Semantic Similarity of Teacher and Student Discourse Linked to Quality Ratings from Classroom Observations

Jessica Boyle
Vanderbilt University
jessica.r.boyle@vanderbilt.edu

Scott Crossley
Vanderbilt University
scott.crossley@vanderbilt.edu

ABSTRACT

Effective classroom communication is critical for students' academic performance. This study investigates the semantic similarity of adjacent teacher-to-student, teacher-to-teacher, and student-to-student utterances using natural language processing (NLP) tools. It explores their relationship with quality ratings from classroom observation measures. Focusing on the cohesiveness of classroom language, the study analyzes transcripts from elementary math classrooms and scores from the Classroom Assessment Scoring System (CLASS) and Mathematical Quality of Instruction (MQI). Linear regression models identified that the semantic similarity between teacher-to-student utterances significantly predicted the CLASS and MQI scores. However, the models explain little variance in the observational scores. The study underscores the complexity of classroom discourse and proposes future analyses. The findings prompt reflection on the practical significance of observed associations and highlight the importance of considering the evolving landscape of educational technology in supporting teacher practice.

Keywords

Semantic similarity, classroom discourse, teacher-student interactions, cohesion

1. INTRODUCTION

Classroom communication is pivotal in influencing students' academic performance [19]. Effective classroom communication relies on teachers using clear language to create shared meaning between themselves and students. Clear language use benefits all students and is particularly advantageous for students with lower language proficiency [2, 7]. Effective communication requires teachers to use concrete, explicit language with appropriate vocabulary and content-related words [3, 11]. Teachers who skillfully use sequencing and scaffolding strategies can effectively break complex skills into manageable units, addressing cognitive overload and accommodating students' working memory. Additionally, teachers' ability to elaborate on student responses using feedback, follow-up explanations, and questioning techniques is critical for enhancing students' learning and engagement [15, 21].

Traditional observational tools used to measure teacher talk often include binary judgments that lack the granularity needed for self-reflection and improvement in teacher practice [20]. This study leverages natural language processing (NLP) tools to analyze the cohesiveness of teacher language within instructional lessons and assess its relationship with established observational measures, extending previous research on NLP's role in evaluating and enhancing teacher discourse.

J. Boyle and S. Crossley. Semantic similarity of teacher and student discourse linked to quality ratings from classroom observations. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 797–801, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12729954

1.1 Purpose Statement and Research Question

This study analyzed the cohesiveness of teacher and student discourse in elementary math classrooms and its relationship with two widely used classroom observational measures [14, 16]. Classroom transcripts from a nationally representative dataset were used to assess classroom discourse [4], with each transcript aligned with quality ratings from the Classroom Assessment Scoring System (CLASS) and the Mathematical Quality of Instruction (MQI) [14, 16]. This study focused on the Instructional Dialogue dimension from the CLASS and the Teacher's Use of Student Contribution item from the MQI. Cohesiveness was assessed using semantic embeddings from SpaCy to compute the semantic similarity across three types of adjacent utterances: teacher-to-student, teacher-to-teacher, and student-to-student. This analysis aims to determine whether semantic similarity between these utterances predicts quality ratings in instructional dialogue and teachers' responses to student contributions.

The following research questions guide this study:

1. Does the semantic similarity of adjacent teacher-to-student, teacher-to-teacher, and student-to-student utterances relate to the quality ratings for Instructional Dialogue scored via Classroom Assessment Scoring System (CLASS) observations?
2. Does the semantic similarity of adjacent teacher-to-student utterances relate to the quality ratings for Teacher Use of Student Contribution scored via Mathematical Quality of Instruction (MQI) observations?

1.2 Literature Review

Analyzing classroom discourse allows for examination for how language is used by both teachers and students in the interactive process of teaching and learning [15, 21]. During teacher-directed instruction, teachers typically talk for two-thirds of the lesson time [17]. During this time, teachers introduce and use academic language to aid students in developing ideas and acquiring knowledge. Understanding the concepts of scaffolding and coherence is essential when exploring how teacher language can effectively support students in knowledge construction.

In an instructional context, scaffolding is the deliberate steps to simplify task complexity, enabling students to focus on acquiring a new skill [10]. It involves teachers gradually fading their guidance and support as students develop new concepts or skills. Scaffolding can be implemented at a macro level, planning an entire lesson, or

at a micro level, focusing on moment-to-moment interactions [9, 21]. Effective micro-level scaffolds are often facilitated through organized, connected, and logical teacher dialogue and feedback [10]. Similarly, coherence in discourse pertains to the organization and connectedness of spoken or written language. Coherent spoken language allows ideas to flow naturally, enhancing comprehension for listeners when it is logical and consistent and creates a unified whole [8]. Teachers are encouraged to use cohesive, scaffolded language, but current observational methods do not often include granular measurements of teachers' language use. Leveraging NLP technology offers a robust and objective approach to studying teachers' spoken language.

Efforts to measure teacher practice using automated tools have predominantly used NLP to assess classrooms through a dialogue-driven instruction process characterized by classroom discussions [21] and focused mainly on the initiate-response-evaluate (IRE) pattern. Researchers have accurately detected and classified focus and funneling questions, various instructional activities, open-ended questions, and teachers' uptake of student contributions [1,5, 6,13,18]. A noticeable gap exists in measuring the clarity and cohesion of teacher language within teacher-student interactions.

2. METHODS

2.1 Data

The open-source National Center for Teacher Effectiveness (NCTE) dataset includes anonymized transcripts of teachers' instruction from the NCTE Main Study [12], conducted between 2010-2013 in 4th and 5th-grade elementary math classrooms across four districts, predominately serving historically marginalized students. The transcripts are linked with various classroom observation scores. The data are accessible at: https://github.com/ddemszky/classroom-transcript-analysis. This analysis uses the transcripts and the linked CLASS and MQI data. Additional details and the associated code used in this analysis can be found at https://github.com/jessicarboyle/semantic_similarity_nctedata.

2.1.1 Classroom Transcripts

This analysis includes 1,325 transcripts from 301 teachers, each with an average of 4 transcripts. Classroom recordings were captured using three cameras, a lapel microphone for teacher talk, and a bidirectional microphone for student talk. Professional transcribers, working under contract for a commercial transcription company, transcribed the recordings.

The transcripts were labeled by speaker turns (teacher, students, multiple students), where each row represented a speaker utterance, which could include one or more speech acts or "sentences." On average, each transcript contains 5,733 words, of which 87.7% are spoken by teachers across 172 utterances.

The transcripts are fully anonymized, with names replaced with labels like "Student J", "Teacher," or "Mrs. H". Inaudible speech and metadata such as [laughter] and [students putting away materials] noted by the transcribers were excluded from this analysis.

2.1.2 Classroom Assessment Scoring System (CLASS) Scores

The CLASS is an observation tool that evaluates teacher-student interactions across 3 broad domains: emotional support, classroom organization, and instructional support. The 3 domains include a total of 11 sub-dimensions. This analysis includes the Instructional Dialogue dimension from the Instructional Support domain. The Instructional Dialogue dimension measures the purposeful use of cumulative content-focused discussion among teachers and students, assessing how teachers actively support students in connecting ideas and fostering a deeper understanding of the content.

Observers scored the dimension using a 7-point scale. Lower scores (1,2) are assigned when there are minimal or no discussions in the classroom and when the teacher seldom acknowledges, repeats, or extends on student comments. Mid-range scores (3,4,5) are given when discussions occur, but they are brief or shift rapidly between topics without subsequent questions or comments. Higher scores (6,7) indicate the presence of frequent, content-driven discussions between teachers and students, fostering cumulative exchanges where teachers actively promote elaborate dialogue through open-ended questions and repetitions.

2.1.3 Mathematical Quality of Instruction (MQI) Scores

The MQI instrument measures the quality of math instruction across five elements: richness of the mathematics, errors and imprecision; working with students and mathematics, student participation in meaning-making and reasoning, and connections between classroom work and mathematics. This analysis includes the Teacher's Use of Student Contribution item from the working with students and mathematics element. The Teacher's Use of Student Contribution item captures how effectively the teacher incorporates student mathematical contributions to advance instruction. Contributions can include students' answers, comments, explanations, and questions posed to the teacher.

Observers scored the item using a 4-point scale. A score of 1 indicates a lesson or overlooked student contributions; a 2 is given when student inputs are acknowledged but only superficially responded to; a 3 signifies some use of student ideas, but not optimally leveraged; and a score of 4 reflects comprehensive integration of student thoughts, demonstrated through comments on student ideas, asking students to comment on other students' ideas, and expanding on or reinforcing student utterances. The MQI protocol involves segmenting lessons into roughly 5–7-minute segments, assigning each segment a score, and then averaging the scores to obtain an overall score for each lesson.

2.2 Semantic Similarity Analysis

The semantic similarity analysis assessed the cohesion between teacher and student discourse by calculating the similarity between adjacent utterances. It included two distinct analyses: one that included all words in the utterances and another focusing only on content words. These comparisons were made for teacher-to-student, teacher-to-teacher, and student-to-student utterances, excluding the other types for each. For example, student utterances were omitted when analyzing teacher-to-teacher utterances.

The semantic similarity scores were calculated by spaCy's large English model, which uses Word2Vec embedding to represent words as vectors in a continuous vector space, capturing their semantic relationships based on their contextual usage in a large corpus. The similarity between words, sentences, or entire documents is determined using spaCy's 'doc.similarity' function that uses pre-trained word embeddings. Each utterance in the transcripts was treated as a document – whether it consisted of one word, one statement, or multiple statements. The resulting scores, ranging from 0 to 1, indicate the degree of similarity between adjacent utterances.

For the analyses including all the words, punctuation was removed from the transcripts to avoid the arbitrary influence in spoken language, where conventional sentence structures are often not adhered to, especially in a context such as a classroom. This ensured the focus remained on teacher and student words without punctuation affecting the scores.

For the analyses including only content words, The Natural Language Toolkit (NLTK) was used to remove stop words (e.g., a, the, is) using a predetermined list of words that typically carry minimal semantic weight. This step isolated the essential content words, providing a focused view of the primary message in the teachers' and students' utterances.

2.3 Statistical Analysis

Each transcript included numerous teacher and student utterances, resulting in multiple semantic similarity scores from pairwise comparison between two adjacent utterances. A mean similarity score was computed for each transcript by summing all semantic similarity values and dividing by the total number of values within that transcript. Mean similarity scores were calculated separately for all words and content words.

The NCTE dataset included 3-4 CLASS and MQI observations per teacher, conducted within a 1–2-month period. To establish generalizable scores of instructional quality for the teacher, separate average scores for CLASS and MQI were calculated for each teacher by summing the scores from all observations and dividing by the total number of observations.

For the CLASS measure, the Instructional Dialogue considers coherence across all classroom discourse (student and teacher talk); thus, the semantic similarity for teacher-to-student, teacher-to-teacher, and student-to-student utterances was included in the CLASS analyses. In contrast, the Teacher's Use of Student Contribution item from the MQI measure focuses on teachers' responses to student utterances; therefore, only the teacher-to-student semantic similarity was included in MQI analyses.

The statistical analyses were conducted in R. First, correlations were examined between the mean semantic similarity scores and the CLASS and MQI scores. Subsequently, a 10-fold cross-validation linear regression model with automatic feature selection was used to determine which semantic similarity variables predicted the CLASS and MQI scores. Multicollinearity ($r \geq .700$) between individual predictor variables was checked. Suppression effects were examined to determine if variables needed to be removed from the model.

3. RESULTS

3.1 Relationship between Mean Similarity Scores and CLASS Scores

The mean semantic similarity scores across teacher-to-student, teacher-to-teacher, and student-to-student utterances for both all words and content words were positively correlated with Instructional Dialogue scores. The scores for all words showed stronger correlations with Instructional Dialogue across all three types of utterances, with teacher-to-student similarity scores showing the highest correlations.

For the semantic similarity with all words, the correlation coefficients indicate statistically significant, yet weak, positive correlations with Instructional Dialogue scores (teacher-to-student $r = 0.14$, teacher-to-teacher $r = 0.12$, student-to-student $r=0.097$). Positive correlations were also noted for content words, though only teacher-to-student scores were statistically significant, with a weak effect size ($r = 0.12$). Scatterplots revealed a modest linear relationship between semantic similarity scores and Instructional Dialogue scores.

The initial multicollinearity assessment for the 10-fold cross-validation linear regression model showed that the content words and all words similarity scores were highly correlated ($r > 0.80$) for each utterance pair, teacher-to-student, teacher-to-teacher, and student-to-student. Due to higher correlations with Instructional Dialogue scores, all word similarity scores were included in the regression analysis. The final linear regression model included similarity scores for teacher-to-student and teacher-to-teacher utterances, with teacher-to-student semantic similarity reaching statistical significance ($p=0.1$). No suppression effects were observed. The model accounted for approximately 3% of the variance in CLASS Instructional Dialogue scores, as indicated by an r-squared value of 0.026, underscoring the weak relationship depicted in Figure 1, showing the predicted and actual values.

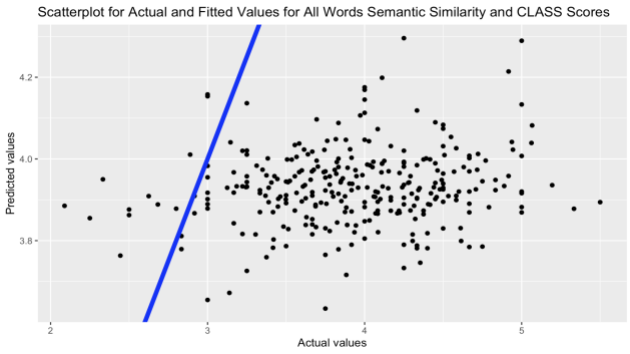

Figure 1. Scatterplot for the 10-fold cross-validation model for teacher-to-student and teacher-to-teacher semantic similarity scores (for all words) and CLASS Instructional Dialogue scores

3.2 Relationship between Teacher-to-Student Mean Similarity Scores and MQI Scores

The mean semantic similarity scores for teacher-to-student utterances, for all words and content words, showed statistically significant positive correlations with the MQI Teacher's Use of Student Contribution scores (all words $r = 0.32$, content words $r = 0.21$). Due to the high multicollinearity ($r=0.84$) between the similarity scores for all words and content words, a simple linear regression model was used, including only all words similarity scores with the MQI Teacher's Use of Student Contribution scores. This model indicated that the semantic similarity between teacher-to-student utterances significantly predicted Teachers' use of Student Contribution scores. With a multiple R-squared value of 0.105, the model accounts for approximately 11% of the variance in these scores. The RMSE of 0.578 indicates the model's predicted values typically deviated about half a point from the observed values. Figure 2 illustrates the weak relationship between the predicted and actual values.

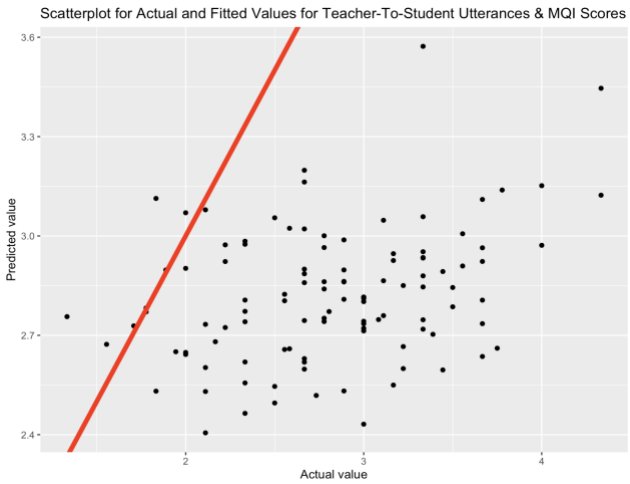

Figure 2. Scatterplot for the linear regression model for the similarity of teacher-to-student utterances (with all words) and MQI Teacher's Use of Student Contribution scores

4. CONCLUSION AND FUTURE WORK

The analyses in this study revealed some statistically significant relationships between the semantic similarity scores derived from NLP tools and the classroom observation scores from the CLASS Instructional Dialogue and the MQI Teacher's Use of Student Contribution measures. However, linear regression models demonstrated that the semantic similarity scores explained only a small amount of variance in the observation scores.

The positive correlations underscore the important role of aligned teacher-student discourse in creating high-quality math instruction. The statistically significant correlations between the semantic similarity of teacher-to-student utterances and the CLASS and MQI highlight the positive impact of cohesive teacher-student communication on instructional quality. Additionally, the significant correlation between the semantic similarity of teacher-to-teacher utterances and the MQI measure reinforces the value of coherent, progressively structured teacher dialogue. When teachers and students engage in similar discourse, it appears to be associated with higher classroom observation scores.

The weak effect sizes observed in the CLASS correlations call for reflection on the practical significance of the results, suggesting that semantic similarity among adjacent utterances of teachers and students does not fully capture the variations seen in CLASS Instructional Dialogue scores. This may be partly because the CLASS Instructional Dialogue dimension evaluates coherence across all types of classroom discourse – teacher-to-student, teacher-to-teacher, and student-student interactions – thereby diluting the impact of any single type of interaction. In contrast, the MQI's Teacher's Use of Student Contribution focuses exclusively on teacher responses to student utterances, which may explain why the teacher-to-student semantic similarity shows strong correlations in the MQI analysis. The low to moderate effect sizes suggest that other classroom practices (e.g., non-verbal communication) may impact human ratings of cohesion across classroom discourse. It also potentially highlights the reliability and accuracy of human observers scoring the cohesion of teacher and student talk with traditional classroom observation procedures.

Higher semantic similarity scores were observed when content and stop words were included compared to only content words, suggesting that stop words may carry essential semantic meaning in math instruction. This potentially reflects the importance of stop words in providing structure to spoken discourse and in discussing abstract concepts like math. However, the weak correlations with the CLASS and MQI scores suggest a potential overestimation of semantic similarity when stop words are included. Given the intricate nature of classroom discourse, future analyses should continue including both content and stop words for a more nuanced understanding.

The linear regression model for CLASS scores showed that teacher-to-student semantic similarity was most predictive of the Instructional Dialogue scores, indicating that this type of cohesion influences observation ratings of instructional dialogue. The persistence of teacher-to-teacher semantic similarity in the model suggests its relevance in observation ratings, whereas the exclusion of student-to-student utterances indicates their lower predictive values, possibly due to fewer student utterances during teacher-directed instruction or the quality of questions posed to students. It is important to note that while the model demonstrated that the teacher-to-student and teacher-to-teacher scores were predictive, it explained a small amount of variance in the CLASS Instructional Dialogue scores.

Similarly, the linear regression model for the MQI Teacher's Use of Student Contribution scores showed that teacher-to-student semantic similarity was a significant predictor and explained a greater amount of variance than the CLASS model. This stronger relationship could be due to the MQI being a math-specific observational measure. These results highlight the impact of coherent and responsive teacher-student discourse on math instructional quality.

Overall, while the semantic similarity scores calculated from NLP tools explained some variance in the cohesion-related classroom observation scores, the divergence from theoretical expectations highlights the complexity of instructional dialogue. This complexity underscores the need for further investigations, which should include additional linguistic features such as the complexity of teacher language, the quality of questions posed, and student engagement. Additionally, the performance of models trained to automate the measurement of teacher practice is often compared to established "gold standard" classroom observation measures; however, there should be considerations for the difficulty of measuring semantic similarity during live observations. The results of this analysis could potentially highlight limitations in humans' ability to evaluate such a construct in real time during classroom instruction.

Furthermore, automating the measurement of cohesion between teacher and student discourse could enable the development of tools that provide teachers with frequent automated feedback, allowing them to reflect on and improve their instructional practices. It is essential to consider the utility of traditional NLP tools trained on textual data. The rapid advancements of NLP technologies, particularly audio and transformer-based models, may be critical for accurately analyzing the complexity of spoken language in a dynamic classroom context, Continued research should focus on the evolving utility of NLP tools in education, offering significant potential for supporting teacher practice as educational technology rapidly progresses.

5. REFERENCES

[1] Alrajhi, L., Alamri, A., Pereira, F. D., & Cristea, A. I. 2021. Urgency Analysis of Learners' Comments: An Automated Intervention Priority Model for MOOC. In *International*

Conference on Intelligent Tutoring Systems (pp. 148-160). Springer, Cham. https://doi.org/10.1007/978-3-030-80421-3_18

[2] Archer, A. L., & Hughes, C. A. 2011. Explicit instruction: Effective and efficient teaching. New York: Guilford.

[3] Brophy, J. 1988. Research linking teacher behavior to student achievement: Potential implications for instruction of Chapter 1 students. *Educational Psychologist*, 23(3), 235–286. https://doi.org/10.1207/s15326985ep2303_3

[4] Demszky, D., & Hill, H. 2022. The NCTE transcripts: A dataset of elementary math classroom transcripts. arXiv preprint arXiv:2211.11772. https://doi.org/10.18653/v1/2023.bea-1.44

[5] Demszky, D., Liu, J., Hill, H. C., Jurafsky, D., & Piech, C. 2023. Can Automated Feedback Improve Teachers' Uptake of Student Ideas? Evidence From a Randomized Controlled Trial in a Large-Scale Online Course. *Educational Evaluation and Policy Analysis*. https://doi.org/10.3102/01623737231169270

[6] Donnelly, P. J., Blanchard, N., Olney, A. M., Kelly, S., Nystrand, M., & D'Mello, S. K. 2017. Words matter: Automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK* '17, 218–227. https://doi.org/10.1145/3027385.3027417

[7] Ernst-Slavit, G., & Mason, M. R. 2011. "Words that hold us up": Teacher talk and academic language in five upper elementary classrooms. *Linguistics and Education*, 22, 430–440.

[8] Foltz, P. W. 2007. Discourse coherence and LSA. Handbook of latent semantic analysis, 167, 184.

[9] Hammond, J. 2001. Scaffolding and language. In J. Hammond (Ed.), Scaffolding: Teaching and learning in language and literacy education (pp. 15–30). Sydney: Primary English Teaching Association.

[10] Hogan, K., & Pressley, M. 1997. Scaffolding student learning. Cambridge, MA: Brookline Books.

[11] Hollo, A., & Wehby, J. H. 2017. Teacher Talk in General and Special Education Elementary Classrooms. *The Elementary School Journal*, 117(4), 616–641. https://doi.org/10.1086/691605

[12] Kane, T., Hill, H., & Staiger, D. 2015. National center for teacher effectiveness main study. icpsr36095-v2.

[13] Kelly, S., Olney, A. M., Donnelly, P., Nystrand, M., & D'Mello, S. K. 2018. Automatically Measuring Question Authenticity in Real-World Classrooms. *Educational Researcher*, 47(7), 451–464. https://doi.org/10.3102/0013189X18785613

[14] Learning Mathematics for Teaching Project. (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education*, 14(1), 25–47. https://doi.org/10.1007/s10857-010-9140-1

[15] Mercer, N. 2002. Developing Dialogues. In G. Wells & G. Claxton (Eds.), Learning for life in the 21st century: Sociocultural perspectives on the future of education (pp. 141–153). Oxford: Blackwell Publishers Ltd. https://doi.org/10.1002/9780470753545.ch11

[16] Pianta, R. C., La Paro, K. M., & Hamre, B. K. 2008. Classroom assessment scoring systemTM: Manual k-3. Paul H Brookes Publishing.

[17] Sinclair, J., & Coulthard, M. 1975. Towards an analysis of discourse: The English used by teachers and pupils. London: Oxford University Press.

[18] Suresh, A., Jacobs, J., Lai, V., Tan, C., Ward, W., Martin, J. H., & Sumner, T. 2021. Using Transformers to Provide Teachers with Personalized Feedback on their Classroom Discourse: The TalkMoves Application. arXiv preprint arXiv:2105.07949.

[19] Thatcher, K. L., Fletcher, K., & Decker, B. 2008. Communication disorders in the school: Perspectives on academic and social success, an introduction. *Psychology in the Schools*, 45(7), 580–581. https://doi.org/10.1002/pits.20310

[20] Titsworth, S., Mazer, J. P., Goodboy, A. K., Bolkan, S., & Myers, S. A. 2015. Two meta-analyses exploring the relationship between teacher clarity and student learning. *Communication Education*, 64(4), 385–418. https://doi.org/10.1080/03634523.2015.1041998

[21] Wells, G. 1999. Dialogic inquiry: Towards sociocultural practice and theory of education. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511605895