES on class mean growth in the LC-test was also small, and a nonlinear effect was not confirmed based on all scaling models.

There were differences in the size of the regression coefficients of the predictors with regard to class mean growth calculated using different scaling models. The largest difference with regard to different scaling models was found for the estimated effect of "Narrow focused monitoring."

On the other hand, a significant linear effect was confirmed for "Teacher speaking time using Vietnamese in transitions", "Student speaking time in mixed languages", "Relative frequency of repeated questions", and "Lesson authenticity" across all imputed datasets and scaling models. For the latter three factors, a nonlinear effect was also found. A nonlinear effect of "Narrow focused monitoring" was also confirmed but not by all imputed datasets. In the joint model with other effects taken into account, the nonlinear effect of "Total teacher speaking time" was negligible, and a nonlinear effect of "Teaching objective: Involvement of as many students as possible" was no longer confirmed. For the last-mentioned factor, an interaction effect with "Teacher speaking time using Vietnamese in transitions" was revealed instead (but negligible if the unidimensional 2PL was used as the scaling model).

For simplicity purposes, to understand the interaction and joint effect between two variables, an OLS regression analysis was applied with their main effects (linear and nonlinear if they were significant, as shown in Table 32) and the interaction between them as predictors. “Teaching objective: Involvement of as many students as possible” and “Teacher speaking time using Vietnamese in transitions” jointly explained R^2 = 18% of variance of the dependent variable calculated based on the Rasch model, R^2 = 20% based on the unidimensional 2PL model, and R^2 = 19% based on the unidimensional 3PL model. The effect of each variable was dependent on the level of the other variable (see Figure 43). Of classes with teachers who used Vietnamese in transitions less often, those with teachers who made an effort to involve as many students as possible in lesson conversations were expected to show higher growth. In contrast, if a teacher often used Vietnamese in transitions, involving as many students as possible did not prove to be advantageous for student progress. Accordingly, using Vietnamese in transitions only had a negative effect on student growth in classes in which teachers involved as many students as possible in class conversations. In classes with a low rating regarding “Teaching objective: Involvement of as many students as possible”, no effect was found regarding using Vietnamese in transitions. Note, however, that these two variables correlated positively with each other ($r = .35, p = .006$): Teachers in this study who used Vietnamese more often in transitions tended to involve more students in lesson conversations and the other way around.

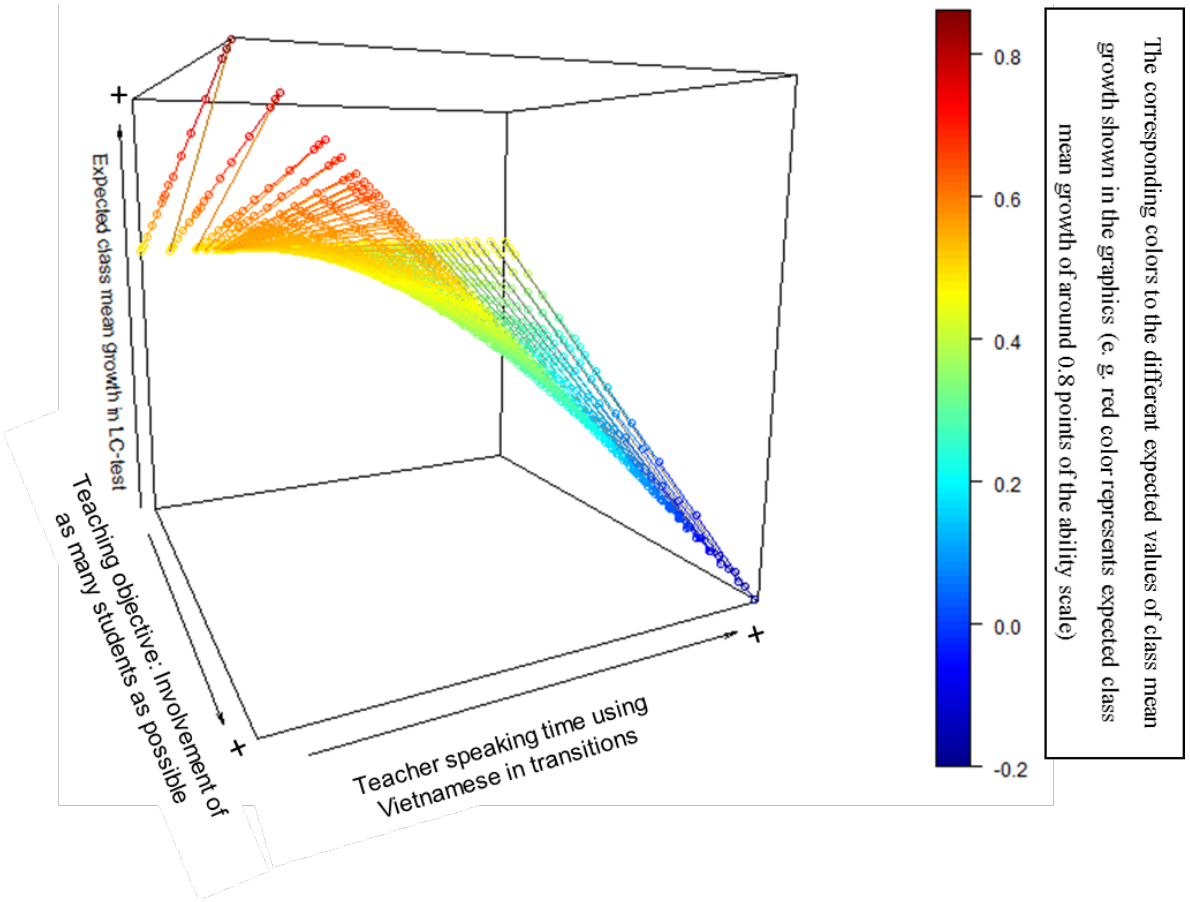

Figure 43: Expected class mean growth in the LC-test (Rasch model) based on the joint effect of "Teaching objective: Involvement of as many students as possible (rating variable)" and "Teacher speaking time using Vietnamese in transitions (time percentage)".

Note: this figure was created using R package plot3D (Soetaert, 2016)

Considerable interaction effects between the instructional factors and context variables (covariates) were not found. The most noteworthy interaction effects on class mean growth in the LC-test were found between "Teacher speaking time using Vietnamese in transitions" and "Narrow focused monitoring" as well as between "Relative frequency of repeated questions" and "Lesson authenticity."

According to the results of the OLS regression model, "Teacher speaking time using Vietnamese in transitions" and "Narrow focused monitoring" jointly explained 29% of variance in class mean growth in the LC-test when using the Rasch model, 31% when using the unidimensional 2PL model, and 23% when using the unidimensional 3PL model. Figure 44 shows that the effect of one variable depended on the level of the other variable. In classes in which teachers used Vietnamese less often in transitions, higher growth was expected in the LC-test for classes with a lower rating regarding "Narrow focused monitoring." In contrast, classes with teachers who frequently used Vietnamese in transitions, higher growth was expected when teachers monitored lessons with a narrow focus.

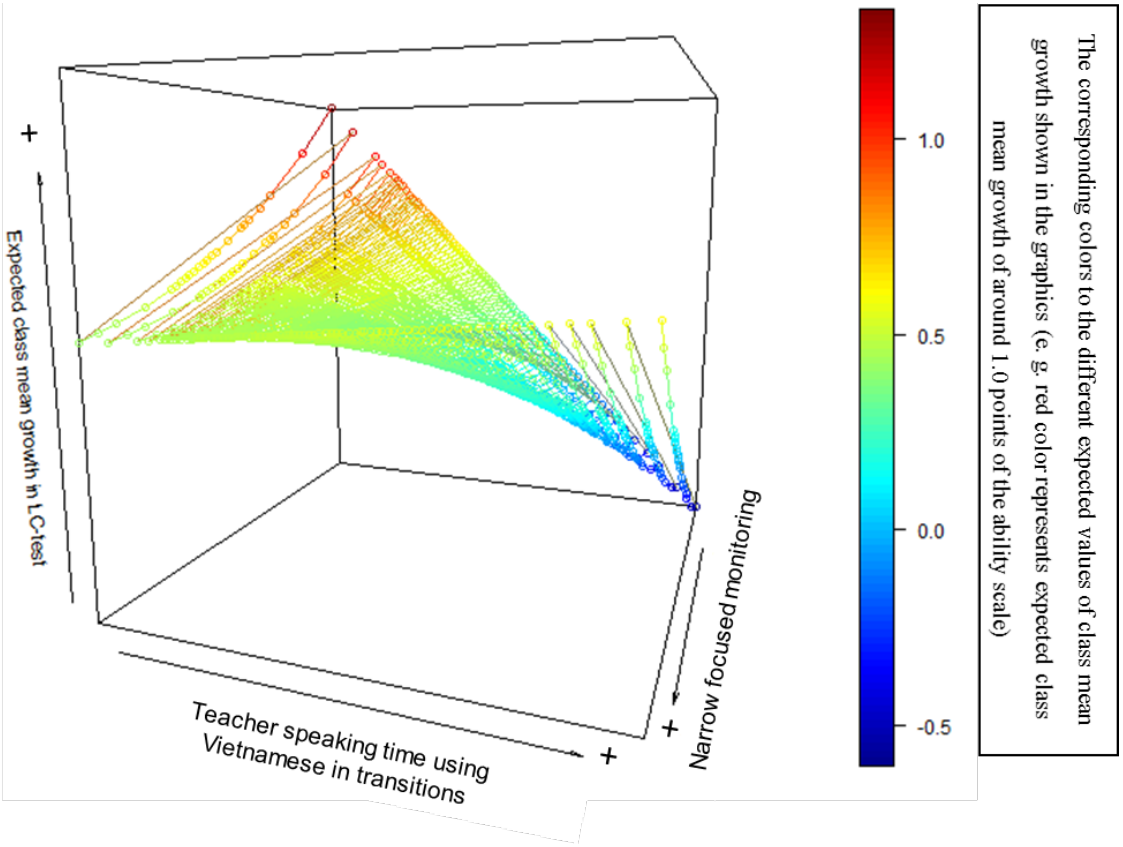

Figure 44: Expected class mean growth in the LC-test (Rasch model) based on the joint effect of "Narrow focused monitoring (rating variable)" (with both the linear and quadratic term) and "Teacher speaking time using Vietnamese in transitions (time percentage)" including interaction.

Note: this figure was created using R package plot3D (Soetaert, 2016)

Correspondingly, in lessons with less narrow focused monitoring, higher class mean growth was expected for teachers who used Vietnamese less often in transitions. The negative effect of using Vietnamese in transitions was not found in classes with the highest rating regarding "Narrow focused monitoring." The highest growth was expected for classes with the least narrow focused monitoring and, simultaneously, classes in which teachers who used Vietnamese in transitions the least often.

Both "Lesson authenticity (rating)" and "Repeated question (relative frequency)" had a nonlinear relationship with student growth in the LC-test (see Figure 45). Jointly, including their interaction and taking into account their nonlinear relationship with the dependent variable, these variables explained 31% of variance in student growth in the LC-test based on the Rasch model, 34% based on the unidimensional 2PL model, and 39% based on the unidimensional 3PL model. Repeated questions were an effective teaching tool, providing that teachers did not overuse them.

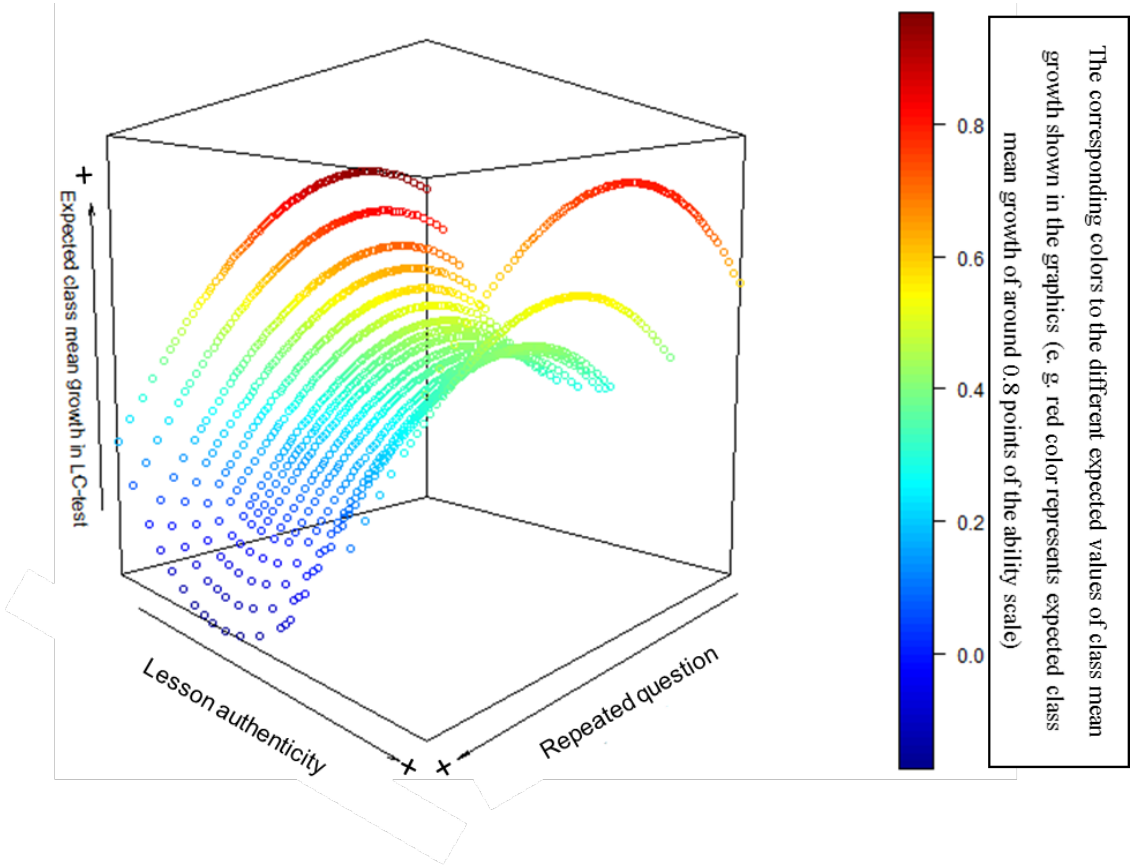

Figure 45: Expected class mean growth in the LC-test (Rasch model) based on the joint effect of "Lesson authenticity (rating variable)" and "Repeated questions (relative frequency)" including their quadratic terms and interaction.

Note: this figure was created using R package plot3D (Soetaert, 2016)

All predictors together explained between 30%–35% of the class mean growth variance in the LC-test depending on the scaling model chosen to calculate the test scores based on the hierarchical lasso regression model.

IX.4.3 Chapter summary and answer to Research Questions 12–14

Analyses based on strong hierarchical lasso regression models were implemented to investigate the hypothesized interaction effects, aptitude treatment effects as well as the compensatory effects of selected instructional factors on student growth in both tests. Based on the analyses' results, Research Questions 12–14 can be answered as follows:

Research Question 12. *Is there empirical evidence supporting the assumption of interaction effects between instructional factors on student growth?*

One possible reason for the observed nonlinear relationship between some instructional effects and student growth might be partly the interaction effect between them. For instance, when students were encouraged and motivated more through affectively stressed positive feedback, it was only beneficial for the learning progress of students in the C-test if the teachers did not often used mixed languages in lessons.

With regard to student growth in the LC-test, "Teacher speaking time using Vietnamese in transitions" was found to be a negative predictor, which seemed to be even more harmful in classes in which teachers made more effort to involve as many students as possible in the lesson conversation. For classes with high ratings of "Narrow focused monitoring" or classes in which teachers did not involve many students in class conversations, the use of the Vietnamese language in transitions (by the teacher) showed no negative effect.

Research Question 13. *Is there empirical evidence supporting the assumption of the aptitude treatment interaction effect of classroom instruction?*

This assumption was also confirmed by the analysis results regarding student growth in the C-test. In general, a teacher's frequent use of mixed languages in lessons was associated with lower class mean growth in the C-test, which was especially harmful for classes with high initial achievement. When students were encouraged to make statements in lessons and motivated through affectively stressed positive feedback, it was found to be beneficial for student growth in the C-test in classes with low initial achievement but not in classes with high initial achievement. For the LC-test, no considerably differentiated instructional effects were found on student growth with reference to initial class achievement and class mean SES.

Research Question 14. *Do the findings confirm the assumption of the compensatory effect of instructional factors on student progress?*

As mentioned earlier, each single instructional factor only explained a small amount of the variance in student growth in both tests. But jointly, the factors explained up to 35% of variance in class mean growth in the C-test and LC-test. Because the effects of different instructional predictors on student growth were all nonzero in the presence of other predictors and covariates, the hypothesized compensatory effect of instructional factors was confirmed. To enhance student growth in the C-test,

for instance, encouraging students and giving positive feedback (with emphasized affective affirmation) could, to some extent, compensate other negative effects. It is worth noting that the joint effect of the instructional factors was not a simple additive formula due to the existence of interaction effects (see above).

Of all the important instructional effects on student growth, none revealed characteristics of instruction that promoted student growth in classes with high initial student ability. All instructional effects on student growth in the LC-test were negative on average; whereas quality of motivation in instruction enhanced student growth in the C-test, these effects held only for classes with low initial ability. One possible explanation could be the limitations of the central curriculum and textbook, which would have been rather easy for high achievers and, thus, could not enable optimal growth of these students/classes, regardless of how good the instruction was. This presumption is supported by the negative correlation (although it was not statistically significant according to the results based on all scaling models in both tests) between initial student ability and growth at individual level in the C-test.

X. Relevance of scaling model selection on study results

In this study , different scaling models were applied to estimate the test results. In Chapters VIII and IX, similarities and differences associated with the different scaling models regarding the estimates of student achievement and growth as well as the instructional effects on student progress were observed. In this chapter, these results will be discussed in the context of the last research question: “To what extent is the estimation of instructional effects on academic student progress independent of the selection of a specific scaling model?”

X.1 Validity of the test scores with regard to different scaling models

In order to gain validity indicators of the test scores (see Chapter VI.2.4.3), we calculated correlations between student ability estimates based on two tests and different scaling models and midterm school marks in the school subjects English, mathematics, and Vietnamese, which are presented and discussed in this section (see Table 33).

The individual test scores of both tests at both MPs correlated more positively and significantly (no overlap between confidence intervals) with the midterm school marks in English than in mathematics and Vietnamese, regardless of the scaling model applied. This supports the assumption that the English tests in this study measure English ability rather than only intelligence (suggested by the school mark in mathematics) or general language competence (suggested by the school mark in Vietnamese).

Furthermore, the correlation coefficients between English school marks and the test scores of each test based on different scaling models did not differ significantly from each other. Hence, although the ability estimates associated with different scaling models do not have the same meanings, there was no evidence in favor of disregarding any scaling model.

Table 33: Correlations between individual test scores and midterm school marks in English, mathematics, and Vietnamese

Correlations between test scores and midterm school marks in…			English		Mathematics		Vietnamese	
Test	MP	Scaling model	r	(95%-CI)	r	(95%-CI)	r	(95%-CI)
C-test	T1	M1	.62	(.56–.68)	.41	(.34–.47)	.39	(.32–.45)
		M2	.62	(.56–.68)	.41	(.34–.47)	.38	(.31–.45)
		M1T	.67	(.61–.73)	.45	(.38–.52)	.42	(.35–.50)
		M2T	.64	(.57–.71)	.44	(.37–.52)	.39	(.32–.46)
	T2	M1	.62	(.57–.68)	.40	(.35–.46)	.38	(.31–.46)
		M2	.62	(.56–.67)	.40	(.34–.46)	.38	(.30–.46)
		M1T	.67	(.61–.74)	.46	(.38–.53)	.42	(.33–.50)
		M2T	.65	(.59–.71)	.44	(.37–.50)	.39	(.31–.47)
LC-test	T1	M1	.55	(.48–.62)	.34	(.26–.42)	.29	(.22–.37)
		M2	.54	(.47–.61)	.33	(.26–.41)	.28	(.20–.36)
		M3	.55	(.48–.63)	.36	(.28–.44)	.30	(.22–.38)
	T2	M1	.56	(.48–.63)	.33	(.26–.40)	.31	(.23–.38)
		M2	.54	(.48–.61)	.32	(.25–.39)	.30	(.23–.38)
		M3	.54	(.47–.61)	.33	(.26–.41)	.30	(.22–.38)

Note: M1 = Rasch model, M2 = unidimensional 2PL model, M1T = testlet 1PL model, M2T = testlet 2PL model, M3 = unidimensional 3PL model. r = correlation coefficients (all are statistically significant with $p < .001$), (95%-CI) = 95% confidence interval of the corresponding correlation coefficient.

The correlations between the English school marks and the C-test test scores were higher than those with the LC-test scores. However, the overlaps between the 95% confidence intervals between them do not point toward significant differences regarding the validity of the test scores of the two tests.

In short, the two tests in this study measured the English ability of the Vietnamese sample of ninth-graders equally well. And no scaling model was superior in estimating student ability with regard to the validity of these estimates (which corresponds to the curriculum-based English achievement of students).

X.2 Relevance of the scaling model selection on study results

Taking the results regarding academic student achievement and growth (Chapter VIII.1) into account, more similarities than differences with regard to the ability estimates based on the different scaling models were found. The estimates of student achievement and growth (*M*, *SD* and *SE*) of both tests generated with the different scaling models did not differ significantly at individual and class level (see Table 6, Table 8, Table 9).

The differences associated with using different scaling models were more obvious when the ICC of class mean growth was considered. Regarding the ICC of class mean growth and academic class composition effects, the results based on ability estimates obtained through unidimensional scaling models were similar to each other. In the C-test, the ICCs of student growth estimates based on the testlet models were similar to each other (M1T: ICC = .44, SE = .08, M2T: ICC = .41, SE = .07) and considerably and significantly higher than the ICCs based on the unidimensional models (M1: ICC = .15, SE = .05, M2: ICC = .15, SE = .05).

Taking the OLS regression coefficients into account, although the estimates associated with the different scaling models did not differ considerably and significantly with regard to the absolute values (considering the confidence intervals of the estimates), there were often differences in terms of the statistical significance of the results (based on the p-value). For example, the academic class composition effect on student achievement at T2 (C-test) was statistically significant based on the test scores obtained with the unidimensional scaling models and not statistically significant based on the test scores obtained with the testlet models (see Table 15, Chapter VIII.2.3).

In particular, using different scaling models leads to the identification of different important linear and nonlinear instructional effects on student growth (see Chapters IX.2, IX.3), regardless of which criterion is chosen for this purpose: statistically significant regression coefficients with $p < .05$, $f^2 \geq$ one chosen cut-off value, or the nonzero lasso regression coefficient.

Table 34 shows differences in the results with regard to the identification of the important linear instructional factors for student growth in the C-test, for instance. If the testlet 2PL model was used to estimate student achievement in the C-test, either none (based on lasso coefficients) or different instructional factors (based on the p-value of the OLS regression coefficients or the highest Cohen's f^2) were identified as important effects on student growth in comparison to the results obtained when using other scaling models. The results regarding the other three scaling models were similar in that the two factors with the highest local effect Cohen's f^2 were the same ("Encouragement of student statements"

and "Teacher speaking time in mixed languages"). However, they differed in terms of the p-value of the OLS regression coefficients (whether $p < 0.5$) or the lasso regression coefficients (whether the lasso coefficient was nonzero or not).

Table 34: Important linear instructional effects of student growth in the C-test identified based on different criteria with regard to different scaling models used to estimate student ability in the C-test

	Scaling model used to estimate student ability in the C-test			
Selection criterion	**M1**	**M2**	**M1T**	**M2T**
$\boldsymbol{p < .05}$	Encouragement of student statements (+) Teacher speaking time in mixed languages (-)	Encouragement of student statements (+) Teacher speaking time in mixed languages (-)	none	Student grammar mistakes (+)
$\boldsymbol{f^2 > .08}$	Encouragement of student statements Teacher speaking time in mixed languages (time percentage) Affectively stressed positive feedback (relative frequency)	Encouragement of student statements Teacher speaking time in mixed languages (time percentage) Affectively stressed positive feedback (relative frequency)	Encouragement of student statements Teacher speaking time in mixed languages (time percentage) Time used for social activities (time percentage)	Time used for social activities (time percentage) Student grammar mistakes (relative frequency)
Nonzero lasso coefficient	Encouragement of student statements (+)	Encouragement of student statements (+) Teacher speaking time in mixed languages (-) Affectively stressed positive feedback (+)	Encouragement of student statements (+) Teacher speaking time in mixed languages (-)	none

Note: (+) = positive regression coefficient, (-) = negative regression coefficient

In Chapter IX.4, which focused on the joint effect of instructional factors on student growth, more similarities than differences were found, with the exception of the results obtained with the testlet 2PL scaling model in the C-test (using this model, all model predictors had zero lasso regression coefficients). The nonzero status (and the number of nonzero coefficients over all 10 imputed datasets), the absolute value, and the sign of regression coefficients were for the most part similar. Nevertheless, differences between the results regarding different scaling models existed, given the same model specification. For instance, there were differences as to whether or not an overall negative effect of "Narrow focused monitoring" on student growth in the LC-test was confirmed (with the unidimensional 1PL and 2PL scaling models: yes; with the unidimensional 3PL model: no) or whether the interaction effect between "Encouragement of student statements" and "Teacher speaking time in mixed languages" on student growth in the C-test was nonzero or not (with the unidimensional models: yes; with the testlet models: no).

In brief, the results of this study show that the selection of a specific scaling model for estimating student ability has a considerable influence on the results regarding important instructional factors and their joint effect on student progress.

XI. Discussion

XI.1 Brief summary

XI.1.1 Classroom instruction

The results of this study show a differentiated view of instructional quality in the EFL lessons recorded in Vietnam. The majority of the lessons received positive ratings with regard to important instructional quality dimensions of general teaching effectiveness, such as classroom management, clarity, structuredness, and supportive learning climate. On the other hand, the quality dimensions teacher support, student orientation, and cognitive activation were judged rather negatively, which could, to some extent, be explained by the central curriculum and the textbook used for EFL in Vietnam. These dimensions are associated with a teacher-centered and textbook-driven teaching and learning culture.

Despite the dissimilarities between many respects in Germany and Vietnam, teachers' inaccuracy in judging their own speaking time in lessons was found in both countries (c.f. A. Helmke et al., 2008), which points to a generalizable problem of unreliable self-reflection by teachers on their own teaching.

From the point of view of EFL didactics, the data showed rather poor instructional quality. Most of all, there was evidence of shortcomings in English competencies not only among students but also among the English teachers, in particular in terms of the speaking mistakes made – which can often be attributed to the segmental as well as suprasegmental interferences from the tonal mother-tongue language Vietnamese (Lightbown & Spada, 2013; Shen, 2009; Tang, 2007).

XI.1.2 Effects of context factors on student achievement and growth

In accordance with previous international findings, differences in student achievement and class achievement in the posttest were, to a large extent, accounted for by initial ability and the social background of students and classes. The effect of initial achievement was larger than the effect of social background. As expected, the relationship at class level was larger than at individual level due to the existence of a class composition effect in addition to the individual effect.

While a rather small SES effect on student outcomes has been repeatedly reported and is expected in other school subjects in Vietnam (A. Helmke & Hesse, 2010; OECD, 2014a, 2016; Rolleston, James, & Aurino, 2013), a large SES effect on student achievement in posttest in EFL was found in this study. After controlling for the effect of initial student ability, SES effects on student achievement in posttest were considerably lower in both tests.

As far as achievement *growth* is concerned, there were differential effects in the C-test and the LC-test. Effects of initial student achievement on student growth in the LC-test at both levels were negligible. As well, there was a negligible effect of student SES on individual student growth in the LC-test. The class social background had a weak positive effect on class mean growth in the LC-test. A possible explanation might be that progress in the LC-test requires learning materials and an environment which are more easily accessible to students and classes with a higher socio-economic background (such as access to media in the English language, opportunities to practice with native speakers). Regarding student growth in the C-test, no SES effects were found, while initial student achievement had a small negative effect at both levels. The relationship between initial class achievement and class mean growth in the C-test was furthermore suggested to be nonlinear.

XI.1.3 Classroom instructional effects on student growth at class level

The most important instructional factors of student growth in the C-test based on the lasso regression results were quality aspects of motivation in instruction (“Encouragement of student statements” and “Affectively stressed positive feedback”) as well as aspects related to the teaching language (“Teacher speaking time in mixed languages”). Regarding the LC-test results, language-related aspects (“Teacher speaking time using Vietnamese in transitions” and “Student speaking time in mixed languages”) together with the relative frequency of repeated questions were the most important predictors of student growth.

The results of the hierarchical lasso regression models gave further insight into the complex effect and interplay between different contextual and instructional factors. The findings not only confirmed the hypothesized nonlinear relationship between student growth and some instructional factors, but also confirmed the hypotheses regarding the compensatory and interactive joint effect of different instructional factors.

One consideration regarding the instructional effects of student growth concerns the general instructional quality dimensions (such as classroom management, structuredness, clarity). None of these factors turned out to be important predictors of student growth in this study. Actually, the results did not indicate that these dimensions were unimportant, but rather that the lessons were all judged positively

and did not differ much, showing a ceiling effect; for instance, all lessons were highly structured – which could partly be attributed to the structuredness of the textbook. On the other hand, it is possible that other quality dimensions of instruction – ones that would not, however, be considered important in Western countries – might be more predictive of student outcomes in Vietnam. The adaptation of research instruments from one country to another country/culture is certainly problematic if the transcultural equivalence regarding the research topic itself is not guaranteed.

XI.1.4 The relevance of the scaling models to study results

The study results summarized above were not confirmed by the student ability estimates based on *all* scaling models. Although the ability estimates of each test at class level using different scaling models were very similar (see Appendix D2), the different scaling models produced significant differences in the results.

Because none of the models was decisively superior based on all reliability and validity indicators, calculating results using the model-averaging approach might have been a possible solution for taking into account the differences produced by the different scaling models when reporting results. However, it is more common to choose and apply only *one* scaling model in one study to facilitate the interpretability and communicability of the results. The relevance of the scaling model, as shown in this study, should therefore be taken seriously when reporting and interpreting results.

XI.2 Limitations of the study

XI.2.1 Restricted reliability and validity of the video-based measures of instructional quality

The first limitation of this study is related to the reliability and validity of the measures of instructional quality, as mentioned in Chapter VI.1.2. Because the measures of classroom instructional quality were based on only *one* lesson, the measurement errors might be large, thus the replicability and generalizability of the results might be restricted.

To examine the validity of these measures, students' perception of instructional quality dimensions over the whole school year was taken into account. For this purpose, six scales of the student questionnaire (T2) with content that corresponded to the available rating variables (see Table 35) were selected. Video ratings as well as student questionnaire data ranged from 1 (minimum) to 4 (maximum). The ICC of these six scales ranged from .09 to .13, which meant that the variation was not only due to differences *within* classes but also due to differences *between* classes.

Before interpreting these results, one aspect should be taken into account: Little consensus has been found between different perspectives (e.g., students, teachers) with regard to rating of the same occasion using the same items, even after reliability correction (Clausen, 2002; G. Pham et al., 2012). Against this background, small positive correlations between ratings and student perception of instructional quality dimensions were expected.

In general, the six quality dimensions of classroom instruction were also judged positively from the student perspective, with a small variation between classes. Raters and students did not agree with regard to their judgement of student orientation of instruction: While the instruction in the recorded lessons was seen as "rather not" student oriented (M = 2.2) by the raters, the students judged the overall instructional quality over one school year as "rather" student oriented (M = 2.9). Likewise, the EFL instruction over one school year was rated as somewhat more structured by the students (M = 3.2) than the rating data based on the quality of one recorded lesson (M = 2.8). In contrast, the clarity of the content as well as classroom management (with reference to discipline) of the recorded lessons were judged more positively by the raters (M = 3.44 – 3.49) than the students' perception based on the instructional quality over one school year (M = 2.97 – 3.15). Actually, this was in line with the results of the student short questionnaire, according to which the recorded lessons were quieter than usual (c.f. Chapter VII.1).

Table 35: Correlation between rating variables and student perception of selected instructional quality dimensions

	Video ratings		*Student questionnaire*			*Correlation*			
Dimension	*M*	*SD*	*M*	*SD*	*ICC*	*r*	*SE*	*p*	R^2
Clarity of content	3.49	.70	3.15	.21	.13	.04	.15	.80	.01
Structuredness (Preview, summary, review, highlight)	2.84	.74	3.23	.24	.12	.18	.16	.27	.04
Motivation (Student commitment)	3.23	.60	3.12	.19	.09	.28	.16	.08	.08
Lesson monitoring (Student orientation)	2.20	.76	2.92	.17	.09	.13	.16	.43	.02
Teacher-student relationship (warm and friendly)	3.12	.64	2.95	.24	.13	.34	.14	.01	.12
Classroom management (Disciplines)	3.44	.76	2.97	.25	.16	.14	.15	.36	.02

Note: M = mean of ratings/mean of class mean of student judgments, *SD* = standard deviation, ICC = intraclass correlation, r = correlation (corrected for the reliability of class mean of student judgments), *SE* = standard error of the correlation, p = significant level, R^2 = the amount of variance between classes based on student judgments that is explained by the ratings

Overall, positive correlations between video ratings and student perceptions of instructional quality were found, but only the correlation with regard to the teacher-student relationship (warm and friendly) was statistically significant (r = .34, p = .01). In total, 12% of the variation between classes regarding the

teacher-student relationship and 8% of the variation between classes regarding the motivational quality of instruction based on student data in terms of instructional quality over one school year can be explained by the video ratings of one lesson. Regarding other quality dimensions, the differences between classes based on student data were not reflected to a large extent by the rating data.

Thus, it can be assumed that, while a reliable overview of the classroom instructional quality (positive/negative) was obtained based on the video data, the results regarding the instructional effects on student outcomes might not be replicable if the indicators of instructional quality were acquired using a different method (e.g., questionnaire) and captured from a different perspective (e.g., student perception). Furthermore, even if recordings and ratings of multiple lessons were available and, as a result, more reliable indicators of classroom instructional quality could be obtained (see Praetorius et al., 2014), it would be difficult to completely avoid positive biases toward factors that could be partially prepared by teachers (e.g., discipline, structuredness).

XI.2.2 Validity of the tests

Another limitation of this study concerns the curricular validity of the tests. Because the tests were designed by researchers in another country with a different cultural background (Germany), teaching and learning conditions as well as a different curriculum, problems might arise regarding the congruence between what the tests measured and the EFL curriculum in Vietnam. This weakness is not, however, specific to this study, but rather a general problem in studies which use one/several test(s) in different countries. Applying tests that are tailored to the curriculum would result in more reliable and valid findings regarding the instructional effects of student outcomes (c.f. Teddlie, Reynolds, & Sammons, 2000).

XI.3 Prospects

In this study, the theoretically hypothesized *indirectness* of classroom instructional effects on student outcomes (see Chapter IV.3) has not yet been investigated. Given the large number of instructional variables together with the availability of numerous student and teacher variables via questionnaires, further research questions and hypotheses regarding the indirectness of classroom instructional effects can be examined with a direct link to the results of this study. For instance, future analyses might explore whether the effect of the video-based ratings of quality of motivation in lessons on student growth in the C-test is direct or whether it is indirect and/or mediated by individual student learning motivation. Furthermore, the inclusion of data from different perspectives (raters, students, teachers) in analyses

within the framework of the model of instructional provision and uptake (see Chapter IV.3) might enable a more in-depth understanding of how success can be achieved in teaching and learning EFL.

More importantly, the test results reflect only *one* (albeit very important) aspect of educational outcomes. Based on student questionnaire data, motivational and affective aspects of student outcomes such as academic self-concept and learning interest could be investigated together with their context and instructional effects. Therefore, further analyses with motivational student outcomes as the dependent variables are currently being prepared.

In addition, a re-analysis of the German DESI-study is planned, using the same methodological approaches and techniques to answer the same research questions as those in this study. This might provide us with a further insight into the mechanisms of the interplay between different contextual and instructional factors as well as the relevance of the scaling model to the research results.

Finally, the quest for understanding the success mechanisms of teaching and learning can also be investigated using different methodological approaches. For the *process-product paradigm* approach, nonparametric regression methods (e.g., regression trees, Breiman, Friedman, Olshen, & Stone, 1984; random forests, Breiman, 2001) can also be used to deal with the "large *p*, small *n*" problem and to identify the most important variables among a large number of variables, taking into account their possible nonlinear and possible interaction effects (Grömping, 2009; Sinharay, 2016; Strobl, Malley, & Tutz, 2009; Strobl, 2013). Analyses using these methods will be executed, and the results will be compared with those found in this study.

Or alternatively, based on the best practice paradigm of teacher effectiveness ("master teacher", cf. Hattie, 2012; Moser & Tresch, 2003; Weinert, Helmke, & Schrader, 1992), initial analyses have been conducted and will be completed soon. These analyses compare and contrast the personal and instructional characteristics of extreme groups, namely of classes with the highest vs. the lowest student growth in order to explore any differences and to build up hypotheses on this basis.

References

Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS regression). *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(1), 97–106.

Agresti, A., & Lang, J. B. (1993). Quasi-Symmetric Latent Class Models, with Application to Rater Agreement. *Biometrics*, *49*(1), 131–139.

Albert, P. S., & Dodd, L. E. (2008). On estimating diagnostic accuracy from studies with multiple raters and partial gold standard evaluation. *Journal of the American Statistical Association*, *103*, 61–73.

Allen, P., Cummins, J., Mougeon, R., & Swain, M. (1983). *The development of bilingual proficiency*. Toronto: Ontario Institute for Studies in Education.

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Asian Development Bank. (2008). *Educating Future Professionals*. Manila: ADB. Retrieved from http://www.adb.org/features/educating-future-professionals

Asian Development Bank. (2010). *Asian Development Bank & Viet Nam: Fact Sheet*. Manila: ADB. Retrieved from https://www.asienhaus.de/public/archiv/ADB_2010_Vietnam_Fact_Sheet.pdf

Asian Development Bank. (2015). *Asian Development Bank & Viet Nam: Fact Sheet*. Manila: ADB. Retrieved from http://www.adb.org/sites/default/files/publication/27813/vie.pdf

Asparouhov, T., & Muthén, B. O. (2006). *Constructing covariates in multilevel regression* (No. 11). Mplus Web Notes. Retrieved from http://www.statmodel.com/download/webnotes/webnote11.pdf

Asparouhov, T., & Muthén, B. O. (2010). *Plausible values for latent variables using Mplus*. Retrieved from https://www.statmodel.com/papers.shtml

Babu, S., & Mendro, R. (2003). Teacher accountability: HLM-based teacher effectiveness indices in a State Assessment Program. *Paper presented at the annual meeting of the American Educational Research Association, Chicago*. Chicago.

Baker, F. J., & Giacchino-Baker, R. (2003). Lower secondary school ciurriculum development in Vietnam. *California State Polytechnic, Pomona Journal of Interdisciplinary Studies*, *16*, 1–11.

Barton, K. (2016). *MuMIn: Multi-Model Inference*. Retrieved from https://CRAN.R-project.org/package=MuMIn

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model*. Princeton, NJ: Educational Testing Service.

Baumert, J., & Kunter, M. (2006). Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft*, *9*, 469–520.

Baumert, J., & Kunter, M. (2013). The COACTIV Model of Teachers' Professional Competence. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive Activation in the Mathematics Classroom and Professional Competence of Teachers* (pp. 25–48). New York: Springer.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., & Tsai, Y. (2010). Teachers' mathematic knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, *47*(1), 133–180.

Baumert, J., Lehmann, R. H., Lehrke, M., Schmitz, B., Clausen, M., Hosenfeld, I., Köller, O., et al. (1997). *TIMSS - Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich. Deskriptive Befunde*. Opladen: Leske + Budrich.

Baumert, J., Stanat, P., & Watermann, R. (2006). Schulstruktur und die Entstehung differenzieller Lern- und Entwicklungsmilieus. In J. Baumert, P. Stanat, & R. Watermann (Eds.), *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit* (pp. 95–188). Wiesbaden: VS Verlag für Sozialwissenschaften.

Bausch, K. R., Christ, H., & Krumm, H. J. (2002). *Handbuch Fremdsprachenunterricht*. Tübingen: Francke.

Beck, B., Bundt, S., & Gomolka, J. (2008). Ziele und Anlage der Studie. In DESI-Konsortium (Ed.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (pp. 11–25). Weinheim: Beltz.

Beck, B., & Dahl, D. (2006). *Sprachliche Kompetenzen von Schülerinnen und Schülern der neunten Jahrgangsstufe in Deutsch. Zentrale Befunde der Studie Deutsch-Schülerleistungen-International in Südtirol*. Deutsches Institut für Internationale Pädagogische Forschung. Retrieved from http://www.schule.suedtirol.it/pi/downloads/desi_bericht_suedtirol.pdf, date accessed: 14.03.2017

Beck, B., & Klieme, E. (2007). *Sprachliche Kompetenzen. Konzepte und Messung. DESI-Ergebnisse Band 1* (Vol. 1). Weinheim: Beltz Pädagogik.

Begley, C. G. (2013). Reproducibility: Six red flags for suspect work. *Nature*, *497*, 433–434.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*(2–3), 62–87.

Berezner, A., & Adams, R. J. (2017). Why LSAs Use Scaling and Item Response Theory (IRT). In P. Lietz, J. Cresswell, K. F. Rust, & R. D. Adams (Eds.), *Implementation of Large-Scale Education Assessments* (pp. 323–356). Chichester, UK: Wiley.

Berk, R., Brown, L., Buja, A., George, E., Pitkin, E., Zhang, K., & Zhao, L. (2014). Misspecified mean function regression: Making good use of regression models that are wrong. *Sociological Methods & Research*, *43*(3), 422–451.

Berk, R., Brown, L., & Zhao, L. (2010). Statistical inference after model selection. *Journal of Quantitative Crimonology*, *26*, 217–236.

Berliner, D. C. (2006). Educational Psychology: Searching For Essence Throughout a Century of Influence. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of Educational Psychology* (2nd ed., pp. 3–27). Mahwah, NJ: Lawrence Erlbaum.

Beuchling, O. (2003). *Vom Bootsflüchtling zum Bundesbürger. Migration, Intergration und schulischer Erfolg in einer vietnamesischen Exilgemeinschaft*. Münster: Waxmann.

Bien, J., Taylor, J., & Tibshirani, R. (2013). A Lasso for Hierarchical Interactions. *Annals of Statistics*, *41*(3), 1111–1141.

Bien, J., & Tibshirani, R. (2014). *hierNet: A Lasso for Hierarchical Interactions*. Retrieved from https://CRAN.R-project.org/package=hierNet

BIFIE. (2015). *BIFIEsurvey: Some Tools for Survey Statistics in Educational Assessment, Developed by BIFIE*. Retrieved from http://CRAN.R-project.org/package=BIFIEsurvey

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Blackwell, M., Honaker, J., & King, G. (2017a). A Unified Approach to Measurement Error and Missing Data: Overview and Applications. *Sociological Methods and Research*, *46*(3), 303–341.

Blackwell, M., Honaker, J., & King, G. (2017b). A Unified Approach to Measurement Error and Missing Data: Details and Extensions. *Sociological Methods and Research*, *46*(3), 342–369.

Bock, G. (2000). Difficulties in implementing communicative theory in Vietnam. *Teachers Edition*, *2*, 24–30.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29–51.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153–168.

Braeken, J. (2011). A boundary mixture approach to violations of conditional independence. *Psychometrika, 76*, 57–65.

Braeken, J., Tuerlinckx, F., & De Boeck, P. (2007). Copula functions for residual dependency. *Psychometrika, 72*, 393–411.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

Brennan, R. L. (2001a). *Generalizability Theory*. New York: Springer.

Brennan, R. L. (2001b). Some Problems, Pitfalls, and Paradoxes in Educational Measurement. *Educational Measurement, 20*(4), 6–18.

Brennan, R. L. (2006). Perspectives on the Evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 1–16). Westport, CT: American Council on Education and Praeger.

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education, 24*, 1–21.

Brock, W. A., Durlauf, S. N., & West, K. D. (2007). Model uncertainty and policy evaluation: Some theory and empirics. *Journal of Econometrics, 136*, 629–664.

Brophy, J. E. (2000). Teaching. (H. J. Walberg, Ed.)Educational Practices Series. Brussels: International Academy of Education & International Bureau of Education (www.ibe.unesco.org).

Brückmann, M., Duit, R., Tesch, M., Fischer, H., Kauertz, A., Reyer, T., Gerber, B., et al. (2007). The potential of video studies in research on teaching and learning science. In R. Pinto & D. Conso (Eds.), *Contributions from science education research* (pp. 77–89). Dordrecht: Springer.

Bruneforth, M., Oberwimmer, K., & Robitzsch, A. (2016). Reporting und Analysen. In S. Breit & C. Schreiner (Eds.), *Large-Scale Assessment mit R: Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung* (pp. 333–362). Wien: facultas.

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality*, *80*, 796–846.

Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Berlin Heidelberg: Springer.

Buhn-Wiggers, J. (2014). Smaller classes - more graduates? Evidence from Vietnamese provinces. *The seventh Vietnam Economists Annual Metting (VEAM 2014)*. Ho Chi Minh City, Vietnam. Retrieved from http://veam.org/papers2014/74_Julie%20Buhl%20Wiggers_smaller_classes.pdf

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference*. New York: Springer.

Butler, Y. G. (2011). The implementation of communicative and task-based Language teaching in the Asia-Pacific Region. *Annual Review of Applied Linguistics*, *31*, 36–57.

Van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton: Chapman & Hall.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45(3)*.

Byun, S., & Park, H. (2012). The academic success of East Asian American youth: The role of shadow education. *Sociology of Education*, *85*(1), 40–60.

Canh, L. V. (2002). Sustainable professional development of EFL teachers in Vietnam. *Teachers Edition*, *10*, 32–37.

Cao, Y., Lu, R., & Tao, W. (2014). Effect of Item Response Theory (IRT) Model Selection on Testlet-Based Test Equating. *ETS Research Report Series* (Vol. 2014, pp. 1–13). Wiley Periodicals, Inc.

Carman, K. G., & Zhang, L. (2012). Classroom peer effects and academic achievement: Evidence from a Chinese middle school. *China Economic Review*, *23*, 223–237.

Celce-Murcia, M., Dörnyei, Z., & Thurrell, S. (1997). Direct approaches in L2 instruction: A turning point in communcative language teaching? *TESOL Quarterly*, *31*, 141–152.

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6), 1–29. Retrieved from http://www.jstatsoft.org/v48/i06/

Chen, Q., & Wang, S. (2013). Variable selection for multiply-imputed data with application to dioxin exposure study. *Statistics in Medicine*, *32*, 3646–3659.

Chen, Y., Li, X., Liu, J., & Ying, Z. (2017). Regularized Latent Class Analysis with Application in Cognitive Diagnosis. *Psychometrika*, *82*(3), 660–692. Retrieved from http://EconPapers.repec.org/RePEc:spr:psycho:v:82:y:2017:i:3:d:10.1007_s11336-016-9545-6

Clausen, M. (2002). *Qualität von Unterricht: Eine Frage der Perspektive?* Münster: Waxmann.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). New York: Routledge.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfield, F. D., & al. (1966). *Equality of educational opportunity (2 vols.)*. Washington, DC: U.S. Goverment Printing Office.

Cook, V. (2001). *Second language learning and language teaching*. London: Arnold.

Creemers, B. P. M., Kyriakides, L., & Sammons, P. (2010a). Background to Educational Effectiveness Research. In B. P. M. Creemers, L. Kyriakides, & P. Sammons (Eds.), *Methodological advances in educational effectiveness research* (1st ed., pp. 3–18). New York: Routledge.

Creemers, B. P. M., Kyriakides, L., & Sammons, P. (2010b). Methodological issues in Educational Effectiveness Research. In B. P. M. Creemers, L. Kyriakides, & P. Sammons (Eds.), *Methodological advances in educational effectiveness research* (1st ed., pp. 19–36). New York: Routledge.

Cunningham, U. (2009). Phonetic correlates of unintelligibility in Vietnamese-accented English. *Paper presented at the FONETIK 2009 The XXIIth Swedish Phonetics Conference*. Retrieved from http://www2.ling.su.se/fon/fonetik_2009/108%20cunningham_fonetik2009.pdf

Cunningham, U. (2010). Quality, quantity and intelligibility of vowels in Vietnamese-accented English. In E. Waniek-Klimczak (Ed.), *Issues in accents of English 2: Variability and Norm* (Vol. 2, pp. 3–22). Newcastle Upon Tyne: Cambridge Scholars Publishing.

Cunningham, U. (2013). Teachability and learnability of English pronunciation features for Vietnamese-speaking learners. In E. Waniek-Klimczak & L. R. Shockey (Eds.), *Teaching and Researching English Accents in native and non-native speakers* (pp. 3–14). Heidelberg: Springer.

Dang, T. N. (1986). *Teaching oral communication skills to trainee interpreters at the University of Hanoi*. Unpublished MA (TESOL) Field Study Report. Faculty of Education, University of Canberra.

Dao, T. M. H. (2007). *A Study on pronunciation of some English consonants by Vietnamese learners*. University of Languagues and International Studies, Hanoi, Vietnam. Retrieved from http://hdl.handle.net/123456789/1889

Darling-Hammond, L. (2015). Can Value Added Add Value to Teacher Evaluation? *Educational Researcher*, *44*(2), 132–137.

Von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful. *IERI monograph series*, *2*, 9–36.

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, *50*, 164–185.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement*, *43*, 145–168.

Denham, P. A. (1992). English in Vietnam. *World Englishes*, *11*(1), 61–69.

DESI-Konsortium. (2006). *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Zentrale Befunde der Studie Deutsch-Englisch-Schülerleistungen-International (DESI)*. Frankfurt: Deutsches Institut für Internationale Pädagogische Forschung.

DESI-Konsortium. (2008). *Unterricht und Kompetenzerwerb in Deutsch und English. DESI-Ergebnisse Band 2* (Vol. 2). Weinheim: Beltz Pädagogik.

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific*, *May*, 116–130.

Dorans, N. J., Pommerich, M., & Holland, P. W. (2007). *Linking and aligning scores and scales*. New York: Springer.

Dörnyei, Z. (2013). Communicative Language Teaching in the twenty-first century: The 'Principled Communicative Approach. *Meaningful action: Earl Stevick's influence on language teaching* (pp. 161–171). Cambridge: Cambridge University Press.

Doyle, W. (2006). Ecological Approaches to Classroom Management. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of Classroom Management. Research, Practice, and Contemporary Issues* (pp. 97–125). Mahwah, NJ: Lawrence Erlbaum.

Dubberke, T., & Harks, B. (2008). Zur curricularen Validität der DESI-Aufgaben: Ergebnisse eines Expertenratings. In DESI-Konsortium (Ed.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (pp. 26–33). Weinheim: Beltz.

Dumont, H., Neumann, M., Maaz, K., & Trautwein, U. (2013). Die Zusammensetzung der Schülerschaft als Einflussfaktor für Schulleistungen: Internationale und nationale Befunde. *Psychologie in Erziehung und Unterricht*, *60*, 163–183.

Eckert, H., & Barry, W. (2005). *The Phonetics and phonology of English and pronunciation: A coursebook*. Trier: WVT Wissenschaftlicher Verlag Trier.

Eckes, T. (2010). Der Online-Einstufungstest Deutsch als Fremdsprache (onDaF): Theoretische Grundlagen, Konstruktion und Validierung. In R. Grotjahn (Ed.), *Der C-Test: Beiträge aus der aktuellen Forschung/The C-Test: Contributions from Current Research* (pp. 125–192). Frankfurt am Main: Lang.

Eckes, T. (2014). Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach. *Language Testing*, *31*(1), 39–61.

Eckes, T. (2015). Lokale Abhängigkeit von Items im TestDaF-Leseverstehen: Eine Testlet-Response-Analyse. *Diagnostica*, *61*(2), 93–106.

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-Tests. *Language Testing*, *23*(3), 290–325.

EF EPI. (2011). EF English Proficiency Index. EF Education First. Retrieved from http://www.ef.com/__/ /media/efcom/epi/pdf/EF-EPI-2011.pdf

EF EPI. (2012). EF English Proficiency Index. EF Education First. Retrieved from http://www.ef.com/__/ /media/efcom/epi/2012/full_reports/ef-epi-2012-report-master-lr-2.pdf

EF EPI. (2013). EF English Proficiency Index. EF Education First. Retrieved from http://www.ef.com/__/ /media/efcom/epi/2014/full-reports/ef-epi-2013-report-master-new.pdf

EF EPI. (2014a). EF English Proficiency Index. EF Education First. Retrieved from http://media.ef.com/__/ /media/centralefcom/epi/v4/downloads/full-reports/ef-epi-2014-english.pdf

EF EPI. (2014b). EF English Proficiency Index: Country fact sheet Viet Nam. EF Education First. Retrieved from http://media.ef.com/sitecore/__/ /media/centralefcom/epi/v4/downloads/fact-sheets/ef-epi-country-fact-sheet-v4-vn-en.pdf

EF EPI. (2015). EF English Proficiency Index. EF Education First. Retrieved from http://media2.ef.com/__/ /media/centralefcom/epi/downloads/full-reports/v5/ef-epi-2015-english.pdf

EF EPI. (2016). EF English Proficiency Index. EF Education First. Retrieved from http://media2.ef.com/__/ /media/centralefcom/epi/downloads/full-reports/v6/ef-epi-2016-english.pdf

Efron, B., & Tibshirani, R. (1986). The Bootstrap Method for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, *1*, 1–35.

Eid, M., Gollwitzer, M., & Schmitt, M. (2010). *Statistik und Forschungsmethoden* (1. Auflage.). Weinheim, Basel: Beltz.

Ellis, G. (1994). *The appropriateness of the communicative approach in Vietnam: An interview study in intercultural communication.* La Trobe University, Bundoora, Victoria, Australia. Retrieved from http://eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED378839

Ellis, G. (1996). How culturally appropriate is the communicative approach? *ELT Journal*, *50*(3), 213–218.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Enders, C. K. (2010). *Applied Missing Data Analysis*. New York: The Guilford Press.

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized Network Psychometrics: Combining Network and Latent Variable Models. *Psychometrika*. Retrieved from https://doi.org/10.1007/s11336-017-9557-x

Erosheva, E. A., & Joutard, C. (2014). Estimating Diagnostic Error without a Gold Standard. In E. M. Airoldi, D. Blei, E. A. Erosheva, & S. E. Fienberg (Eds.), *Handbook of Mixed Membership Models and Their Applications* (1st ed., pp. 141–157). Boca Raton, FL: Chapman & Hall/CRC.

Evertson, C. M., & Weinstein, C. S. (2006). Classroom Management as a Field of Inquiry. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of Classroom Management. Research, Practice, and Contemporary Issues* (pp. 3–15). Mahwah, NJ: Lawrence Erlbaum.

Fahrmeir, L., Kaufmann, H., & Kredler, C. (1996). Regressionsanalyse. In L. Fahrmeir, A. Hamerle, & G. Tutz (Eds.), *Multivariate statistische Verfahren* (2nd ed., pp. 93–168). Berlin: Walter de Gruyter.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.

Fend, H. (1998). *Qualität im Bildungswesen. Schulforschung zu Systembedingungen, Schulprofilen und Lehrerleistung*. Weinheim: Juventa.

Fischer, G. H. (2007). Rasch Models. In C. R. Rao. & S. Sinharay (Eds.), *Handbook of Statistics, Volume 26: Psychometrics* (pp. 515–585). North Holland: Elsevier.

Fondel, E., Lischetzke, T., Weis, S., & Gollwitzer, M. (2015). Zur Validität von studentischen Lehrveranstaltungsevaluationen Messinvarianz über Veranstaltungsarten, Konsistenz von Urteilen und Erklärung ihrer Heterogenität. *Diagnostica*, *61*(3), 124–135.

Fox, J.-P. (2010). *Bayesian Item Response Modeling: Theory and Applications*. New York: Springer.

Foy, P., Brossman, B., & Galia, J. (2013). Scaling the TIMSS and PIRLS 2011 Achievement Data. In M. O. Martin & I. V. S. Mullis (Eds.), *TIMSS and PIRLS methods and procedures*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA). Retrieved from https://timssandpirls.bc.edu/methods/pdf/TP11_Scaling_Achievement.pdf

De Fraine, B., Van Damme, J., Van Landeghem, G., & Opdenakker, M. C. (2003). The effects of schools and classes on language achievement. *British Educational Research Journal*, *29*(6), 841–859.

Francis, B., & Archer, L. (2005). British-Chinese pupils' and parents' constructions of the value of education. *British Educational Research Journal*, *31*(1), 89–108.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. Retrieved from http://www.jstatsoft.org/v33/i01/

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information bi-factor analysis. *Psychometrika*, *57*, 423–436.

Gignac, G. E. (2014). On the inappropriateness of using items to calculate total scale score reliability via coefficient alpha for multidimensional scales. *European Journal of Psychological Assessment*, *30*, 130–139.

Van Ginkel, J. R., & Kroonenberg, P. M. (2014). Analysis of Variance of Multiply Imputed Data. *Multivariate Behavioral Research*, *49*(1), 78–91.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology, 33*, 234–246.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement, 20*, 369–377.

Goldstein, H., & McDonald, R. (1988). A general model for the analysis of multilevel data. *Psychometrika, 53*, 455–467.

Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. *Seminars in hematology, 45*(3), 135–140.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research, 26*, 499–510.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology, 31*, 37–350.

Gröhlich, C., Guill, K., Scharenberg, K., & Bos, W. (2010). Differenzielle Lern- und Entwicklungsmilieus beim Erwerb der Lesekompetenz in den Jahrgangsstufen 7 und 8. In W. Bos & Gröhlich (Eds.), *KESS 8 – Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 8* (pp. 100–106). Münster: Waxmann.

Grömping, U. (2009). Variable importance assessment in regression: Linear regression versus random forest. *The American Statisticican, 63*(4), 308–319.

Grotjahn, R. (1995). Der C-Test: State of the Art. *Zeitschrift für Fremdsprachenforschung, 6(2)*(2), 37–60.

Grotjahn, R. (2006). *Der C-Test: Theorie, Empirie, Anwendungen/The C-Test: Theory, empirical research, applications.* Frankfurt am Main: Peter Lang.

Grotjahn, R. (2010). *Der C-Test: Beiträge aus der aktuellen Forschung/The C-Test: Contributions from Current Research*. (R. Grotjahn, Ed.). Frankfurt am Main: Peter Lang.

Gruehn, S. (2000). *Unterricht und schulisches Lernen*. (D. H. Rost, Ed.). Münster: Waxmann.

Guilford, J. P. (1954). *Psychometric methods* (2nd. ed.). New York, NY: McGraw-Hill.

Guo, P., Zeng, F., Hu, X., Zhang, D., Zhu, S., Deng, Y., & Hao, Y. (2015). Improved Variable Selection Algorithm Using a LASSO-Type Penalty, with an Application to Assessing Hepatitis B Infection Relevant Factors in Community Residents, *10*(7), 1–23.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29–48.

Gwet, K. L. (2010). *Handbook of inter-rater reliability. The Definitive Guide to Measuring the Extent of Agreement Among Raters* (4th ed.). Gaithersburg, MD: Advanced Analytics, LLC.

Ha, C. T. (2005). Common pronunciation problems of Vietnamese learners of English. *Journal of Science – Foreign Languages*, *21*(1), 35–46. Retrieved from http://tapchi.vnu.edu.vn/Ngoaingu_1/Bai3

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (pp. 65–110). Westport, CT: American Council on Education and Praeger.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park: Sage.

Han, K. T. (2012). Fixing the c Parameter in the Three-Parameter Logistic Model. *Practical Assessment, Research & Evaluation*, *17*(1), 1–24.

Hao, N., Feng, Y., & Zhang, H. H. (2016). Model Selection for High Dimensional Quadratic Regression via Regularization. *Journal of the American Statistical Association*.

Harris, D. N., & Herrington, C. D. (2015). Editors' Introduction: The Use of Teacher Value-Added Measures in Schools: New Evidence, Unanswered Questions, and Future Prospects. *Educational Researcher*, *44*(2), 71–76.

Harsch, C., & Hartig, J. (2010). Empirische und inhaltliche Analyse lokaler Abhängigkeiten im C-Test. In R. Grotjahn (Ed.), *Der C-Test: Beiträge aus der aktuellen Forschung/The C-Test: Contributions from Current Research* (pp. 193–204). Frankfurt/Main: Lang.

Harsch, C., & Schröder, K. (2007). Textrekonstruktion: C-Test. In B. Beck & E. Klieme (Eds.), *Sprachliche Kompetenzen. Konzepte und Messung* (pp. 212–225). Weinheim: Beltz.

Hartig, J. (2007). Skalierung und Definition von Kompetenzniveaus. In B. Beck & E. Klieme (Eds.), *Sprachliche Kompetenzen: Konzepte und Messung* (pp. 83–99). Weinheim: Beltz.

Hartig, J., & Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modeling*, *54*, 418–431.

Hartig, J., Jude, N., & Wagner, W. (2008). Methodische Grundlagen der Messung und Vorhersage sprachlicher Kompetenzen. In DESI-Konsortium (Ed.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (pp. 34–54). Weinheim: Beltz.

Hasebrook, J. (2006). Aptitude-Treatment-Interaktion. In D. H. Rost (Ed.), *Handwörterbuch Pädagogische Psychologie* (3. überarb. u. erw., pp. 20–26). Weinheim: Beltz Psychologie Verlags Union.

Hastie, T., Tibshirani, R., & Wainwright, M. J. (2015). *Statistical Learning with Sparsity: the Lasso and Generalizations*. Monographs on Statistics and Applied probability. Boca Raton, FL: Chapman and Hall and CRC Press. Retrieved from https://web.stanford.edu/ hastie/StatLearnSparsity_files/SLS_corrected_1.4.16.pdf

Hastings, A. J. (2002). Error analysis of an English C-Test: Evidence for integrated processing. In R. Grotjahn (Ed.), *Der C-Test: Theoretische Grundlagen und praktische Anwendungen* (pp. 53–66). Bochum: AKS Verlag.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

Hattie, J. (2012). *Visible learning for teachers: maximizing impact on learning*. New York: Routledge.

Hattie, J., Beywl, W., & Zierer, K. (2013). *Lernen sichtbar machen*. Schneider Verlag.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, *44*, 1–12.

Helmke, A. (2002). Kommentar: Unterrichtsqualität und Unterrichtsklima - Perspektiven und Sackgassen. *Unterrichtswissenschaft*, *30*(3), 261–277.

Helmke, A. (2004). *Unterrichtsqualität: Erfassen, Bewerten, Verbessern* (3rd ed.). Seelze: Klett-Kallmeyer.

Helmke, A. (2014a). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts* (5. Aufl.). Seelze: Klett-Kallmeyer.

Helmke, A. (2014b). Forschung zur Lernwirksamkeit des Lehrerhandelns. In E. Terhart, H. Bennewitz, & M. Rothland (Eds.), *Handbuch der Forschung zum Lehrerberuf* (2nd ed., pp. 807–821). Münster: Waxmann.

Helmke, A., Göbel, K., Hosenfeld, I., Schrader, F.-W., Helmke, T., & Wagner, W. (2007). *Die Videostudie im DESI-Projekt: Anlage, Ziele, Kameramanual*. Landau: Universität Koblenz-Landau.

Helmke, A., Helmke, T., Heyne, N., Hosenfeld, A., Hosenfeld, I., Schrader, F.-W., & Wagner, W. (2008). Zeitnutzung im Grundschulunterricht: Ergebnisse der Unterrichtsstudie - Gute Unterrichtspraxis. *Zeitschrift für Grundschulforschung*, *1*, 23–36.

Helmke, A., Helmke, T., Lenske, G., Pham, G., Praetorius, A.-K., Schrader, F.-W., & AdeThurow, M. (2014). Unterrichtsdiagnostik mit EMU. In M. Ade-Thurow, W. Bos, A. Helmke, T. Helmke, N. Hovenga, M. Lebens, G. Lenske, et al. (Eds.), *Aus- und Fortbildung der Lehrkräfte in Hinblick auf Verbesserung der Diagnosefähigkeit, Umgang mit Heterogenität und individuelle Förderung* (pp. 149–163). Münster: Waxmann.

Helmke, A., Helmke, T., Lenske, G., Pham, G., Praetorius, A.-K., Schrader, F.-W., & Ade-Thurow, M. (2017). EMU – Evidenzbasierte Methoden der Unterrichtsdiagnostik und -entwicklung. Landau: KMK-Unterrichtsdiagnostik Team, Universität Koblenz-Landau. Retrieved from www.unterrichtsdiagnostik.info

Helmke, A., & Hesse, H. G. (2002). Kindheit und Jugend in Asien. In H.-H. Krüger & C. Grunert (Eds.), *Handbuch Kindheits- und Jugendforschung* (pp. 439–471). Opladen: Leske + Budrich.

Helmke, A., & Hesse, H.-G. (2010). Kindheit und Jugend in Asien. In H.-H. Krüger & C. Grunert (Eds.), *Handbuch Kindheits- und Jugendforschung* (2nd ed., pp. 479–514). Wiesbaden: VS Verlag für Sozialwissenschaften.

Helmke, A., & Klieme, E. (2008). Unterricht und Entwicklung sprachlicher Kompetenzen. In DESI-Konsortium (Ed.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (pp. 301–312). Weinheim: Beltz.

Helmke, A., & Schrader, F.-W. (1998). Entwicklung im Grundschulalter. Die Münchner Studie. *Pädagogik, 6*, 25–30.

Helmke, A., Schrader, F.-W., Vo, T. A. T., Le, D. P., & Tran, T. B. T. (2003). Selbstkonzept und schulische Leistungen im Kulturvergleich: Ergebnisse der Grundschulstudie SCHOLASTIK in München und Hanoi. In W. Schneider & M. Knopf (Eds.), *Entwicklung, Lehren und Lernen: Zum Gedenken an Franz Emanuel Weinert* (pp. 187–206). Göttingen: Hogrefe.

Helmke, A., & Weinert, F. E. (1997). Bedingungsfaktoren schulischer Leistungen. In F. E. Weinert (Ed.), *Psychologie des Unterrichts und der Schule*, Enzyklopädie der Psychologie, Pädagogische Psychologie (Vol. 3, pp. 71–176). Göttingen: Hogrefe.

Helmke, T., Helmke, A., Schrader, F.-W., Wagner, W., Nold, G., & Schröder, K. (2008). Die Videostudie des Englischunterrichts. In DESI-Konsortium (Ed.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (pp. 345–363). Weinheim: Beltz.

Helsper, W., & Böhme, J. (2004). Einleitung in das Handbuch der Schulforschung. In W. Helsper & J. Böhme (Eds.), *Handbuch der Schulforschung* (pp. 11–31). Wiesbaden: VS Verlag für Sozialwissenschaften.

Hesse, H.-G. (2007). Lernen innerhalb und au{sserhalb der Schule aus interkultureller Perspektive. In G. Trommsdorff & H. J. Kornadt (Eds.), *Anwendungsfelder der kulturvergleichenden Psychologie*, Enzyklopädie der Psychologie, Serie VII: Kulturvergleichende Psychologie, Band 3 (pp. 187–277). Göttingen: Hogrefe.

Heuer, H., & Klippel, F. (1993). *Englischmethodik. Problemfelder, Unterrichtswirklichkeit, Handlungsempfehlungen*. Berlin: Cornelsen.

Heyneman, S. P., & Loxley, W. A. (1983). The effect of primary-school quality on academic achievement across twenty-nine high- and low-income countries. *The American Journal of Sociology*, *88*(6), 1162–1194.

Hiebert, J., Gallimore, R., Garnier, H., Bogard Givvin, K., Hollingsworth, S., Jacobs, J., Chui, A. M. Y., et al. (2003). *Teaching Mathematics in Seven Countries: Results from the TIMSS 1999 Video Study*. National Center for Education Statistics, U.S. Department of Education.

Hiep, P. H. (2007). Communicative language teaching: Unity within diversity. *ELT Journal*, *61*(3), 193–201.

Hill, H. C., Charalambous, C. Y., Blazar, D., McGinn, D., Kraft, M. A., Beisiegel, M., Humez, A., et al. (2012). Validating Arguments for Observational Instruments: Attending to Multiple Sources of Variation. *Educational Assessment*, *17*, 1–19.

Ho, D. Y. F. (1994). Cognitive socialization in Confucian Heritage cultures. In P. M. Greenfield & R. R. Cocking (Eds.), *Cross-cultural roors of minority child development* (pp. 284–312). Hillsdale, NJ: Erlbaum.

Ho, D. Y. F., Peng, S. Q., & Chan, S. F. (2001). An investigative research in teaching and learning. In C. Y. Chiu, F. Salili, & Y. Y. Hong (Eds.), *Multiple Competencies and Self-Regulated Learning: Implications for Multicultural Education* (Vol. 2, pp. 215–244). Greenwich, CT: Information Age Publishing.

Ho, D. Y. F., Peng, S. Q., & Chan, S. F. (2002). Authority and Learning in Confucian-heritage Education: A Relational Methodological Analysis. In F. Salili, C. Y. Chiu, & Y. Y. Hong (Eds.), *Multiple Competencies and Self-Regulated Learning: Implications for Multicultural Education* (Vol. 2, pp. 29–47). Greenwich, CT: Information Age Publishing.

Ho, W. K., & Wong, R. Y. L. (2004). *English language teaching in East Asia today*. Singapore: Eastern Universities Press.

Hoang, V. V. (2010). The Current Situation and Issues of the Teaching of English in Vietnam. *Ritsumeikan Studies in Language and Culture*, *22*(1), 7–18.

Holland, P. W. (1990). The Dutch identity: A new tool for the study of item response models. *Psychometrika*, *55*, 5–18.

Holland, P. W. (2007). A framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 5–30). New York: Springer.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (pp. 187–220). Westport, CT: American Council on Education and Praeger.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning: Theory and practice*. Hillsdale, NJ: Erlbaum.

Holzberger, D., Kunter, M., Praetorius, A.-K., & Seidel, T. (2016). Individuelle Schwerpunkte im Mathematikunterricht? Eine latente Profilanalyse zu unterschiedlichen Mustern der Unterrichtsqualität. In N. McElvany, W. Bos, H. G. Holtappels, M. M. Gebauer, & F. Schwabe (Eds.), *Bedingungen und Effekte guten Unterrichts* (pp. 135–146). Münster: Waxmann.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A Program for Missing Data. *Journal of Statistical Software*, *45*(7), 1–47. Retrieved from http://www.jstatsoft.org/v45/i07/

Honey, P. J. (1987). Vietnamese speakers. In M. Swan & B. Smith (Eds.), *Learner English: A teacher's guide to interference and other problems* (1st ed., pp. 243–248). London: Cambridge University Press.

Hoyt, W. T. (2000). Rater bias in psychological research: when is it a problem and what can we do about it? *Psychological Methods*, *5*(1), 64–86.

Hu, L., & Bentler, P. (1999). Cutoff criteria for fit indices in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Huang, P. H., Chen, H., & Weng, L. J. (2017). A penalized likelihood method for structural equation modeling. *Psychometrika*, *82*(2), 329–354.

Hunter, D., & Smith, R. (2012). Unpacking the past: 'CLT' through ELTJ keywords. *ELT*, *66*(4), 430–439.

Hutchison, D., & Schagen, L. (2007). Comparisons between PISA and TIMSS – Are We the Man with Two watches? In T. Loveless (Ed.), *Lessons Learned. What International Assessments tell us about Math Achievement* (pp. 227–262). The Brookings Institution.

IEA. (1971). Six Subject Survey: English as a Foreign Language. IEA. Retrieved from http://www.iea.nl/six-subject-survey-english-foreign-language

IMD. (2014). *IMD World competitiveness yearbook 2014*. Lausanne, Switzerland: IMD World Competitiveness Center.

Ip, E. H. (2000). Adjusting for information inflation due to local dependency in moderately large item clusters. *Psychometrika*, *65*, 73–91.

Ip, E. H. (2010). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology*, *63*, 395–416.

Itzlinger-Bruneforth, U., Kuhn, J.-T., & Kiefer, T. (2016). Testkonstruktion. In S. Breit & C. Schreiner (Eds.), *Large-Scale Assessment mit R: Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung* (pp. 21–50). Wien: facultas.

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural equation modeling: a multidisciplinary journal*, *34*(4), 555–566.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. New York: Springer.

Janík, T., Seidel, T., & Najvar, P. (2009). Introduction: On the Power of Video Studies in Investigation Teaching and Learning. In T. Janík & T. Seidel (Eds.), *The Power of Video Studies in Investigating Teaching and Learning in the Classroom* (pp. 7–19). Münster: Waxmann.

Jenkins, J. (2000). *The Phonology of English as an International Language*. Oxford: Oxford University Press.

Jerrim, J. (2014). *Why do East Asian children perform so well in PISA? An investigation of Western-born children of East Asian descent* (No. 14–16). London: Department of Quantitative Social Science, Institute of Education University of London. Retrieved from http://repec.ioe.ac.uk/REPEc/pdf/qsswp1416.pdf

Johnson, V. E. (2013). Uniformly Most Powerful Bayesian Tests. *Annals of Statistics*, *41*, 1716–1741.

Jonsson, J. O., & Rudolphi, F. (2011). Weak Performance – Strong Determination: School Achievement and Educational Choice among Children of Immigrants in Sweden. *European Sociological Review*, *27*(4), 487–508. Retrieved from http://esr.oxfordjournals.org/content/27/4/487.abstract

Judd, E. L., Tan, L., & Walberg, H. J. (2001). Teaching additional languages. (H. J. Walberg, Ed.)Educational Practices Series. Brussels: International Academy of Education & International Bureau of Education (www.ibe.unesco.org).

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kang, C. (2007). Classroom peer effects and academic achievement: Quasi-randomization evidence from South Korea. *Journal of Urban Econometrics*, *61*(3), 458–495.

Kao, G., & Thompson, J. S. (2003). Racial and Ethnic Stratification in Educational Achievement and Attainment. *Annual Review of Sociology*, *29*, 417–442.

Kelley, K., & Maxwell, S. E. (2003). Sample Size for Multiple Regression: Obtaining Regression Coefficients That Are Accurate, Not Simply Significant. *Psychological Methods*, *8*(3), 305–321.

Kennedy, M. M. (2010). Attribution error and the quest for teacher quality. *Educational Researcher*, *39*, 591–598.

Kiefer, T., Robitzsch, A., & Wu, M. (2016). *TAM: Test Analysis Modules*. Retrieved from http://CRAN.R-project.org/package=TAM

Klieme, E. (2008). Systemmonitoring für den Sprachunterricht. In DESI-Konsortium (Ed.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (pp. 1–10). Weinheim: Beltz.

Klieme, E. (2016). TIMSS 2015 and PISA 2015 – How are they related on the country level? German Institute for International Educational Research (DIPF). Retrieved from http://www.dipf.de/de/publikationen/pdf-publikationen/Klieme_TIMSS2015andPISA2015.pdf

Klieme, E., & Baumert, J. (2001). TIMSS als Startpunkt für Qualitätssicherung und Qualitätsentwicklung im Bildungswesen. In B. für Bildung und Forschung (Ed.), *TIMSS - Impulse für Schule und Unterricht. Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente* (pp. 5–11). Bonn: Bundesministerium für Bildung und Forschung (BMBF).

Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras Study: Investigating Effects of Teaching and Learning in Swiss and German Mathematics Classrooms. In T. Janík & T. Seidel (Eds.), *The Power of Video Studies in Investigating Teaching and Learning in the Classroom* (pp. 137–160). Münster: Waxmann.

Klieme, E., & Reusser, K. (2003). Unterrichtsqualität und mathematisches Verständnis im internationalen Vergleich - Ein Forschungsprojekt und erste Schritte zur Realisierung. *Unterrichtswissenschaft*, *31*(3), 194–205.

Knofczynski, G. T., & Mundfrom, D. (2008). Sample Sizes When Using Multiple Linear Regression for Prediction. *Educational and Psychological Measurement*, *68*(3), 431–442.

Kohavi, R. (1995). A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95 (pp. 1137–1143). Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc. Retrieved from http://dl.acm.org/citation.cfm?id=1643031.1643047

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 18*, 1–11.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York: Springer.

Krippendorff, K. (1980). *Content Analysis: An Introduction to its Methodology*. Beverly Hills, CA: Sage.

Krippendorff, K. (2004). *Content Analysis: An Introduction to its Methodology*. Thousand Oaks, CA: Sage.

Krippendorff, K. (2011). Computing Krippendorff 's Alpha-Reliability. University of Pennsylvania. Retrieved from http://repository.upenn.edu/asc_papers/43

Krüger, H.-H., & Pfaff, N. (2004). Triangulation quantitativer und qualitativer Zugänge in der Schulforschung. In W. Helsper & J. Böhme (Eds.), *Handbuch der Schulforschung* (pp. 159–182). Wiesbaden: VS Verlag für Sozialwissenschaften.

Language Teaching Review Panel. (2008). Replication studies in language learning and teaching: Questions and answers. *Language Teaching, 41*(01), 1–14.

Le, H. T., & Booth, A. L. (2014). Inequality in Vietnamese Urban–Rural Living Standards, 1993–2006. *Review of Income and Wealth, 60*(4), 862–886.

Le, T. S. (2011). *Teaching English in Vietnam: Improving the provision for private sectors*. Victoria University.

Le, V. (2011). *Form-focused instruction: A case study of Vietnamese teachers' beliefs and practices*. The University of Waikato, Waikato. Retrieved from http://core.ac.uk/download/pdf/29199049.pdf

Le, V. C. (1999). Language and Vietnamese pedagogical contexts. *Paper presented at the Fourth International Conference on Language and Development*. Retrieved from www.languages.ait.ac.th/hanoi_proceedings/canh.htm

Le, V. C. (2002). Sustainable professional development of EFL teachers in Vietnam. *Teachers Edition, 10*, 32–37.

Le, V. C., & Barnard, R. (2009). Curricular innovation behind closed classroom doors: A Vietnamese case study. *Research Publications in Prospect: An Australian Journal of TESOL, 24*(2), 20–33.

Lee, J. D., Sun, Y., & Saunders, M. A. (2014). Proximal Newton-Type Methods for Minimizing Composite Functions. *SIAM Journal on Optimization*, *24*(3), 1420–1443.

Levin, A. T., & Williams, J. C. (2003). Robust monetary policy with competing reference models. *Journal of Monetary Economics*, *50*, 945–975.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative testlet models. *Applied Psychological Measurement*, *30*, 3–21.

Lightbown, P., & Spada, N. (2013). *How Languagues are Learned*. (P. Lightbown & N. Spada, Eds.) (4th ed.). Oxford: Oxford.

Lim, M., & Hastie, T. (2015). Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics: A Joint Publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, *24*(3), 627–654.

Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. *Handbook of modern item response theory*. New York: Springer.

Lipowsky, F. (2006). Auf den Lehrer kommt es an. Empirische Evidenzen für Zusammenhänge zwischen Lehrerkompetenzen, Lehrerhandeln und dem Lernen der Schüler. In C. Allemann-Ghionda & E. Terhart (Eds.), *Zeitschrift für Pädagogik* (Vol. 51. Beiheft, pp. 47–70). Weinheim and Basel: Beltz.

Livingston, S. A. (2006). Item analysis. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 421–441). Mahwah, NJ: Lawrence Erlbaum.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585.

Loken, E., & Rullison, K. L. (2010). Estimation of a four-parameter item response model. *British Journal of Mathematical and Statistical Psychology*, *63*, 509–525.

Longford, N. T., & Muthén, B. O. (1992). Factor analysis for clustered observations. *Psychometrika*, *57*(4), 581–597.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lüders, M., & Rauin, U. (2004). Unterrichts- und Lehr-Lernforschung. In W. Helsper & J. Böhme (Eds.), *Handbuch der Schulforschung* (1st ed., pp. 691–719). Wiesbaden: VS Verlag für Sozialwissenschaften.

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*(3), 203–229.

Lüdtke, O., & Robitzsch, A. (2010). Umgang mit fehlenden Daten in der empirischen Bildungsforschung. In S. Maschke & L. Stecher (Eds.), *Enzyklopädie Erziehungswissenschaft Online. Fachgebiet Methoden der empirischen erziehungswissenschaftlichen Forschung, Quantitative Forschungsmethoden*. Weinheim: Juventa.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modelling. *Methodology*, *1*(3), 85–91.

Magis, D. (2013). A note on the Item information function of the four-parameter logistic model. *Applied Psychological Measurement*, *37*(4), 304–315.

Major, R. (2001). *Foreign accent: The ontogeny and phylogeny of second language phonology*. Mahwah, NJ: Lawrence Erlbaum.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York.

Martin, M. O., Mullis, I. V. S., Foy, P., & Stanco, G. M. (2012). *TIMSS 2011 International Results in Science*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Marzano, R. J., Gaddy, B. B., Foseid, M. C., Foseid, M. P., & Marzano, J. S. (2005). *A Handbook for Classroom Management that Works*. Alexandria, Virginia: Association for Supervision and Curriculum Development.

Marzano, R. J., Marzano, J. S., & Pickering, D. J. (2003). *Classroom Management that works. Research-Based Strategies for Every Teacher*. Alexandria, VA: ASCD.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

Mayr, J. (2014). Der Persönlichkeitsansatz in der Lehrerforschung. In E. Terhart, H. Bennewitz, & M. Rothland (Eds.), *Handbuch der Forschung zum Lehrerberuf* (pp. 189–215). Münster: Waxmann.

McDonough, S. (2000). Psychology. In M. Byram (Ed.), *Routledge encyclopedia of language teaching and learning* (pp. 491–498). London and New York: Routledge.

McNeish, D. M. (2015). Using Lasso for Predictor Selection and to Assuage Overfitting: A Method Long Overlooked in Behavioral Sciences. *Multivariate Behavioral Research*, *50*, 471–483.

Mei, B., & Wang, Z. (2016). An efficient method to handle the 'large p, small n' problem for genomewide association studies using Haseman–Elston regression. *Journal of Genetics, 95*(4), 847–852.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, (2), 105–118.

Mevik, B.-H., & Wehrens, R. (2007). The pls package: Principal component and partial least squares regression in R. *Journal of Statistical Software, 18(2)*, 1–24.

Meyer, H. (2004). *Was ist guter Unterricht?* Berlin: Cornelsen.

Ministry of Education. (1990). *45 years of educational development in Vietnam*. Hanoi: Education Publishing House.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177–196.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133–161.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in the National Assessment for Educational Progress. *Journal of Educational Statistics*.

Monahan, P. O., McHorney, C. A., Stump, T. E., & Perkins, A. J. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics, 32,* 92–109.

Moore, D. S., McCabe, G. P., & Craig, B. A. (2009). *Introduction to the practice of statistics* (6th ed.). New York, NY: W. H. Freeman and Company.

Morgan, G. B., Hodge, K. J., Trepinski, T. M., & Anderson, L. W. (2014). The stability of teacher performance and effectiveness: Implications for policies concerning teacher evaluation. *Education Policy Analysis Archives*, *22*, 1–21. Retrieved from http://files.eric.ed.gov/fulltext/EJ1050120.pdf

Morganstein, D., & Wasserstein, R. (2014). ASA Statement on Value-Added Models. *Statistics and Public Policy*, *1*(1), 108–110.

Moser, U., & Tresch, S. (2003). *Best Practice in der Schule. Von erfolgreichen Lehrerinnen und Lehrern lernen*. Buchs: Lehrmittelverlag des Kantons Zürich.

Moss, P. A., & Haertel, E. H. (2016). Engaging methodological pluralism. *Handbook of Research on Teaching* (pp. 127–248). Washington: AERA.

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 International Results in Mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Mullis, I. V. S., Martin, M. O., Foy, P., & Drucker, K. T. (2012). *PIRLS 2011 International Results in Reading*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York: Springer.

Musoro, J. Z., Zwinderman, A. H., Puhan, M. A., Riet, G. ter, & Geskus, R. B. (2014). Validation of prediction models based on lasso regression with multiply imputed data. *Medical research methodology*, *14*(116), 1–13.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557–585.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, *29*(1), 81–117.

Muthén, B. O., & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. In J. Hox & J. K. Roberts (Eds.), *Handbook of Advanced Multilevel Analysis* (pp. 15–40). New York: Taylor and Francis.

Muthén, B. O., & Satorra, A. (1989). Multilevel aspects of varying parameters in structural Models. In R. D. Bock (Ed.), *Multilevel analysis of educational data*. San Diego: Academic PIress.

Nachtigall, C., & Suhl, U. (2002). Der Regressionseffekt. Mythos und Wirklichkeit. *Schriftenreihe des Lehrstuhls für Psychologische Methodenlehre und Evaluationsforschung am Insitut für Psychologie der Friedrich-Schiller-Universitäet Jena*, *4*(2). Retrieved from https://www.metheval.uni-jena.de/materialien/reports/report_2002_02.pdf

NAEP. (2016). NAEP Technical documentation. NCES National Center for education statistics. Retrieved from https://nces.ed.gov/nationsreportcard/tdw/analysis/

Nagengast, B., & Trautwein, U. (2015). The prospects and limitations of latent variable models in educational psychology. In E. M. Anderman & L. Corno (Eds.), *Handbook of Educational Psychology* (pp. 41–58). New York: Routledge.

National Audit Office. (2003). *Making a Difference: Performance of mantained secondary schools in England*. London: The Stationery Office.

Nelder, J. A. (1977). A Reformulation of Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, *140*(1), 48–77. [Royal Statistical Society, Wiley]. Retrieved from http://www.jstor.org/stable/2344517

Nesselroade, J. R., Stigler, S. M., & Baltes, P. B. (1980). Regression toward the mean and the study of change. *Psychological Bulletin*, *88*(3), 622–637.

Ng, H. L., & Koretz, D. (2015). Sensitivity of School-Performance Ratings to Scaling Decisions. *Applied Measurement in Education*, *28*, 330–349.

Nguyen, D. L. (1970). A contrastive phonological analysis of English and Vietnamese. *Pacific Linguistics Series*, Pacific Linguistics, Series C, *8*.

Nguyen, N. Q. (1993). English Teaching and Learning in the System of Continuing Education in Vietnam. *International TESOL Conference*. Ho Chi Minh, Vietnam.

Nguyen, P.-M. (2008). *Culture and Cooperation: Cooperative Learning in Asian Confucian Heritage Cultures – The case of Viet Nam*. Institute of Education of Utrecht University, Utrecht.

Nguyen, T. A. T., Ingram, C. L. J., & Pensalfini, J. R. (2008). Prosodic transfer in Vietnamese acquisition of English contrastive stress patterns. *Journal of Phonetics*, *36*(1), 158–190.

Nguyen, T. T. T. (2007). *Difficulties for Vietnamese when pronouncing English: Final Consonants*. Dalarna University, School of Languages and Media Studies, English, Falun, Sweden.

Nold, G., & Rossa, H. (2007). Hörverstehen. In B. Beck & E. Klieme (Eds.), *Sprachliche Kompetenzen. Konzepte und Messung* (pp. 178–196). Weinheim: Beltz.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: a research synthesis and quantitative meta-analysis. *Language Learning*, *50*(3), 417–528.

Nunan, D. (2003). The impact of English as a global language on educational policies and practices in the Asia-Pacific region. *TESOL Quarterly*, *37*(4), 589–613.

Nussbeck, & Eid, M. (2015). Multimethod latent class analysis. *Frontiers in Psychology*, *6*, 1332.

Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature*, *506*(7487), 150–152.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How Large Are Teacher Effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237–257.

OECD. (2012). *PISA 2009: Technical Report*. Paris: OECD Publishing.

OECD. (2014a). *PISA 2012 Results in Focus: What 15-year-olds know and what they can do with what they know*. OECD. Retrieved from http://www.oecd.org/pisa/keyfindings/pisa-2012-results.htm

OECD. (2014b). *PISA 2012 Technical Report*. Paris: OECD Publishing.

OECD. (2016). *PISA 2015 Results in Focus*. Paris: OECD. Retrieved from http://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf

OECD. (n.d.). *PISA 2015 Technical Report*. Paris: OECD Publishing. Retrieved from http://www.oecd.org/pisa/data/2015-technical-report/

Olson, J. F., Martin, M. O., & Mullis, I. V. S. (2009). *TIMSS 2007 Technical Report*. *TIMSS 2007 Technical Report* (pp. 281–338). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Oser, F., & Baeriswyl, F. J. (2002). Choreographies of teaching: Bridging instruction to learning. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 1031–1065). Washington: American Educational Research Association.

Osterlind, S. J. (1998). *Constructing Test Items: Multiple-Choice, Constructed-Response, Performance and Other Formats* (2nd ed.). Norwell, MA: Kluwer.

Padilla, A. M. (2006). Second Language Learning: Issues in Research and Teaching. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of Educational Psychology* (2. ed., pp. 571–591). Mahwah, NJ: Lawrence Erlbaum.

Pennington, M., & Richards, J. (1986). Pronunciation revisited. *TESOL Quarterly*, *20*(2), 207–225.

Pham, G., Koch, T., Helmke, A., Schrader, F.-W., Helmke, T., & Eid, M. (2012). Do teachers know how their teaching is perceived by their pupils? *Procedia-Social and Behavioral Sciences Journal*, *46*, 3368–3374.

Pham, G., Robitzsch, A., George, A. C., & Freunberger, R. (2016). Fairer Vergleich in der Rückmeldung. In S. Breit & C. Schreiner (Eds.), *Large-Scale Assessment mit R: Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung* (pp. 295–332). Wien: facultas.

Pham, H. H. (2001). Teacher development: A real need for English departments in Vietnam. *English Teaching Forum*, *39*(4), 36–40.

Pham, H. H. (2005). Imported' communicative language teaching: Implications for local teachers. *English Teaching Forum*, *43*(4), 2–9.

Phan, H., & Vo, S. (2012). Pronunciation errors and perceptual judgements of accented speech by native speakers of English. *TESOL in Context, Special Edition S3*. Retrieved from http://www.tesol.org.au/files/files/264_hoa_phan.pdf

Polikoff, M. S. (2015). The stability of observational and student survey measures of teaching effectiveness. *American Journal of Education*, *121*, 183–212.

Praetorius, A.-K., Lenske, G., & Helmke, A. (2012). Observer ratings of instructional quality: Do they fulfill what they promise? *Learning and Instruction*, *6*, 387–400.

Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, *31*, 2–12.

Prenzel, M., Seidel, T., Lehrke, M., Rimmele, R., Duit, R., Euler, M., Geiser, H., et al. (2002). Lehr-Lern-Prozesse im Physikunterricht - eine Videostudie. *Zeitschrift für Pädagogik*, *45*, 139–156.

R Core Team. (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Raatz, U., & Klein-Braley, C. (1981). The C-test: A modification of the cloze procedure. In T. Culhane, C. Klein-Braley, & D. K. Stevenson (Eds.), *Practice and Problems in Language Testing* (pp. 113–138). Colchester: University of Essex.

Raatz, U., & Klein-Braley, C. (1985). How to develop a C-Test. In C. Klein-Braley & U. Raatz (Eds.), *C-Test in der Praxis. Fremdsprachen und Hochschule, AKS-Rundbrief 13/14* (pp. 20–22). Bochum, West Germany: Arbeitskreis Sprachenzentrum [AKS].

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Nielsen & Lydiche.

Rauch, W. A., & Moosbrugger, H. (2011). Klassische Testtheorie. Grundlagen und Erweiterungen für heterogene Tests und Mehrfacettenmodelle. In L. F. Hornke, M. Amelang, & M. Kersting (Eds.), *Enzyklopädie der Psychologie: Themenbereich B Methodologie und Methoden, Serie II Psychologische Diagnostik, Band 2, Methoden der psychologischen Diagnostik* (pp. 1–87). Göttingen: Hogrefe.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models*. Thousand Oakes: Sage.

Ray, A., McCormack, T., & Helen. (2009). Value Added in English schools. *Education Finance and Policy*, *4*(4), 415–438.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*, 667–696.

Reiter, J. P. (2007). Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. *Biometrika*, *94*, 502–508.

Reusser, K., & Pauli, C. (1999). *Unterrichtsqualität: multideterminiert und multikriterial*. Universität Zürich.

Reusser, K., & Pauli, C. (2004). Bericht über Ergebnisse einer internationalen und schweizerischen Video-Unterrichtsstudie zum Mathematikunterricht in der Schweiz und in weiteren sechs Ländern. *SEMINAR*, *10*(4), 102–104.

Rimmele, R. (2007). *Videograph - Multimedia-Player zur Kodierung von Videos* (4.1 ed.). Kiel: PN-Leibniz-Institut für die Pädagogik der Naturwissenschaften.

Rindermann, H., Hoang, Q. S. N., & Baumeister, A. E. E. (2013). Cognitive ability, parenting and instruction in Vietnam and Germany. *Intelligence*, *41*(5), 366–377.

Rivkin, S., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometria*, *73*(2), 417–458.

Robitzsch, A. (2009). Methodische Herausforderungen bei der Kalibrierung von Leistungstests. In A. Bremerich-Vos, D. Granzer, & O. Köller (Eds.), *Bildungsstandards Deutsch und Mathematik* (pp. 42–106). Weinheim: Beltz.

Robitzsch, A. (2014). *sirt: Supplementary Item Response Theory Models*. Retrieved from http://CRAN.R-project.org/package=sirt

Robitzsch, A. (2016). *Essays zu methodischen Herausforderungen im Large-Scale Assessment*. Humboldt-Universität zu Berlin, Berlin.

Robitzsch, A. (2017). *LAM: Some Latent Variable Models*. Retrieved from https://CRAN.R-project.org/package=LAM

Robitzsch, A., Dörfler, T., Pfost, M., & Artelt, C. (2011). Die Bedeutung der Itemauswahl und der Modellwahl für die längsschnittliche Erfassung von Kompetenzen: Lesekompetenzentwicklung in der Primarstufe. *Zeitschrift für Entwicklungspsychologie und pädagogische Psychologie*, *43*, 213–227.

Robitzsch, A., Freunberger, R., Itzlinger-Bruneforth, U., Breit, S., & Schreiner, C. (2015). *Ein Kommentar zu Vohns: M8 - Wie kommen die offiziellen Zahlen zustande und was sagen sie (nicht) aus?* Salzburg: BIFIE. Retrieved from https://www.bifie.at/node/2842

Robitzsch, A., Grund, S., & Henke, T. (2016). *miceadds: Some additional multiple imputation functions, especially for 'mice*. Retrieved from http://CRAN.R-project.org/package=miceadds

Robitzsch, A., Lüdtke, O., Köller, O., Kröhne, U., Goldhammer, F., & Heine, J.-H. (2017). Herausforderungen bei der Schätzung von Trends in Schulleistungsstudien. Eine Skalierung der deutschen PISA-Daten. *Diagnostica*, *63*(2), 148–165.

Robitzsch, A., Pham, G., & Yanagida, T. (2016). Fehlende Daten und Plausible Values. In S. Breit & C. Schreiner (Eds.), *Large-Scale Assessment mit R: Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung* (pp. 259–293). Wien: facultas.

Rogosa, D. (1995). Myths and Methods: Myths about longitudinal research. In J. M. Gottman (Ed.), *The analysis of change* (pp. 3–66). Mahwah: Erlbaum.

Rolff, H.-G., Leucht, M., & Rösner, E. (2008). Sozialer und familiärer Hintergrund. In DESI-Konsortium (Ed.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (pp. 283–300). Weinheim: Beltz.

Rolleston, C., James, Z., & Aurino, E. (2013). *Exploring the effect of educational opportunity and inequality on learning outcomes in Ethiopia, Peru, India, and Vietnam* (No. 2013/4). UNESCO. Retrieved from http://www.younglives.org.uk/node/6740

Rolleston, C., James, Z., Pasquier-Doumer, L., & Tran, N. T. M. T. (2013). *Making progress: Report of the Young Lives School Survey in Vietnam* (No. 100). London: Young Lives. Retrieved from http://www.younglives.org.uk/node/7348

Rosenshine, B., & Furst, N. (1971). Research on teacher performance criteria. In B. O. Smith (Ed.), *Research in teacher education: A symposium*. Englewood Cliffs, NJ: Prentice-Hall.

Rossi, N., Wang, X., & Ramsay, J. O. (2002). Nonparametric Item Response Function Estimates with the EM Algorithm. *Journal of Educational and Behavioral Statistics*, *27*(3), 291–317.

Rost, J. (2004). *Testtheorie - Testkonstruktion* (2nd ed.). Bern: Verlag Hans Huber.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Hoboken, NJ: Wiley.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement*, *66*, 63–84.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*(2), 413–428.

Saito, E., & Tsukui, A. (2008). Challenging common sense: Cases of school reform for learning community under an international cooperation project in Bac Giang province, Vietnam. *International Journal of Educational Development*, *28*(5), 571–584.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, (Monograph Supplement No. 17).

San Martin, E., del Pino, G., & De Boeck, P. (2006). IRT Models for Ability-Based Guessing. *Applied Psychological Measurement*, *30*(3), 183–203.

Saunders, L. (1998). *Value added' measurement of school effectiveness: an overview*. Slough: NFER.

Saunders, L. (1999). *Value added' measurement of school effectiveness: a critical view*. Slough: NFER.

Savignon, S. J. (2000). Communicative language teaching. In M. Byram (Ed.), *Routledge Encyclopedia of Language Teaching and Learning* (pp. 124–129). London: Routledge.

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed.). Westport, CT: American Council on Education and Praeger Publishers.

Schroeders, U., Robitzsch, A., & Schipolowski, S. (2014). A comparison of different psychometric approaches to modeling testlet structures: An example with c-tests. *Journal of Educational Measurement*, *51*, 400–418.

Scott, W. A. (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly*, *XIX*, 321–325.

Seidel, T. (2014). Lehrerhandeln im Unterricht. In E. Terhart, H. Bennewitz, & M. Rothland (Eds.), *Handbuch der Forschung zum Lehrerberuf* (2. Aufl., pp. 605–629). Münster: Waxmann.

Seidel, T., & Prenzel, M. (2004). Muster unterrichtlicher Aktivitäten im Physikunterricht. In J. Doll & M. Prenzel (Eds.), *Bildungsqualität von Schule: Lehrerprofessionalisierung, Unterrichtsentwicklung und Schülerförderung als Strategien der Qualitätsverbesserung* (pp. 177–194). Münster: Waxmann.

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.

Selya, A. S., Rose, J. S., Dierker, L. C., Hedeker, D., & Mermelstein, R. J. (2012). A Practical Guide to Calculating Cohen's *f*2, a Measure of Local Effect Size, from PROC MIXED. *Frontiers in Psychology*, *3*, 111.

Shen, X. (2009). Ability of learning the prosody of an intonational language by speakers of a tonal language: Chinese speakers leraning French prosody. *International Review of Applied Linguistics in Language Teaching*, *28*(2), 119–134.

Sijtsma, K. (2016). Playing with Data–Or How to Discourage Questionable Research Practices and Stimulate Researchers to Do Things Right. *Psychometrika*, *81*(1), 1–15.

Simon, M. K. (2008). *Comparison of concurrent and separate multidimensional IRT linking of item parameters*. University of Minnesota, Minneapolis, Minneapolis.

Sinharay, S. (2016). An NCME Instructional Module on Data Mining Methods for Classification and Regression. *Educational Measurement: Issues and Practice, 35*(3), 38–54.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237–247.

Slavin, R. E. (1994). Quality, appropriateness, incentive, and time: A model of instructional effectiveness. *International Journal of Educational Research, 21*, 141–157.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London: SAGE.

Soetaert, K. (2016). *plot3D: Plotting Multi-Dimensional Data*. Retrieved from https://CRAN.R-project.org/package=plot3D

Spada, N. (2007). Communicative Language Teaching. In J. Cummins & C. Davison (Eds.), *International Handbook of English Language Teaching* (pp. 271–288). Boston, MA: Springer.

Spada, N., & Fröhlich, M. (1995). *The communicative orientation of language teaching (COLT) observation scheme: coding conventions and applications*. Sydney: National Centre for Englisch Language Teaching and Research.

Stanat, P., Schwippert, K., & Gröhlich, C. (2010). Der Einfluss des Migrantenanteils in Schulklassen auf den Kompetenzerwerb: Längsschnittliche Überprüfung eines umstrittenen Effekts. *Zeitschrift für Pädagogik, 55*, 147–164.

Stelzl, I. (1982). *Fehler und Fallen der Statistik: für Psychologen, Pädagogen, Sozialwissenschaftler*. Bern: Huber.

Steyer, R. (2002). *Wahrscheinlichkeit udn Regression*. Berlin: Springer.

Stigler, J. W., Gallimore, R., & Hiebert, J. (2000). Using video surveys to compare classrooms and teaching across cultures: Examples and lessons from the TIMSS Video studies. *Educational Psychologist, 35*, 87–100.

Stigler, J. W., Gonzales, P., Kawanaka, T., Knoll, S., & Serrano, A. (1996). *The TIMSS videotape classroom study: Methods and preliminary findings*. Los Angeles, CA: National Center for Educational Statistics, U.S. Department of Education.

Stigler, J. W., & Hiebert, J. (1999). *The Teaching Gap. Best Ideas from the World's Teachers for Improving Education in the Classroom*. New York: Free Press.

Strobl, C. (2013). Data mining. *The Oxford handbook of quantitative methods in psychology* (Vol. 2, pp. 678–700). New York, NY: Oxford University Press.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods, 14*, 323–348.

Sun, J., Chen, Y., Liu, J., Ying, Z., & Xin, T. (2016). Latent Variable Selection for Multidimensional Item Response Theory Models via *L*1 Regularization. *Psychometrika, 81*(4), 921–939.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement, 27*(4), 361–370.

Tanaka, J. S. (1987). How big is big enough?": Sample size and goodness of fit in structural equation models with latent variables. *Child Development, 58*, 134–146.

Tang, G. (2007). Cross-Linguistic analysis of Vietnamese and English with implications for Vietnamese language acquisition and maintenance in the United States. *Journal of Southeast Asian American Education and Advancement, 2*(1).

Teddlie, C., Reynolds, D., & Pol, S. (2000). Current topics and approaches in school effectiveness research: The contemporary field. In C. Teddlie & D. Reynolds (Eds.), *The international handbook of school effectiveness research* (pp. 26–51). New York: Routledge.

Teddlie, C., Reynolds, D., & Sammons, P. (2000). The methodology and scientific properties of school effectiveness research. *The international handbook of school effectiveness research* (pp. 55–133). New York: Routledge.

Teddlie, C., & Sammons, P. (2010). Applications of mixed methods to the field of educational effectiveness research. In B. P. M. Creemers, L. Kyriakides, & P. Sammons (Eds.), *Methodological advances in educational effectiveness research*, Quantitative methodology series (pp. 115–152). New York: Routledge.

The World Bank. (2005). *Expanding Opportunities and Building Competencies for Young People: A New Agenda for Secondary Education*. Washington, D.C.: World Bank.

The World Bank. (2010). Education in Vietnam. Development history, challenges and solutions. Retrieved from http://siteresources.worldbank.org/EDUCATION/Resources/278200-1121703274255/1439264-1153425508901/Education_Vietnam_Development.pdf

The World Bank. (2011). *Vietnam Urbanization Review: Technical Assistance Report*. Hanoi, Vietnam: World Bank.

The World Bank. (2013). *Vietnam Development Report 2014 – Skilling up Vietnam: Preparing the WORKforce for a modern market economy*. Washington, DC: World Bank.

The World Bank. (2015). Vietnam Overview - World Bank. Washington: The World Bank. Retrieved from www.worldbank.org/en/country/vietnam/overview

The World Factbook. (2015). The World Factbook 2014–2015. Washington, DC: Central Intelligence Agency. Retrieved from https://www.cia.gov/library/publications/the-world-factbook/geos/vm.html

Thinh, D. H. (2006). The Role of English in Vietnam's Foreign Language Policy: A Brief History. *19th Annual EA Education Conference 2006.*

Thompson, J. (2009). *Changing chalk and talk: The reform of teaching methods in Vietnamese higher education.* The George Washington University. Retrieved from http://digitalcollections.sit.edu/cgi/viewcontent.cgi?article=1711&context=isp_collection

Thompson, L. C. (1965). *A Vietnamese Grammar*. Seattle: University of Washington Press.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society B*, *58*(1), 267–288.

Timm, J.-P. (1998). *Englisch lernen und lehren*. Berlin: Cornelsen.

Tomlinson, B., & Bao, D. (2004). The contributions of Vietnamese learners of English to ELT methodology. *Language Teaching Research*, *8*(2), 199–222.

Trendtel, M., Pham, G., & Yanagida, T. (2016). Skalierung und Linking. In S. Breit & C. Schreiner (Eds.), *Large-Scale Assessment mit R: Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung* (pp. 185–224). Wien: facultas.

Trendtel, M., Schwabe, F., & Fellinger, R. (2016). Differenzielles Itemfunktionieren in Subgruppen. In S. Breit & C. Schreiner (Eds.), *Large-Scale Assessment mit R: Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung* (pp. 111–147). Wien: facultas.

Tutz, G. (2012). *Regression for Categorical Data*. Cambridge University Press.

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika*, *80*(1).

Uebersax, J. S. (1993). Statistical modeling of expert ratings on medical treatment appropriateness. *Journal of the American Statistical Association*, *88*, 421–427.

Uebersax, J. S. (1999). Probit Latent Class Analysis: Conditional Independence and Conditional Dependence Models. *Applied Psychological Measurement*, *23*(4), 283–297.

Uebersax, J. S., & Grove, W. M. (1993). A latent trait finite mixture model for the analysis of rating agreement. *Biometrics*, *49*, 823–835.

UNESCO. (2007). *Secondary education regional information base: country profile – Viet Nam*. (Y. Sato, Ed.) (1st ed.). Bangkok: UNESCO Asia and Pacific Regional Bureau for Education.

UNESCO Institute for Statistics. (2015). Education database. Retrieved from http://data.uis.unesco.org/Index.aspx?queryid=144

UNESCO-IBE. (2011). World Data on Education. VII Ed. 2010/11. Viet Nam. Retrieved from http://www.ibe.unesco.org/fileadmin/user_upload/Publications/WDE/2010/pdf-versions/Viet_Nam.pdf

Vockrodt-Scholz, V., & Zydatiß, W. (2010). Sprachproduktive Faktoren und die Konstruktvalidität von C-Tests: Kompetenzniveaus und Fehlerquotient in textsortengebundenen Schreibaufgaben. In R. Grotjahn (Ed.), *Der C-Text: Beiträge aus der aktuellen Forschung./The C-Text: Contributions from Current Research* (pp. 1–40). Frankfurt am Main: Lang.

Wagner, W., Helmke, A., & Rösner, E. (2009). *Deutsch Englisch Schülerleistungen International. Dokumentation der Erhebungsinstrumente für Schülerinnen und Schüler, Eltern und Lehrkräfte*. Frankfurt: Deutsches Institut für Internationale Pädagogische Forschung (DIPF). Retrieved from http://www.pedocs.de/volltexte/2010/3252/pdf/MatBild_Bd25_1_D_A.pdf

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245–270). Boston, MA: Kluwer-Nijhoff.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge, MA: Cambridge University Press.

Wainer, H., & Kiely, G. L. (1987). Item Clusters and Computerized Adaptive Testing: A Case for Testlets. *Journal of Educational Measurement*, *24*, 185–201.

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, *37*, 203–220.

Wang, W.-C., & Wilson, M. (2005). The Rasch Testlet Model. *Applied Psychological Measurement*, *29*(2), 126–149.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, *70*(2), 129–133. Retrieved from amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108

Waxman, H. C., Hilberg, R. S., & Tharp, R. G. (2004). Future Directions for Classroom Observation Research. In H. C. Waxman, R. G. Tharp, & R. S. Hilberg (Eds.), *Observational Research in U.S. Classrooms: New Approaches for Understanding Cultural and Linguistic Diversity* (pp. 266–277). Cambridge: Cambridge University Press.

Weigel, A. P., Knutti, R., Liniger, M. A., & Appenzeller, C. (2010). Risks of model weighting in multimodel climate projections. *Journal of Climate*, *23*, 4175–4191.

Weinert, F. E., Helmke, A., & Schrader, F.-W. (1992). Research on the model teacher and the teaching model: Theoretical contradiction or conglutination? In F. Oser, A. Dick, & J. L. Patry (Eds.), *Effective and responsible teaching: The new synthesis* (pp. 249–260). San Francisco: Jossey-Bass.

Wellenreuther, M. (2004). *Lehren und Lernen - aber wie? Empirisch-experimentelle Forschungen zum Lehren und Lernen im Unterricht*. Grundlagen der Schulpädagogik, Band 50. Baltmannsweiler: Schneider Verlag Hohengehren.

Wells, C., Subkoviak, M., & Serlin, R. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, *26*, 77–87.

Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität*. Göttingen: Hogrefe.

Wu, M. (2005). The role of plausible values in large-scale assessments. *Studies in Educational Evaluation*, *31*, 114–128.

Wu, M. (2010). Comparing the Similarities and Differences of PISA 2003 and TIMSS. Paris: OECD.

Wu, W. (2008). Misunderstandings of Communicative Language Teaching. *English language teaching*, *1*(1), 50–53.

Yen, W. M. (1984). Effects of local dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*, 125–145.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (pp. 111–154). Westport, CT: American Council on Education and Praeger.

Yuan, M., Joseph, V. R., & Zou, H. (2009). Structured variable selection and estimation. *Annals of Applied Statistics*, *3*, 1738.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B, 68*, 49–67.

Zhao, P., Rocha, G., & Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, *37*(6A), 3468–3497. The Institute of Mathematical Statistics.

Zimmer, R. W., & Toma, E. F. (2000). Peer effects in private and public schools across countries. *Journal of Policy*, *19*(1), 75–92.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R. (2012). *A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement*. Princeton: ETS.

Appendices

Appendix A. The adapted basic coding guides and rating sheets

Appendix A1. The adapted basic coding guides

Codebook

DESI-VN

Video Study Basic Coding

Basic categories: Explanation of the categories

Preliminary remarks:

(1) The coding is performed based on the transcript. The turn in question is primarily decisive for coding. In case of doubt, take into account the context, i.e., in particular the earlier turns. When coding, ***interpret as little as possible****.*

(2) All Vietnamese words or sentences in the transcript have been translated into English and are provided in brackets […]

Instructions:

*(1) To reduce the workload for coding, the characteristic that occurs most frequently (e.g., English for language) among the different variables is no longer coded (categories). Instead, they have been left out of the videograph window, but are still included in these explanations with the note "****not in the videograph window." If this characteristic applies, skip the entire category.***

Based on the explanations, decide whether the characteristic left out of the videograph window (e.g., English) applies. If it applies (and only if it applies), you can skip the entire category (in this case: language).

(2) To understand all marks in the transcript (which have been used to facilitate coding many categories), read the section "Specifics regarding the transcription" carefully (which follows these explanations on page 16). It is also very important to pay attention to all of the steps in the section "Saving of coding" (page 17).

N.B.:

(1) The numbering of the categories in these coding instructions is ***not*** *identical to the numbering (i.e., to the value) of the categories in the videograph coding window.*

(2) The first turn to be coded is the turn linked to the beginning of instruction (e.g., the greeting in the first lesson or an explicit reference to it in the second lesson). This means that the preparation before the beginning of the lesson (often under C turn) is not coded.

1.	**To whom**	**All turns are coded (i.e., also AV and C turns).**
	[komm0]: no verbal production [komm1]: S speaks [komm2]: T speaks	Information is needed on *who speaks to whom*. ***Who*** speaks (speaker/actor**)** can be seen in the transcript. It is noted in the transcript as follows: T = teacher S = student SS = students (group) E = entire class C = comment (turns during which there is no speaking; also during silent work) AV = audio-video You therefore only need to code **to whom** the person is speaking.
	[to the class or group (E)] *not in videograph window: if this characteristic applies, skip the entire category 1* [wzw13]: AV to E/SS [wzw12]: C to E/SS [wzw4]: S to E/SS [wzw7]: SS to E/SS [wzw11]: T to E/SS	**If in doubt, assume that teacher statements** are directed at the class; this also applies to *teacher self-commentary* which is coded under *type/function of teacher activity* as *not assignable* **also for C turns** in ***silent work phases*** **also when** *a non-verbal presentation (writing on the board) or audio-visual presentation* (audio / video recording, etc.; AV is noted in the transcript) is directed at the entire class (which is generally the case) **also *reading aloud*** (regardless of whether reading a "normal" text or a *dialog text)* ***in a role play*** is always coded as ***to the class*** because reading aloud (also a *dialog text)* is not a natural conversation. (**N.B.**: ***Reading aloud* as a task/upon instruction by the T** and ***not in a role play*** is coded as ***to T***) *Remark:* to others (in exceptional cases, e.g., to the caretaker or to AV...) is coded as "to the class" *N.B.* In a student work phase (e.g., partner work), only code elements as to the class/group if it is not possible to tell whether the teacher is talking to a specific student.
1	***to an individual student (S)*** [wzw3]: S to S [wzw6]: SS to S [wzw10]: T to S	An individual S (who did *not* speak in the previous turn) is spoken to or called upon to speak. (a) is coded if it clearly recognizable that the statement is directed at one individual S: • T addresses the S by his/her name, ends the question with the student's name • T points at a specific S, visibly turns to a specific S, looks in the direction of the S • T speaks so quietly that it can be concluded that the statement was intended for only one single S (b) is **also coded in silent work phases** if the T (or S) interacts with an individual S (providing this has been recorded in the transcript): If interaction is continued with the same S, code it as Sg.

2	***to the same student (Sg)*** *Interaction is continued with the same S* [wzw9]: T to Sg	An individual S (who already spoke in the previous turn) is spoken to again, called upon, is given another opportunity to answer, etc. If the **discussion situation** indicates that it is a longer conversation sequence between the T and a single student, assume that the T is speaking *to the same student (Sg)*, as long as there are no clear signs that this is not the case. If in doubt, assume that the T is talking to a different student (which is coded as *to an individual student*). <u>*Remark*</u>: (a) The situation does not change with break turns (a C turn is introduced between two statements if there is a break of 3 seconds or more) if it is Sg and no new S is involved in the meantime. **In principle, all turns and break turns/C turns are coded *under to whom*.** (b) In principle, S or **Sg is only related to a S-T interaction** and **not to a S-S interaction**. This means that **S is always coded in a S-S interaction** and not Sg (c) **If in doubt, always code S** (not Sg) <u>*N.B.*</u> <u>Contrary to</u> a discussion situation between the teacher and a student, a discussion situation between two students is <u>always coded as to an individual student (S)</u> (i.e., it is not coded to the same student even if it is a longer conversation sequence).
3	***to the teacher (T)*** *Statement is directed at the teacher* [wzw8]: E to T [wzw2]: S to T [wzw5]: SS to T	*(a)* **Responsive S statements** or student answers to teacher questions ***are generally always directed at the T*** (and not at the class/group); this also applies to ***reading aloud when instructed to do so by the T*** <u>*N.B.*</u>: **Learning game** - ***take into account the context in a learning game*** (to S, T, or E, although *learning games* are more frequently directed at the *class/group E* than at the *T*). Here the category *"(who) to whom"* is clarified again later under "Episode: learning game" - ***Reading aloud in a learning game*** is ***always*** coded ***to the class (E)***, ***even if a dialog text is read aloud***, because it is <u>not a natural conversation</u>. (b) **Absence of response to T**: If a linguistic statement is expected from the S (after a question), but the S does not make this statement for at least 3 seconds (recorded as C in the transcript), please code ***T under to whom***. (c) **Speaker/actor** can be the entire class (E) or a group of students (SS) or an individual S (providing he/she did *not* already speak in the previous turn).
4	***Same student to teacher (Sg/T)*** [wzw1]: Sg to T	(a) Is coded when (1) a ***student who is called upon*** responds to a teacher instruction, (2) the ***same S*** continues talking without an interruption, and (3) the ***T-S interaction*** continues with the *same* S. If it is not clear whether the same student or a different student is speaking in the next turn, if in doubt always code *to individual student* (b) Absence of response to T: code ***T*** under ***to whom*** (c.f. 1.3) <u>*Remark*</u>: - When a S is called upon by the T, assume that the S being called upon answers, unless there are clear indications that this is not the case - Student answers to teacher questions are generally addressed to the teacher (and not to the class/group)

		Examples of the "to whom" category: T calls upon an individual S: S S who is called upon answers: Sg/T A different S to the S who is called upon answers: T **Examples of an interaction sequence:** T to class: E T to Hans: S (regardless of verbal or non-verbal) Hans to T: Sg/T T to Hans: Sg T to Maria: S T to Hans: S *Instructions*: (1) You therefore do not take into account the fact that (1) Hans already talked before Maria and (2) that T interacted with Maria only briefly. (2) An interaction sequence between T-S is also coded in a non-verbal situation or in a student work phase. *N.B.* Contrary to an interaction sequence between T-S, an interaction sequence between S-S is not defined as Sg, i.e., always code an interaction sequence between S-S as to an individual student (S)!

2.	**Language**	**All teacher turns (T) and student turns (S, SS, E) are coded.** **N.B.: contrary to before, do not code C turns and AV turns**
	English ***Statement completely in English*** not in videograph window: if this characteristic applies, please skip the entire category 2 [sprst4]: entire [sspraches4]: S [lsprachet4]: T	(a) The naming of a *student's name* in an own turn is always coded as *not assignable*. If the naming of the *student's name is <u>part</u> of an entire sentence*, however, the language of the sentence is coded <u>without taking into account</u> the pronunciation of the name (b) For translation tasks, the Vietnamese word (if relevant, also for the part/sentence to be translated) is not taken into account: "What is the English word for [bac si rang]? " “The English word for [bac si rang] is dentist” *Remark*: only the part of the statement ***to be translated*** is not taken into account when coding the language. <u>*Remark*</u>: The Vietnamese word [teacher] or the word [student] used as a ritual at the beginning of every S statement or every T statement is <u>*not*</u> taken into account.
1	***Vietnamese*** [sprst3]: entire [sspraches3]: S [lsprachet3]: T	Statement completely in Vietnamese
2	***Mixed*** [sprst2]: entire [sspraches2]: S [lsprachet2]: T	Statement partly in Vietnamese, partly in English is coded when only one single word of Vietnamese is spoken within a sentence, except for translations tasks (see above)
3	***Not assignable*** [sprst1]: entire [sspraches1]: S [lsprachet1]: T	- Student is called upon by *only saying his/her name* - The statement is too quiet to make an assessment - Expressions such as "hm", "eh", or expressions that are technically difficult to understand (marked with (p)) - Statements made in silent work phases when the language spoken cannot be identified <u>*Remark*</u>: (1) When entire statements cannot be understood (marked with (p)), the turn is coded as "not assignable." When <u>only part</u> of a turn cannot be understood, the rest of the statement is coded. (2) The Vietnamese word [teacher] or the word [student] used as a ritual at the beginning of every S statement or every T statement is <u>*not*</u> taken into account.

	STUDENT	
3.	**SGeschl (gender)**	**Only S turns**
1	***Girl*** [geschl3]	
2	***Boy*** [geschl2]	

3	***Not recognizable*** [geschl1]	• Class answer • Based on name, voice not recognizable; statement cannot be assigned, too quiet; S not in picture *Instruction:* When a S is called upon by the T, assume that the S being called upon answers, unless there are clear indications that this is not the case. In this case, you can determine the **gender** based on the S name.
4.	**SAeusser (type of student statement)**	**Only student turns (S, SS, E) are coded.**
	Speaking freely Statement is independently formulated by S; S is not bound by any instructions regarding the way he/she answers not in videograph window: if this characteristic applies, please skip the entire category 5 [saeuss7]: entire [sesaeuss7]: only in E	Also for very brief answers, e.g., "yes", "no" It is important that the statement is made by the student himself and not based on a model statement given by the teacher
1	***Speaking based on instructions*** S adheres to the T instructions when speaking; statements are not freely formulated, but formed based on a model statement given by the T [saeuss6]: entire [sesaeuss6]: only in E	- When answering, specific words, parts of sentences, grammatical forms are used; sentences are transformed into another tense; statements are transformed into questions; sentences or gap-fill texts are completed; sentences are searched for in a text provided by the teacher. *N.B.* - For "Speaking based on instructions" it is important that the S statement is made based on a linguistic model statement (grammar-, pronunciation-related). - If S search for words or sentences when doing a translation task, code as *Speaking based on instructions*. Differentiation between *Speaking based on instructions* and *Repetition*: T speaks (as model), e.g., when learning new words or correcting pronunciation S repeats after T (*Speaking based on instructions*) T speaks (as model again) S repeats after T (*Speaking based on instructions*) T instructs S to repeat (e.g., "again") S speaks again (*Repetition*) It is important that a statement is repeated immediately after a model statement for *Speaking based on instructions* in pronunciation practice.
2	***Reading out own text*** S reads out texts that he/she produced himself beforehand (e.g., during silent work, as homework) [saeuss5]: entire [sesaeuss5]: only in E	Was recorded in the transcription in quotation marks and additionally marked with **(w)** at the end of the text *Remark*: When S complete tasks **following grammatical rules** and are asked to read out their answers, code as ***Speaking based on instructions*** (i.e., also when it is read out) *Instruction:* A text is only defined as *Own text* when it is produced without a model.

3	***Reading out text of others*** S reads aloud from a book, reads the instructions from his/her worksheet [saeuss4]: entire [sesaeuss4]: only in E	Is recorded in the transcription in quotation marks and additionally marked with **(r)** at the end of the text *Remark*: - Reading out a gap-fill text provided by the T without solution (i.e., a pure reading task) is coded as *"Reading out text of others."* - Reading out a gap-fill text provided by the T with a solution (i.e., reading out result) is coded as *"Speaking based on instructions.*" - Reading out (e.g., a new word on the board) for practice purposes (e.g., in the practice phase) or for correction purposes is either coded as *"Repetition" or as "Speaking based on instructions"* depending on the context. *Instruction*: In the practice phase (e.g., when learning new words), the S statement is always either coded as *"Repetition" or "Speaking based on instructions"* depending on the context, even if the word is written on the board.
4	***Statement of non-knowledge*** [saeuss3]: entire [sesaeuss3]: only in E	S expresses verbally (e.g., "don't know") that he/she does not know the answer. *N.B.:* non-verbal behavior is no longer assessed.
5	***Repetition*** A statement is repeated (virtually) verbatim [saeuss2]: entire [sesaeuss2]: only in E	e.g., A word, a sentence has to be repeated by the S *Remark*: Repetition for the purpose of **practicing or correcting pronunciation** is coded as ***"Speaking based on instructions***" (see 4.1.)
6	***Not assignable*** [saeuss1]: entire [sesaeuss1]: only in E	- The statement is too quiet to make an assessment - Expressions such as "hm", "eh", or expressions that are technically difficult to understand (marked with (p))

5.	**SALaenge (Length of student statement)**	**Only student turns (S, SS, E) are coded.**
1	***One-word statement*** [salaenge5]: entire [salaengese5]: only in E	Individual names are always one-word statements even if they consist of several parts (John F. Kennedy) *Instruction*: Saying letters or parts of words is coded as a one-word statement.
2	***Sentence fragment*** [salaenge4]: entire [salaengese4]: only in E	Lists of several names, objects that are not embedded in a full sentence are coded as a sentence fragment Word + article *Instruction*: Compound words are coded depending on how they are written, i.e., if two parts of a word are not written together, they are coded as a *sentence fragment.*
3	***Full sentence*** [salaenge3]: entire [salaengese3]: only in E	A full sentence consists of at least a subject and a predicate and, in some cases, an object *Instruction:* - *A full sentence* is evaluated based on its grammar, regardless of its length. - If part of a statement cannot be understood (marked with (p)), the rest of it is evaluated regardless. *Exception*: Greeting ("Good morning") is considered to be a full sentence

4	***Sentence interruption by teacher (or by another student)*** [salaenge2]: entire [salaengese2]: only in E	Student statements that are interrupted by a brief T statement are coded as 2 turns: • Turn 1: First part of the sentence is coded as *Sentence interruption* (marked by ...//) • Turn 2: Interruption by teacher is not taken into account (T turn is not assessed) or interruption by another student is also coded as *Sentence interruption* (marked by //...//) • Turn 3: Second part of the sentence is also coded as *Sentence interruption* (marked by //...). *N.B.:* if the sentence does not contain a second part, the new statement is assessed as usual. *Instruction:* Sentence interruption **always has priority** over the other categories (i.e., when a statement is interrupted, it is always coded as *Sentence interruption*, regardless of the length of the statement)
5	***Not assignable*** [salaenge1]: entire [salaengese1]: only in E	e.g., cannot be understood, technical flaws in the recording - The statement is too quiet to make an assessment - Statements like "hm", "eh"

6.	**SUBJECT (Gegenst)**	***All turns are coded (i.e., also AV and C turns).*** *N.B.*: All S turns also have to be assessed.
	Related to the syllabus not in videograph window: if this characteristic applies, please skip the entire category 7 [gegenst5]	**Guiding question** Is it about stimulating, steering, securing, and consolidating syllabus-related learning processes? Content of statement is related to the topic of the lesson (e.g., a topic related to culture of the country) or to the language/form (grammar, pronunciation, vocabulary, pragmatics); the subject is syllabus-related content and learning processes Teacher statements that encourage S to engage with the subject matter (presentations, questions, instructions) or support engagement with the subject matter (teacher responses to S statements)

1	***Related to discipline*** Activity/statement is related to ***discipline issues/disruptions***; it is frequently related to violations of classroom rules which lead to instruction being disrupted, interrupted, or hampered (e.g., "nobody talks without being asked to") [gegenst4]	**Guiding question** Is it a disruption due to a violation of discipline or of classroom rules? Please code violation of discipline or of classroom rules that result in disruptions, interruptions, hindrance of instruction (e.g., "it is quiet in the class"; "nobody is disturbing his/her neighbor"; "nobody is eating during the lesson") "Please be quiet" "It is too noisy for me"
2	***Procedural*** (a) purely organizational activities which have nothing to do with subject-based instruction	**Guiding question** Are they purely organizational activities and conditions, transitions between teaching phases, or activities that solely serve to prepare and support teaching of learning content? **Distinction:** If the subject-related learning process is fostered, steered, secured, or consolidated by a teacher activity, please code it as **syllabus-related**. Code it as **procedural** if it prepares and supports the subject-related learning process.

	(b) <u>preparatory and supporting activities</u> which prepare and support subject-based and syllabus-related activities (criterion: the statement or activity is directly related to subject-based learning / learning contents) [gegenst3]	Purely ***organizational*** activities: T collects money for a class trip; T writes an organizational note on the board (e.g., bring money for a class trip) Behavior that is related to the organizational ***conditions of instruction*** "Open the window!" "Wipe the board!" "Speak more loudly" (providing that this statement is not caused by restlessness in the class; if it is caused by restlessness, code it as *Discipline-related*) ***Transitions*** between teaching phases: Handing out worksheets; moving around tables and chairs; wiping the board "Put your chairs in a circle" ***Preparatory and supporting*** activities: Statements such as the following are used to prepare, support, and wrap up a syllabus-related activity, but are not actually a syllabus-related activity as such and should therefore be coded as *procedural*: "Open up your books!" "Turn to p. 83 of your book" "Close your exercise books!" "Come to the board" "Get out your books" *S do not have to think about subject-related matters for these activities.* ***Instructions to wrap up or interrupt a task****:* "Stop" "Who can get started now", meaning "Are you finished?" "Stop doing the task now!" *(this is an indication of a transition: the task has come to an end and something else is going to happen)* "Now, all of you listen" "Thank you" *(when it is used to only signal that a task, an activity such as reading aloud, or an answer has come to an end, without it being understood as feedback or a social activity)* ***Reference to homework without an explanation of content*** (explanation of content is coded as *Presentation/explanation*)
3	***Social*** Activity/statement is related to social or personal matters with **no function related to the syllabus** (not about the subject or topic of instruction); it can, however, have a function related to instruction (influence of motivation and atmosphere in the class); it is about caring, showing empathy (possibly also rejection) [gegenst2]	**Guiding question** Does it only have a <u>social purpose or is there a recognizable subject-related function</u>? • Greeting ("Good morning") • Personal statements / statements related to private life ("Is your mother feeling better again?") • Expressing sympathy, comfort; apology • Reference to aspects of social relationships ("It's such fun talking with you"; "You're rather slow today") • Jocular remarks, jokes, etc. (sometimes recognizable because they are followed by laughter) • Also laughing and clapping by students (as own turn in the transcript) <u>*Instruction*</u>*:* If the statement has a recognizable **function related to the syllabus**, is related to the subject of instruction or the topic of instruction, code it as *syllabus-related*. *Example*: asking for an opinion, assessment, evaluation, preference, etc. regarding the **subject of instruction** is ***syllabus-related*** and not social ("Do you like spring?")

4	***Not assignable*** [gegenst1]	e.g., Cannot be understood due to technical reasons; this coding is also used for ***teacher self-commentary*** such as "ok", "yes", "okay", "very good" providing these statements do not constitute *feedback* or have a *procedural* purpose. *Instruction:* All statements that cannot be understood due to technical reasons but where it is clear from the context that it is a syllabus-related, procedural, discipline-related, or social activity, please code accordingly.
		If procedural, discipline-related, social, or not assignable is coded under 6 (subject), please go to the next turn
	TEACHER	**Code all syllabus-related teacher turns** ***(also for C turns if applicable, such as non-verbal interactions, non-verbal presentation, or within a silent work phase)***
7.	**LAktivit (teacher activity)**	(Type/function of teacher activity)
	Presentation and instruction regarding structure	
1	***Other:*** ***Presentation/instruction/explanation/repetition/request (both verbal and non-verbal)*** ... [lakt15]	**Presentations** are syllabus-oriented and initiated by the teacher; i.e., they originate from the T and have the purpose of teaching the teaching content • T presents the facts, explains, illustrates, or comments on something; T talks about something; T reads something aloud • T writes something on the board, demonstrates something using the overhead projector • Other media-assisted presentation T provides subject-related information related to the syllabus. The focus is on **conveying information**. T activity does not serve to confirm or correct an answer provided previously by a S. **Presentation** is also coded if the T expresses his/her own opinion or appraisal of a syllabus-related fact or provides a comment on it. **Presentation** is also coded when the T answers a S question about the ***subject matter*** (T gives information). ***Rhetorical questions*** (to which no answer is expected) are usually coded as *Presentation* ***Instruction:*** Any teacher behavior in a S work phase that does not belong in the categories listed below is coded under this category (*other*) (e.g., teacher walks around in the S work phase...)
2	***Instruction regarding structure / structuring aid*** ***All statements that serve to*** • ***structure thinking,*** • ***to make clear the links between subjects or relations*** • ***emphasize important points*** This includes:	**Requirement:** T expects ***a purely cognitive*** student response which is aimed at **linking, classifying, structuring (e.g., also stressing) information**, but not at solving a task. T activity does ***not*** serve to correct, confirm, secure or consolidate an answer provided previously by a S. An instruction regarding structure can be ***prompted*** by a S answer, e.g., can place the answer in the larger picture.

	• structuring the subject matter • stressing points; linking with other points; showing relationships; references to preceding / previous content • preview, overview, summary, review [lakt14]	T gives information that serves to help the S to better understand, classify, process, or retain the subject matter or to provide the S with a better overview of the sequence of the learning process (teaching procedure). The focus is not on information transfer as such, but on ***steering and guiding the way the S processes the information***. The goal is to provide or activate a cognitive framework, scaffold, model, or reference system. ***Instructions related to structure*** provide cognitive support for learning by specifying and elucidating structures, relationships, and processes. The purpose of ***instructions related to structure*** is to ***better organize and structure either available knowledge or new knowledge that S have to acquire***. (e.g., statements, mnemonics such as: "keep in mind that..."; "that's important"; "always remember that..."
	Types of questions	T questions are primarily *syllabus-oriented* and *initiated by the teacher*; i.e., they originate from the T and have the purpose of teaching the teaching content
3	***Question*** T statement aims to prompt a linguistic answer from the S; T statement does <u>not</u> have to have the linguistic form of a question; what matters is that **<u>a linguistic answer is expected from the S</u>** ***Code additionally: Category 3.E - question dimensions!*** [lfrage]: Question as a whole	**Requirement:** T expects a ***linguistic*** response from the S (other than reading aloud, translating). It can be recognized from the T statement what the T wants the S to do. "Tell us..."; "Explain..."; these are considered to be questions because a linguistic statement is expected. <u>*Remark*</u>: If **the focus is not on an oral response**, but S are initially asked to work on a written assignment or are given a task for a S work phase or an extended cognitive activity ("Think about it first for a couple of minutes ..., and then report ... "), code ***Presentation***/***instruction***. <u>*N.B.:*</u> A **reworded or simplified question** resulting from either an ***absence*** of S answer or a clearly ***insufficient*** S answer (with the goal that the S finds an answer or a sufficient answer <u>to the initial question</u>) is coded as **Assistance**. **Translation tasks** are always coded ***under "Presentation/request"*** (even if they are assigned **in the form of a question**)
4	***Repetition of question*** [lakt11]	<u>*Remark*</u>: Please differentiate between a question asked for the first time and a repeated question. <u>*Instruction:*</u> (1) When a question is repeated, it is essential that the content remains <u>*identical*</u> regardless of how it is worded. (2) There can be several turns between repeated questions. It is only coded as a new *question* if a new or different question has been asked in the meantime. (3) It is always coded as a repeated question, if the original question was formulated in another language (mother tongue).

5	***Comprehension question*** Questions that serve to ensure comprehension a) Questions by the T asking the S if he/she has correctly understood his/her question/statement or a statement made by the others a) Questions by the T verifying that he/she has correctly understood the S answer/statement *Instruction:* The purpose is to ensure comprehension (i.e., to avoid a misunderstanding), **not** to ensure the accuracy of the question/statement. Question about the accuracy of an answer/statement is coded as *question* [lakt10]	**Requirement:** possible both after a previous student response and without a previous student response • T asks whether the S have understood something correctly "Is that clear to you?" "Are you aware of that? Do you know that?" "Are you sure?" • T asks whether he/she has correctly understood the S answer "Have I understood correctly that...?" "Do you mean the following?" *Remark:* Comprehension questions are coded both when the T and the S clarify understanding. A comprehension question should be differentiated from an instruction to find a mistake This is coded as correction (T asks S to find the mistake). The wording might be the same in both cases, e.g., "Is that right?" It is therefore necessary to assess which category applies depending on the situation *Instruction:* A T statement is coded as a ***Comprehension question, Assistance, or Error handling*** depending on the situation Examples: *"Do you mean it like this?"* - as in "Have I understood correctly" as a ***Comprehension question*** - as in "Think about it again!" as ***Assistance*** when the S answer is **insufficient** - as in "Is that really right? as ***Error handling*** when the S answer is **incorrect** (and also under "Dealing with mistakes" when the ***T gets the S to find and/or correct mistakes***)
	Teacher responses	T responses are prompted by S responses; they are *student-oriented*
6	***Assistance*** T provides assistance, further information, explains or rewords a question with the aim of enabling the S to answer or to improve an insufficient answer [lakt9]	**Requirement:** The starting point is a S statement that is recognizably **insufficient or the absence** of an answer. T expects an (improved) answer. Assistance is only coded when the answer is • recognizably **insufficient** (e.g., incomplete; question is only partially answered; quality of the answer is insufficient) • when **no answer** has been given not, however, when the T only wants a different answer. *Instruction*: Many questions aim to elicit suggestions, solutions, answers, etc.; when the T asks for another suggestion, it is not as a rule assistance but a new question (T does not generally expect the S being asked to name all possibilities); assistance is, e.g., when the S does not give an answer and the T gives him a tip When the T signals that an answer is ***incorrect***, code it as ***Error handling***. • For *Assistance*, an answer is expected, like for a question, (contrary to a *Presentation*, where only listening is expected): • *Assistance* is coded when it is recognizable that the T is not satisfied with the S answer and intends to get a better answer.

		• *Assistance* must contain an explicit instruction that shows that the T wants to get an answer or a better answer. • *Assistance* can range from *very general* ("Try and remember"; "We've already talked about that") to *very specific* (for example an explanation). • *Assistance* can be provided by rewording the original question (with the intention of making it easier). • Assistance can also be provided when the T repeats a S statement with a questioning/doubting tone of voice (c.f. category 4). • The purpose of *Assistance* is to obtain a better answer or even to obtain an answer to a concrete question in the first place. General explanations that go beyond answering a concrete question are coded as *Presentation*. • *Assistance* can also have a motivating nature (encouragement), e.g., "I know that it's difficult"... The difference to *Social*: is it only a <u>social matter or does it have a recognizable subject-related function</u>? e.g., (1) "I know that it's difficult but you can do it" as *Social* (2) "I know that it's difficult but if you read the first part again..." as (subject-related) *Assistance* **No assistance**: • T simply asks another S or passes the question on to the class without giving any tips (change of addressee), code as *Calling upon* or *Question/repetition of question* • Statements that appear to be *"Assistance*" but which in fact are used to <u>steer</u> the class discussion into another direction (T wants to hear something specific) are generally coded as a (new) question (or presentation) • Teacher questions that are linked to student responses but bring up a new aspect (asking for additional or complementary information) are coded as *Question* **Silent work phases** • In this context, it is often impossible to determine whether the student activity (doing a writing task) is *insufficient*. • Questions by the S suggest that his/her *comprehension* is still *insufficient*. If the T gives tips and explanations and expects the S to do the task better or to understand better, code *Assistance*. • Code as *Presentation/instruction* when the T assigns the original task again *without giving any additional tips* or assigns a new/additional task that is different to the original task.

7	***Feedback/sanction*** T response to S answer: T gives feedback/sanctions (verbal and/or non-verbal) in response to the S answer ***Additional category 7.E.: Code the content of the feedback*** [lakt4_5]+ [lakt6]+[lakt7_8]	Requirement: it is preceded by a S statement. Praise or reprimand, e.g., yes/no, good/wrong, also non-verbal (nodding or shaking head). If the T frequently makes comments like "okay", please assess the situation to decide whether it really is a T reaction or not merely a habit or a linguistic quirk.
8	***Error handling*** ***Additionally code the entire category 8 "Dealing with mistakes"*** [lakt2]: Dealing with artificial mistakes (see item 8 below) [lakt3]: Dealing with real mistakes	**Requirement:** it is recognizable that the T ***finds*** the student statement ***incorrect***. This can be established based on the fact that • T makes an ***explicit reference to the mistake*** • T ***manifestly corrects the mistake*** himself (e.g., interrupts the S and recognizably corrects a word/pronunciation/the grammatical form, etc.) • T responds to a mistake that has ***indeed been made*** (see dealing with mistakes) rather than simply continuing with the lesson Frequently – but not always – preceded by a student mistake Frequently linked to statements like "is that right?" "that wasn't quite right" "that was wrong" "who knows what the mistake was?" *Instruction*: It is not enough when the T only finds the S statement suboptimal or is aiming at something else. *Remark*: If a mistake is made when listing the results of a learning game, it is not considered to be a mistake because it is not a serious linguistic mistake. The feedback for it is coded accordingly.
9	***Answering himself/herself*** T answers himself/herself, provides the answer to the question himself/herself [lakt1]	**Requirement:** S does not answer or his/her answer does not meet the teacher's expectation This occurs occasionally when the T is aiming at something else.
		Example of teacher behavior: T question A S answers → correctly → T feedback/explanation/ new question → wrongly → T feedback/explanation/ answering himself new question (does not deal with the mistake) → T error handling S no answer/ insufficient answer → T feedback/explanation/ answering himself/ new question → T assistance (so that the S can finally answer the original question A)

3.E	Characteristics of the question	
	LFrage2: Linguistic complexity of the answer required from the S	*Instruction**:*** It is important whether the answer can be seen as sufficient
1	***Low complexity*** Answer only requires the S to say one or several words or a list [lakt13]	- The S has been asked about characteristics, components, etc.: It is enough to name individual words to answer the question appropriately - A sentence does not have to be formulated independently, but is formed based on input from the T or is completed
2	***High complexity (independent formulation)*** Answer requires the S to independently formulate a full sentence (at least a simple main clause) [lakt12]	Question requires S to explain, justify, elucidate, describe, comment on, evaluate, assess, express an opinion, summarize. *Instruction*: To suitably answer the question, the S has to independently formulate a full sentence. *Special case:* When it is not possible to differentiate between *high* and *low* (e.g., 50 to 50), code as *high*.

7.E	LGehaltR: Content of teacher feedback	This concerns the information-related and affective content of feedback (is the focus only on information or is the feedback supported affectively?) *Instructions:* To determine the strength of feedback (e.g., positive or stressed positively), it is necessary to observe the teacher's style of responding. It is often not possible to determine it right at the beginning, but only after the first minutes of the lesson. If this is the case, the turns that have already been coded need to be changed accordingly, if necessary.
1	***Affectively stressed positive feedback*** T praises S in addition to merely confirming that the answer is correct, shows positive feelings and that he/she cares, expresses his/her appreciation, expresses himself favorably; feedback is made with particular emphasis [lakt8]	Great; super (providing it is not merely an empty phrase); I have seldom heard a good answer; I'm happy that you translated it so well; Sometimes also (if observable) especially positive non-verbal signals (vehemently nodding head with delight)
2	***Affectively neutral positive feedback*** T confirms the S answer, tells the student that the answer is right, but remains affectively neutral [lakt7]	Correct; yes; good; the answer is right; non-verbal confirmation (nodding)

3	***Mixed feedback*** T signals that part of the S answer is correct, that he/she is only partially satisfied with the S answer; the S answer has correct and incorrect aspects [lakt6]	So/so; the answer wasn't completely correct; I don't fully agree
4	***Affectively neutral negative feedback*** T signals that the S answer is not correct, tells the student that the answer is wrong, but remains affectively neutral [lakt5]	T only states that an answer is wrong (e.g., a verbal reaction "that was wrong", "wrong", "no", or a non-verbal response "merely head shaking"), but <u>does not focus on the mistake in any other way</u>.
5	***Affectively stressed negative feedback*** T admonishes S, shows negative feelings, disapproval, expresses himself dismissively, makes fun of the S, expresses aversion with regard to the S, shows irony, sarcasm; feedback is made with particular emphasis [lakt4]	Nonsense; you've been sleeping again You've really excelled yourself again (in a sarcastic voice) Occasionally also particularly negative non-verbal signals (providing they are observable), such as rolling eyes, groaning, grimacing as if pain
8.	**Dealing with mistakes** [lakt2]: Dealing with artificial mistakes [lakt3]: Dealing with real mistakes	***<u>Instruction:</u>*** Dealing with mistakes in an artificial situation (e.g., dealing with correction tasks students have been set in the lesson) is coded as "Dealing with artificial mistakes" (new category in the videograph); any other secondary categories (type of correction and type of corrected mistake) should not be coded (contrary to real mistakes: see under 8.1)
8.1	**LUmFeKor: Type of correction (How)**	*<u>Instruction:</u>* Differentiate between "T points out mistake" (1) and T gets S to deal with mistakes (4, 5, 6); if it is **<u>unclear</u>**, determine who is actually dealing with the mistake depending on the context.
	T gets active himself	
1	***T points out mistakes*** Instruction: Is also coded when T expects the S to correct the mistake himself [lfekor6]	**Requirement:** It is recognizable that the T ***finds*** the student statement ***incorrect*** and wants ***to deal with the mistake***. T often refers to mistake using the question form, meaning: is that right? *Distinction from negative feedback:* A mere reference to the fact that something is wrong is coded as negative feedback (see requirement) *Distinction from T explaining the mistake:* T explains/justifies **which mistake** is made (e.g., a grammar mistake) or **where** the mistake is made (e.g., "The following word is wrong in this sentence") or **why** something is wrong.
2	***T corrects mistake*** (correction by teacher) [lfekor5]	T gives the right answer "The following answer is right"

3	***T explains mistake or makes some explanatory remarks*** [lfekor4]	Is also coded when the T ***both corrects and explains*** the mistake "That's wrong because... you didn't apply the following rule" "Let me explain that to you again"
	S have to get active	
4	***T gets the S to find the mistake*** [lfekor3]	T asks the S to find the mistake, asks the S (who made the mistake/other S/whole class) whether something is right or wrong, what the correct version is, wants the S to find the mistake "Is that right?", "Something is wrong there", "Who can find the mistake?", "Where is the mistake?", "What is wrong with that?", "Think again!"
5	***T gets the S to correct the mistake (correction by student)*** [lfekor2]	T explicitly asks S (who made the mistake/other S/whole class) to correct the mistake "What is the correct version?", "Say it correctly now, please", "Who can correct that?"
6	***T gets the S to explain the mistake*** [lfekor1]	T asks S (who made the mistake/other S/whole class) to explain the mistake Is also coded when the T gets the S to ***both correct and explain*** the mistake "Why was that wrong?", "Which rule did you break there?", "Which rule applies here?"
8.2	**LUmFeArt: Type of mistake dealt with by teacher (which)**	Code the type of mistake the T deals with. *Remark*: A statement can refer to one or several types of mistake → it is essential to distinguish between a situation in which the T only deals with only one type of mistake or one in which the T deals with several types of mistake at the same time.
1	***Content-related*** [lufeart7]	Teacher points out a content-related mistake.
2	***Language-related*** [lufeart4] [lufeart5] [lufeart6]	Teacher points out a language-related mistake. • Phonological mistake (pronunciation) • Lexical mistake (vocabulary) • Grammar mistake
3	***Situational / contextual*** [lufeart3]	The T deals with a mistake that is relevant for the specific situation and context (e.g., formal speech in a particular situation / a specific context).
4	***Several at the same time*** [lufeart2]	Different types of mistake are dealt with simultaneously (e.g., both content-related and language-related mistakes). *N.B.:* If it is not possible to distinguish which type of mistake the T is dealing with among different types of mistake, code it as *several at the same time*.
5	***Not assignable*** [lufeart1]	e.g., Correction in a S work phase where the type of mistake is not found out from pictures or non-verbal behavior.

Appendix A2. The adapted coding guides for coding lesson episodes

Codebook

DESI-VN

VIDEO STUDY – Formation of Episodes

Formation of episodes (according to teaching method / class arrangement)

Instruction:

Episodes are longer phases in instruction (class discussion, individual work, etc.) which can mostly be clearly differentiated from each other. They generally last for at least several minutes. This does **not** apply to *transitions*, which can also be very short and should always be coded when they occur. **The teaching method and class arrangement are decisive, not the content.**

It might occasionally be difficult to distinguish one episode from another when relatively short activities are inserted into a teaching phase (e.g., a brief teacher presentation within a class discussion). In this case, it is necessary to determine whether the inserted activity ***actually ends*** the current teaching phase ***or merely serves to support the phase*** (e.g., subsequent explanation of a task within a group work phase):
The formation of episodes is **based on the transcript,** i.e.**, it starts and ends with the transcript.**

Episode	**Explanation**
1. Class discussion	• The focus is on *linguistic exchange* in the instruction. • The linguistic exchange is "public", i.e., *intended for the entire class.* This also applies when only individual students are spoken to (e.g., instructions for a student reading out loud). ***Remark:*** Discussions between the T and one or several students during a student work phase do not constitute a class discussion, but part of the student work phase (see item 2).
1.1 Teacher lecture [episode13]	• A lecture is a long presentation which is generally initiated by the person speaking and is not a response (e.g., a longer explanation in the event of a mistake). • A subject-related lecture is generally prepared and structured (e.g., split into an introduction, body, conclusion). • A lecture can also be non-subject specific (e.g., explanations about a class trip), as is sometimes the case at the beginning of a lesson (warm-up phase), for example. • It can be an oral lecture, a non-verbal activity (writing on the board), or a media-assisted presentation (demonstration using a projector). ***Distinction from AV usage:*** In a teacher lecture, media are only used to support the T activity. When, however, media **replace** T activity (e.g., playing an audio recording), it is coded as *AV usage*.

1.2 Teacher-centered discussion [episode12]	• Class discussion, discussion, talk that is ***led by the T***. • The discussion can be between the T and an individual S, a S group, or the entire class, although the *discussion is intended for the entire class.* • T expects the S to participate (even if in actual fact they don't, they have the opportunity to do so and can ask to speak). ***Distinction from teacher lecture:*** A (usually brief, possibly media-assisted) presentation can occur during a class discussion. *Teacher lecture* is only coded if this presentation is prepared and structured.
1.3 Student lecture [episode11]	• A lecture (which, like a teacher lecture, is prepared and structured; see above) can also be held by a S. • *One or several* S can be involved. • S presents/present solutions to tasks (prepared at home, in the lesson). ***Distinction from learning game:*** A *student lecture* can also occur in a *learning game* (see below). The fact that the actions of one or several persons (*role players*) are simulated is decisive; there is generally interaction between the persons, i.e., the role players respond to each other.
1.4 Student-centered discussion [episode10]	• Class discussion, discussion, talk that is ***led by a S*** *(or possibly also by several S*). • *T can either participate or not participate in the discussion; if participating, the T does not take on a leading role.* ***Remark:*** • The situation "S calls on S" can occur **both** as part of a *student-centered* and as part of a *teacher-centered* discussion. What matters is who has the "*responsibility*" for the class discussion, e.g., who asks questions or provides feedback. • The situation "S calls on S" can also occur during a learning game (e.g., as part of a role play).
1.5 Learning game [episode9]	**Learning games** are games that convey knowledge on given topics or train specific skills alongside the *activities of the game*, thus giving rise to implicit learning. In a learning game, the interaction of the S follows *specific rules* in addition to the usual classroom management rules (e.g., only one person talks; S raise their hands, etc.). ***Instruction***: • As a rule, the learning game stops when the interaction of S according to specific rules comes to an end. • If the learning game is followed by a discussion of the game's results, please observe whether... - *content-related/subject-related matters are discussed* (in this case the follow-up discussion belongs to discussion) - if no content-related/subject-related matters are discussed, but *procedural and social matters are discussed*, e.g., only who won by getting more points (in this case the follow-up discussion still belongs to the learning game).

1.6 AV usage [episode8]	• AV is used in the lesson to present facts. • AV is not only used as a teaching aid (e.g., to support a teacher presentation), but as an independent presentation form (e.g., showing a film, playing an audio recording).
1.7 Beginning and end phase [episode7]	Discussions at the beginning and end of the lesson, in which the *subject-related purpose* of the lesson is not yet discussed or no longer discussed, although the focus is on *linguistic exchange*. *Remark*: Up until now, this phase was recorded under "transition."

2. Student work	
2.1 Individual work [episode6]	• All S work alone on a task, without interacting with each other. • The individual S can either work on the same or on different tasks. • Simply reading texts or copying something off the board, etc., during a break in the class discussion is not considered to be individual work.
2.2 Partner work [episode5]	• Student work in pairs on a task together. • The different pairs can either work on the same or on different tasks.
2.3 Group work [episode4]	• The class is split into (at least two) groups. • The group members should interact or are allowed to interact with each other. • The different groups can either work on the same or on different tasks.
	Instruction: A (mostly brief) discussion can occur between the T and the individual group or individual S in group, partner, or individual work phases.
2.4 Student work with an unclear teaching method or switch of teaching methods / mixed methods [episode3]	At least two cases are possible: 1. It is not clear which type of student work should actually be taking place. 2. Change in teaching method (e.g., first silent work, then group or partner work) without the T explicitly instructing S to change at any point in time. • Indirect instruction at the beginning (e.g., "when you've finished working alone, then discuss the task with your neighbor."). • One teaching method merges into another, without an explicit T instruction. • Several methods at the same time.

3. Several at the same time [episode2]	Two or more teaching methods occur within one phase. This is mainly the case when T deliberately implement differentiation. e.g., Part of the class is participating in a "teacher-centered class discussion" and another part is preparing a learning game.

4. Transition [episode1]	A transition is characterized by the fact that one teaching method/class arrangement or episode comes to an end, but the next does not begin *immediately*. • The transition is always a ***non-syllabus related teaching phase*** (i.e., there are *no specific and systematic subject-related activities*); during a transition, both procedural and *isolated subject-related instructions can occur* or corresponding activities. • The following activities can take place in a transition: changing places, moving around tables or chairs, handing out materials, clearing up desks, fetching things (***a focus on procedural activities***; *T can also still "add" subject-related explanations*). • *A teaching phase can directly follow the previous one without there being a transition* (e.g., T immediately starts with a typical activity for the following teaching phase, ends a presentation with a question which leads into a classroom discussion). • A transition can also *consist of only one turn* (e.g., under C turn, T hands out worksheets). • Transition is *the **time period** between the end of the previous class arrangement and the beginning of a new class arrangement:* - A teaching phase is wrapped up (often explicitly announced "we're stopping now..."). - The following teaching phase is often announced with an explicit instruction or with a task for the S ("Now work on..."). ***Remark**:* If the T does not give an explicit instruction, the T's intention to start the following teaching phases has to be deduced (e.g., T ends a class discussion; based on an earlier instruction or a general rule, it is clear that the S should now start to work in silence). ***Instruction:*** *- a subsequent subject-related instruction to do a new task after having announced the **end of a phase** generally still belongs to the current / preceding teaching phase.* *- an additional subject-related instruction to do a new task after having announced the **beginning of a new phase** generally belongs to the transition. The new phase only starts* when all S are actually concentrating on the new task and are getting started (they no longer have to listen).

Newly formed variables

ebspr_1	Use of native language in classroom discussion (episodes=1-6)
ebspr_2	Use of native language in student work (episodes=8-11)
ebspr_3	Use of native language in transition (episodes=12)
ebsprs_1	Student: Use of native language in classroom discussion (episodes=1-6)
ebsprs_2	Student: Use of native language in student work (episodes=8-11)
ebsprs_3	Student: Use of native language in transition (episodes=12)
ebsprt_1	Teacher: Use of native language in classroom discussion (episodes=1-6)
ebsprt_2	Teacher: Use of native language in student work (episodes=8-11)
ebsprt_3	Teacher: Use of native language in transition (episodes=12)
eblaeu_1	T statement in classroom discussion (episodes=1-6)
eblaeu_2	T statement in student work (episodes=8-11)
eblaeu_3	T statement in transition (episodes=12)

Appendix A3. The adapted rating sheet

Codebook

VIDEO STUDY RATING

RATING SHEET (DESIVN Video Study)

Instructions: In the following, you are required to assess the extent to which the specific characteristic of instruction applies. For each item, there are four possible ratings. It is possible that some characteristics of instruction cannot be assessed in some lessons (e.g., it is only possible to assess how errors are dealt with if errors occur in the lesson). The category ***not assessable*** has been included for such cases.

Characteristic	Description	Degree				Not assessable
1. Goal orientation		does not apply	does rather not apply	applies somewhat	applies	
Teaching objective: Involvement of as many students as possible **[EinbezS]**	Please assess the extent to which this goal was achieved. It is necessary to observe the number of S ***who are actually*** called on to speak in class For the evaluation, it is necessary to assess the number of S who spoke in class, who had the opportunity to say something, or who actually said something • Applies: most S, at least three-quarters of all S (approx. 75-100%) spoke in class • Applies somewhat: at least half, but fewer than three-quarters of all S (approx. 50-<75%) spoke in class • Does rather not apply: at least a quarter, but fewer than half of all S (approx. 25-<50%) spoke in class • Does not apply: only a few S, fewer than a quarter of all S (approx 0-<25%) spoke in class Remarks: *It makes sense to assess the number of S called on to speak in class using the class camera.*					
Teaching objective: Communication **[Kommuni]**	• Applies: The lesson largely consists of discussions in which the teacher and students communicate with each other in an every-day manner and conduct dialogs. The T does not play the role of the teacher, but behaves like a communication partner. • Applies somewhat: frequent and/or longer discussions. • Does rather not apply: only a few discussions. • Does not apply: hardly any discussions. Remarks: – *The conversations do not have to be about subjects that play a role in the every-day life of the students* – *The conversations can also be prepared*					

Teaching objective: Fostering learning and thinking strategies *N.B.*: This refers to the teacher ***explicitly*** fostering, supporting, teaching learning strategies and ***not*** to how capable the S are of implementing strategies **[AnregStr]**	Teaching and fostering strategies, cognitive and meta-cognitive activities. Strategies are systematic approaches for dealing with a task or for achieving a learning objective. Meta-cognitive activities encompass planning, supervising, testing, checking, monitoring, regulating (achieving the objective in another way). Examples: – How to deduce the meaning of words in a foreign language – How to use dictionaries, reference books, and other aids – How to learn and practice vocabulary – How to comprehend grammar rules – How to use mnemonic aids – How to comprehend what one hears – How to deal with texts (summarizing, structuring, underlining, highlighting, asking questions, making notes) – How to approach reading (e.g., scanning, anticipatory reading) – How to verify one's knowledge and understanding – How to manage one's time – How to improve one's oral skills • Applies: T gives instructions and suggestions related to strategy, asks questions about strategy or provides explanations of strategies <u>several</u> times during the lesson • Applies somewhat: T gives <u>several</u> instructions and suggestions <u>related to strategy</u>, asks questions about strategy or provides explanations of strategies at <u>one</u> point in time during the lesson; also, when this only happens once, but lasts longer • Does rather not apply: T gives <u>a brief</u> instruction or suggestion <u>related to strategy</u>, asks one question about strategy or briefly provides an explanation of a strategy at <u>one</u> point in time during the lesson • Does not apply: never <u>Remarks</u>: – *T can explain strategies, ask the S about them, or elicit them from S with questions* – *Fostering strategies is more than just providing assistance for a specific task; the strategies should be generally applicable beyond the actual task in question* – *T does not have to refer to strategies in general, but can also refer to them only in relation to a specific task (general applicability has to be assessed by the rater)*					
Teaching objective: Addressing learning and comprehension difficulties **[EingSchw]**	T addresses learning and comprehension difficulties. He/She tries to detect them, determine them more precisely, and clarify them in a targeted manner. If relevant, he/shetries to eliminate difficulties by teaching the content to the class or to individual S again in a different manner, by explaining the content again including the basics, or by deploying suitable auxiliary remedial measures. T pays attention and monitors (e.g., by asking specific questions) whether learning or comprehension difficulties, gaps in learning, misunderstandings, or systematic errors arise. He/She strives to get to the bottom of them, dig deeper when S lack skills or knowledge. He/She strives to clarify misunderstandings. He deals specifically with S who have difficulties, e.g., during work in silence or group work, pays special attention to them, gives them tips or assistance. • Applies: T always clarifies difficulties that arise exactly, teaches the content again in another way, makes use of auxiliary and remedial measures • Applies somewhat: T frequently clarifies difficulties that arise, addresses them partially, varies his/her approach or deploys special measures • Does rather not apply: T occasionally addresses difficulties that arise again, however without clarifying them more exactly; he/she does not change his/her approach and does not deploy any special measures • Does not apply: T does not clarify difficulties that arise, does not address difficulties, does not deploy any special measures • Not assessable: no learning or comprehension difficulties arise N.B. – *Not merely assistance with random mistakes, but more targeted and systematic clarification and handling of mistakes* – *Not merely asking whether something has been understood*					

Teaching objective: Proposal of demanding, advanced topics *N.B.:* This refers to demanding topics that are ***suitable for high achievers***; it is not necessary to know the actual proficiency level of the S for this **[AnregThe]**	T signals that proposing demanding, difficult, advanced topics is important to him. He/She takes advantage of the opportunity to introduce advanced questions and tasks designed to foster transfer of learning. T assigns appropriate additional tasks and exercises to S who have already achieved the learning objective; T gives S the opportunity to introduce and present demanding, advanced topics • Applies: Demanding, advanced topics, tasks, and questions are systematically included in the instruction and take up a considerable proportion of the entire lesson time • Applies somewhat: Demanding, advanced topics, tasks, and questions are occasionally proposed; if necessary they are broached and dealt with more deeply • Does rather not apply: Demanding, advanced topics, tasks, and questions are rarely proposed; they are broached and dealt with only briefly • Does not apply: no advanced questions; T does not pursue demanding topics. Remark: *Also check "Does not apply" if T stops good S when they explain content too quickly or at a level that is too high, thus preventing the rest of the class from keeping up*					
Attention to accuracy/precision **[Korrekt]**	L signals that *accuracy/precision* is important to him; L pays attention to the accuracy/precision of S statements; L makes it clear when mistakes are made, corrects them, or asks the S to correct them • Applies: L addresses almost all mistakes; when mistakes are made, the T either addresses them himself/herself or asks the S to do so, also when this compromises the flow of the lesson • Applies somewhat: L frequently addresses mistakes • Does rather not apply: L only occasionally addresses mistakes • Does not apply: L does not address mistakes; L does not take mistakes into account, almost never corrects them (one time at most) Remark*:* *The evaluation should be related to the overall tendency throughout the entire lesson, regardless of the stage of the lesson (e.g., for both language learning stages and communicative stages)*					
Attention to fluency **[Fluss]**	T signals that he/she finds *free communication* important; T ensures that the S **have the opportunity to speak fluently** and that the S's speech flow is not disturbed by interruptions, corrections, etc.; T refrains from correcting mistakes during communicative lesson stages; he/she makes corrections in such a way that the flow of the lesson and speech is not interrupted (e.g., by repeating the statement made by the S in the correct form); if necessary, T delays discussion of mistakes until a later point in time; T pays attention to fluent speech, encourages the S to speak freely, gives the S the opportunity to speak as fluently and continuously as possible, if necessary at the expense of accuracy. • Applies: The flow is almost never interrupted; T encourages fluent speech, in particular when it does not occur; he/she also tolerates mistakes and imprecision if they arise; T always allows the S to finish speaking, waits with corrections until a suitable opportunity arises • Applies somewhat: Good flow with only a few, rather short interruptions; occasional encouragement; T waits with corrections until a suitable opportunity arises • Does rather not apply: T makes frequent corrections, interrupts the lesson flow more strongly • Does not apply: T frequently interrupts and frequently makes corrections; the lesson appears to lack flow. T neglects this aspect, emphasizes accuracy, corrects mistakes too frequently or in an unsuitable manner, makes corrections immediately in such a way that the flow of the lesson and speech is interrupted; the S are rarely encouraged to speak freely in the lesson; T frequently interrupts the S's free statements, stifles free speaking Remarks: – *The evaluation should be related to the overall tendency throughout the entire lesson, regardless of the stage of the lesson (e.g., for both language learning stages and communicative stages)* – *The S are free to decide whether they answer in full sentences or not*					

2. Clarity						
Clarity/coherence (with regard to content) **[KlarKo]**	The contents are communicated in a clear and understandable manner; coherent, logical presentation that is easily understandable; presentation focused on key points; suitable examples; clearly understandable questions Clarity/coherence is related to the coherence of the explanations and their logical structure; it is about presenting specific facts, specific sections of the instruction and not the structure of the entire lesson; coherence can be related to the consistency of several questions on one topic or subject matter. • Applies: almost all the T's explanations are clear, understandable, coherent, logical, clearly structured, and comprensible • Applies somewhat: mostly, but occasionally not • Does rather not apply: mostly not, but occasionally • Does not apply: T's explanations are almost never clear, understandable, coherent, logical, and comprehensible; unintelligible explanations; contradictory statements; unsuitable examples; digressions: T jumps from one subject to another Remarks: - *A possible indicator of a lack of clarity is when S frequently ask questions or the T answers the questions he/she has asked the S himself/herself* - *Questions asked by S are an indicator of a lack of clarity when they are related to tasks assigned by the T that are not clear or understandable (questions initiated by S to bring up new points of view are not relevant here)*					
Conciseness of the T's language (language aspect) **[PraegL]**	Clear diction, full, clear, well-planned sentences, cohesive and fluent formulations; T tries to make clear what is important to him with his/her speech (word stress, etc.); clear articulation Conciseness is related to the clarity of the individual linguistic formulations, mostly individual sentences (clear sentence structure) • Applies: T almost always formulates full, clear, well-planned sentences which express the essence; T has clear diction, uses cohesive and fluent formulations • Applies somewhat: mostly, but occasionally not • Does rather not apply: often not • Does not apply: T frequently formulates incomplete sentences that are difficult to understand, sentences in which it is hard to identify the essence; diction is unclear; formulations are halting, lack coherence; frequent "errs", frequent use of expletives and words reflecting awkwardness such as "ok"; incomplete or convoluted sentences, empty phrases reflecting uncertainty Remarks*:* – *This point is not related to content, but exclusively to the T's use of language* – *Emphasis of main points: It is about whether the T tries to make clear with his/her language what is important to him/her. It is irrelevant whether it is "in actual fact" important (according to the rater's opinion).*					
3. Structuredness						
Systematic, logically structured lesson plan **[SysStund]**	The lesson plan is well-structured, systematically and logically consistent, follows a clear logic (moving from easy to difficult, from presentation and processing to autonomous application); the central thread is easily recognizable; the key instruction stages necessary for language teaching are present (language input, language processing, and language application) • Applies: well-structured, systematic, and logical lesson plan; the central thread is clearly recognizable • Applies somewhat: mainly the case, but with minor inconsistencies; the central thread is still recognizable • Does rather not apply: bigger inconsistencies; difficult to recognize the central thread • Does not apply: unstructured, unsystematic lesson plan and teaching process, no recognizable logic or contrary to logic; no central thread; impression that the lesson contents are randomly strung together					

Previews, summaries, reviews; Emphases (cues) **[PreReVi]**	***At the beginning,*** T gives an overview of the lesson plan (related to the content of the lesson), summarizes the key points ***at the end*** of the lesson; ***during the lesson***: summaries; emphases (cues); advanced organizers; emphasis of the importance of specific aspects related to goals or content; formulation of mnemonics; remarks to regulate attention • Applies: present at all three points in time (at the ***beginning*** preview, at the ***end*** summary; ***during the lesson*** summaries, emphases, mnemonics) • Applies somewhat: present at two points in time • Does rather not apply: present at one point in time • Does not apply: not present at any of the three points in time					
Establishment of links; activation or creation of previous knowledge for the following subject matter **[Vorwiss]**	T links old and new subject matter within the same subject; preview of future subject matter; T establishes links to the homework, creates transparency with regard to the importance of lesson contents both for the subject itself and from a more general perspective Checking and, if necessary, activation of previous knowledge for the next subject matter: through vocabulary work, by discussing vocabulary that is essential for understanding the subject matter; the lesson is linked to students' previous knowledge • Applies: linking of content to subject matter that preceded it or will follow it; the subject matter is clearly situated within the context of the subject itself and within a more general context • Applies somewhat: linking is partially recognizable • Does rather not apply: only a few indications (e.g., checking and creation of previous knowledge) • Does not apply: no linking					
4. Monitoring						
Narrow-focused monitoring **[Engfueh]**	Class discussion is steered by the T toward a goal (e.g., frequent expectation of an exact, specific answer or word); blocking out or ignoring of any statements that do not fit in with the lesson goal • Applies: T almost always expects a specific answer, a specific word, frequently blocks out statements that do not fit in with the lesson goal • Applies somewhat: Often, but not always • Does rather not apply: occasionally, but relatively seldom • Does not apply: almost never					
Student orientation **[Sorient]**	S are involved; T considers their questions and suggestions, responds to them • Applies: T constantly involves S; T almost always takes into account their questions and suggestions • Applies somewhat: frequently, but not always • Does rather not apply: occasionally, but relatively seldom • Does not apply: almost never <u>Remarks</u>: *It is irrelevant whether the S actually have an influence on the subject, contents, activities of the lesson*					

5. Classroom management and use of time						
Task orientation **[AuOrient]**	Lesson time is used for subject-related tasks: • Use of lesson time to achieve subject-related objectives and minimization of time devoted to matters that are not expedient and not conducive to teaching success • The lesson is organized such that little time is devoted to matters that are not relevant for the subject • The lesson situation suggests that the T applies a suitable *system of rules* which ensures that the lesson runs smoothly Examples: – T tries to keep discussions about topics that are not expedient for the lesson short and ends the discussion as soon as the goal has been reached – T urges S to be brief when they talk about topics that are not expedient for the lesson, questions whether the discussion makes sense at this particular point in time, or postpones discussing the topic until after the lesson – When S are working in groups or in silence or during discussions, T ensures that the class adheres to the planned time frame – Topics not related to the subject are used to **pursue subject-related objectives, if possible** Remarks – An indication of the presence of rules is that classroom interactions run smoothly and that class activities are initiated and executed without any special explanations, instructions, and justifications – The transitions between different stages of instruction are brief and smooth; there are no unnecessary breaks, no idling; the lesson runs in a rapid, organized manner without any avoidable interruptions; the materials and handouts needed for the lesson are available; handing out and taking in of material is well organized and done quickly; the S know what they have to do at the beginning of a teaching stage or lesson transition; if tables have to be moved around, groups formed, or an open circle of chairs set up, etc., the T ensures that these activities are performed rapidly. – Routines (such as handing out, taking in), activities involving S raising their hands, the beginning of group or silent work are performed smoothly in a standardized manner, no questions • Applies: T uses lesson time almost exclusively for subject-related tasks; no idling; no time wasted • Applies somewhat: T uses almost all of the lesson time; only brief and rare activities not related to the subject; occasional idling; transitions sometimes somewhat ineffective • Does rather not apply: occasional lesson phases that are not used for subject-related tasks or a longer phase; apparent idling; few effective transitions • Does not apply: frequent lesson phases that are not used for subject-related tasks or several longer phases; wasting of time, delayed lesson start; time wasted with transitions between different activities <u>Instructions</u>: - S ask (what they have to do, e.g., whether they are allowed to say something in Vietnamese or have to say it in English, or a lack of organization indicates that rules have not been established at all or not such that they effectively influence S behavior - *This point is related to the classroom management skills required to optimize **subject-related** learning processes (and not to disciplinary matters)*					
Lesson planning and time management **[Uplan]**	The lesson situation reflects the fact that instruction and time have been suitably planned; there is sufficient time for the different parts of the lesson; no additional explanations are required for the class during group and silent work phases and no information needs to be added later, which indicates that the T is well prepared; teaching phases that have started and the lesson itself are concluded in an organized way; no indication that lesson phases have been curtailed or aborted; no rush at the end of the lesson • Applies: good lesson and time planning • Applies somewhat: isolated, minor planning errors (e.g., subsequent explanations; poor use of board) • Does rather not apply: isolated signs of more major planning errors (e.g., individual teaching phases aborted or concluded in a rushed manner) • Does not apply: several indications of more major planning errors <u>N.B.</u>: *Raters should <u>not</u> evaluate how well balanced the didactics are in the individual phases* <u>Remark:</u>					

	Raters should also pay attention to planning errors that are not related to timing (e.g., poor use of board which is disadvantageous for the remainder of the lesson; missing materials; search for materials)					
Free of disruptions; control **[Stfrei]**	Free of disruptions – The lesson takes place without any disruptions; instruction is organized and structured such that disruptions do not occur due to discipline challenges; there are clear classroom management rules which prevent discipline challenges and disruptions from taking place – The noise level is suitable during the lesson. What is considered suitable depends on the respective teaching phase (it should be relatively low during class discussions; during group work it can in some cases be relatively high). An indication of an inappropriate noise level is when T or S complain about the noise level or apparently feel disturbed Control: When disruptions occur, the T has them under control and deals with them appropriately; he/she reacts effectively and intervenes minimally (only intervenes as much as is required to stop the disruptions; he/she does not make an issue of disruptions) – T is always well informed about what is going on in the class, notices everything, and communicates this to the S by, for example, establishing eye contact or briefly talking to potential trouble-makers (T is ever-present, "withitness"), no overreactions; no misjudgments – T takes visible measures to prevent disruptions – T ensures that silent work phases can be carried out without disruptions – T notices immediately when S do not abide exactly by his/her instructions or do not follow them and intervenes • Applies: lesson largely free of disruptions; suitable interventions in the event of disruptions • Applies somewhat: rare minor disruptions or occasional slightly unsuitable interventions in the event of disruptions • Does rather not apply: isolated massive disruptions or frequent smaller disruptions; often inappropriate interventions • Does not apply: frequent massive disruptions; occurrence of discipline challenges resulting in disruptions, misconduct, noise, chaotic situations; necessity to take disciplinary measures; more frequent or extremely unsuitable interventions Remarks: *Inappropriate forms of intervention in the event of disruptions are, for example:* – *overreaction* – *time-related misjudgment (wrong, unfavorable point in time)* – *addressee error (misjudgment of person responsible for the disruption)* – *This is related to classroom management skills required to prevent **discipline challenges which result in disruptions***					
6. Supportive classroom climate						
Social environment; warmth, cordiality **[Sozklima]**	T considers the S's personal and private matters, shows an interest in personal matters; T shows an interest in the S's feelings; T and S get on well, have a good relationship, trust each other The T treats his/her S in a friendly, empathetic, cordial, warm manner. Raters should also pay attention to non-verbal aspects, e.g., smiling, patting S on the shoulder in an encouraging way, etc. The T's general tone of interaction is friendly and respectful. The T treats the S with respect (e.g., greeting) • Applies: consistently friendly interaction; signs of personal interest, warmth, and cordiality • Applies somewhat: mainly friendly, interested, and warm; occasionally dismissive, etc., occasional signs of low respect • Does rather not apply: frequently dismissive, distanced, not very approachable, but occasional signs of interest, warmth, etc. • Does not apply: dismissive, distanced, impersonal, not very approachable, no signs of warmth, cordiality, only factual, condescending, signs of lack of respect; students are put down; contemptuous gestures (e.g., dismissive hand gestures)					

Humor **[Humor]**	T likes to have fun, often banters, jests, or makes jokes, reacts to discipline problems humorously, gives funny examples, also sometimes makes fun of himself/herself, does not always take everything so seriously, spreads a cheerful atmosphere • Applies: T often jests, tells jokes, gives funny examples; cheerful atmosphere; T knows how to joke and also can take jokes at his/her own expense • Applies somewhat: likes to have fun, responds to S jokes, often tells jokes himself/herself; funny examples • Does rather not apply: rarely tells jokes himself/herself, leaves S to jest but does not respond to their jokes, makes it clear where the fun ends • Does not apply: matter-of-fact, serious, humorless; no fun at all, cannot take a joke; no signs of humor; possibly ironic, sarcastic, cynical, or tells jokes that are exclusively at the expense of the S Remark: *Humor is not the same thing as being friendly, showing a personal interest, etc., but relatively independent of these aspects*					
7. Mistake-making environment						
Positive treatment of mistakes (on the T side) **[PosFehl]**	T treats mistakes positively, productively uses mistakes T talks about mistakes in a constructive manner without putting the S down, uses mistakes to make connections clearer T deals with the causes of and logic behind mistakes T gives S space to correct mistakes T emphasizes the benefit of mistakes for the learning process • Applies: mistakes treated in a constructive (stimulating and supporting the learning process) and motivating (boosting willingness to learn) manner • Applies somewhat: favorable treatment of mistakes or at least an affectively neutral treatment of mistakes; the occurrence of mistakes is accepted; mistakes are not however used constructively • Does rather not apply: occasional indication of mistakes being dealt with negatively • Does not apply: mistakes frequently dealt with negatively; dismissive treatment of mistakes; T reacts negatively, reprovingly, disparagingly to mistakes • Not assessable: there are no mistakes or the T does not address mistakes Remark: *Raters should pay attention to **how** the T deals with mistakes*					
Positive mistake-making environment (on the S side) **[PosKlima]**	Positive attitude of class to mistakes Class does not react negatively or even reacts supportively to mistakes made by individuals • Applies: positive attitude of class to mistakes; positive treatment of mistakes; fellow S are supported when they make mistakes; assistance • Applies somewhat: favorable treatment of mistakes or at least an affectively neutral reaction to mistakes, but no support or assistance • Does rather not apply: occasional negative reactions to mistakes • Does not apply: negative attitude of the class to mistakes; frequent negative (e.g., contemptuous) reactions of class to mistakes made by fellow S; laughing at S who make mistakes • Not assessable: there are no mistakes or the S do not react to mistakes					
8. Quality of motivation						
Teacher commitment enthusiasm **[Lengag]**	T seems to be very stimulating, energetic, and active, mentally involved. You get the impression that the T loves the subject English and the English language and is enthusiastic about it. He/She clearly enjoys teaching. The T expresses the fact that he/she is interested in the teaching goals and contents • Applies: great enthusiasm; T clearly communicates his/her positive attitude to the subject • Applies somewhat: T is partly very stimulating, shows a positive attitude and interest in the subject, but without coming over as enthusiastic • Does rather not apply: T is rather reserved and undemonstrative • Does not apply: T comes over as dull, not particularly mentally involved, bored, routined, uninterested, indifferent, negative					

Student commitment **[Sengag]**	Dedicated and interested cooperation and involvement of S; rapid completion of tasks assigned; S show interest, energy, mental involvement • Applies: most or all S appear to be interested in the instruction; they seem interested and mentally involved; all S join in, raise their hands to speak frequently and spontaneously • Applies somewhat: a considerable proportion of the S appear to be interested in the instruction, seem interested and mentally involved • Rather does not apply: only a few S appear to be interested in the instruction, seem interested and mentally involved • Does not apply: most or all S seem bored by the instruction or seem not to be interested in the lesson; low involvement; very few students spontaneously raise their hands to speak					
Relevance to every-day life and authenticity **[Authent]**	T shows the relevance to the every-day life of the S, uses tasks, examples, materials, etc., that come from the every-day life of S, treats the subject matter in such a way that it is relevant for the every-day life of S T uses authentic and illustrative tasks, materials, examples The tasks are highly challenging due to the illustrative materials, examples Reference to the usefulness and importance of the English language for life, travelling, career, school, other subjects, surfing in the Internet T incorporates the sphere of experience of the S; T establishes a link to the interests of the S • Applies: T frequently tries to integrate authentic elements, uses authentic tasks if possible, often incorporates the sphere of experience of the S • Applies somewhat: T tries to integrate authentic elements here and there, often uses authentic tasks • Does rather not apply: T only rarely uses authentic tasks • Does not apply: The lesson comes over as being abstract, remote from reality, textbook-like; materials are stilted, didactically prepared					
Encouragement/stimulation of S statements **[Ermut]**	T encourages and stimulates S statements, shows respect when S express themselves, encourages S to speak, gives S the feeling that they are competent • Applies: T frequently encourages S in a convincing manner, strikes the right tone • Applies somewhat: T encourages S, but not always in a convincing manner, sometimes does not strike the right tone • Does rather not apply: occasional encouragement; frequent, but rather mechanical and routined encouragements • Does not apply: hardly any encouragement. Remark: *This point does not relate to feedback, but to encouraging S to make statements*					
9. Variation/suitability						
Variation of instruction; adaptivity **[Adaptiv]**	T visibly tries to adapt instruction to the different S and to the varying difficulty of the subject matter Variation of the difficulty of questions, instructions, or response depending on the level of proficiency/skills/other personality characteristics of the S Variation of volume and speed of speech depending on the demands/difficulty/importance of the subject matter; more repetition when presenting difficult subject matter (more detailed presentation of subject matter; explanation of subject matter one more time in other words) T uses measures to temporarily differentiate within the class; this involves groups of S or individual S being assigned different tasks; e.g., group work, partner work, individual work with differentiated tasks • Applies: T frequently varies his/her measures depending on the subject matter and the students, uses different tasks when differentiating within the class • Applies somewhat: several signs • Does rather not apply: rare signs • Does not apply: no sign of variation of instruction; no sign of different tasks being assigned to differentiate within the class Remark: – *Merely organizing group, partner, or individual work does not constitute differentiation; differentiation only occurs when different tasks are assigned* – *Raters should not evaluate whether the T has adapted instruction suitably or successfully*					

Level of difficulty: difficulty and teaching speed **[Niveau]**	Suitable level of difficulty; a good fit of teaching contents (cognitive level, required knowledge, life situation) with the preconditions of the target group: – S are met on their level; S are not under- or over-challenged – Suitable: neither too fast nor too slow; a rather fast speed of instruction, which motivates S to join in without over-challenging them, is mostly favorable • Applies: good fit; S join in; no signs of them being under- or over-challenged; no indications that they are having difficulties understanding or are bored; no questions by S reflecting that they are having problems understanding; fast yet not too high teaching speed • Applies somewhat: rarely signs of students being under- or over-challenged • Does rather not apply: frequent signs of students being under- or over-challenged • Does not apply: massive signs of students being under- or over-challenged; many indications that S are not keeping up or are bored Remark: *Raters should evaluate whether the degree of difficulty and speed of instruction are successfully adapted to the proficiency and skills of the students*					
10. Teacher language	Instructions: • *Overall impression throughout the entire lesson* • *This only applies to statements made in English (not in Vietnamese or when saying names)*					
Suitable way of speaking; formulations that correspond with the way the English / Americans speak **[Sprechw]**	• Applies: formulations and idioms fully correspond with the way the English / Americans speak • Applies somewhat: formulations and idioms do not always correspond with the way the English / Americans speak • Does rather not apply: formulations and idioms often deviate from the way the English / Americans speak • Does not apply: substantial deviations from the way the English / Americans speak; T often uses formulations and idioms that deviate considerably from the way the English / Americans speak N.B.: *This aspect should be evaluated independently of pronunciation.*					
Suitability of pronunciation (including accent-free pronunciation) **[Ausspra]**	• Applies: pronunciation is accent-free, largely corresponds with the way the English / Americans speak • Applies somewhat: slight deviations from the way the English / Americans speak; slight accent which is not very noticeable and subtle; slight indications of a regional dialect, some sounds with a regional accent • Does rather not apply: considerable deviations from the way the English / Americans speak; pronounced (Vietnamese) accent • Does not apply: significant deviations from the way the English / Americans speak; strong accent that is disturbing and hampers comprehensibility					
Vocabulary: reliability **[Wortscha]**	• Applies: extensive and reliable English vocabulary; rapid recall of words; fluent speech that is not hampered by difficulties recalling words • Applies somewhat: occasionally slightly halting speech because words do not come automatically; occasional need to search for the right word • Rather does not apply: occasionally does not find the right word, has to reword • Does not apply: temporary use of Vietnamese words because he/she obviously cannot reliably recall the right English word (but not when Vietnamese words are used because the S do not understand the English words); uses incorrect words, struggles to find the right words					
Grammar **[Gramm]**	• Applies: reliable availability of English grammar; competent command of the language with no noticeable deviations • Applies somewhat: occasionally uncertain about more complex and less common grammatical rules; T occasionally self-corrects • Does rather not apply: occasional visible difficulties that are not corrected and apparently go unnoticed • Does not apply: noticeable / frequent grammar errors and/or uncertainty occur					
Accuracy of content **[InhKor]**	• Applies: factually correct explanations; the contents presented by the T are factually correct; there are no content errors • Applies somewhat: occasionally minor inaccuracies and inconsistencies • Does rather not apply: explanations sometimes not entirely correct; occasional minor inaccuracies and inconsistencies • Does not apply: T makes noticeable mistakes, frequently makes statements that are not factually correct					

Verbal tics and quirks **[Tics]**	• Applies: frequent use of stereotyped or unsuitable, meaningless phrases, words that have taken on an own meaning ("Oh", "at the end of the day", so-called therapeutic grunts, etc.) that are often disturbing • Applies somewhat: frequent minor language tics (permanent use of "ok"); occasional obviously disturbing quirks • Does rather not apply: occasional use of minor language quirks • Does not apply: nothing striking; no tics and quirks Remark: *Frequent repetition of single words such as ok is only considered to be a quirk if the words do not recognizably have a teaching purpose (e.g., feedback)*					

11. CHECK LIST		Put a cross next to materials used (yes/no)
Use of following materials and technologies in the lesson	§ Board	**[Tafel]**
	§ Overhead projector	**[Overh]**
	§ Video, TV	**[Video]**
	§ Audio materials (audio tape, MC, CD, MD)	**[Audio]**
	§ Notebook, PC	**[PC]**
	§ Projector	**[Beamer]**
	§ Internet, email	**[Internet]**
	§ Language laboratory	**[Labor]**
	§ Objects brought to class	**[Gegenst]**

Appendix B. Item analysis

In order to enhance the reliability and validity of the tests and test scores, test items were analyzed according to different item selection criteria in the LSAs (Eid et al., 2010; Itzlinger-Bruneforth, Kuhn, & Kiefer, 2016; Osterlind, 1998; Rost, 2004; Schmeiser & Welch, 2006). In the following sections, item statistics are reported, which were used for selecting items for/eliminating items form subsequent calibrating and scaling processes.

Appendix B1. Coding and scoring student responses

The student responses in this study were coded as follow: 1 = *correct*, 0 = *incorrect*, 8 = *incomplete, unreadable or unassignable*, 9 = *missing* (*non-response*). All codes other than 1 (*correct*) were treated as 0 (*incorrect*). With each correct response, a student scored 1 point.

Appendix B2. Item difficulties

The difficulty (or easiness) of an item denotes the correct response rate:

$$p = \frac{N_{\text{correct}}}{N_{\text{test}}}$$

An item with p = 0.3 has a correct response rate of 30%, an item with p = 0.7 has a correct response rate of 70%. The second item is easier than the first item.

B2a. C-test item difficulties

The item difficulties ranged from .05 to .87 at T1, and from .04 to 0.94 at T2. The mean difficulty of all items at T1 was .40, at T2 was .48.

The item difficulty at T2 was higher than at T1 on average, that means the items became easier for the sample over one school year. Assuming that the mean item difficulty did not change over time, this change could be attributed to the growth in student ability. This trend was observed in all subsamples by school location.

Table 36: Item difficulty by school location at each MP

	Hanoi	Ho Chi Minh city	Bac Ninh
p_{mean} at T1	.48	.32	.27
p_{mean} at T2	.56	.45	.39
p_{min} at T1	.09	.03	.01
p_{min} at T2	.07	.03	.01
p_{max} at T1	.90	.84	.86
p_{max} at T2	.96	.94	.92
$N_{p < 0.05}$ at T1	0	5	12
$N_{p < 0.05}$ at T2	0	2	10

Note: p_{mean} = average item difficulty, p_{min} = minimum item difficulty, p_{max} = maximum item difficulty, $N_{p < 0.05}$ = number of items with $p < .05$

The item difficulty statistics for each subsample by location are shown in Table 36. The test items were easier for the students in the two cities Hanoi and Ho Chi Minh city than for students in the rural province Bac Ninh.

Considering items with difficulty $p < .05$ as too difficult (cf. Itzlinger-Bruneforth et al., 2016), no test items were too difficult for the students in Hanoi at both MPs, while 12 items at T1 and 10 items at T2 were too difficult for the students in Bac Ninh.

Items, which only few or almost all students can solve correctly, reveal no helpful information about differences between students (Livingston, 2006). Among the too difficult items, there were words and terms, which were not found in the English textbooks of the 9th grade as well as of other lower secondary grades (6th to 8th grades) (see Table 37). It was assumed that the students in Bac Ninh had less out-of-school learning opportunities than students in the urban areas (Hanoi and Ho Chi Minh city). Thus, (too difficult) items which were not part of the official textbooks were eliminated to enhance the curricular validity of the test.

Table 37: Themes, topics and grammars in the 9th grade English textbook

Unit	Theme	Topic	Grammar
1	A visit from a pen pal	Make & response to introduction Scan for specific information Write a personal letter	The past simple The past simple with *wish*
2	Clothing	Ask and respond to questions on personal preferences Ask for and give information Write an exposition	The present perfect The passive (review)
3	A trip to the countryside	Ask for and give information Complete summary Write a passage	Modal *could* with *wish* The past simple with *wish* (review) Prepositions of time Adverb clauses of results
4	Learning a foreign language	Seek information Express opinions Scan for specific information Write a letter of inquiry	Modal verb with if Direct and reported speech: - *here* and *now* words in reported speech - reported questions
5	The media	Agree and disagree Ask for and give opinions Write a passage	Tag questions Gerunds after some verbs
6	The environment	Persuade Complete a questionnaire Write a letter of complaint	Adjectives and adverbs Adverb clauses of reason: *as, because, since* Adjective + that clause Conditional sentences: type 1
7	Saving energy	Show concern Give and respond to suggestions Seek information Write a speech	Connectives: *and, but, because, or, so, therefore, however* Phrasal verbs Make suggestions: *suggest + verb-ing, suggest (that) + S + should*
8	Celebrations	Give and respond to compliments Describe events Express opinions Write a letter to a pen pal	Relative clauses Adverb clauses of concession
9	Natural disasters	Make predictions Talk about the weather forecast Describe events Write a story	Relative pronouns Relative clauses (continued)
10	Life on other planets	Talk about possibility Seek information, write an exposition	Modals: *may, might* Conditional sentences: type 1 and type 2

All items with $p < .05$ at any MP by any subsample by school location were analyzed in this regard. In total, 8 items were eliminated (see Table 38).

Table 38: Too difficult items, decision and explanation regarding elimination

Item	Coding guide(s)	Hanoi		Ho Chi Minh city		Bac Ninh		Explanation	Decision
		p_1	p_2	p_1	p_2	p_1	p_2		
ct112	th*ey*	.30	.22	.12	.12	.07	.04	found in the lower-secondary textbooks	
ct207	a*nd*	.28	.38	.20	.23	.04	.13	found in the lower-secondary textbooks	
ct209	atten*tion*	.21	.35	.10	.16	.02	.06	found in the lower-secondary textbooks	
ct218	patie*ntly*	.13	.15	.05	.08	.01	.02	*not* found in the lower-secondary textbooks.	eliminated
ct221	car*rots*	.20	.23	.07	.07	.05	.02	*not* found the lower-secondary textbooks	eliminated
ct223	any*way*	.27	.30	.21	.20	.01	.03	*not* found in the lower-secondary textbooks	eliminated
ct309	a*nd*|a*lso*	.34	.52	.21	.36	.05	.19	found in the lower-secondary textbooks	
ct310	i*ts*	.18	.10	.05	.03	.01	.03	*not* found in the lower-secondary textbooks	eliminated
ct312	hair-*drier*|hair-*dryer*	.22	.36	.05	.15	.02	.04	*not* found in the lower-secondary textbooks	eliminated
ct325	posi*tion*	.16	.18	.09	.08	.06	.04	found in the lower-secondary textbooks	
ct614	w*ho*	.14	.18	.07	.11	.01	.08	Relative pronouns including *who* was found in Unit 9 (at the end of the school year) in the 9th grade textbook	
ct618	o*ut*	.09	.07	.03	.03	.01	.01	*not* found in the lower-secondary textbooks	eliminated
ct622	trave*llers*	.14	.11	.04	.07	.01	.02	*not* found in the lower-secondary textbooks	eliminated
ct623	instr*ucted*	.18	.17	.14	.13	.01	.03	*not* found in the lower-secondary textbooks	eliminated

B2b. LC-test item difficulties

All 18 LC-test items had a difficulty between .14 and .78 at T1, and between .15 and .82 at T2. The mean difficulty of the items was .44 at T1, and .50 at T2. None of the items was eliminated.

Appendix B3. Missing value analysis

B3a. Non-responses in the C-test

Table 39 shows the proportions of the non-responses (missing values) and not-reached responses (all consecutive non-responses clustered at the end of a booklet except the first value of the missing series, OECD, 2012) at both MPs.

Table 39: Non-response and not-reached rates

			Non-response rate				Not-reached rate			
	N_p	N_i	*mean*	*SD*	*min*	*max*	*mean*	*SD*	*min*	*max*
T1	2036	100	.36	.17	.04	.80	.04	.05	.00	.25
T2	2029	100	.24	.14	.02	.57	.01	.03	.00	.15

Note*:* N_p is the number of participants (students); N_i is the number of items in each test/booklet; *mean*, *SD*, *min*, *max* are arithmetic mean, standard deviation, minimum and maximum of the respective category over all the items at each MP.

The proportion of all item non-responses was $p_{missing\_T1}$ = .36 at T1, and lower at T2 ($p_{missing\_T2}$ = .24). Differences between items with regard to the item non-response rate were observed. At T1: minimum value of item non-response rate was .04, maximum value was .80; at T2: minimum value = .02, maximum value = .57.

At both MPs, the proportion of the not-reached responses was low: $p_{not\text{-}reached\_T1}$ = .04, $p_{not\text{-}reached\_T2}$ = .01. Thus, it can be assumed that the main reason for item non-responses in this study was not because that the C-test was too long.

B3b. Non-responses in the LC-test

The missing rates p_{miss} of LC-test items ranged from .01 to .23 at T1 (mean p_{miss_T1} = .09), and from .01 to .11 at T2 (mean p_{miss_T2} = .04), less than of C-test items. Due to the special test administration of the LC-test, nearly no not-reached responses of all items at both MPs were found.

Appendix B4. Item discrimination

Item discrimination (r_{pb}) denotes the point biserial correlation coefficient between the responses of an item with the sum score of the test at each MP:

$$r_{pb} = \frac{M_1 - M_0}{s}\sqrt{\frac{n_1 n_0}{n(n-1)}}$$

where *s* is the sample standard devidation, M_1 is the average of the sum scores of all students who response correctly (group 1) to an item, M_0 is the average of the sum scores of all students who do not response correctly (group 2) to that item, n_1 is the size of group 1, n_0 is the size of group 2, and *n* is the total sample size. The possible range of r_{pb} is from –1 to 1. A high value of r_{pb} indicates high discriminatory power of the item, that means it can well differentiate between low- and high-achievers based on the total test scores. An item with negative or small discrimination (r_{pb} < .10) should be eliminated from the test, since it does not contribute to what the test measures as a whole (Itzlinger-Bruneforth et al., 2016).

None of the C-Test items had an item discrimination r_{pb} < .20 at both MPs, except item `ct315` with low discrimination at T2 (r_{pb} = *0.08*). No C-test items were eliminated only due to their discriminatory power.

One LC-test item had a negative discrimination at both MPs (r_{pb} = –.12 at T1, and r_{pb} = –.11 at T2), and was eliminated from the test. All other LC-test items had good item discrimination at both MPs.

Appendix B5. Differential item functioning (DIF)

One important analysis in the item selection process is the differential item functioning analysis (DIF, see Holland & Wainer, 1993; Embretson & Reise, 2000; Trendtel, Schwabe, & Fellinger, 2016). DIF occurs when an item functions differently for different groups which have the same overall ability estimates (e.g. regarding school location, MP). The results might be misleading when a specific group has systematic advantages/disadvantages due to construct-irrelevant reasons. In this study, the *uniform*-DIF (Mellenbergh, 1982) of the items was analyzed, which implies whether one item is easier/more difficult for one group than it is for other groups over all ability levels. An item was only eliminated from the test based on DIF analysis results if construct-irrelevant reasons were strongly suspected for a large DIF effect.

B5a. DIF-analysis method and ETS classification

Among many available methods to analyse DIF effects (Trendtel, Schwabe, et al., 2016), the method suggested by Swaminathan & Rogers (1990, cf. Zumbo, 1999) based on logistic regression was applied in this study, which can be calculated via the R function dif.logistic.regression() in the R packages **sirt** (Robitzsch, 2014). The logistic regression equation is:

$$P(Y = 1|x, g) = \frac{\exp(\beta_0 + \beta_1 x + \beta_2 g + \beta_3 xg)}{1 + \exp(\beta_0 + \beta_1 x + \beta_2 g + \beta_3 xg)} \qquad (5)$$

The probability $P(Y = 1|x, g)$ of having the correct response to an item given an overall ability *x* and group membership g (0 = reference group, 1 = focal group) is formulated as in equation (5). The exponential of $\hat{\beta}_2$ provides the reference-to-focal odds ratio for endorsing the item, conditional on the overall ability, and ranges from 0 to ∞. If $\hat{\beta}_2$ is significantly different from 0, uniform DIF is indicated.

To identify items with substantial uniform DIF effect, the ETS[2] classification (Monahan, McHorney, Stump, & Perkins, 2007; Zwick, 2012) was applied, in which not only the significant test is taken into account but also the effect size LR-D-DIF:

$$\text{LR-D-DIF} = -2.35(\hat{\beta}_2)$$

According to ETS classification system, the items can be assigned into three categories:

A. Category A. Items with negligible or nonsignificant DIF. Defined by LR-D-DIF not significantly differenz from zero or absolute value less than 1.0
B. Category B. Items with slight to moderate magnitude of statistically significant DIF. Defined by LR-D-DIF significantly different from zero and absolute value of at least 1.0 and either less than 1.5 or not significantly greater than 1.0
C. Category C. Items with moderate to large magnitude of statistically significant DIF. Defined by absolute value of LR-D-DIF of at least 1.5 and significantly greater than 1.0

Items with DIF effects of category C were regarded as having substantial DIF effect. They were further analyzed with regard to whether there were construct-irrevant causes for that, and only in this case they should be eliminated from the test.

To perform DIF analysis of the C-Test items, the weighted likelihood estimates (WLEs) (Warm, 1989) based on the Rasch model (see chapter VI.2.3) were used as measures of the student general ability *x*. The following grouping variables were considered: gender (female = 1, male = 0), school location (Hanoi, Ho Chi Minh city and Bac Ninh, each location is reprented by a dummy coded variable), and MP (T1 = 0, T2 = 1).

B5b. DIF effects of C-test items

No item with substantial gender DIF effect at both MPs were found.

Regarding school location, there was a long list of items with substantial DIF effect. Items with substantial location DIF effects were of all testlets: 10 of testlet C01, 9 of testlet C02, 11 of testlet C03 and 7 of testlet C06.

[2] Educational Testing Service, Princeton, NJ 08541: https://www.ets.org/

Table 40: Location DIF analysis results

		Location x MP						
		Hanoi		**Ho Chi Minh**		**Bac Ninh**		
Text	item	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**	N DIF effects
C01	ct101				C-			1
C01	ct105	C+	C+					2
C01	ct106		C+		C-			2
C01	ct108					C+	C+	2
C01	ct110					C-		1
C01	ct112	C+						1
C01	ct113	C-					C+	2
C01	ct114		C+				C-	2
C01	ct118					C-	C-	2
C01	ct125						C-	1
C02	ct201		C+					1
C02	ct207					C-		1
C02	ct209		C+			C-	C-	3
C02	ct212	C-						1
C02	ct217	C+			C-			2
C02	ct219		C-					1
C02	ct221		C+		C-		C-	3
C02	ct222						C-	1
C02	ct223					C-	C-	2
C03	ct305					C-		1
C03	ct306					C-		1
C03	ct309					C-		1
C03	ct310	C+				C-		2
C03	ct312	C+	C+	C-				3
C03	ct315		C-		C+			2
C03	ct316						C+	1
C03	ct319				C+	C-	C-	3
C03	ct321				C+			1
C03	ct323		C-				C+	2
C03	ct324					C+	C+	2
C06	ct606		C-	C-	C+	C+		4
C06	ct608				C-			1
C06	ct610	C+		C-				2
C06	ct615					C-		1
C06	ct622	C+				C-		2
C06	ct623			C+		C-		2
C06	ct624	C-				C+		2

Construct-irrelevant reasons for only one item with substantial location DIF effects were suspected: A the change of sign of the location DIF effect of item ct606 (orb*it*) was observed, it was more difficult for students in Ho Chi Minh city at T1 (p_{1_HCM} = .03 vs. p_{1_HN} = .11, p_{1_BN} = .09), but easier for them at T2 (p_{2_HCM} = .40 vs. p_{2_HN} = .28, p_{2_BN} = .30). In the textbook, the word *orbit* was found in a reading text in the last Unit (Unit 10) at the end of the school year, the same time at the second MP in Ho Chi Minh city, while the second MP for students in Hanoi and Bac Ninh was about two or three weeks earlier. It was assumed that students in Ho Chi Minh city had learnt this word before the test, while students in other locations had not learnt it at the time of the test. Hence, this item was eliminated.

B5c. DIF-effects of LC-test items

There were six items with substantially location DIF effects regarding results of the LC-test. However, no construct-irrelevant reasons were found, hence no items were eliminated due to location DIF effects.

B5d. Item parameter drift

Item parameter drift (IPD) refers to differential change in item difficulties over time (Goldstein, 1983). An item with substantial IPD becomes either relatively much easier or much more difficult over time in comparison to the change in difficulties of all other test items. For comparing test results at both MPs and estimating student growth, it is necessary to examine IPD (Goldstein, 1983; Trendtel, Pham, et al., 2016; Wells, Subkoviak, & Serlin, 2002). IPD analysis was done via DIF analysis with MP as grouping variable (Rupp & Zumbo, 2006).

IPD of C-test items

Overall, there were 14 items with substantial IPD, two of them were comparably much easier, and twelve were relatively more difficult at T2.

Among them, six items were already eliminated from the C-test due to other problems involving item difficulty and substantial location DIF effect.

Two other items also had substantial location DIF effect: item `ct118` was comparably much more difficult for students in Bac Ninh at both MPs, item `ct315` was absolutely and relatively more difficult for students in Hanoi at T2; in addition, `ct315` had low discriminatory power at T2 (see above). Since these items involved multiple problems, they were eliminated from the C-test.

Other six items with substantial IPD were: `ct112` (th*ey*, $p_1 = .18$, $p_2 = .15$), `ct202` (lo*ved*, $p_1 = .41$, $p_2 = .27$), `ct206` (ope*ned*, $p_1 = .22$, $p_2 = .17$), `ct225` (Nob*ody*, $p_1 = .68$, $p_2 = .87$), `ct601` (stud*ents*, $p_1 = .28$, $p_2 = .23$), and `ct619` (experi*ment*, $p_1 = .31$, $p_2 = .26$). Their item difficulties at T1 and T2 together with the difficulties at both MPs of the remaining (not eliminated) C-test items are shown in Figure 46, these six are marked in red rectangles. Among them, item `ct202` was noticeably much more difficult at T2, but no construct-irrelevant reasons were found for it.

Since these six items involved no other problems, and no construct irrelevant reasons were found for their IPD effect, they were not eliminated from the C-test.

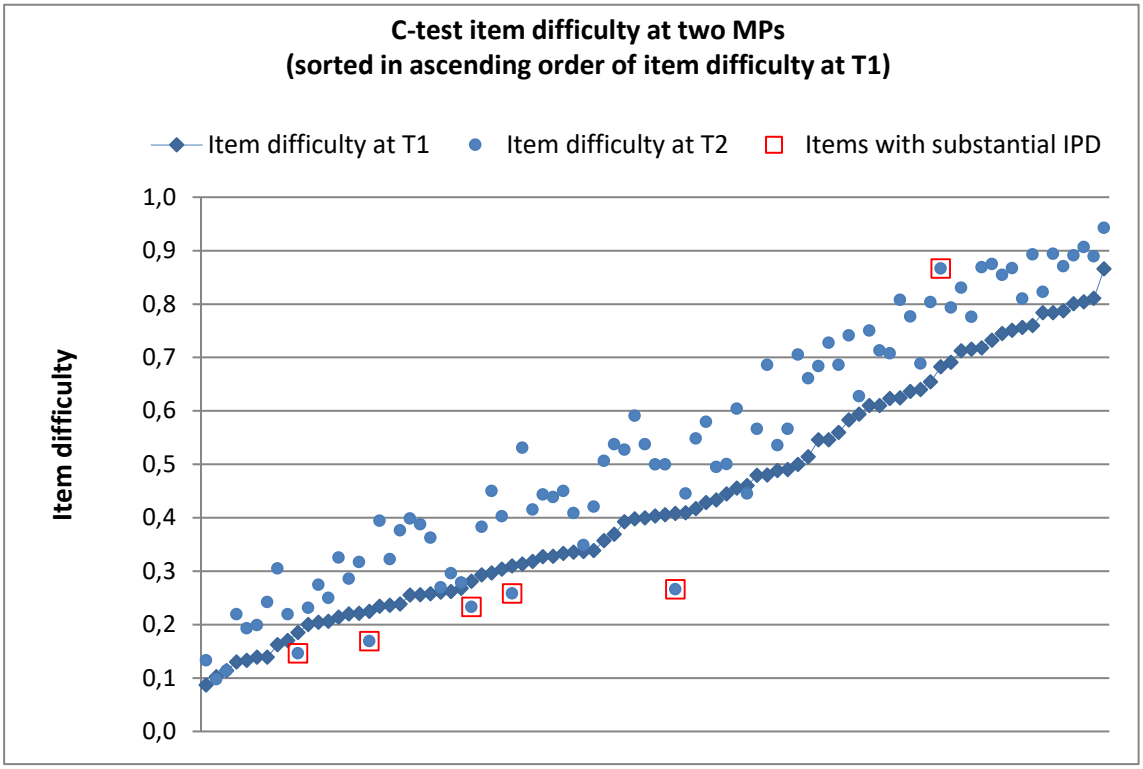

Figure 46: C-test item difficulties at two MPs

IPD of LC-test items

All LC-test items (excluding item `LC17` with negative discrimination) had negligible to moderate IPD. Hence, none of them were eliminated from further analyses. The difficulties of 17 LC-items at both MPs are depicted in in Figure 47.

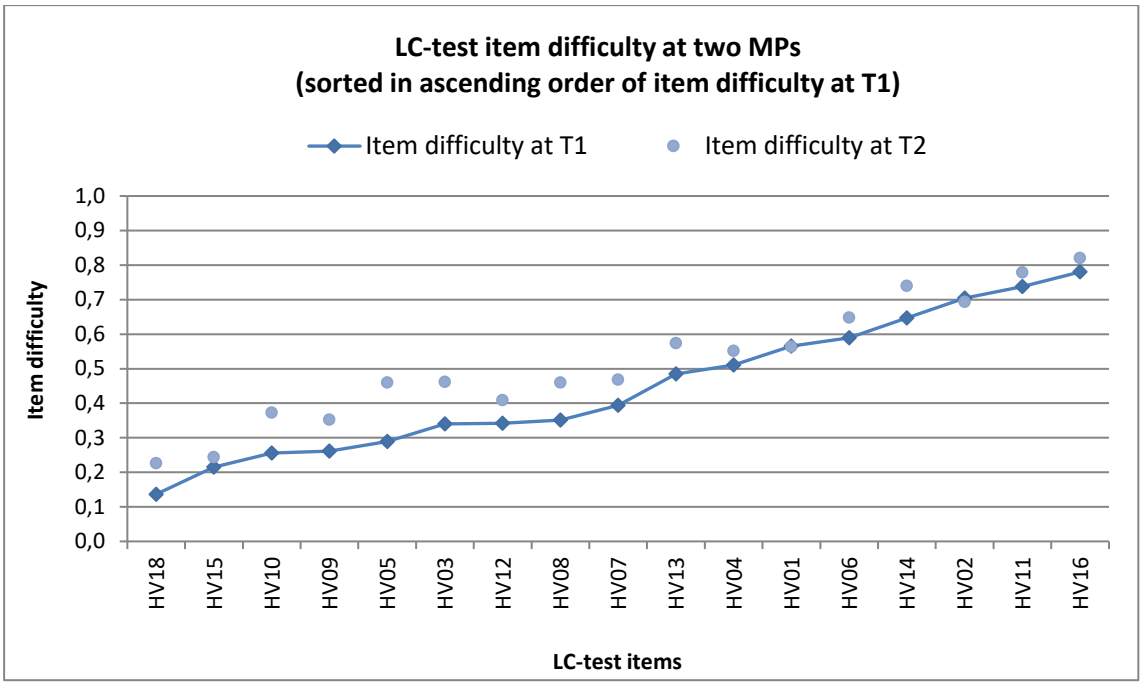

Figure 47: LC-test item difficulties at two MPs

Appendix C. Some preliminary checks

Appendix C1. C-test booklet effect

No significant C-test booklet effect was found. On average, the ability estimates of two groups of students with regard to the C-test booklet version they received did not differ significantly from each other. Given a normal standard a priori ability distribution of student sample at T1 $N(0, 1)$, the difference in ability estimates of these two groups at T1 was .04 (*SE* = .03) based on the Rasch model.

Appendix C2. Differences in student ability regarding participation in video study

On average, students, whose classes participated in the video study, and students, whose classes did not participate in the video study, did not differ significantly in their ability estimates at T1 (average difference *M* = .03, *SE* = .04 in the C-test, and *M* = .05, SE = .05 in the LC-test based on the Rasch model).

Appendix C3. Local dependencies of test items

C3a. Local dependencies of C-test testlet items

The Q3 statistics of C-test items (based on test data of both MPs) are shown below:

```
Test of Global Model Fit (Maximum Chi Square)

     maxX2 Npairs p.holm

1 722.6765   3916      0


MADaQ3 Statistic and Test of Global Model Fit (Maximum aQ3)

  MADaQ3 maxaQ3 p

1 0.0329 0.5325 0


Summary of Q3 and adjusted Q3 statistics (based on posterior distribution)

  type         M      SD      min    max   SGDDM wSGDDM

1   Q3   -0.0104  0.0461  -0.1491 0.5221 0.0349 0.0349

1  aQ3    0.0000  0.0461  -0.1387 0.5325 0.0329 0.0329
```

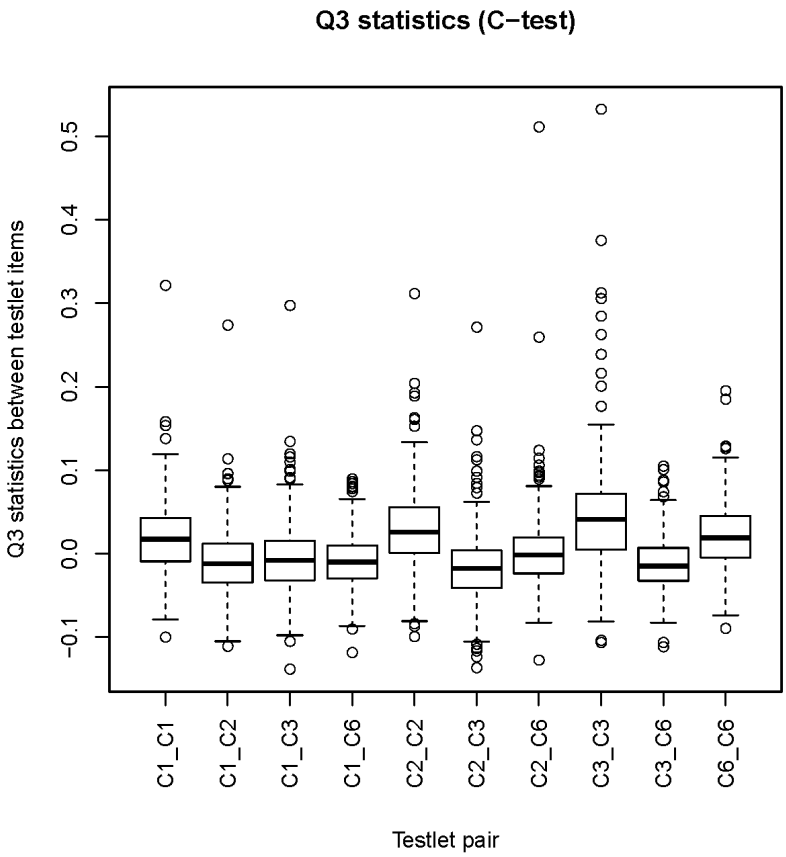

Figure 48: Q3 statistics of C-test testlet items

C3b. Local dependencies of LC-test items

The Q3 statistics of LC-test items (based on test data of both MPs) are shown below:

```
Test of Global Model Fit (Maximum Chi Square)

     maxX2 Npairs p.holm

1 200.3059    136      0


MADaQ3 Statistic and Test of Global Model Fit (Maximum aQ3)

  MADaQ3 maxaQ3 p

1 0.0346  0.214 0
```

```
Summary of Q3 and adjusted Q3 statistics (based on posterior distribution)

  type       M     SD     min    max  SGDDM wSGDDM

1   Q3 -0.0427 0.0454 -0.1520 0.1713 0.0526 0.0526

2  aQ3  0.0000 0.0454 -0.1093 0.2140 0.0346 0.0346
```

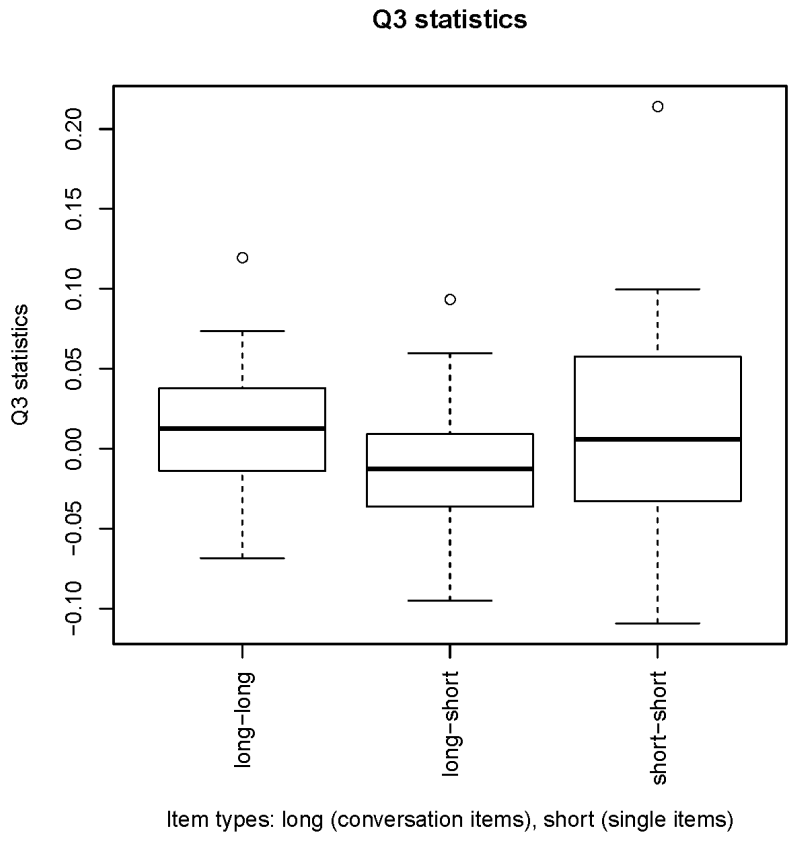

Figure 49: Q3 statistics of LC-test items

Note: long-long = Q3 statistics between long conversation items, long-short = Q3 statistics between conversation items and short dialogue items, short-short = Q3 statistics between items of the short dialogues

Appendix D. Model fits and person ability estimates in comparison

Appendix D1. Model fits in comparison

Table 41: Model fits in comparison

Test	Model	loglike	Deviance	Npars	AIC	BIC	RMSEA	SRMSR
C	1PL	–178746	357491	92	357675	358256	.090	.071
	2PL	–175947	351894	180	352254	353389	.046	.035
	Testlet 1PL	–177113	354225	100	354425	355056	.032	.072
	Testlet 2PL	–173630	347259	203	350319	351600	.025	.045
LC	1PL	–40957	81914	20	81954	82080	.070	.064
	2PL	–40544	81088	36	81160	81387	.039	.031
	3PL	–40356	80713	52	80817	81145	.023	.029

Appendix D2. Person ability estimates in comparison

Table 42: Person ability estimates in comparison (C-test, T1)

	Rohscore	M1_WLE	M2_WLE	M1T_MAP	M2T_MAP
Rohscore	**318.24**	.998	.998	.998	.993
M1_WLE	.991	**1.66**	.999	1.000	.992
M2_WLE	.986	.996	**1.12**	.998	.991
M1T_MAP	.992	.997	.990	**.86**	.991
M2T_MAP	.975	.977	.976	.975	**.88**

Note: Rohscore = sum score, M1_WLE = WLE ability estimates based on Rasch model, M2_WLE = WLE ability estimates based on unidimensional 2PL model, M1T_MAP = MAP ability estimates based on Rasch testlet model, M2T_MAP = MAP ability estimates based on testlet 2PL model. Bold values on the diagonal = variance of corresponding person ability estimates, values below the diagonal = correlations between individual person ability estimates based on different scaling models, values above the diagonal = correlations between class mean ability estimates based on different scaling models.

Table 43: Person ability estimates in comparison (C-test, T2)

	Rohscore	M1_WLE	M2_WLE	M1T_MAP	M2T_MAP
Rohscore	**308.62**	.998	.996	.999	.997
M1_WLE	.992	**1.56**	.999	1.000	.998
M2_WLE	.984	.996	**1.11**	.998	.999
M1T_MAP	.994	.998	.991	**.84**	.998
M2T_MAP	.989	.994	.994	.992	**.83**

Note: Rohscore = sum score, M1_WLE = WLE ability estimates based on Rasch model, M2_WLE = WLE ability estimates based on unidimensional 2PL model, M1T_MAP = MAP ability estimates based on Rasch testlet model, M2T_MAP = MAP ability estimates based on testlet 2PL model. Bold values on the diagonal = variance of corresponding person ability estimates, values below the diagonal = correlations between individual person ability estimates based on different scaling models, values above the diagonal = correlations between class mean ability estimates based on different scaling models.

Table 44: Person ability estimates in comparison (LC-test, T1)

	Rohscore	M1_WLE	M2_WLE	M3_WLE
Rohscore	**11.33**	.999	.993	.990
M1_WLE	.994	**1.00**	.995	.991
M2_WLE	.968	.977	**1.51**	.992
M3_WLE	.944	.955	.966	**.76**

Note: Rohscore = sum score, M1_WLE = WLE ability estimates based on Rasch model, M2_WLE = WLE ability estimates based on unidimensional 2PL model, M3_WLE = WLE ability estimates based on unidimensional 3PL model. Bold values on the diagonal = variance of corresponding person ability estimates, values below the diagonal = correlations between individual person ability estimates based on different scaling models, values above the diagonal = correlations between class mean ability estimates based on different scaling models.

Table 45: Person ability estimates in comparison (LC-test, T2)

	Rohscore	M1_WLE	M2_WLE	M3_WLE
Rohscore	**11.65**	.999	.994	.992
M1_WLE	.995	**1.12**	.996	.992
M2_WLE	.966	.977	**1.94**	.995
M3_WLE	.954	.963	.974	**.88**

Note: Rohscore = sum score, M1_WLE = WLE ability estimates based on Rasch model, M2_WLE = WLE ability estimates based on unidimensional 2PL model, M3_WLE = WLE ability estimates based on unidimensional 3PL model. Bold values on the diagonal = variance of corresponding person ability estimates, values below the diagonal = correlations between individual person ability estimates based on different scaling models, values above the diagonal = correlations between class mean ability estimates based on different scaling models.

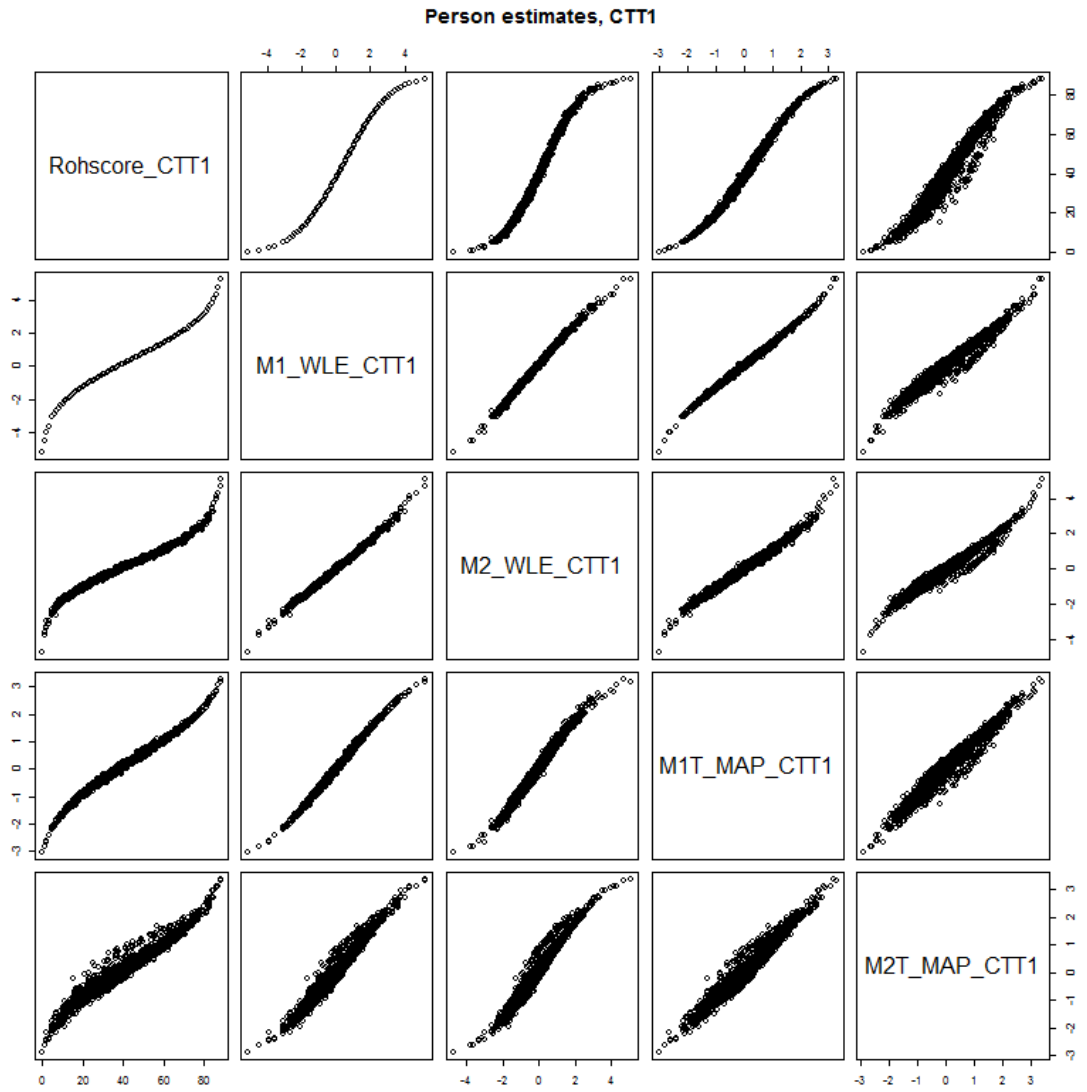

Figure 50: Person ability estimates at individual level in comparison (C-test, T1)

Note: Rohscore = sum score, M1_WLE = WLE ability estimates based on Rasch model, M2_WLE = WLE ability estimates based on unidimensional 2PL model, M1T_MAP = MAP ability estimates based on Rasch testlet model, M2T_MAP = MAP ability estimates based on testlet 2PL model, CTT1 = individual results of the C-test at T1.

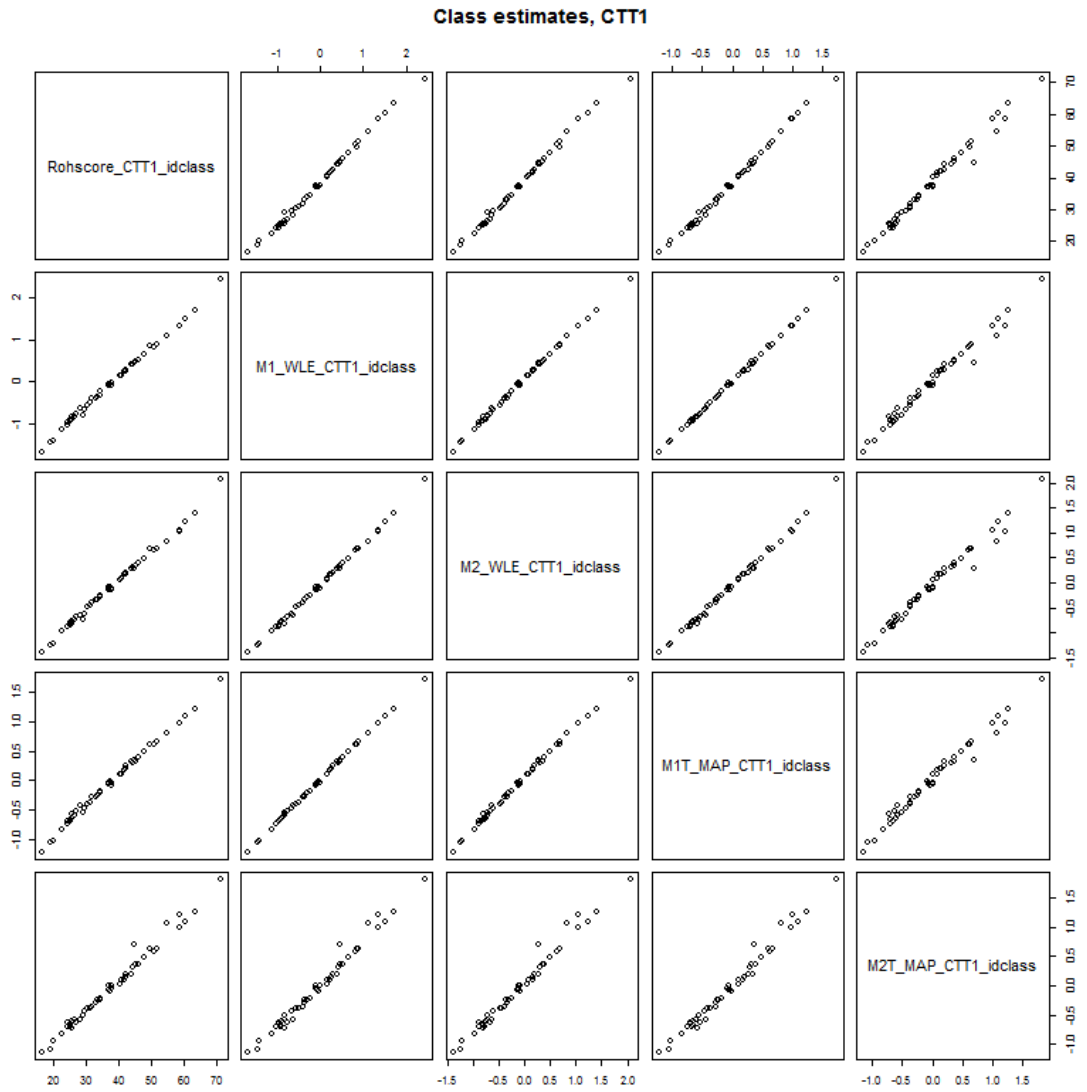

Figure 51: Class mean ability estimates in comparison (C-test, T1)

Note: Rohscore = sum score, M1_WLE = WLE ability estimates based on Rasch model, M2_WLE = WLE ability estimates based on unidimensional 2PL model, M1T_MAP = MAP ability estimates based on Rasch testlet model, M2T_MAP = MAP ability estimates based on testlet 2PL model, CTT1_idclass = class mean ability estimates in the C-test at T1.

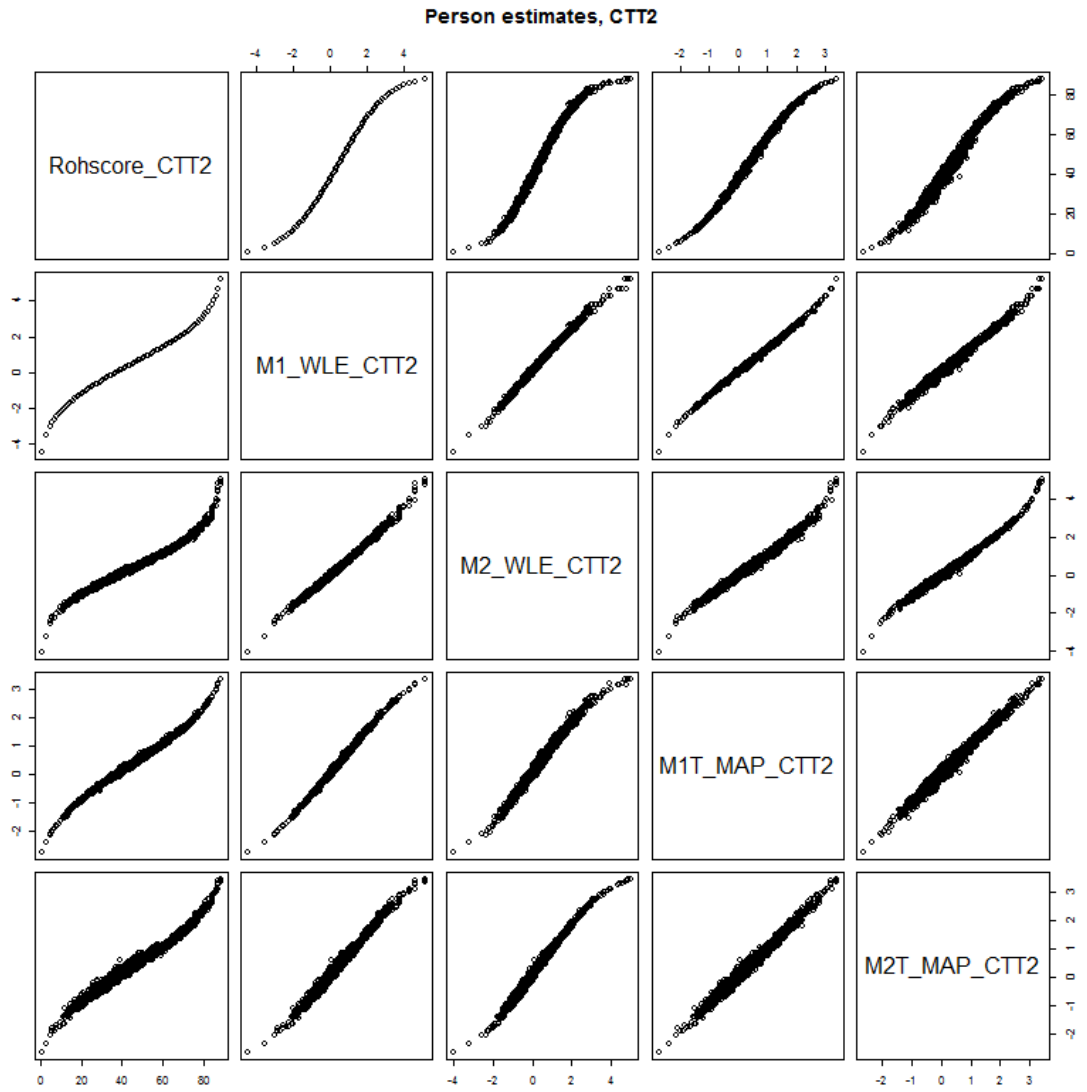

Figure 52: Person ability estimates at individual level in comparison (C-test, T2)

Note: Rohscore = sum score, M1_WLE = WLE ability estimates based on Rasch model, M2_WLE = WLE ability estimates based on unidimensional 2PL model, M1T_MAP = MAP ability estimates based on Rasch testlet model, M2T_MAP = MAP ability estimates based on testlet 2PL model, CTT2 = individual results of the C-test at T2.

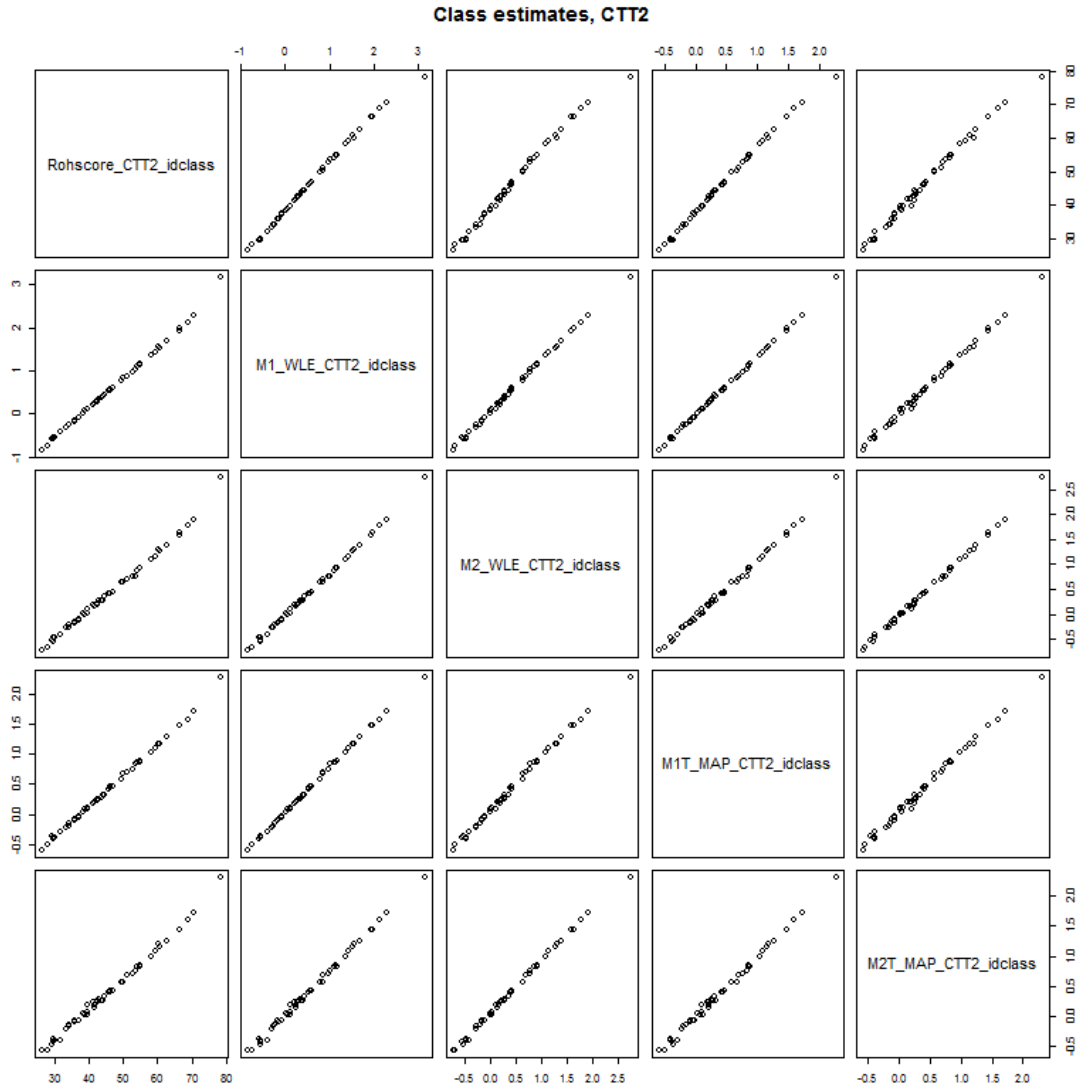

Figure 53: Class mean ability estimates in comparison (C-test, T2)

Note: Rohscore = sum score, M1_WLE = WLE ability estimates based on Rasch model, M2_WLE = WLE ability estimates based on unidimensional 2PL model, M1T_MAP = MAP ability estimates based on Rasch testlet model, M2T_MAP = MAP ability estimates based on testlet 2PL model, CTT2_idclass = class mean ability estimates in the C-test at T2.

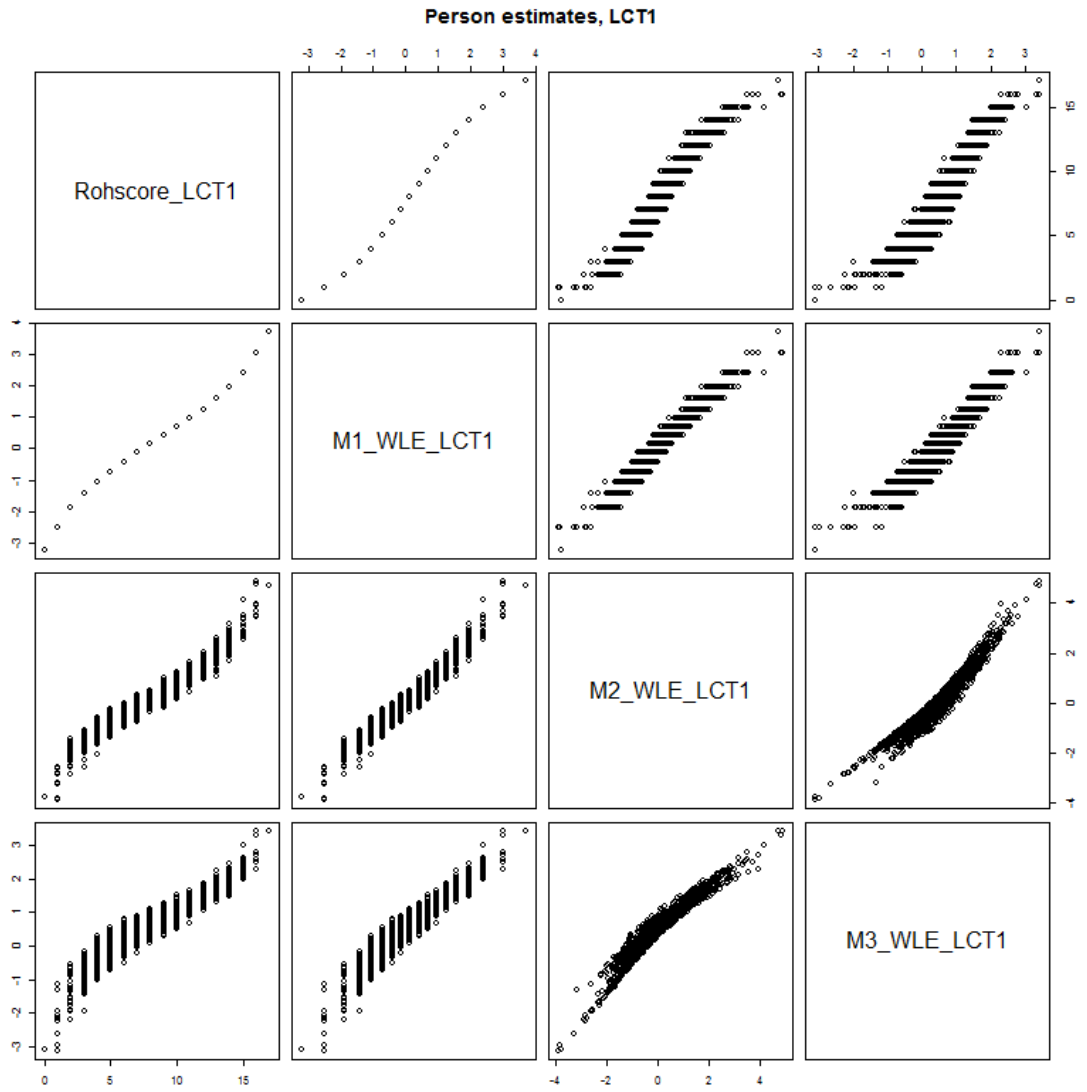

Figure 54: Person ability estimates at individual level in comparison (LC-test, T1)

Note: Rohscore = sum score, M1_WLE = WLE ability estimates based on Rasch model, M2_WLE = WLE ability estimates based on unidimensional 2PL model, M3_WLE = WLE ability estimates based on unidimensional 3PL model, LCT1 = individual results of the LC-test at T1.

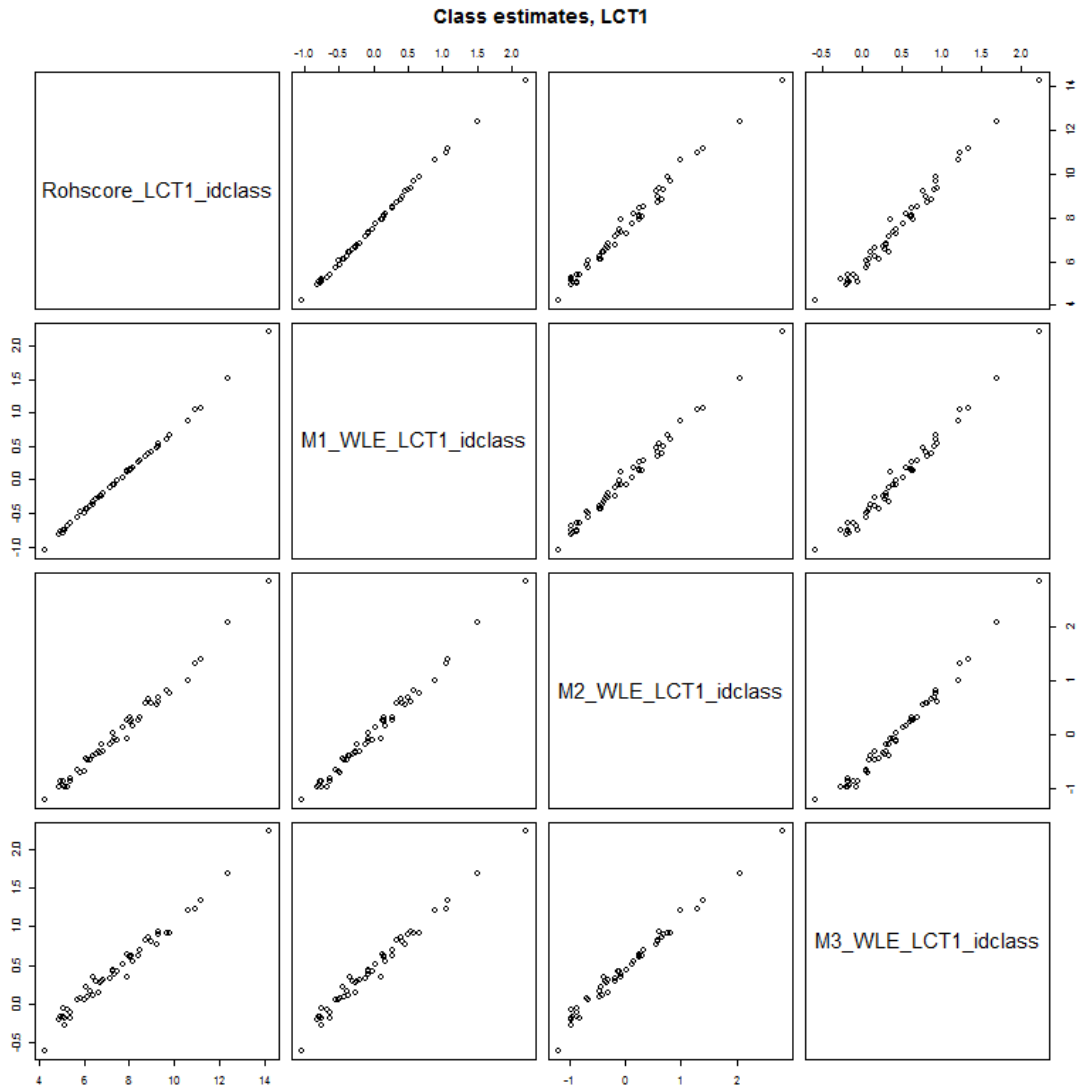

Figure 55: Class mean ability estimates in comparison (LC-test, T1)

Note: Rohscore = sum score, M1_WLE = WLE ability estimates based on Rasch model, M2_WLE = WLE ability estimates based on unidimensional 2PL model, M3_WLE = WLE ability estimates based on unidimensional 3PL model, LCT1_idclass = class mean ability estimates in the LC-test at T1.

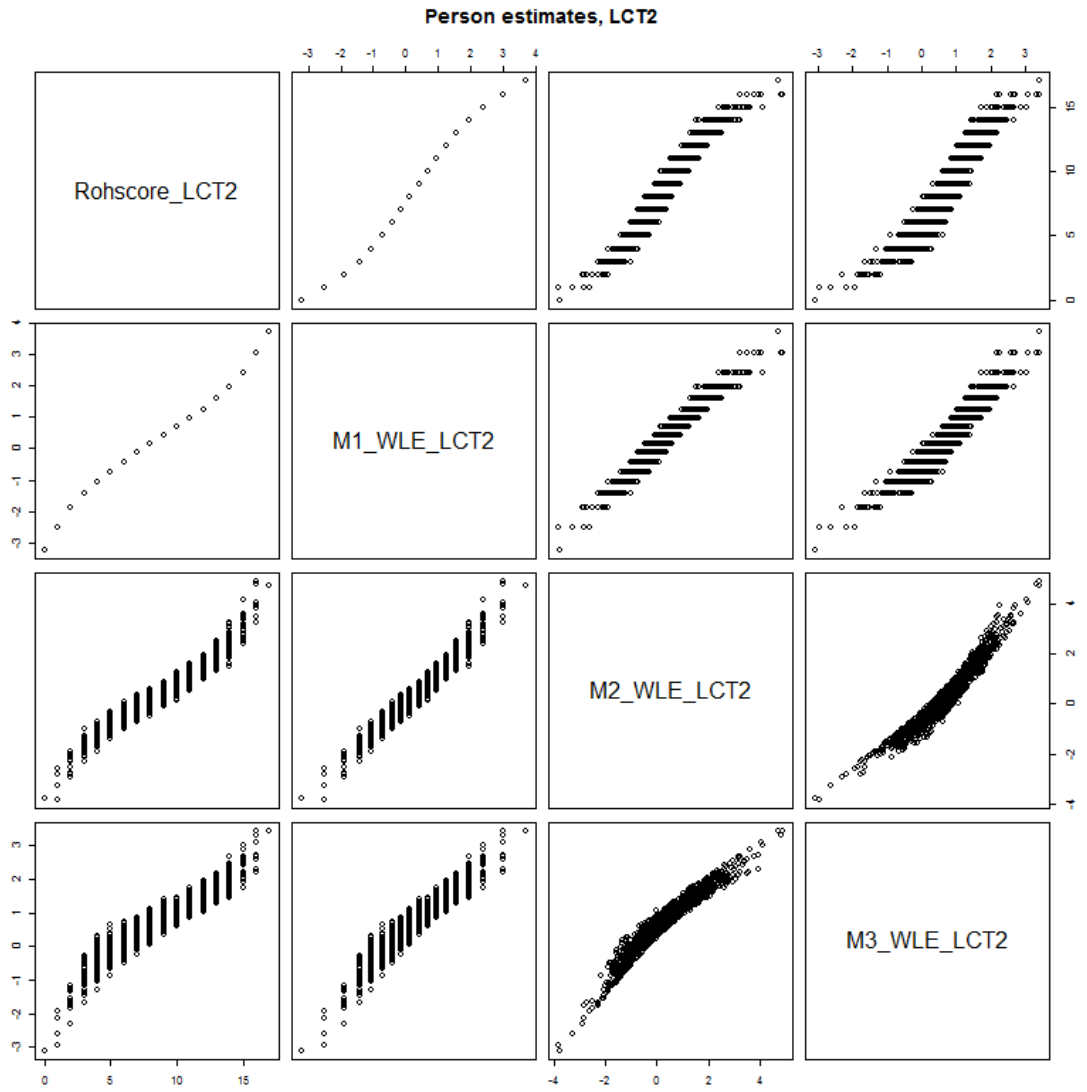

Figure 56: Person ability estimates at individual level in comparison (LC-test, T2)

Note: Rohscore = sum score, M1_WLE = WLE ability estimates based on Rasch model, M2_WLE = WLE ability estimates based on unidimensional 2PL model, M3_WLE = WLE ability estimates based on unidimensional 3PL model, LCT2 = individual results of the LC-test at T2.

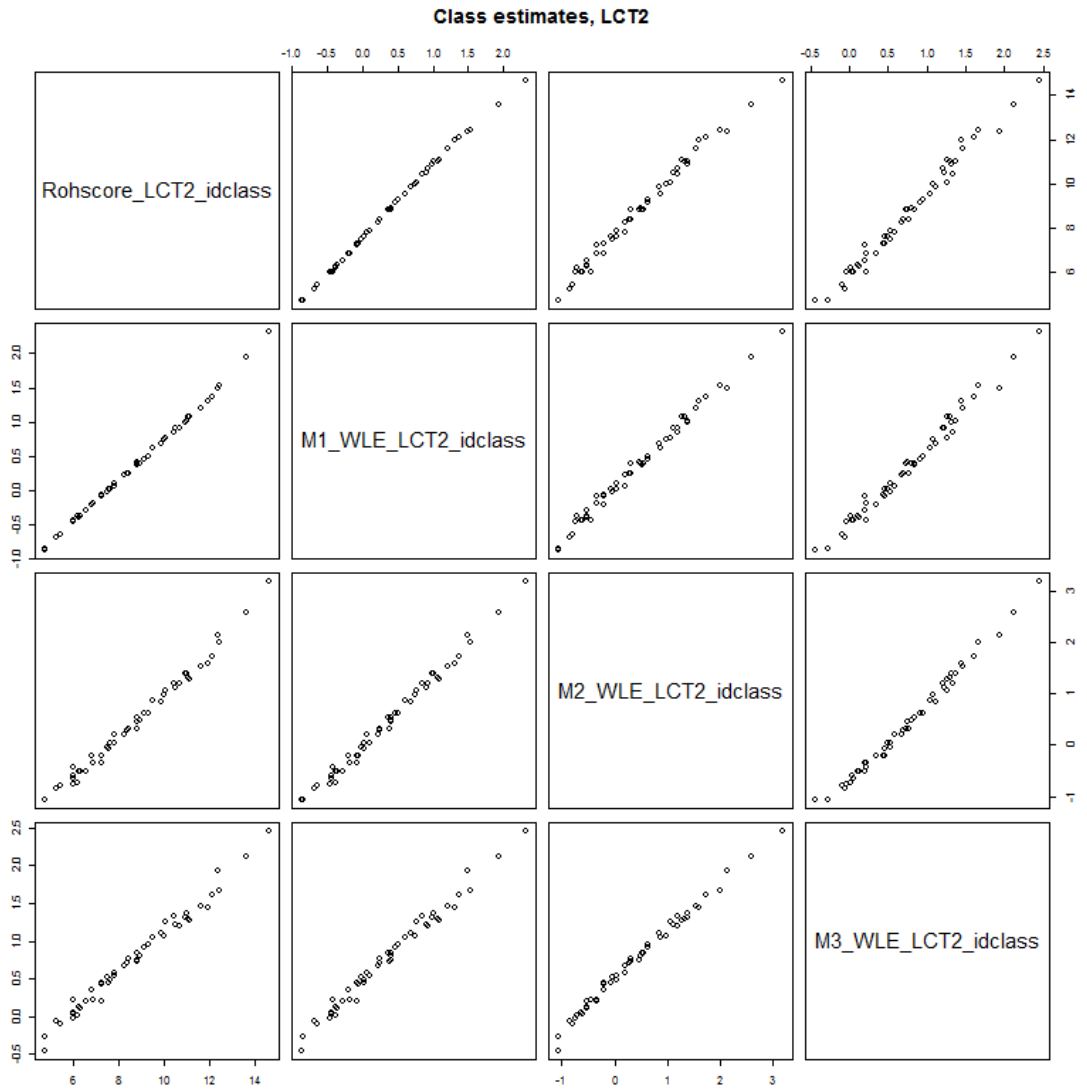

Figure 57: Class mean ability estimates in comparison (LC-test, T2)

Note: Rohscore = sum score, M1_WLE = WLE ability estimates based on Rasch model, M2_WLE = WLE ability estimates based on unidimensional 2PL model, M3_WLE = WLE ability estimates based on unidimensional 3PL model, LCT2_idclass = class mean ability estimates in the LC-test at T2.

Appendix E. Socioeconomic status (SES)

After the data imputation process, the socioeconomic status (SES) of the students was estimated based on the following indicators based on data of the student questionnaire at T1:

- parental occupation status (2: employee, 1: other),
- highest parental school leaving certificate (1: none, 2: primary school, 3: lower-secondary school, 4: higher-secondary school),
- highest occupational qualification/graduation (1: no university degree, 2: graduate degree, 3: PhD),
- material disposal of the family (telefon, mobile phone, internet connection, literature, dictionary, work of art),
- number of computers, music instruments, bath rooms at home (1: 0, 2: 1, 3: 2, 4: 3 or more),
- number of books at home (1: 0, 2: 1–10, 3: 11–50, 4: 51–100, 5: 101–250, 6: 251–500, 7: more than 500),
- and things at the student's disposal (room, work table, PC, internet connection, mobile phone, 0: no, 1: yes).

As in the DESI study, the SES was estimated via the partial credit model (PCM) (Masters, 1982), in which the probability $P(x_{pi} = k \mid \theta_p)$ of response category k (k = 0,1,…,j,l,…m) of person p to item i (with highest possible response category m) given the latent individual SES value θ_p is defined as follows:

$$P(x_{pi} = k|\theta_p) = \pi_{pik} = \frac{exp\sum_{j=0}^{k}(\theta_P - b_{ij})}{\sum_{l=0}^{m} exp\sum_{j=0}^{l}(\theta_p - b_{ij})} \; for\; k = 0, m,$$

with $\sum_{j=0}^{0}(\theta_p - b_{ij}) \equiv 0 \; and \; \sum_{j=0}^{l}(\theta_p - b_{ij}) \equiv \sum_{j=1}^{l}(\theta_p - b_{ij})$

The calibration and scaling processes were done using the R package **TAM** (Kiefer et al., 2016). The individual scale scores were then transformed to have the pooled sample mean of 0 and standard deviation of 1 over 10 imputed datasets.

Appendix F. Additional descriptive results

Appendix F1. Teaching materials and using multimedia in lessons

There was a black board in all classes, only 3 teachers did not use it in lesson. Video, television, internet, and language lab were not observed in all classes. Overhead projector was used in only six lessons. PC and beamer were more common, and in use in 15 lessons. In 19 lessons, teachers used CD-player for

listening exercises. 29/41 teachers prepared and used additional teaching materials such as pictures, posters in lessons.

Appendix F2. Individual effects on student achievement and growth

Table 46: Differences in student achievement and growth regarding student demographic factors

Test results		Group 1	Group 2	$\bar{d}$	min (d–$2SE$)	max (d+$2SE$)
C	T1	girls	boys	.22	.06	.36
	T2	girls	boys	.25	.11	.39
	Growth	girls	boys	.06	–.11	.26
LC	T1	girls	boys	.13	.00	.25
	T2	girls	boys	.15	.04	.26
	Growth	girls	boys	.06	–.10	.18
C	T1	Birth year ≥ 1992	Birth year ≤ 1991	.99	.59	1.42
	T2	Birth year ≥ 1992	Birth year ≤ 1991	.89	.54	1.31
	Growth	Birth year ≥ 1992	Birth year ≤ 1991	–.17	–.60	.33
LC	T1	Birth year ≥ 1992	Birth year ≤ 1991	.74	.41	1.10
	T2	Birth year ≥ 1992	Birth year ≤ 1991	.71	.41	.99
	Growth	Birth year ≥ 1992	Birth year ≤ 1991	.05	–.39	.44
C	T1	No repetition	Class repetition	1.28	.84	1.80
	T2	No repetition	Class repetition	1.09	.71	1.53
	Growth	No repetition	Class repetition	–.29	–.89	.30
LC	T1	No repetition	Class repetition	.99	.61	1.46
	T2	No repetition	Class repetition	.80	.43	1.15
	Growth	No repetition	Class repetition	–.07	–.66	.48

Note: C = C-test, LC = listening comprehension test. $\bar{d}$ = average Cohen's d between corresponding student group 1 and group 2 based on estimates of different scaling models, min (d–$2SE$) and max (d+$2SE$) are minimum and maximum estimates over all scaling models.

Wissenschaftlicher Bildungsgang

2000	Abitur in Hanoi, Vietnam
2000 – 2004	Studium der Psychologie an der Universität der Sozial- und Humanwissenschaft, Hanoi
06/2004	Bachelor-Abschluss (B.Sc. Psych.)
2005 – 2008	Master-Studium der Psychologie an der Ruhr-Universität Bochum;
10/2008	Abschluss des Masterstudiums (M.Sc. Psych.)

Selbstständigkeitserklärung

Hiermit versichere ich, Giang Hong Pham, geb. am 14.10.1982 in Hanoi, Vietnam, dass ich die vorliegende Dissertation sebstständig verfasst, ohne unzulässige Hilfe Dritter und ohne Nutzung anderer als der angegebenen Hilfsmittel angefertigt habe.

Die aus fremden Quellen direkt oder indirekt übernommenen Inhalte sind unter Angaben der Quellen kenntlich gemacht. Dies gilt auch für bildliche Darstellungen sowie für Quellen aus dem Internet.

Kein Teil dieser Arbeit ist in einem anderen Promotions- oder Habilitationsverfahren verwendet worden.

Zug, 04 Oktober 2017

Giang Hong Pham