The Omitted Variable: Could DuoTest Enable a New Way to Assess Team Performance in Team-Based Learning?

RICCARDO BONAZZI[1]& YVIANE ROUILLER[2]

[1] University of Applied Sciences and Arts Western Switzerland (HES-SO), Institute of Entrepreneurship and Management , Sierre, Switzerland, e-mail: riccardo.bonazzi@hevs.ch

[2] Haute Ecole Pédagogique du Valais, St-Maurice, Switzerland, e-mail: yviane.rouiller@hepvs.ch

Abstract This article is part of an ongoing project to develop a method for team-based learning named Testudo. We present an assessment technique called DuoTest, which allows students to do their final exam twice in a row: the first time, participants do their exam individually (Exa01); the second time, they solve the same exam in groups (Exa02). By comparing individual and group exams, the system induces the positive (or negative) effect of each team over the individual performances. Empirical results collected from 70 students show that individual exams are a reliable, although weak, predictor of the group scores ($p<0.10$, Adj R2= 0.02). Instead, by measuring the fixed effect of each team, we obtain a better predictor of Exa02 (Adj R2= 0.71). Although additional testing is required, our guidelines address a current gap in the literature for techniques that rigorously assess the individual and team dimensions, and that are easy to implement.

Ključne besede: team-based learning, project-based learning, team assessment, individual assessment, immediate feedback assessment technique, design science research.

University of Maribor Press

DOI https://doi.org/10.18690/978-961-286-362-3.39
ISBN 978-961-286-362-3

1 Introduction

> *"The town had a low wall of no great extent on one side, and to attack this the Romans employed three picked maniples. [...] The men of the first held their shields over their heads, and closed up, so that, owing to the density of the bucklers, it became like a tiled roof [...] in the shape of a tortoise (testudo)".*
> *Polybius, The Histories – Book 28.11*

The Roman Testudo is a well-known example of a military formation, where soldiers put together their shields to achieve a common goal, such as to protect themselves against a threat or to let other soldiers walk upon it whenever they come to a narrow ravine. Nonetheless, such powerful feature came at a price, since Roman Testudo were said to be advancing slowly in combat, since soldiers had to coordinate themselves. Accordingly, the Roman Testudo and its trade-off could be used as a metaphor for a situation, where students are expected to work together and solve a problem as a team.

There are still mixed evidences on whether working in teams is an appropriate method to prepare students for the challenges of a constantly changing business environment: on the one hand, some teachers prefer to give instruction via teacher-centered methods (lectures with little text reading and student discourse), under the belief that the best way to ensure content learning is for the instructor to present all necessary information to students (McKeachie and Svinicki, 2013). On the other hand, some scholars claim that traditional teaching methods do not enable all students to appropriately engage with the types of academic literacy constitutive to higher education (Hake, 1998; Lea and Street, 2006). Hence, this article starts with a simple intuition to bridge the two viewpoints: if we assume that the team itself is an important outcome of a team project, could we assess, at the end of the course, if the students would have been more/less effective without it? Indeed, there is a consensus on the difficulty of correctly assessing the performance of each student in a team project (Brazhkin & Zimmerman, 2019), and most educators lack a simple tool to do it. Nonetheless, most of the previous works have considered the team as noise to be cancelled to assess the individual, whereas we consider it as the most important artefact of a course, which asks students to work in teams to solve real-world projects and reflect on what they learned by doing so.

According to Kolb (2015) *learning* is the process whereby knowledge is created through the transformation of experience. *Group-based learning* is seen as a form of experiential learning and it has been termed differently through the years: (a) *small group learning* (Springer et al., 1999) include activities where the teacher lectures for 15–20 minutes and then asks students to pair with the student beside them to discuss a question, (b) *collaborative learning* involves carefully planned and structured group activities that are infused into a course of learning, whereas (c) *Team-based learning (TBL)* makes intense use of small groups in that it changes the structure of the course, in order to develop and then take advantage of the special capabilities of high-performance learning teams (Michaelsen et al., 2004). According to its authors, TBL is an important opportunity for teamwork skill development, experiential learning, and learning from peers. However, TBL presents many challenges and is most appropriate in courses that meet two conditions: (1) students are required during the course to understand a significant body of information and (2) a primary goal of the course is to apply this content by solving problems, answering complex questions and resolving issues (Swanson et al., 2019).

Accordingly, our research question is: **"how can we design a summative assessment of individual and team performance in a team-based learning scenario?"**

The rest of the paper proceeds as it follows. Section 2 briefly reviews the existing body of knowledge to answer our research question. Section 3 describes design science as our chosen methodology, highlights the relevant elements of the course which applies the Testudo method and then describes how to create and test the DuoTest prototype. Section 4 presents our preliminary findings, whereas section 5 concludes the paper by discussing the contribution and shortcomings of our work.

2 Literature review

In this section, we briefly assess the existing body of knowledge and define three constructs to avoid the *jingle fallacy* (constructs with the same name referring to different phenomena): (a) *team health*, which can be used to assess how well individuals work together in a team, (b) *transactivity*, to assess how each individual in a team can build on previous works from team members and (c) *immediate feedback* *assessment technique,* a tool used for summative evaluation in team-based learning that could be used to assess transactivity.

2.1 I2T: Individual contributions for the Team health

Recent work from (O'Neill et al., 2020) presents a set of 18 questions to rapidly and reliably assess the *team health* by asking team members to describe their perception of team communication, adaptability, relationships and education. Other scholars have suggested that assessment in TBL should take into account the cognitive, affective and behavioral dimensions (Brazhkin & Zimmerman, 2019). Indeed, students have multiple goals and motivations, which influence the team performance: mastery goals ("I want to learn new things") and social responsibility goals ("I want help my peers") prevail in effective teams, whereas belongingness goals (e.g., "I want my peers to like me") were more important than mastery goals in ineffective teams (Hijzen et al., 2007).

2.2 T2I: Team effect on the Individual performance

To some degree, the *group product* will be codified in an artifact (e.g., group report, dialogue, diagram, etc.), but the individual experience of that collaborative learning event will be transposed to future collaborative learning events. (Strijbos, 2010). Accordingly, the team effect can be associated to *transactivity*, that is the extent to which students refer and build on each other's' contributions and it can be measured by reflected in collaborative dialogue or individual products, or the extent to which students transform a shared artifact (e.g., a group report) (Weinberger et al., 2007).

2.3 Gap in the literature: how to assess transactivity

The *immediate feedback assessment technique* (IF-AT) form has (a) a series of boxes covered by an opaque, waxy coating similar to that found on scratch-off lottery tickets corresponding to the alternatives, with only one correct alternative having with a small star in it (Maurer & Kropp, 2015). The athours found that students who did the final exam with the Immediate Feedback Assessment Technique (IF-AT) scored 10% more on average when they got partial credit for iterative responding (they could scratch more then one box). Although, this approach is already used in team-based learning scenarios (Mazur, 1999), there is not a simple way to use it and assess how team transactivity influence individual performance.

3 Chosen methodology to develop and test the artefact

We position our study in the field of design science research (Hevner et al., 2004) and we developed an artefact in the shape of a prototype (March & Smith, 1995), following the guidelines of Peffers et al. (2007).

Identify problem and motivate. We describe an example of course of organization design, which would like to assess transactivity. At the beginning of the semester, students play a multi-round business simulation game (Martin-Rios & Erhardt, 2019). In this phase, students are assigned to a new random group every week, to learn how to rapidly work together and take decision under uncertainty. After four weeks, students form a group of max 5 team members. In this phase, students are assigned to a real project done with an external firm for eight weeks. All projects respect the five criteria for a project-based learning activity (Thomas, 2000): (a) projects are central to the curriculum, since the score given to the students reports will count as their midterm exam, (b) they are focused on problems that 'drive' students to encounter/struggle with the central concepts of a discipline, (c) they involve students in a constructive investigation, since students have to help the firm make sense of its data to find the solution, (d) they are student-driven to a significant degree, and (e) they are realistic and not school-like. Every week, students are asked to fill in a new section of the report and to submit it on a Moodle Workshop activity (Moodle, 2019a), where it will be assessed by their peers. During each class, the teacher briefly clarifies the required activities and facilitates discussions among team members. Slides are seldomly presented in class, since they are available to students in advance, together with check-up questions, as Moodle Lessons (Moodle, 2019b).

Define objectives of the solution. We wanted to improve the immediate feedback assessment technique (IF-AT) by developing an online solution, which could allow students to do the final exame by themselves and then to get partial credits if they managed to correct their mistakes, by discussing with their team members. This way, we could measure the degree of transactivity in each team. Accordingly, we state three hypotheses, which we would like to test:

- H1: the individual performance of Exa01 has a positive and statistically significant effect over the individual performance of Exa02. This statement is supported by all the reviewed literature on team-based learning
- H2: the team performance (*transactivity*) has a statistically significant effect over the individual performance of Exa02. If this hypothesis is correct, we should be able to see different improvement in different teams, depending on their degree of transactivity
- H3: the team performance has positive and statistically significant effect over the indivdual performance of Exa02. H3 extends H2. Based on previous results from (Maurer & Kropp, 2015) on IF-AT with partial credit, we could assume that a student having the possibility to correct his mistakes by discussing with his team will improve his final score.

Design and development of the artefact: the DuoTest prototype. The underlying idea of DuoTest is simple: to allow students to do their final exams twice in a row: the first time, participants do their exam individually (Exa01); the second time, they solve the same exam in groups (Exa02). By comparing individual and team performances, the system induces the positive (or negative) effect of each group over the individual performances.

Demonstration. Before the exam, we create a Moodle Quiz activity (Moodle, 2019) with ten questions: five theoretical questions and five questions about a case study. The type of the ten questions is Short Answer (Moodle, 2020): this will be relevant when we explain how to analyze the data after the exam. In the parameters of the Moodle Quiz activity, hereinafter referred to as Exa01, we set the duration at 35 minutes. Then, we copy the Quiz activity a second time, hereinafter referred to as Exa02. This way, the questions of Exa02 are the same of Exa01. In the parameters of Exa02, we set the beginning of the activity 5 minutes after the end of Exa01, to allow students the logistical time to setup their teams in the class. The duration of Exa02 is set at 20 minutes, which brings the total to 60 minutes. Finally, in the Moodle Gradebook (Moodle, 2019), we set the score of the final exam as the average between Exa01 and Exa02.

During the exam, students are expected to do Exa01 without additional material and by themselves. When Exa01 is over after 35 minutes, each student assembles with the team members, with whom he has been working between week 5 and 12.

Students can talk among them during Exa02 and they have access of any type of material. Indeed, Exa02 recreates the conditions that the team has lived during the semester and allows educators to assess in detail the dynamics of each team.

After the test, each answer is corrected by using a special feature of Short-answer questions: the educator defines a set of rules in the parameters of each question, and the answers of all students are corrected automatically by Moodle. This assures a coherent assessment all along and it increases the rigor of the overall process.

Evaluation. We tested our prototype with three classes of undergraduate students undertaking the same course, for a total of 71 students attending the final exam in Sierre (Switzerland) the 20th of January 2020. We claim that the exam was (a) valid, since chosen questions provide useful information about the concepts seen in class, (b) reliable, thanks to the rule-driven correction of each question, and (c) recognizable, since it fully replicated the way students work during the semester.

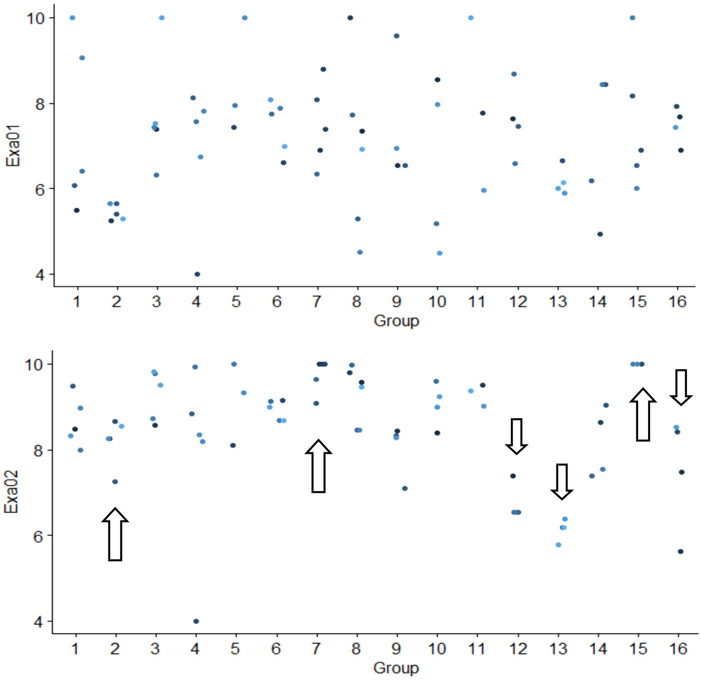


Figure 1: Students scores in the individual exam (Exa01) and group exam (Exa02)

4 Preliminary findings

This section analyses the results of the individual and the group exams, which are shown in Figure 1. One could expect the results of the second exam to be better than the first one, such as in the case of group G02, which had a strong concentration of scores below 6/10 and shifted up above 8/10. Some team performed better than other, with team G15 bringing all team members up to 10/10 and group G07 bringing a dispersed set of points in the first exam up above 9/10 in the second exam. Nonetheless, some teams performed worse in the second exam, the groups G12 and G13 being the most evident example of individuals, who decided to change some correct answers into wrong answers after discussing with the rest of the team. Finally, Group G04 had a student who attended the exam, but did not do it (row 18 in the table of Appendix A). To assign some quantitative data to our assessment, with start by scaling the raw data presented in Appendix A, in order to properly compare the coefficients of each variables. Moreover, after having looked for outliers with a large residual, we identify and remove the outlier in the row 18. Table 1 illustrates that the performance of the first exam (*Exa01*) positively effects the score of the second exam (*Exa02*), with a coefficient of 0.20 (hence *Exa02 = 0.20*Exa01*). The value of p = 0.09 shows that the relationship between the two variables is statistically significant. Therefore, ***we confirm the hypothesis H1***, and affirms that there is a causal effect between the first exam (done individually) and the second exam (done in group). Nonetheless, the Adjusted R2 = 0.03 suggests that the explanatory power of this model is fairly low. Hence, we add 15 binary variables for the 16 groups (the first group G01 will have 0 for each group variable). The Adjusted R2 of the new model is very good (0.71) and the coefficient of the first Exam (0.06) is not statistically significant anymore (p = 0.41), leading us to ***confirm the hypothesis H2***, which states that the team effect increases the explanatory power of our model.

Indeed, one could assume that the increase in the value of the R2 would be the consequence of using more variables; but the Adjusted R2 automatically adjusts the R2 of the model to take this effect into account. Moreover, the regression diagnostics in Appendix B does not indicate any further issues. Nonetheless, the analysis of the coefficients shows that ***we cannot confirm nor reject hypothesis H3***, which state that the team has a positive effect on the individual performance. The quantitative analysis rejoins the insights already visible from Figure 1: the coefficient of some groups (e.g. G07 and G15) is greater than the one of Exa01, whereas some other groups have a negative coefficient (G12 G13 and G16).

Table 1: Exa02 as a function of individual exam(model 01) and team *transactivity* (model 02)

Variable	Model 01: Individual	Model 02: Group Effect
Intercept	0.00 (1.00)	0.02 (0.92)
Exa01	0.20 (0.09)	0.06 (0.41)
Group 02		-0.31 (0.39)
Group 03		0.53 (0.13)
Group 04		0.15 (0.69)
Group 05		0.38 (0.35)
Group 06		0.24 (0.49)
Group 07		**0.94 (0.01)**
Group 08		0.57 (0.09)
Group 09		-0.53 (0.15)
Group 10		0.39 (0.30)
Group 11		0.54 (0.18)
Group 12		**-1.65 (0.00)**
Group 13		**-2.12 (0.00)**
Group 14		-0.41 (0.26)
Group 15		**1.16 (0.00)**
Group 16		**-0.99 (0.01)**
Adjusted R2 of the model	*0.03*	*0.71*

A final remark should be done for G02, and its surprising negative coefficient. Figure 1 shows that the score Exa02 of everyone increased from Exa01. Nonetheless, the quantitative analysis shows that students of group G02, who got the best Exa01 results, are those who got the worse Exa02 results afterwards.

5 Discussions and conclusions

This article started by using the metaphor of the Roman Testudo to describe how students learn to cooperate in order to deal with problems in their future careers. Our study suggests that what seems to be a single phenomenon (*team performance*) is in reality composed of assorted heterogeneous elements (Davis, 1971): *team health*, which depends on each team member, and *transactivity*, which influences the future performance of each team member and that we called "the omitted variable" in the title of the article. Accordingly, we wanted to look for new ways to design a final exam to assess individual and team performance in a team-based learning (TBL) course. Such objective is relevant and persisting in the field of study of information systems, since TBL is increasingly used to teach university students how to work together and solve complex problems in a growing number of fields, and we were missing of a structured and simple way to perform summative assessment. Although our approach might be biased towards TBL as a form of teaching, our intent is to bridge forms of experiential learning with classic testing techniques such as written exams. We have selected and reviewed previous works from the fields of team-based learning, project-based learning and software solution to assess students. Although such works are complementary, a paper that combines these three views to develop an artefact is missing. Therefore, we have decided to create a theory of design and action (Gregor, 2006), which explains how to do something and gives explicit prescriptions for teachers to construct a new type of final test for TBL classes, which we called DuoTest. Our preliminary findings show promising results that needs to be replicated in other classes and other topics. So far, DuoTest extends existing solutions for immediate impact assessments (Maurer & Kropp, 2015), since it allows to obtain deeper insights on the effect of the team on the individual performance and on the effect of such individuals on the team, at a fraction of its cost. Nonetheless, future work should try to categorize the different types of transactivity performance, and to explain how to predict the coefficients of each team by using data collected during the semester to link together *team health* and *transactivity*.

References

Brazhkin, V., & Zimmerman, H. (2019). Students' Perceptions of Learning in an Online Multiround Business Simulation Game: What Can We Learn from Them? Decision Sciences Journal of Innovative Education, 17(4), 363–386.

Davis, M. S. (1971). That's interesting! Towards a phenomenology of sociology and a sociology of phenomenology. Philosophy of the Social Sciences, 1(2), 309–344.

Gregor, S. (2006). The nature of theory in information systems. MIS Quarterly, 611–642.

Hevner, A., March, S., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. Management Information Systems Quarterly, 28(1). http://aisel.aisnet.org/misq/vol28/iss1/6

Hijzen, D., Boekaerts, M., & Vedder, P. (2007). Exploring the links between students' engagement in cooperative learning, their goal preferences and appraisals of instructional conditions in the classroom. Learning and Instruction, 17(6), 673–687.

Kolb, D. A. (2015). Experiential learning: Experience as the source of learning and development (2nd Edition). Pearson Education, Inc.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. Decision Support Systems, 15(4), 251–266.

Martin-Rios, C., & Erhardt, N. (2019). Organizational Design Simulation: Evolving Structures. Harvard Business School Publishing. https://hbsp.harvard.edu/product/7140-HTM-ENG

Maurer, T. W., & Kropp, J. J. (2015). The Impact of the Immediate Feedback Assessment Technique on Course Evaluations. Teaching & Learning Inquiry: The ISSOTL Journal, 3(1), 31–46. JSTOR. https://doi.org/10.2979/teachlearninqu.3.1.31

Mazur, E. (1999). Peer instruction: A user's manual. American Association of Physics Teachers.

Michaelsen, L. K., Knight, A. B., & Fink, L. D. (2004). Team-based learning: A transformative use of small groups in college teaching.

Moodle. (2019). Quiz activity—MoodleDocs. https://docs.moodle.org/38/en/Quiz_activity

Moodle. (2019, July 16). Grader report—MoodleDocs. https://docs.moodle.org/38/en/Grader_report

Moodle. (2019a, September 16). Workshop activity—MoodleDocs. https://docs.moodle.org/38/en/Workshop_activity

Moodle. (2019b, December 30). Lesson activity—MoodleDocs. https://docs.moodle.org/38/en/Lesson_activity

Moodle. (2020, January 14). Short-Answer question type—MoodleDocs. https://docs.moodle.org/38/en/Short-Answer_question_type

O'Neill, T. A., Pezer, L., Solis, L., Larson, N., Maynard, N., Dolphin, G. R., Brennan, R. W., & Li, S. (2020). Team dynamics feedback for post-secondary student learning teams: Introducing the "Bare CARE" assessment and report. Assessment & Evaluation in Higher Education, 0(0), 1–15. https://doi.org/10.1080/02602938.2020.1727412

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of Management Information Systems, 24(3), 45–77.

Springer, L., Stanne, M. E., & Donovan, S. S. (1999). Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis. Review of Educational Research, 69(1), 21–51.

Strijbos, J.-W. (2010). Assessment of (Computer-Supported) Collaborative Learning. IEEE Transactions on Learning Technologies, 4(1 (Jan.-March 2011)), 59–73.

Swanson, E., McCulley, L. V., Osman, D. J., Scammacca Lewis, N., & Solis, M. (2019). The effect of team-based learning on content knowledge: A meta-analysis. Active Learning in Higher Education, 20(1), 39–50.

Thomas, J. W. (2000). A Review of Research on Project-based Learning. 49.

Weinberger, A., Stegmann, K., & Fischer, F. (2007). Knowledge convergence in collaborative learning: Concepts and assessment. Learning and Instruction, 17(4), 416–426.

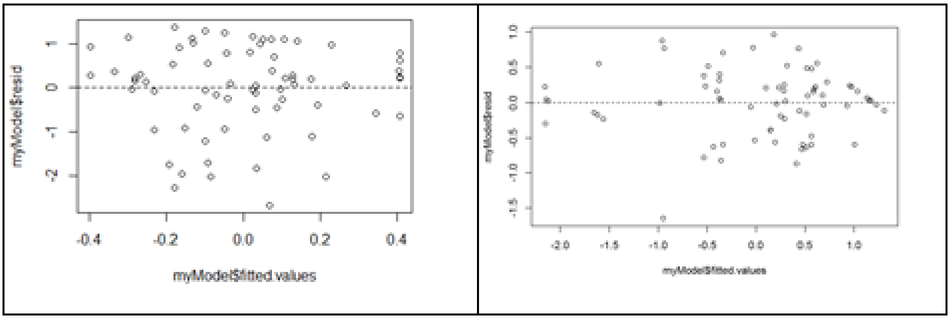
Appendix A: Complete dataset with the raw data

UID	Group	Class	Exa01	Exa02
1	1.00	2.00	5.50	8.48
2	10.00	3.00	8.54	8.39
3	8.00	1.00	7.34	9.57
4	12.00	3.00	7.64	7.39
5	16.00	3.00	6.89	7.47
6	9.00	3.00	6.54	8.44
7	3.00	1.00	7.39	8.57
8	16.00	3.00	7.69	5.62
9	7.00	2.00	6.90	10.00
10	2.00	1.00	5.25	8.25
11	11.00	3.00	7.77	9.52
12	8.00	1.00	10.00	9.79
13	7.00	2.00	8.79	10.00
14	14.00	2.00	4.94	8.64
15	7.00	2.00	7.39	10.00
16	5.00	2.00	7.44	8.09
17	6.00	3.00	6.60	9.14
18	4.00	2.00	4.00	4.00
19	16.00	3.00	7.92	8.42
20	15.00	3.00	6.89	10.00
21	2.00	1.00	5.40	8.65
22	13.00	1.00	6.64	6.19
23	15.00	3.00	8.18	10.00
24	14.00	2.00	8.43	9.04
25	7.00	2.00	8.09	9.09
26	9.00	3.00	6.55	7.10
27	9.00	3.00	9.58	8.34
28	2.00	1.00	5.65	7.25
29	1.00	2.00	6.08	9.48
30	15.00	3.00	6.55	10.00
31	12.00	3.00	7.45	6.54
32	4.00	2.00	8.12	8.84
33	8.00	1.00	5.29	8.47
34	14.00	2.00	6.19	7.39
35	8.00	1.00	7.72	9.97
36	12.00	3.00	8.69	6.54
37	7.00	2.00	6.34	9.64
38	6.00	3.00	7.74	9.14
39	10.00	3.00	5.18	9.59
40	3.00	1.00	6.32	9.77
41	4.00	2.00	7.57	9.94
42	12.00	3.00	6.59	6.54
43	1.00	2.00	6.40	7.98
44	5.00	2.00	7.95	10.00
45	14.00	2.00	8.44	7.54
46	6.00	3.00	7.88	8.69
47	15.00	3.00	10.00	10.00
48	3.00	1.00	7.43	8.72
49	15.00	3.00	6.00	10.00
50	8.00	1.00	4.50	8.47
51	1.00	2.00	9.05	8.98
52	4.00	2.00	7.82	8.19
53	2.00	1.00	5.65	8.25
54	5.00	2.00	10.00	9.34
55	4.00	2.00	6.74	8.34
56	9.00	3.00	6.95	8.29
57	10.00	3.00	7.97	8.99
58	11.00	3.00	5.97	9.02
59	13.00	1.00	5.90	6.39
60	1.00	2.00	10.00	8.33
61	16.00	3.00	7.44	8.52
62	10.00	3.00	4.50	9.24
63	6.00	3.00	8.09	8.99
64	13.00	1.00	6.00	5.79
65	3.00	1.00	7.52	9.82

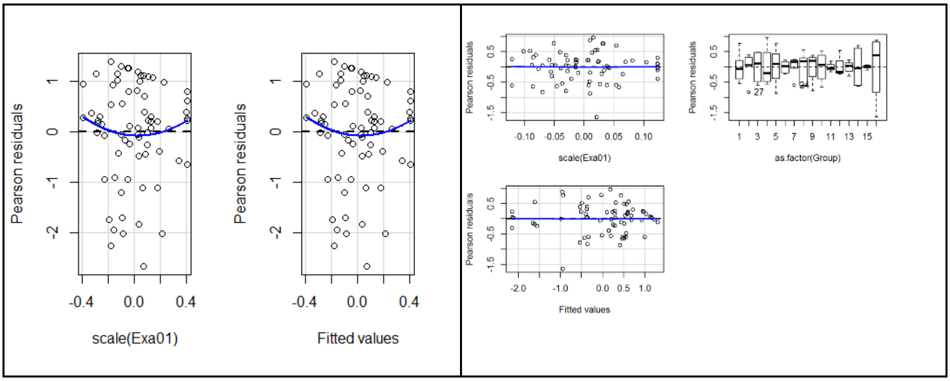
66	3.00	1.00	9.99	9.52
67	2.00	1.00	5.30	8.55
68	11.00	3.00	10.00	9.37
69	8.00	1.00	6.92	9.47
70	6.00	3.00	6.99	8.69
71	13.00	1.00	6.14	6.19

Appendix B: Regression diagnostic for model 01 (left) and model 02 (right)

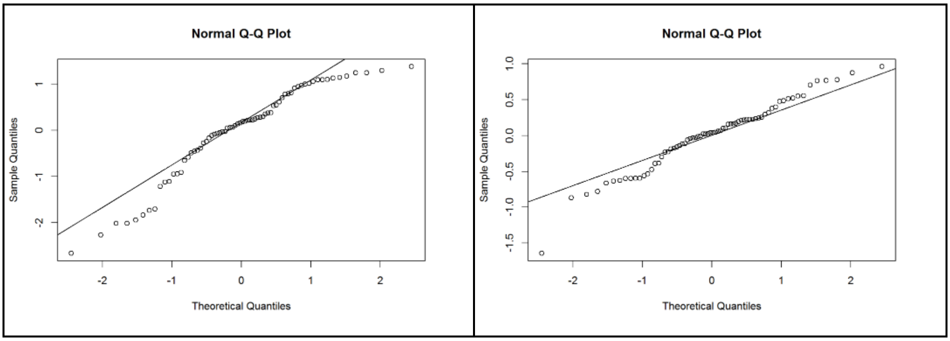
Homogeneity of variance: The error variance seems constant in the two models



Linearity: the relationships predictors and Exam02 becomes linear in model 02



Normality: the errors are normally distributed; hypotheses testing is reliable



Multicollinearity: when VIF > 10 a variable merits further investigation

	VIF	Df	GVIF^(1/(2*Df))
scale(Exa01)	1.32	1	1.15
as.factor(Group)	1.32	15	1.01