

Making Sense of Learning Analytics with a Configurational Approach

Ilias O. Pappas¹, Michail N. Giannakos¹ and Demetrios G. Sampson²

¹ Norwegian University of Science and Technology (NTNU), Trondheim, Norway
michailg@idi.ntnu.no

² Curtin University Perth, Australia
demetrios.sampson@curtin.edu.au

Abstract. This paper is an attempt to provide the basic guidelines on how to implement configurational analysis in the context of learning analytics. In detail, we offer a step by step approach on the fuzzy set qualitative comparative analysis (fsQCA). Learning analytics gain increased popularity, however studies use traditional symmetric statistical methods to analyze them. Building on the theory of complexity and configuration theory we suggest on using fsQCA in order to gain a deeper understanding of the data, which may lead to understanding different learning phenomena as well as to the creation of new theories. We further describe the steps on how to perform a contrarian case analysis, which will help in identifying asymmetric relations among the data. Finally, testing for predictive validity with fsQCA is explained. Many of the steps described here may be implemented in various contexts, however we tried to provide examples and instructions for learning analytics oriented research.

Keywords: learning analytics, fuzzy-set qualitative comparative analysis, fsQCA, complexity theory, configuration theory

1. Introduction

Studies in the area of learning analytics have been grounded mainly on symmetric tests and regression based models (RBM), such as multiple regression analysis (MRA) and structural equation modelling (SEM). Symmetric tests assume that a change on the predictor variable will result in the same change on the outcome variable. These methods estimate the significance of the effects between two variables in a model or compare the effects among the variables between two or more models [1-3]. Further, regression based models build on variance theories, which suggest that a predictor variable needs to be both necessary and sufficient condition in order to achieve the desired outcome [4, 5]. However, focusing on symmetric and net effects may be misleading, usually because the observed net effects do not apply to all of the cases in a dataset [6], and most relationships in real life are not symmetrical [1, 7].

Qualitative comparative analysis (QCA) has recently been applied in social sciences [8], including education and learning [9-11]. QCA has three main variations, that is crisp set QCA (csQCA), multi-value QCA (mvQCA), and fuzzy set QCA (fsQCA) [12]. Although fsQCA is able to address various limitations of the other QCA variations

[5], recent studies in the context of learning have not chosen to employ it [11]. However, fsQCA and configurational analysis have been applied primarily in the last decade in organizational research, and lately in the area of IS and business management in order to examine user behavior [1, 5-7]. It is thus evident that configurational analysis may offer valuable insights in the context of learning analytics. Nonetheless, there is still little work on this area and many researchers are still unfamiliar with this method.

This study aims to increase awareness and offer a step by step approach of fsQCA in the context of learning analytics. fsQCA identifies patterns between independent and dependent variables, which leads to outcomes and goes a step further from analyses of variance, correlations and multiple regression models. Similarly, it is important to extend the present application of QCA on learning analytics by employing fsQCA in this area.

2. Benefits and limitations of configurational analysis

As we have already mentioned, the majority of the studies in learning analytics (and even in the wider area of educational technology) research apply regression based methods (e.g., least squares, linear regression) in order to examine and predict learner behavior and the learning outcome (e.g., [13, 14]). A variable that affects the outcome in only a small subset of cases cannot be identified by regression analysis. In the area of learning analytics, where you always have different subsets (e.g., different learning styles, competences, demographics) researchers' capacity to investigate different subsets of learners is of great importance. Thus, applying configurational analysis may complement and extend the findings from RBMs. The benefits of configurational analysis and fsQCA mainly occur from the limitations of RBMs [1, 3-6]. In detail, RBMs take a net effect approach in examining the effects among the factors of interest and the variables are examined in a competing environment. The covariance among the variables in a model indicates that the presence or absence of a certain variable will influence their effect on each other as well as on the expected outcome, adding to the importance of applying configurational analysis, which is based on this notion [15].

Configurational analysis focuses on the asymmetric relations that exist among the examined variables and the outcome of interest, while at the same time the outcome of interest may be achieved with various ways. For instance, students' activity (e.g., materials views) and background knowledge (e.g., results from previous tests) can predict the future learning outcome or dropouts only if they are examined in combination. It is not possible to predict the learning outcome based only on students' activity or background knowledge. Finally, configurational analysis may be more robust than RBMs mainly as it is not sensitive to outliers. Employing fsQCA to analyze the data, the sample is divided into multiple subsets, thus creating multiple combinations of configurations. In effect, the outliers will not have influence all solutions (i.e., configurations) but only on specific ones. To this end, every configuration represents only a subset of the sample, hence the representativeness of the sample is not able to affect all the configurations [5, 16]

Nonetheless, configurational analysis has certain limitations, which should be taken into account when implementing fsQCA [1, 2, 5, 6]. In detail, in order to apply fsQCA the researcher is required to have substantial knowledge on the conditions and the outcome of interest, which will be used to calibrate (i.e., transform variables into fuzzy sets) the data, to simplify the solutions, as well as to interpret the results. This necessary knowledge however may lead to a subjective bias on the results. Also, it is not able to identify the unique contribution of every variable on every solution, but this is not the case because the goal of fsQCA is to identify complex solutions and combinations of the independent variables. Finally, fsQCA does not account for the validity and reliability of the latent variables, as it was designed to be used with single-item variables. To address this issue, before applying fsQCA the measurement model is tested for its reliability and validity applying the traditional SEM techniques [1, 5]. Once reliability and validity have been established, configurational analysis may be employed by transforming the variables into fuzzy sets.

3. Conceptual model and formulation of propositions

In order to conceptualize all the possible relationships among the examined factors (i.e., independent variables) and the outcome of interest (i.e., dependent variable), the researchers may use a Venn diagram [1-2]. Since, multiple relationships exist among variables, depending on how they combine with each other they may predict high level of learning outcomes. Figure 1 presents an example of a Venn diagram illustrating the conceptual model.

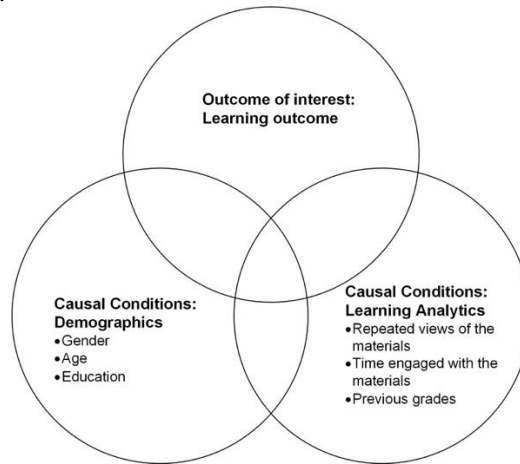


Fig. 1. A Venn diagram illustrating a conceptual model adopted from Pappas et al. [1]

When performing a configurational analysis the researchers need to present the propositions based on which they will proceed to implement fsQCA. Typically, studies that employ regression based models *make hypotheses* in order to examine the relations among the variables of interest. However, in the case of configurational analysis, the

formulation of propositions is more appropriate. On the one hand, a *proposition* is defined as a logically and theoretically valid statement, which explains relations among constructs/parameters/concepts. On the other hand, a *hypothesis* is a logical statement, based one or more propositions, and is to be tested for validity [17]. Building on the assumptions of configuration theory and the theory of complexity the researchers may create propositions that can be later verified through fsQCA. Configurational analysis is theory driven and is up to the researchers to present the research questions, formulate propositions and interpret the findings based on their knowledge.

Complexity theory and configuration theory incorporate the principle of *equipfinality*, based on which the outcome of interest can be explained equally by alternative sets of causal conditions that combine in sufficient configurations for the outcome [6, 16]. For example, including different learning analytics (log files from learners' activity, demographics, knowledge, attitudes) in our investigation it is possible to identify different combinations of learning analytics that will explain the same outcome (e.g., dropouts, students' learning). A proposition example in the area of learning technology, consistent with figure 1, can be as follows: *"No single best configuration of learning analytics from students' activities and demographics leads to high learning outcomes"*. Thus, researchers do not look for a single solution. Further, configuration theory proposes the occurrence of causal asymmetry. Causal asymmetry means that for an outcome to occur, the presence and absence of a causal condition depend on how this condition combines with one or more others [6]. For instance, in order to have high learning outcome or low dropouts, the presence and absence of various learning analytics depend on how these learning analytics combine together. Similarly, building on the principle of causal asymmetry a proposition example can be the following *"Single causal conditions may be present or absent within configurations for high learning outcomes, depending on how they combine with other causal conditions"*.

4. Methodology (Concepts and analysis)

This paper provides basic steps on how to employ fuzzy set qualitative comparative analysis, using fs/QCA 2.5 [18]. fsQCA was developed by integrating fuzzy set and fuzzy logic with QCA [19]. fsQCA offers two types of configurations: necessary and sufficient. Such configurations may be marked by their presence, their absence, or a "do not care" condition. The necessary and the sufficient conditions create a distinction among core and peripheral elements. Core elements are those with strong causal relationships with the outcome, and peripheral elements are those with weaker ties [16].

4.1. Data Collections

In order to support the propositions made in the previous chapter the researcher need to gather the appropriate data. The data that are typically used in the regression based methodologies can be also used to perform configurational analysis with fsQCA. Further, the data may be based on either single- or multi-item constructs. The constructs may be both categorical (e.g., gender) or continuous. Regarding their values there is no

specific limitation for fsQCA, since all values need to be transformed into fuzzy sets (see section Data calibration). Data gathering may be performed through the classical tools, such as surveys, interviews, observations. Big data from various sectors (e.g., learning analytics) may be used as well to perform configurational analysis.

4.2. Reliability and validity

As we have already mentioned, fsQCA does not address the reliability and validity of measures. In order to overcome this issue, we suggest on applying the traditional techniques applied on RBMs and SEM before proceeding to the implementation of configurational analysis with fsQCA [1, 5].

4.3. Contrarian case analysis

When examining main relations between two variables, stating that a variable positively or negatively affects the other, indicates that most cases in the sample verify this relationship. However, the opposite relationship will occur for some of the cases in the same sample; hence, researchers should test their data for such contrarian cases [1, 6]. Two variables may have positive, negative and no effect in the same dataset, regardless of the significance of main effect of one on the other, thus, studies should employ contrarian case analysis to identify such opposite relations [6].

First, the sample should be divided in order to investigate the relations among the examined variables. Due to the fact that splitting methods, such as median split, may reduce statistical power and lead to false results when the variables are correlated [20], a different approach should be taken when splitting continuous variables. This can be avoided by creating quintiles (i.e., dividing the sample into five equal groups) by ranking the cases using the SPSS Rank Cases corresponding function with the Ntiles option. Next, a cross-tabulation across the quintiles should be performed, using the SPSS Crosstabs function, between every independent variable and the dependent variable. This will create a 5x5 table for every set of variables, that represents all combinations between the two variables for the whole sample. Thus, it is made clear the existence of the cases that present an opposite relation to the main effects with the outcome variable, supporting the importance of configurational analysis for explaining these relationships [6]. An example on how to clearly present the contrarian case analysis is offered by Pappas et al. [1].

4.4. Data calibration

After gathering the data, the first step in fsQCA is to define the outcome and the independent variables. Next, all variables need to be transformed into fuzzy sets with values ranging from 0 to 1 [7]. This procedure is called *data calibration* and its steps may vary depending both on the data as well as on the researchers' knowledge of the relevant theory and context [7]. Various studies describe this process [5, 7, 21, 22]. Data calibration may be either direct or indirect. In the direct method, the researcher

chooses three qualitative breakpoints, whereas in the indirect method, the measurements require rescaling based on qualitative assessments. The researcher may choose either method depending on the data and the underlying theory [5, 7]. The direct method of setting three values that correspond to full-set membership, full-set non-membership and intermediate-set membership is recommended, unless there is substantive reason to choose otherwise. For the present tutorial we choose to describe the direct method of data calibration as performed by Pappas et al. [1].

The value of 1 stands for full-set membership and that of 0 stands for non-set membership. Thus, all variables are continuous from 0 to 1, which defines the level of their membership. Variables are transformed into calibrated sets with the fsQCA software (using the “Calibrate” function) by setting the three thresholds. These represent a full set membership threshold value (fuzzy score = 0.95), a full non-membership value (fuzzy score = 0.05), and the crossover point (fuzzy score = 0.50) [3]. The three thresholds depend on the values of the variable in question (i.e., the variable to be calibrated). When the researcher has limited knowledge of the variable (e.g., big data coming from learners’ activity), a direct calibration method may be performed by choosing as thresholds the variables 1, 0, and 0.5, with the rest of the values being calibrated based on a linear function [5]. For example, on a five point Likert scale the thresholds may be 1, 5 and 3 respectively. Following the procedure employed by [22], for a seven point Likert scale, the thresholds may be set as 6, 2, and 4 respectively as described on Pappas et al. [1]. In the case of LAs a straightforward direct calibration method would be to set the thresholds at the minimum, median and maximum values. Nonetheless it is always up to the researcher to choose the thresholds based on prior knowledge and theory [1, 5-7, 23].

4.5. Obtaining the configurations

Following the calibration, the researcher is ready to run the fsQCA algorithm on the menu “Analyze” and choose “Fuzzy Truth Table Algorithm”. At this point the researcher chooses the outcome of interest (i.e., dependent variable) and all the causal conditions (i.e., independent variables). Regarding the outcome, the researcher may choose to examine the presence of the outcome, and choose “Set”, or the absence of the outcome “Set Negated”.

Next, the fsQCA algorithm produces a truth table of 2^k rows, with k representing the number of outcome predictors and each row representing each possible combination. For example, a truth table between two variables (i.e., conditions) would provide four possible logical combinations between them. For every combination, the minimum membership value is calculated; that is, the degree to which every case supports the specific combination. fsQCA uses the threshold of 0.5 to identify the combinations that are acceptably supported by the cases. Thus, all combinations that are not supported by at least one case with membership over the threshold of 0.5 are automatically removed from further analysis.

The final step is to sort the truth table based on frequency and consistency (Ragin 2008). Frequency describes the number of observations for each possible combination.

Consistency refers to “*the degree to which cases correspond to the set-theoretic relationships expressed in a solution*” [16]. A frequency cut-off point needs to be set in order to ensure that a minimum number of empirical observations is obtained for the assessment of subset relationships. For small and medium-sized samples, a cut-off point of 1 is appropriate, but for large-scale samples (e.g., 150 or more cases), the cut-off point should be set higher [7], and maybe set at 3. The lowest acceptable consistency should be higher than the recommended threshold of 0.75 [23]. Thus, after removing the combinations with low frequency using the option on the “*Edit*” menu, the truth table should be sorted based on their “*raw consistency*”. The final step is to insert the value of 1 or 0 on the column with the outcome variable. Choosing 1 or 0, depends on the consistency threshold that has been chosen. For example, for a consistency threshold of 0.75, all combinations with consistency larger than 0.75 should be set at 1 and the rest at 0. It is up to the researcher to choose how large this threshold will be. Once this is complete, the researcher may proceed with the option of “*Standard Analyses*”

4.6. Obtaining the solutions

Following the sorting of the truth table, the researcher is presented with the option to choose if a single independent variable should be present or absent at all times on the solutions. Unless otherwise needed, we suggest choosing “Present or Absent” in order to be obtain with all the possible combinations. Next, fsQCA provides the following three sets of solutions: complex, parsimonious, intermediate. The complex solution presents all the possible combinations of conditions when traditional logical operations are applied. Complex solutions are simplified into parsimonious and intermediate solutions, which are simpler and up for interpretation. The parsimonious solution is a simplified version of the complex solution and presents the most important conditions which cannot be left out from any solution. These are called “*core conditions*” [16] and are identified automatically but fsQCA. Finally, the intermediate solution is obtained when performing counterfactual analysis on the complex and parsimonious solution [5, 7]. In essence, the intermediate solution depends on simplifying assumptions that are applied by the researcher, which at all times should be consistent with theoretical and empirical knowledge. The intermediate solution is part of the complex solutions and includes the parsimonious solution. The conditions that are part of the intermediate solution and not part of the parsimonious, are called “*peripheral conditions*” [16].

4.7. Interpreting the solutions

FsQCA presents the complex and parsimonious solution regardless of any simplifying assumptions employed by the researcher, while the intermediate solution depends directly on these assumptions. A combination of the parsimonious and intermediate solution is recommended as the main point of reference for interpreting the fsQCA results. In detail, the researchers should create a table that will include both core and peripheral conditions [1, 16]. In order to do this, the researcher should identify the conditions of the parsimonious solution in the intermediate solution. This will lead to a

combined solution, which will clearly present all core and peripheral conditions, thus helping the interpretation of the findings. Typically, the presence of a condition is presented with a black circle (●), the absence with a crossed-out circle (⊗), and the “do not care” condition with a blank space [16]. The distinction between core and peripheral is made by using large and small circles respectively. The researchers should also present the overall solution consistency as well as the overall solution coverage. The overall coverage describes the extent to which the outcome of interest may be explained by the configurations, and may be compared with the R-square reported on RBMs [3]. The next figure offers an example of how findings from fsQCA should be presented.

Configurations for achieving high level of learning outcomes

Configuration	Solution								
	1	2	3	4	5	6	7	8	9
Learning analytics									
Repeated views of the materials	●	●	●	●	●	●			⊗
Time engaged with the materials	⊗	•	•	●	●	⊗		•	⊗
Previous grades		•	•	•			●	●	⊗
Demographics									
Gender (<i>Male</i>)	●	●	●		⊗	⊗	⊗	•	⊗
Age (<25)		⊗	•		⊗	⊗	⊗	•	●
Education (<i>Postgraduate</i>)	⊗		•	⊗	⊗	●	⊗	⊗	⊗
	⊗	⊗	•	⊗	⊗	⊗	⊗	⊗	⊗
Consistency	0.932	0.956	0.959	0.918	0.896	0.877	0.837	0.950	0.863
Raw Coverage	0.261	0.337	0.133	0.690	0.471	0.161	0.535	0.337	0.118
Unique Coverage	0.018	0.005	0.023	0.052	0.002	0.007	0.051	0.007	0.004
Overall solution consistency	0.841								
Overall solution coverage	0.840								
Note: Black circles indicate the presence of a condition, and circles with “x” indicate its absence. Large circles indicate core conditions; small ones, peripheral conditions. Blank circles indicate “don’t care”.									

Fig. 2. An example of fsQCA findings adopted from Pappas et al. [1]

4.8. Predictive validity

After obtaining the fsQCA findings researchers should test for predictive validity, which examines how well the model predicts the outcome in additional samples [1, 6, 24]. Predictive validity is important because achieving only good model fit does not necessarily mean that the model offers good predictions. In order to test for predictive validity, the first step is to divide the sample into two subsamples and ran the same analysis for both subsamples, as it was described in the previous sections. Thus, the second step is to run the fsQCA for the first sample, and then the obtained findings should be tested against the second sample.

After obtaining the findings from the first subsample, the researcher must use the second sample to proceed with the predictive validity testing. From the findings of the first subsample, each solution, which contains the various combinations of present and absent variables, should be modeled as one variable by using “*Compute*” from the “*Variable*” menu. Thus, the fsQCA function “*fuzzynot(x)*” is used for every variable

that is absent (~) in the solution. This function computes the negation (1-x) of a variable (fuzzy set). Next, in order to model each solution, the function “*fuzzyand(x,...)*” is used, which takes as input all the variables that are present in each configuration and the new variables that occurred as the outcome of the “*fuzzynot(x)*” function. The “*fuzzyand(x, ...)*” function returns a minimum of two variables (fuzzy sets).

Finally, the new variable is plotted against the outcome of interest using the second subsample, from the fsQCA menu (“*Graphs*” – “*Fuzzy*” – “*XY Plot*”). Consistency and coverage values are presented here, which they should not contradict the consistency and coverage of the solution. The next figure offers an example on how to present the findings from predictive validity.

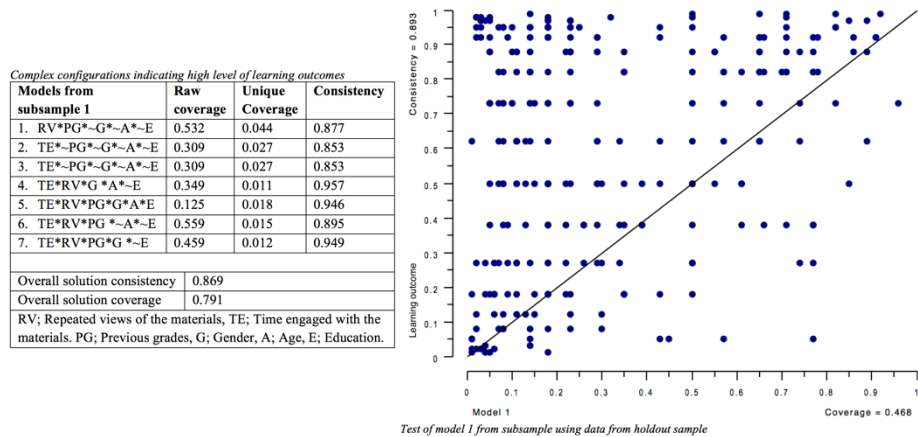


Fig. 3. An example of predictive validity testing using fsQCA adopted from Pappas et al. [1]

5. Discussion

Learning analytics are rapidly implemented in various educational settings, and the majority of the published work in the area are based on traditional tools to analyze such data (e.g., MRA, SEM) [13]. The goal of this paper is to offer a step by step approach on how to perform fuzzy set qualitative comparative analysis in the context of learning analytics and try make sense of diverse learning phenomena happening simultaneously. This approach is of particular interest on heterogeneous learning analytics, coming from datasets consisted of learners with different learning styles, backgrounds and so on. fsQCA can help us to better understand and further develop teaching and learning approaches enhancing learners’ dynamics and personalized needs in a ubiquitous learning era. The implementation of configurational analysis depends on the researchers’ previous knowledge of theory and empirical work, thus, it is not possible to offer a highly detailed analysis in this paper. However, this is not the case here, since our goal is to introduce fsQCA to researchers working with learning analytics, and provide a springboard for them. fsQCA has received increased attention lately in various fields (e.g., management, business), and despite the great potential there are no learning analytics studies utilizing this promising technique.

Acknowledgments

Our thanks to thank the Norwegian Research Council for its financial support under the projects FUTURE LEARNING (number: 255129/H20) and SE@VBL (number: 248523/H20).

References

1. Pappas, I.O., Kourouthanassis, P.E., Giannakos, M.N., Chrissikopoulos, V.: Explaining online shopping behavior with fsQCA: The role of cognitive and affective perceptions. *Journal of Business Research* 69, 794-803 (2016)
2. Pappas, I.O., Mikalef, P., Giannakos, M.N.: Video-Based Learning Adoption: A typology of learners. In: Proceedings of the workshop on Smart Environments and Analytics in Video-Based Learning (SE@VBL), LAK2016. (2016)
3. Woodside, A.G.: Moving beyond multiple regression analysis to algorithms: Calling for adoption of a paradigm shift from symmetric to asymmetric thinking in data analysis and crafting theory. *Journal of Business Research* 66, 463-472 (2013)
4. El Sawy, O.A., Malhotra, A., Park, Y., Pavlou, P.A.: Research Commentary-Seeking the Configurations of Digital Ecodynamics: It Takes Three to Tango. *Information Systems Research* 21, 835-848 (2010)
5. Liu, Y., Mezei, J., Kostakos, V., Li, H.: Applying configurational analysis to IS behavioural research: a methodological alternative for modelling combinatorial complexities. *Information Systems Journal* (2015)
6. Woodside, A.G.: Embrace• perform• model: Complexity theory, contrarian case analysis, and multiple realities. *Journal of Business Research* 67, 2495-2503 (2014)
7. Ragin, C.C.: *Redesigning social inquiry: Fuzzy sets and beyond*. Wiley Online Library (2008)
8. Schneider, C.Q., Wagemann, C.: *Set-theoretic methods for the social sciences: A guide to qualitative comparative analysis*. Cambridge University Press (2012)
9. Glaesser, J., Cooper, B.: Gender, parental education, and ability: their interacting roles in predicting GCSE success. *Cambridge Journal of Education* 42, 463-480 (2012)
10. Bakker, R.M., Cambré, B., Korlaar, L., Raab, J.: Managing the project learning paradox: A set-theoretic approach toward project knowledge transfer. *International Journal of Project Management* 29, 494-503 (2011)
11. Mavroudi, A., Hadzilacos, T., Kalles, D., Gregoriades, A.: Teacher-led design of an adaptive learning environment. *Interactive Learning Environments* 1-15 (2015)
12. Schneider, C.Q., Wagemann, C.: Standards of good practice in qualitative comparative analysis (QCA) and fuzzy-sets. *Comparative Sociology* 9, 397-418 (2010)
13. Baker, R.S., Inventado, P.S.: Educational data mining and learning analytics. *Learning Analytics*, pp. 61-75. Springer (2014)
14. Elbadrawy, A., Studham, R.S., Karypis, G.: Collaborative multi-regression models for predicting students' performance in course activities. In: Proceedings of the Fifth International Conference on Learning Analytics And Knowledge, pp. 103-107. ACM, (Year)
15. Fiss, P.C.: A set-theoretic approach to organizational configurations. *Academy of management review* 32, 1180-1198 (2007)
16. Fiss, P.C.: Building better causal theories: A fuzzy set approach to typologies in organization research. *Academy of Management Journal* 54, 393-420 (2011)
17. Reynolds, P.D.: *Primer in Theory Construction: An A&B Classics Edition*. Routledge (2015)
18. Ragin, C.C., Davey, S.: fs/QCA [Computer Programme], version 2.5. Irvine, CA: University of California (2014)
19. Ragin, C.C.: *Fuzzy-set social science*. University of Chicago Press (2000)
20. Fitzsimons, G.J.: Death to dichotomizing. *Journal of Consumer Research* 35, 5-8 (2008)

21. Mendel, J.M., Korjani, M.M.: Charles Ragin's fuzzy set qualitative comparative analysis (fsQCA) used for linguistic summarizations. *Information Sciences* 202, 1-23 (2012)
22. Ordanini, A., Parasuraman, A., Rubera, G.: When the recipe is more important than the ingredients: a Qualitative Comparative Analysis (QCA) of service innovation configurations. *Journal of Service Research* 1094670513513337 (2013)
23. Ragin, C.C.: Set relations in social research: Evaluating their consistency and coverage. *Political Analysis* 14, 291-310 (2006)
24. Gigerenzer, G., Brighton, H.: Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science* 1, 107-143 (2009)