

Collaborative Problem Solving Skills versus Collaboration Outcomes: Findings from Statistical Analysis and Data Mining

Jiangang Hao
Educational Testing Service
ETS Rosedale Road, MS 02-T
Princeton, NJ 08541
jhao@ets.org

Lei Liu
Educational Testing Service
ETS Rosedale Road
Princeton, NJ 08541
lliu001@ets.org

Alina A von Davier
Educational Testing Service
ETS Rosedale Road
Princeton, NJ 08541
avondavier@ets.org

Patrick Kyllonen
Educational Testing Service
ETS Rosedale Road
Princeton, NJ 08541
pkyllonen@ets.org

Christopher Kitchen
Educational Testing Service
ETS Rosedale Road
Princeton, NJ 08541
ckitchen@ets.org

ABSTRACT

With the aid of educational data mining and statistical analysis, we investigate the relationship between collaboration outcomes and collaborative problem solving (CPS) skills exhibited during the collaboration process. We found that negotiation skill contributes positively to the collaboration outcomes while purely sharing information does the opposite.

Keywords

collaborative problem solving, simulation-based assessment, random forest

1. INTRODUCTION

Collaborative problem solving (CPS) is widely considered as one of the critical skills for academic and career success in the 21st century [9]. However, assessing CPS, particularly in a large-scale and standardized way, is very challenging, as one must take into account the forms of collaboration, the size of teams, and assessment contexts. Among the existing studies on assessing CPS, most of them are not designed from the perspective of a standardized assessment, but more from the perspective of revealing some important aspects of CPS [6, 16, 5, 22]. A recent review can be found in [21]. The first large-scale and standardized assessment for CPS was the international Assessment and Teaching of 21st century skills project (ATC21S) carried out by Griffin and colleagues [9, 4]. In this assessment, two students collaborate via text chat to solve computer-based CPS tasks and their communications as well as some other features (such as the response time) were coded automatically according

to a CPS framework [1]. Another large-scale assessment for CPS was carried out by the Programme for International Student Assessment (PISA) in its sixth survey in 2015 [17]. In this assessment, students collaborate with different number of virtual partners (avatars) on a set of computer-based collaborative tasks and they communicate with their virtual partners by choosing from a list of predefined texts. Both ATC21S and PISA 2015 consider the CPS as skills across different domains and the tasks used in their assessments are not confined into a specific domain.

In this paper, we report our findings on the relationship between the CPS skills and the collaboration outcomes in the domain of science, as we think CPS is more likely to be domain dependent. We developed a simulation-based task, in which two participants collaborate via text chat to complete a set of questions and activities on volcanoes [10]. We choose a simulation-based task because it provides students with opportunities to demonstrate proficiencies in complex interactive environments that traditional assessment formats cannot afford [14], which is especially suitable for measuring the complex skills such as CPS.

In the simulation task, for each item, we ask each member of a dyadic team to respond individually first (initial response). Then, after collaboration, each of them will be given a chance to submit a revised response. The difference between the initial and revised responses directly encodes the effect due to collaboration. Based on the data collected using Amazon Mechanical Turk, we introduce two variables, “number of changes” and “score change”, to characterize the collaboration outcomes. The “number of changes” is the total number of attempts by the team members to change the initial responses after the collaboration. Some of the attempts change the responses from correct to incorrect while some change the responses from incorrect to correct. This number reflects the willingness to make a change after the collaboration. On the other hand, the “score change” is the sum of the score changes between the initial and revised responses, which quantifies the results of the changes. Based on these two variables, we classify the teams into “effective

collaboration” (e.g., teams that have positive “score change”) and “ineffective collaboration” (e.g., teams that have negative “score change” or zero “number of changes”).

In addition to quantifying the collaboration outcomes, we introduced a “CPS profile” to characterize the CPS skills exhibited by each team during the collaboration process. The CPS profile is defined as the frequency distribution of CPS skills (unigram) and the consecutive CPS skill pairs (bigram). Random forest classification analysis [12, 3] is used to analyze the relationship between collaboration outcomes and the CPS skills. Random forest is a decision tree-based binary classifier, with increased robustness by using multiple trees rather than a single tree. It is mainly used as a classifier to map the features (independent variables) to labels (dependent variables). When training a random forest classifier, the relative importance of the feature variables for determining the labels can be obtained as a by-product. In our case, the feature variables are the CPS profile and the labels are the two classes of collaboration outcomes, e.g., effective and ineffective collaborations. By training a random forest classifier on the data, we found that negotiation skill is more important for a successful collaboration outcome.

2. METHOD

2.1 Assessment Instruments

We designed a research study to explore the relationship between CPS skills and the collaboration outcomes. In this large-scale study, we focused on the domain of science and limited the number of members of each team to two. We used text chat as the collaboration medium. There were two major assessment instruments: 1) A standalone test for general science knowledge consisting of 37 multiple-choice items adapted from the Scientific Literacy Measurement (SLiM) instrument [18]; 2) A web-based collaborative simulation task on volcanoes that require two participants collaborate to complete.

The simulation task was modified from an existing simulation, Volcano Trialogue [23]. In this simulation task, two participants worked together via text chat to complete the tasks. All of the turn-by-turn conversations and time-stamped responses to the questions were recorded in a carefully designed log file [11]. These conversations were used to measure CPS skills, while the responses to the in-simulation science items were used to measure science inquiry skills [23]. Figure 1 shows screenshot of the simulation task.

To capture the evidence for the outcomes of the collaboration, we designed a four-step response procedure for each item in the task: 1) Each participant was prompted to respond to the item individually before any collaboration; 2) Each participant was prompted to discuss the item with her partner; 3) Each participant was prompted to revise her initial response if she wanted; 4) A representative was randomly chosen to submit a team answer.

In this way, the changes in the responses before and after the collaboration reflect how effective the collaborations were and allow us to probe directly what CPS skills are more important for better collaboration outcomes.

2.2 Participants and Data



Figure 1: Screenshots from the collaborative simulation task.

We collected data through Amazon Mechanical Turk, a crowdsourcing data collection platform [13]. We recruited 1,000 participants with at least one year of college education to take the general science test. Then, they were teamed randomly into dyads to take the collaborative simulation task.

After removing incomplete responses, we had complete responses from 493 dyads. However, a further scrutiny of the data showed that many of the teams started some conversations even before the system prompted them to discuss. This means that they started conversations before or during the period that they are supposed to make initial responses individually. Different teams had nonprompted conversations for a different subset of the items, which complicates the analysis. Of the teams, 82 did not have nonprompted conversations while the other teams had nonprompted discussions for a varying number of items. We compared the scores of the general science knowledge test for participants from the 82 teams with the scores for the rest of the teams via a two-tailed t-test for independent samples, and the resulting p-value is 0.38. This indicates that participants from the 82 teams are not different in a statistically significant way from the rest of the participants in terms of the general science knowledge. To make our analysis clean, we will stick to the data from this 82 teams throughout this paper.

The data from the simulation task for each team include the responses to the items in the simulation and the text chat communications between the dyads around each item. There are 7 multiple-choice equivalent items. Around each item, there are about 5 turns of conversations.

2.3 Analysis

The focus of this paper is to investigate the relationship between the CPS skills and the collaboration outcomes. As such, our analysis focuses on the responses and communications in the collaborative simulation task.

2.3.1 Scoring and Annotating

Students’ responses to the seven multiple-choice equivalent items were scored based on the corresponding scoring rubrics as presentend in [23]. In addition to the outcome response

data, we also applied a CPS framework to annotate the chat communications during the collaboration [15]. This CPS framework was developed based on the findings from computer-supported collaborative learning (CSCL) research [2, 7, 9, 21] and the PISA 2015 Collaborative Problem Solving Framework [17].

The framework outlines the four specific categories of the CPS construct (skills) we would like to focus on: *sharing ideas*, *negotiating ideas*, *regulating problem-solving activities*, and *maintaining communication*. Each of these major categories had some subcategories and the total number of subcategories amounted to 33 and a summary of the coding rubrics can be found in Table 1. All the coding was done at the subcategory level, based on which of the four major categories were assigned at a later point.

Two human raters were trained on the CPS framework, and they double-coded a subset of the discourse data (15% of the data). The unit of analysis was each turn of a conversation, or each conversational utterance. The raters had two training sessions before they started independent coding. In the first session, the author of the CPS framework (the second author) trained both raters on the 33 subcategories of CPS skills using the skills definitions and coding examples for each subcategory. In the second training session, the trainer and two raters coded data from one dyad together to practice the application of specific codes and address issues specific to classifying utterances using the CPS framework. After the training sessions, the two raters independently coded discourse data from about 80 dyads.

We used the unweighted kappa statistic to measure the degree of agreement between the human raters' coding. The unweighted kappa was 0.61 for all 33 subcategories and 0.65 for the four major categories. According to Fleiss and Cohen [8], a kappa value of 0.4 is an acceptable level of agreement for social science experiments.

2.3.2 Quantifying the Collaboration Outcomes

The difference between the revised response and initial response is a direct measure of the collaboration outcomes. If we treat each dyad as the unit of analysis, we need to define variables to quantify the answer changes for each item. We first introduce the "number of changes" (denoted as n) to quantify how many revised responses are different from initial responses from both members of each dyad for each item. The possible values for n are $\{0, 1, 2\}$: n is zero when nobody makes any changes, one when only one person makes changes, and two when both members make changes. Next, we introduce "score change" (denoted as s) to quantify the total score changes between the revised response and the initial response from both members of each dyad for each item. The definition of s is the sum of the score difference between initial responses and revised responses for the two members of each dyad. The possible states for s are $\{-2, -1, 0, 1, 2\}$. One should note that for the state $s = 0$, there are two different possibilities. The first is that both members do not change their responses. The second is that one member changes a response from incorrect to correct and the other changes from correct to incorrect. Therefore, to have a complete description of the changes at a dyadic level, we introduce the vector "item collaboration effect" for each

item, $\delta_k = (s_k, n_k)$, with δ_k defined at the item level and subscript k denoting the item number. At the task level, we simply sum all items, which gives $\Delta = (S, N)$, where $S = \sum_k s_k$ and $N = \sum_k n_k$. By convention, we use the lowercase n and s to denote the item level changes and the uppercase N and S to denote the task-level changes.

2.3.3 Quantifying the CPS Skills

Each turn-by-turn conversations was classified in one of the four categories of CPS skills (e.g., share ideas, negotiate ideas, regulate problem solving, and maintain communication). We introduce a "CPS profile" as a quantitative representation of the CPS skills of each dyad. The profile was defined by the frequency counts of each of the four CPS-skill categories or their combinations and had two levels, unigram and bigram. The unigram, bigram, or even ngram levels are used in natural language processing to represent text. We borrow this idea here to represent CPS skills and limit us to the unigram and bigram as the frequency count is too low for other ngram. The frequency counts of the different CPS skills were used at the unigram level, while the frequency counts of consecutive pairs of CPS skills in the conversations were used at the bigram level. As such, each dyadic team's communications can be represented by the corresponding CPS profile.

It is worth noting that though we consider only unigram and bigram of the CPS skills, other collaboration-related information can also be appended to the profile. For example, the number of turns, the total number of words, etc. Such a profile is essentially a vector representation of collaboration skills exhibited by each team. The vector nature of this representation allows us to easily calculate "similarity" or "dissimilarity" among the teams, which is the foundation of cluster analysis.

3. FINDINGS

We have introduced two variables, N and S , to quantify the collaboration outcomes. We also introduced the CPS profile to quantify the CPS skills. Now, we investigate the relationship between the CPS skills and the collaboration outcomes.

3.1 Effective versus Ineffective Collaboration

Based on the N and S variables, we define the effective collaboration and ineffective collaboration as follows

- Effective collaboration: $N > 0 \cap S > 0$.
- Ineffective collaboration: $(N > 0 \cap S \leq 0) \cup N = 0$.

We need to point out that the criteria for effective collaboration is not necessarily a fixed one. In the current study, we considered the collaboration as effective as long as at least one member made at least a total net change from incorrect to correct. If nobody in the team made at least one total net correct change, we thought of the collaboration as ineffective. Figure 2 shows how the 82 teams were distributed in the space spanned by S and N .

Table 1: Coding rubric of CPS skills used in this paper was developed based on a review of CSCL research findings [2, 7, 9], and the PISA 2015 Collaborative Problem Solving Framework [17], with a focus on CPS in the domain of science. More details about the CPS framework can be found in [15].

CPS skills	Student performance (subcategories)
Sharing ideas	<ol style="list-style-type: none"> 1. Student gives task-relevant information (e.g., individual response) to the teammate. 2. Student points out a resource to retrieve task-relevant information. 3. Student responds to the teammate's request for task-relevant information.
Negotiating ideas	<ol style="list-style-type: none"> 4. Student expresses agreement with the teammates. 5. Student expresses disagreement with teammates. 6. Student expresses uncertainty of agree or disagree. 7. Student asks the teammate to repeat a statement. 8. Student asks the teammate to clarify a statement. 9. Student rephrases/complete the teammate's statement. 10. Student identifies a conflict in his or her own idea and the teammate's idea. 11. Student uses relevant evidence to point out some gap in the teammate's statement. 12. Student elaborates on his or her own statement. 13. Student changes his or her own idea after listening to the teammate's reasoning
Regulating problem solving	<ol style="list-style-type: none"> 14. Student identify the goal of the conversation. 15. Student suggests the next step for the group to take. 16. Student expresses confusion/frustration or lack of understanding. 17. Student expresses progress in understanding. 18. Student reflects on what the group did. 19. Student expresses what is missing in the teamwork to solve the problem. 20. Student checks on understanding. 21. Student evaluates whether certain group contribution is useful or not for the problem solving. 22. Student shows satisfaction with the group performance. 23. Student points out some gap in a group decision. 24. Student identifies a problem in problem solving.
Maintaining communication	<ol style="list-style-type: none"> 25. Student responds to the teammate's question (using texts and text symbols). 26. Student manages to make the conversation alive (using texts and text symbols, using socially appropriate language). 27. Student waits for the teammate to finish his/her statement before taking turns. 28. Student uses socially appropriate language (e.g., greeting). 29. Student offers help. 30. Student apologizes for unintentional interruption. 31. Student rejects the teammate's suggestions without an accountable reason. 32. Student inputs something that does not make sense. 33. Student shows understanding of the teammate's frustration.

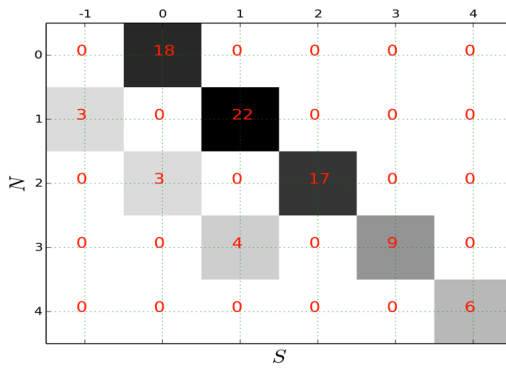


Figure 2: The distribution of the teams in space spanned by N and S.

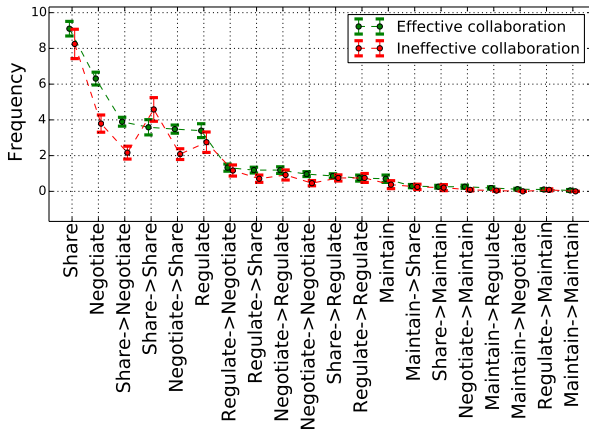


Figure 3: Unigram and bigram profile of CPS skills for the teams corresponding to effective and ineffective collaborations.

Next, we compare the mean CPS profiles of the teams from the effective and ineffective collaborations and the results are shown in Figure 3.

From these results, one can readily see that at the unigram level, the teams with effective collaboration show statistically significantly more negotiating skills than the teams with ineffective collaboration. At the bigram level, teams with effective collaboration exhibited statistically significantly more of the following consecutive CPS skill pairs: share-negotiate, negotiate-share, regulate-share, and negotiate-negotiate. However, the teams with ineffective collaboration showed many more share-share skill pairs.

3.2 Relative Importance of CPS Skills

Figure 3 shows certain CPS skills exhibit more different frequency for effective and ineffective collaborations, which means they have more weight in determining the collaboration outcomes. To get a more quantitative measure of the relative importance of each CPS skills (or skill pairs), we used two methods as follows.

First, we perform a t-test for each of the CPS skills (or skill

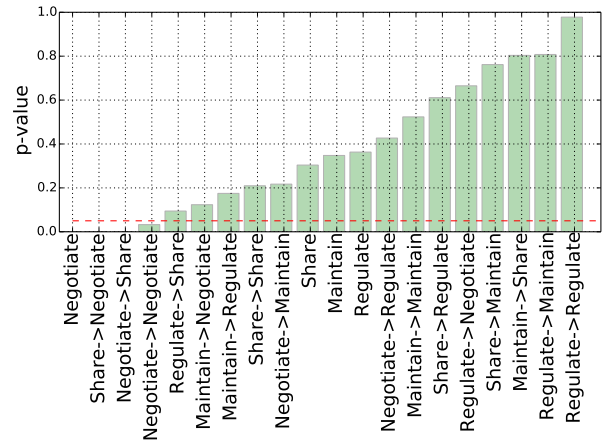


Figure 4: P-value of t-test on the frequency of different CPS skills corresponding to effective and ineffective collaborations. The red horizontal dashed line corresponds to a significant level of 0.05.

pairs) for the effective collaboration and ineffective collaboration groups. We use the corresponding p-value to tell which skills or skill pairs show more distinction. The p-value for each component of the CPS profile was shown in Figure 4. If we choose 0.05 as the significance level, negotiate, share-negotiate, negotiate-share and negotiate-negotiate stand out immediately.

A second method we used to find out the relative importance of the CPS skills or skill pairs (feature variables) is random forest classifier [12, 3]. We choose the collaboration outcomes as label variables. During the training of the classifier, a set of decision cuts were made on each feature variable. The relative depth of a feature used as a decision node in a decision tree represents the relative importance of that feature with respect to the predictability of the target labels. Generally speaking, features used at the top level of the decision tree will affect a larger fraction of the sample in terms of the final prediction. Therefore, the expected fraction over the trees in the forest can be used as an estimate of the relative importance of the features. Figure 5 shows the relative importance of the CPS skills and skill pairs based on such an analysis. The results show that negotiation-related skills top the ranking.

The results from these two different analyses converge nicely on that negotiation is a very critical skill for successful collaboration. This finding is consistent with the findings in the literature on knowledge-building discourse [19, 20], as knowledge is often built upon its use and negotiation includes interpretive process of making meaning of exchanged ideas.

4. CONCLUSIONS AND IMPLICATIONS

In this paper, we introduced a CPS profile approach to quantify the CPS skills of each team and found that the negotiation skill at the unigram level is important for better collaboration outcomes. At the bigram level, we found that more negotiation-related skill pairs, such as share-negotiate,

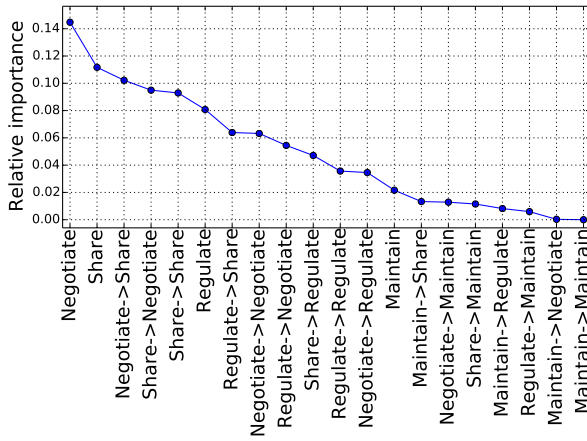


Figure 5: Relative feature importance based on a random forest classifier.

negotiate-share, regulate-share, and negotiate-negotiate, leads to more effective collaboration outcomes. However, purely sharing information with each other (share-share) is associated with poorer collaboration outcomes. This empirical finding may also inform the development of an outcome-oriented scale for CPS skills.

The current study also has limitations. For example, the items in the task are all relatively easy so that there are few turns for each item. There are not many items in the task, which limits the effect of the collaboration outcomes. All these issues will be resolved in our next round of data collection and analysis.

5. ACKNOWLEDGMENTS

Funding for this project is provided by Educational Testing Service through the game, simulation and collaboration initiative.

6. REFERENCES

- [1] R. Adams, A. Vista, C. Scoular, N. Awwal, P. Griffin, and E. Care. Automatic coding procedures. *Assessment and teaching of 21st century skills*, 2, 2015.
- [2] B. Barron. When smart groups fail. *The journal of the learning sciences*, 12(3):307–359, 2003.
- [3] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [4] E. Care and P. Griffin. An approach to assessment of collaborative problem solving. *Research & Practice in Technology Enhanced Learning*, 9(3):367–388, 2014.
- [5] E. G. Cohen, R. A. Lotan, B. A. Scarloss, and A. R. Arellano. Complex instruction: Equity in cooperative learning classrooms. *Theory into practice*, 38(2):80–86, 1999.
- [6] L. A. DeChurch and J. R. Mesmer-Magnus. The cognitive underpinnings of effective teamwork: a meta-analysis. *Journal of Applied Psychology*, 95(1):32, 2010.
- [7] P. Dillenbourg and D. Traum. Sharing solutions: Persistence and grounding in multimodal collaborative problem solving. *The Journal of the Learning Sciences*, 15(1):121–151, 2006.
- [8] J. L. Fleiss and J. Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 1973.
- [9] P. Griffin, B. McGaw, and E. Care. *Assessment and teaching of 21st century skills*. Springer, 2012.
- [10] J. Hao, L. Liu, A. von Davier, and P. Kyllonen. Assessing collaborative problem solving with simulation based tasks. *proceeding of 11th international conference on computer supported collaborative learning*, 2015.
- [11] J. Hao, L. Smith, R. Mislevy, A. von Davier, and M. Bauer. Taming log files from game and simulation based assessment: data model and data analysis tool. *ETS Research Report*, in press.
- [12] T. K. Ho. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 278–282. IEEE, 1995.
- [13] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456. ACM, 2008.
- [14] E. Klopfer, S. Osterweil, J. Groff, and J. Haas. Using the technology of today, in the classroom today. *The Education arcade*, 2009.
- [15] L. Liu, J. Hao, A. A. von Davier, P. Kyllonen, and D. Zapata-Rivera. A tough nut to crack: Measuring collaborative problem solving. *Handbook of Research on Technology Tools for Real-World Skill Development*, page 344, 2015.
- [16] H. F. O’Neil. *Workforce readiness: Competencies and assessment*. Psychology Press, 2014.
- [17] Organization for Economic Co-operation and Development [OECD]. Pisa 2015 draft collaborative problem solving assessment framework. *OECD Publishing*, 2013.
- [18] C.-J. Rundgren, S.-N. C. Rundgren, Y.-H. Tseng, P.-L. Lin, and C.-Y. Chang. Are you slim? developing an instrument for civic scientific literacy measurement (slim) based on media coverage. *Public Understanding of Science*, 21(6):759–773, 2012.
- [19] M. Scardamalia and C. Bereiter. Computer support for knowledge-building communities. *The journal of the learning sciences*, 3(3):265–283, 1994.
- [20] G. Stahl. *Group Cognition: Computer Support for Building Collaborative Knowledge (Acting with Technology)*. The MIT Press, 2006.
- [21] A. A. Von Davier and P. F. Halpin. Collaborative problem solving and the assessment of cognitive skills: Psychometric considerations. *ETS Research Report Series*, 2013(2):i–36, 2013.
- [22] A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, and T. W. Malone. Evidence for a collective intelligence factor in the performance of human groups. *science*, 330(6004):686–688, 2010.
- [23] D. Zapata-Rivera, T. Jackson, L. Liu, M. Bertling, M. Vezzu, and I. R. Katz. Assessing science inquiry skills using trialogues. In *Intelligent Tutoring Systems*, pages 625–626. Springer, 2014.