Statistica Applicata – ITALIAN JOURNAL OF APPLIED STATISTICS

A.S.A CONTACTS
Principal Contact
Francesco Palumbo (Editor in Chief)
editor@sa-ijas.org

Support Contact
Domenico Vistocco (Editorial Manager)
ijas@sa-ijas.org

JOURNAL WEBPAGE
https://www.sa-ijas.org/ojs/index.php/sa-ijas

Statistica Applicata – Italian Journal of Applied Statistics is a four-monthly journal published by the Associazione per la Statistica Applicata (A.S.A.), Largo Gemelli 1 – 20123 Milano, Italy (phone + 39 02 72342904). Advertising:Cleup sc, via G. Belzoni, 118/3 – 35128 Padova, Italy (phone +39 049 8753496 – Fax +39 049 9865390), email: info@cleup.it.
Rules for manuscript submission: https://www.sa-ijas.org/ojs/index.php/sa-ijas/about/submissions
Subscription: yearly €103.30; single copy €40.00; A.S.A. associates €60.00; supporting institutions: €350.00. Advertisement lower than 70%. Postal subscription Group IV, Milan. Forum licence n. 782/89. CLEUP SC on behalf of ASA, 7 March 2023.

Statistica Applicata – Italian Journal of Applied Statistics is associated to the following Italian and international journals:

QTQM – Quality Technology & Quantitative Management (http/web.it.nctu.ew/~qtqm/)

SINERGIE – Italian Journal of Management

Pubblished: April 2024

CLEUP sc
‘Coop. Libraria Editrice Università di Padova’
via G. Belzoni, 118/3 – Padova Italy
Phone +39 049 8753496 Fax +39 049 9865390
info@cleup.it – www.cleup.it – www.facebook.com/cleup

Summary

Statistica Applicata
ITALIAN JOURNAL OF APPLIED STATISTICS

Vol. 36, Number 1

doi.org/10.26398/IJAS.0036-001

HOW STUDENT CHARACTERISTICS AFFECT MOBILITY CHOICES AT THE UNIVERSITY LEVEL: INSIGHTS FROM TWO SURVEYS IN THE CAMPANIA REGION

Roberto Rondinelli, Valeria Policastro, Concetta Scolorato
Department of Political Sciences, Federico II University of Naples , Naples, Italy

Corresponding author
Roberto Rondinelli, roberto.rondinelli@unina.it

ORCID
Roberto Rondinelli: 0000-0002-8510-8162
Valeria Policastro: 0000-0003-1016-460X

Acknowledgements

This study was supported by the Ministero dell'Università e della Ricerca, Italy. PRIN 2017. From high school to job placement: Micro-data life course analysis of university student mobility and its impact on the Italian north-south divide, n. 14 2017HBTK5P- CUP B78D19000180001.

Funding

Italian Ministerial grant PRIN 2017 CUP: B78D19000180001. Recipient: Giancarlo Ragozini.

HOW STUDENT CHARACTERISTICS AFFECT MOBILITY CHOICES AT THE UNIVERSITY LEVEL: INSIGHTS FROM TWO SURVEYS IN THE CAMPANIA REGION

Abstract *In the context of intellectual migration, student mobility is an important and increasingly studied phenomenon. The Campania region represents a special case among the regions of Southern Italy, as it maintains a certain attractiveness for students. In this regard, a focus on the region's student mobility is proposed through two lines of research looking at the transition from high school to university and the continuation of studies after the bachelor's degree obtained at Federico II University of Naples . The aim is to investigate the effects of individual characteristics, socio-family background, geographic aspects, school/university experience and future prospects on mobility decisions. The logistic models for the probability to move show the importance of the family background in both cases. Geographical aspects seem to be important in the transition from high school to university. Future job prospects offered by Federico II University of Naples seem to be not enough to retain all the bachelor students. Finally, the perceived quality of the bachelor's study experience is not an important determinant in the decision to leave* this *University, but students who decided to move have high expectations for the new university.*

Keywords: *Student mobility, university experience, individual factors, survey.*

1. INTRODUCTION

In the context of migration, student mobility is a crucial social, economic, and decision-making issue. To analyse this phenomenon, many works use secondary data – often available at different levels of aggregation – leaving out individual student characteristics and the specific characteristics of certain territories. This strand of research is part of the broader debate on internal migration from Southern to Northern Italy (Genova et al., 2019), where student mobility - and first labour movements - are almost unidirectional. Motivations behind student mobility are both exogenous and endogenous and are often traced back to individual decision-making (Tosi et al., 2019), whereas movement from Southern to Northern Italy (Attanasio and Priulla, 2020), both to study and work, often has historical motivations. It is interesting to note that the push factor of labour migration is the motivation to improve one's economic condition, while intellectual migration starts from the possibility of making a basic economic investment for future improvement but it can also depend on the local labour market conditions (Dotti et al., 2013).

The Italian region of Campania is an interesting case study because it represents an area where local universities can retain students, yet there also exists incentives to migrate to other regions (Dal Bianco et al., 2010; Giambona et al., 2017). The region also has interesting socio-cultural characteristics worthy of investigation (Ragozini et al., 2016; Santelli et al., 2019). For these reasons, we chose the Campania region as the focus of our study.

To analyse the peculiarities of this phenomenon, following Bacci and Bertaccini (2021), we pursued two lines of research. Specifically, we constructed two ad hoc surveys for Campania's high school and bachelor's degree students with the aim of investigating the individual factors that determine two specific moments of transition when student migration can occur. Since these two mobilities refer to different periods in life, they show differences, but they may reveal some common motivations linked to the individual sphere, geographical aspects (Vittorietti et al., 2023), family, secondary school background (Usala et al., 2023), previous study experience, etc. From this perspective, it is possible to make some comparative reflections.

The logistic regression models performed on the data from the two surveys analyse i) the effect of student characteristics on the probability of being movers from the Campania region to attend university (high school students' survey) and ii) the effect of student characteristics and university experience on the probability to move from the Federico II University of Naples[1] (the most important university in Campania) to pursue a second bachelor's degree or master's degree after their bachelor's degree (university students' survey).

Considering the relevant literature, we hypothesize that the two students' migrations can have different determinants. In the case of high school students, we expect an important effect of geographic location and socio-economic background (related to the family), while for bachelor students, the field of study, quality of university and future job opportunities can be important determinants, even though the two migrations may have similar effects.

The paper is organized as follows. Section 2 gives a general overview of the theoretical background. Section 3 describe the characteristics of the two sample surveys. Section 4 sets out the methods used for the analysis. Finally, Section 5 shows the results.

[1] From now on we refer to it as University of Naples without specifying the name Federico II when it is not necessary.

2. THEORETICAL BACKGROUND

The movement of high school and university students in Italy has been described as a phenomenon associated with intellectual migration. It is seen as a distinct subset of Italy's longstanding internal migration pattern, which has traditionally followed a south-north trajectory (Attanasio and Priulla, 2020).

The phenomenon of intellectual migration has been explored by a large body of literature since the 1990s. These studies have tried to understand how human capital has been distributed within the different geographical areas of the country with consequences for the development of the territories involved (Affuso and Vecchione, 2012). The large proportion of young people from the South who decided to move to the north to start or continue their university studies is part of the more general movement from the south to the centre-north which occurs because they hope it will result in an overall improvement in their quality of life.

This type of migration from south to north, primarily determined by economic reasons, has always characterised Italy. Due to the persistent economic divide in the country, in fact, the south of Italy has maintained its role of subalternity (Bonifazi, 2015).

This migration trajectory characterising Italy was essentially a labour migration and became more substantial between the 1950s and 1970s, i.e. during the so-called economic boom. This phase saw workers from the agricultural areas of the south move to the industrialised regions and cities of Northern Italy in search of a job, which was almost always factory work (Pugliese, 2002). Already at that time, an elite part of the population began to move for educational reasons, giving rise to a new migratory phenomenon alongside the labour one. This migration trend is part of the broader phenomenon of intellectual migration. And although this trend decreased at the end of the 1970s, it began to increase again from the 1990s onwards (Attanasio et al., 2020). From the 2000s until now, the shape of this migration pattern repeatedly changed due to many events, such as reforms of the Italian university system, financial crises and the birth of several online universities (Minerva et al., 2022). The current migration is the product of the succession of events that occurred in Italy in this period.

As the literature on the topic indicates, both exogenous and endogenous factors influence this type of mobility. The exogenous factors include, in decreasing order of importance, the degree of accessibility to the educational facilities both in terms of the cost and quality of the transport system, the cultural environment, amount of leisure time, the cost of rent and the quality of life. Endogenous factors, again in decreasing order of importance, include the presence of faculty members in line with the chosen educational and professional pathway, the quality of the teaching provided and the quality of the services offered to students (Columbu et al., 2021b; Lombardi and Ghellini, 2019).

In addition to structural factors, which concern the characteristics of the territorial context and those of the university system (Columbu et al., 2021a), there are also the more strictly psychological/personal and social factors that determine the decision to be stayer or mover[2]. Within these factors, we can recognize the predisposition to change the living context in search of a more favourable one with the desire to improve quality of life, the so-called brain drain phenomenon (see Beine et al. (2008)). Another determinant is given by the family's socio-cultural context, as they try to ensure better educational and future employment opportunities for younger generations (Impicciatore and Tosi, 2019). However, to have more profitable long-term results and make young people more inclined towards an independent life, usually, the family needs to make an economic investment. Finally, other factors are attributable to the perception that students have of their overall university experience, their aspirations, and their willingness to travel (Biancardi and Bratti, 2019; Bratti and Verzillo, 2019; Ciriaci, 2014). It is also important to consider that intellectual migration, like any kind of migration, has effects on demographic and economic aspects both in the contexts of departure and arrival.

Recently, researchers have become interested in the specificities of the internal mobility of students in Italy as a process that contributes to reproducing social inequalities and widening the disparities in opportunities between students from the south and the north or between those who come from families with higher levels of education or lower. For example, according to studies by Impicciatore and Tosi (2019), it is evident that parental education represents one of the determinants of students' university choices and that the major cultural resources of families who consider the investment in education as the main opportunity to strengthen one's social status, favour the south-north migratory flow. Conversely, internal mobility in the southern regions is not associated with parental background.

Together with these determinants, even geographical (Vittorietti et al., 2023) and post-graduation aspects can affect the choice of students to be stayers or movers.

In this context, the case of the Campania region has interesting attributes

[2] Movers are the students who have enrolled in universities located in a region different from their residence and who take more than 90 minutes to reach the university, following the definition provided in Silvia et al. (2021), otherwise, they are defined as stayers (Attanasio and Enea, 2019).

compared to other southern regions (Ragozini et al., 2016; Santelli et al., 2019). Campania is, indeed, the most densely populated region in the south and is home to seven universities, two of which (Federico II University of Naples and University of Salerno) attract bachelor graduates from smaller universities. With these contextual characteristics, Campania seems to be the only one among the southern regions to counteract and in some cases reverse the mobility flows of students towards the centre-north by managing to be an attractive pole for university students both from Campania and from other regions. For these reasons, this work explores student mobility in Campania but without specifically investigating south-north movements.

3. SURVEYS

Two major flows of student mobility occur within regions and between regions. In this regard, the Campania region is an interesting case, as compared to other southern regions, it manages to retain a higher percentage of university students (see Santelli et al. 2022, the 2014-2015 cohort has a percentage of stayers equal to 85.8). Only Sardinia shows similar values due to its status as an island (in Santelli et al. 2022, the percentage of stayers is 81.3). For this reason, we implemented two surveys. The first analyses the decision of students to move from the Campania region in the transition from high school to university, and the second the propensity of University of Naples' bachelor students to enrol in another university after their degree. The survey of high school students was conducted in May 2022 until the end of school activities (mid-June), while the survey of bachelor students started in May 2022 and ended in July 2022. In this section, we describe how we sampled for the two surveys.

3.1. HIGH SCHOOL STUDENTS' SURVEY

We considered the national students register ('Anagrafe Nazionale Studenti', in Italian), and through the institute code, we identified 775 schools of the Campania region which include 28 thousand students. The statistical units of our analysis are the secondary school students; to sample them, we considered the schools that are involved in the national project to promote the enrolment in STEM university degree ('Piano Lauree Scientifiche', in Italian). To accomplish the aim of our analysis, we had to sample schools with students more likely to enrol at the university and with a propensity to move to another region. For this reason, we selected only the schools with more than 20 students enrolled at the university and more than 5 students who moved to another region. Finally, we obtained 112 schools with 25 thousand students, which is our reference population.

To be confident about the representativeness of the sample of students, the number of considered schools is calibrated in order to cover 10% of the total population (2500 students).

After the initial skimming, we applied a quota sampling of the schools by type (lyceum, vocational and technical) and province (Naples, Salerno, Caserta, Avellino and Benevento), where their combinations define each stratum, e.g. lyceum - Naples, vocational - Naples, etc. The relative joint distribution of enrolled students was used to calculate the number of students we had to sample from each stratum, while the number of schools that we needed to sample within each stratum to obtain the desired number of students was calculated by the ratio between the number of students to be sampled and the average number of enrolments. The resulting sample of schools did not include any vocational and technical institutes for the provinces of Avellino and Benevento. For this reason, we oversampled to have at least one school for each stratum. The final number of schools included in the sample is 43 (see Figure 1 for an overview of the sampling strategy and Section 7.1.1 in the Appendix for the quota computation).

Data were collected by means of a questionnaire sent by e-mail to the teachers, who administered it to the students. Given this procedure, the negative attitude of students to filling out the questionnaires, the submission of them at the end of the school year, and considering that we only collected data for students in the last two years of school (which number is also affected by school dropout), we obtained an acceptable response rate of 644 answers (26%).

3.2. BACHELOR STUDENTS' SURVEY

Due to privacy issues in accessing the entire list of bachelor students from the seven universities in the Campania region, the survey has been limited to the students studying at the University of Naples Federico II. From them, we selected only the students who had at least 160 ECTS at the time of the survey. The data for the survey were collected through a questionnaire sent by e-mail to the students. The final sample is composed of 1,048 students representing 11.64% of the entire population (9,003), which can be considered a good proportion.

4. METHOD

The purpose of this study is to test the effect of different individual characteristics on Campania students' mobility. Given the literature described above to

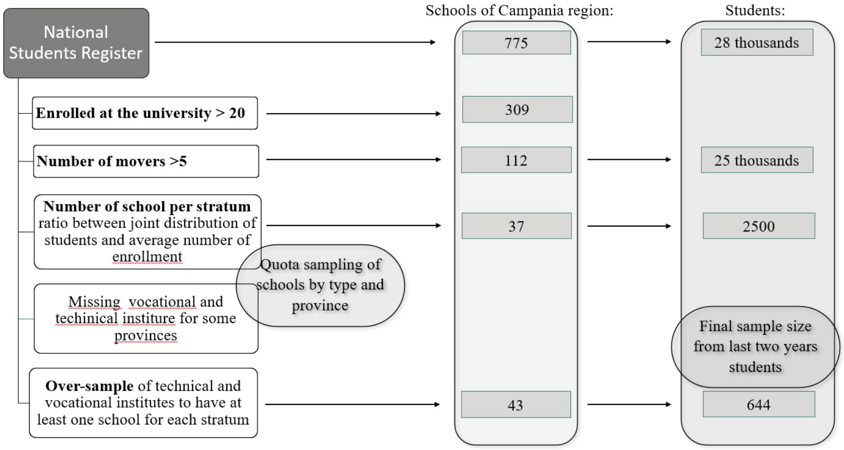

Schools of Campania region:
Students:
National Students Register
775
28 thousands
Enrolled at the university > 20
309
Number of movers >5
112
25 thousands
Number of school per stratum ratio between joint distribution of students and average number of enrollment
37
2500
Quota sampling of schools by type and province
Missing vocational and techinical institure for some provinces
Final sample size from last two years students
Over-sample of technical and vocational institutes to have at least one school for each stratum
43
644

Figure 1: **Sampling procedure**

explain this phenomenon, we identified four principal aspects: geographical distance, family, quality of the university and future perspectives.

For both cases (high school and bachelor students), we verified the interestingness of the variables describing the above four factors by making use of the information value (IV) (Shannon, 1948) which is designed mainly for variable selection in binary logistic regression. The computation of the information value for each variable j is the following:

$$IV_j = \int ln \frac{f(X_j|Y=1)}{f(X_j|Y=0)} |f(X_j|Y=1) - f(X_j|Y=0)| \; dx. \tag{1}$$

where the first part of the formulation can be defined by the weight of evidence (WOE), which is calculated for each individual i by using the probability as the function:

$$WOE_i = ln \frac{P(X_j|Y=1)}{P(X_j|Y=0)}. \tag{2}$$

where $P(X_j|Y=1)$ and $P(X_j|Y=0)$ are respectively the percentage of a generic characteristic given the individual belongs to the group of movers or stayers. For this reason, IV is essentially a weighted sum of all the individual WOE_i, where the weights are the absolute differences between the numerator and the denominator.

Generally, an IV value less than 0.1 refers to those predictors with a weak relationship to the movers/stayers' odds ratio; an IV value of between 0.1 and 0.3 identifies a medium relationship, IV between 0.3 and 0.5 denotes a strong relationship, while a IV value higher than 0.5 represents a suspicious relationship that needs to be verified.

The variables pertaining to the four investigated aspects and with the highest IV are included in the logistic regressions (Agresti, 2012) which we computed for the two cases (high school and bachelor students). In the case of high school students, the response variable is the probability of moving from the region to attend university; in the case of bachelor students, the response variable is the probability of moving from the University of Naples after finishing a bachelor's degree. The following equation identifies the model:

$$P(Y=1) = \frac{e^{\mathbf{X}\beta}}{1-e^{\mathbf{X}\beta}}. \tag{3}$$

where $\mathbf{X}$ is the matrix of the covariates we selected through IV , while β is the vector of parameters to be estimated.

5. ANALYSIS AND RESULTS

In this section, we show the results of our analysis for both case studies. Section 5.1 focuses on the empirical evidence for high school students, while Section 5.2 focuses on bachelor students. For both, we briefly discuss the composition of the sample, the variable selection via IV and the interpretation of the statistical models.

5.1. HIGH SCHOOL STUDENTS

To perform the analysis, we selected only students who intended to go to the university. To the question concerning the region of the university the students intended to enrol, they answered 'Campania' or 'Other region'. We excluded the students who were uncertain.

After this filtering, the final sample consisted of 389 students: 304 (78.1%) stayers and 85 (21.9%) movers. Approximately 66.3% were female, and 31.4% were male, while the remaining indicated 'Other' or 'Do not want to declare'. A majority of the students attended a lyceum (scientific 36.8%, classic 11.8%, and other 21.1%), 16.2% were in technical institute, but only 0.3% attended vocational

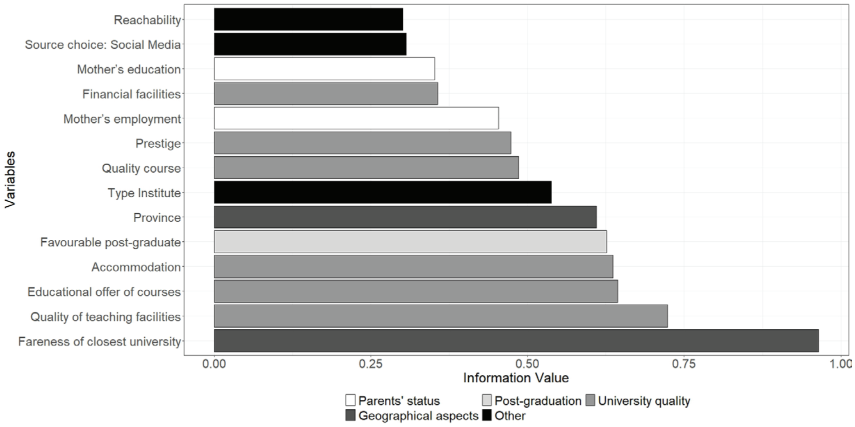


Figure 2: Variables with an *IV* value higher than 0.3

institutes. The rest of the students were from other institutes. Finally, 34.7% were in their fourth year of study, and 65.3% were in their fifth.

The variables[3] depicted in Figure 2 have the highest *IV* values (more than 0.3), including variables strongly (*IV* between 0.3 and 0.5) and suspiciously related (*IV* higher than 0.5) to the response variable. Since the characteristics identifying the four investigated dimensions have a very high *IV*, the logistic regression model is useful to detect whether they statistically affect the probability of students moving from their home region to attend university.

Table 5[4] shows the results for the seven different models. Together with the four dimensions of interest described in Section 4, we added 'gender' as a control variable. The first column (1) refers to the complete model (*Model 1*), while the second (2) describes the selection of the important variables from *Model 1* concerning the four dimensions. The remaining five columns concern respectively the models of geographical aspects (*Model 3*), parents' status (*Model 4*), university quality (*Model 5*), post-graduation (*Model 6*) and other variables of interest (*Model 7*). According to the Bayesian information criterion (*BIC*), the geographical aspect denoted by the variables included in the *Model 3* is the most important dimension to explain the phenomenon. Overall, *Model 2*, including at least one variable for each of the four dimensions, turns out to be the best model (with the lowest value of *BIC*).

Focusing on this model, the geographical aspect seems to be important for

[3] See the Appendix (Section 7.1.2) for their description.

[4] See the Appendix (Section 7.2) for the model estimation.

the decision to move towards another region. When the distance to the closest university in the region of the student increases, the probability of being a mover increases: for ‘more than 1 hour’, the probability to move is 0.84, while for ‘until 1 hour’ and ‘fewer than 30 minutes’, the probability to move is low (respectively 0.26 and 0.23), denoting a propensity to remain in the region. This is also evident from the ‘Province’ variable, which highlights the low propensity to be a mover (0.2) for students who live in the Naples province. However, not in all cases does living in this province ensure being close to the university, so this result can be linked to the perception of the student to be near the university. This is due to the importance of Naples city and to the number of universities based in it (4 of the 7 universities in the Campania region are in Naples). To conclude, when students live far away, or they perceive to live far from the closest university, they prefer to leave the region.

For what concerns the family effect, we found a statistically significant contribution of the mother’s employment status. When the mother is employed, the probability to be a mover is 0.82. This finding can be explained by two different reasons: the family can afford the economic investment aforementioned in Section 2 and the student may be more accustomed and inclined to an independent life. When only the family dimension is analysed in *Model 4*, the mother’s educational status becomes more important in explaining the response variable.

Generally speaking, the post-graduation perspective is not yet driving the decision of high school students to move out of their region, probably because they feel the work world is something very distant. However, this is not verified when this variable is included in the simple regression *Model 6*, which is also the one with the second highest *BIC*, denoting a spurious explanatory power.

The quality of the university is only important for the following variables: accommodation, quality of courses on offer and quality of teaching. For all three the probability to move is slightly higher than 0.5 (respectively 0.543, 0.541, 0.562). For *Model 5*, this denotes that aspects related to the university facilities and subject of the courses are a bit more important than economic and prestige factors (‘Financial support’, ‘Prestige’, and ‘Quality of course’).

The covariates of *Model 7* do not include any variables of the four dimensions, and, although we found some significant effects, this is the model with the highest *BIC*.

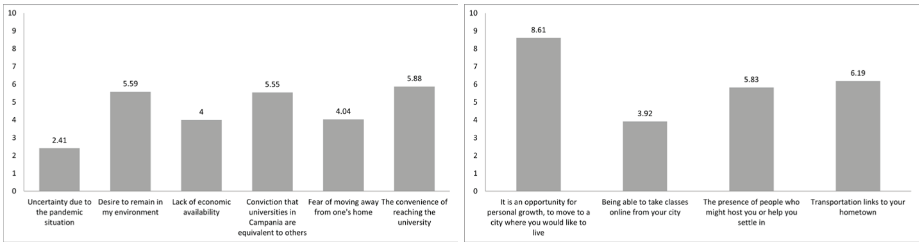
The last thing to underline is that both *Models 1* and *2* do not show any significant ‘Gender’ effect on the probability of being a mover from the Campania region to attend university.

To provide a more straightforward result, we constructed a typical profile for a student who has a high propensity to move outside the region. A student who lives outside the Naples province, who has a commute of over one hour to the nearest university, who has an employed mother and who believes that the quality of the facilities, accommodation and quality of education are important has a probability to move of 0.97.

To conclude this section, we will discuss some of the descriptive statistics that the models did not take into account. In the sample used for the construction of the models, the majority of students were enrolled in the fifth grade (65.4%). For the students who intended to move to a different region, this percentage decreased significantly (47.7%). This finding shows that the choice to stay in the region is influenced by contextual factors. The decision to move in this regard is more characteristic of fourth-year students. They will face the choice of the university to enrol in the year after, so their answers are driven by their aspirations.

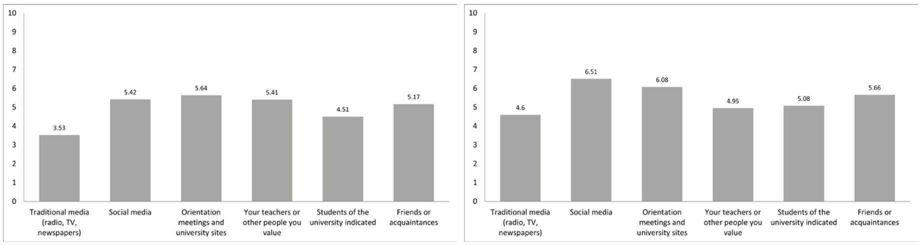
Comparing the personal motivations of students who choose to stay in the region with those who decide to leave, it is once again evident that the decision to stay is not significantly influenced by personal factors but by environmental ones.

Figure 3: Answers to the question: 'How much do the following factors influence your choice?' The left plot depicts the stayers, while the right plot depicts the movers.



Indeed, on a scale from 0 to 10, none of the modalities reported in the left panel of Figure 3 show a clear predominance in the decision, except to a slight extent for 'Desire to stay in my environment', 'Conviction that universities in Campania are equivalent to others' and 'Convenience of reaching the university' (the latter attributed to the importance of the geographic aspect). In other words, those who choose to stay, in part, seem to do so passively, partly being aware of the context they come from and in which Campanian universities are situated. Yet students who intend to go outside the region clearly recognize it as 'an opportunity for personal growth, to move to a city where you would like to live', but at the same time, they remain anchored to their roots ('Transportation links to your hometown').

Figure 4: Answers to the question: 'Can you indicate how important the following sources of information were in your decision to enroll in the university you indicated?' The left plot depicts the stayers, while the right plot depicts the movers.



These two different approaches to decision-making are also evident in the usage of information sources (Figure 4). Indeed, for all the sources considered, on average, students who decide to go outside the region tend to give them more consideration than the stayers.

5.2. BACHELOR STUDENTS

As in the previous case study, we selected only the students who intended to continue their university studies after their bachelor's degree (another bachelor's, master's or professional master's degree). To the question about where they would undertake future studies, they knew which university to enrol (whether University of Naples or another).

After this filtering, the final sample consisted of 469 bachelor students who had at least 160 ECTS: whose 332 (70.8%) were stayers, and 137 (29.2%) were movers. Approximately 55.3% were female, and 42.2% were male; the remaining indicated 'Other'. A majority of them were from the Campania region (94%). Before their bachelor's, they mainly studied at a lyceum (scientific 46.3%, classic 18.1%, and other 12.8%); 11.1% of them were from a technical institute, and only 2.3% attended vocational institutes. The rest were from other institutes. The most common areas studied for the bachelor's degree were humanistic disciplines (26.8%), engineering (25.5%) and economics and statistics (17%).

In contrast to Figure 2, the variables[5] depicted in Figure 5 include the ones

[5] See the Appendix (Section 7.3.1) for their description.

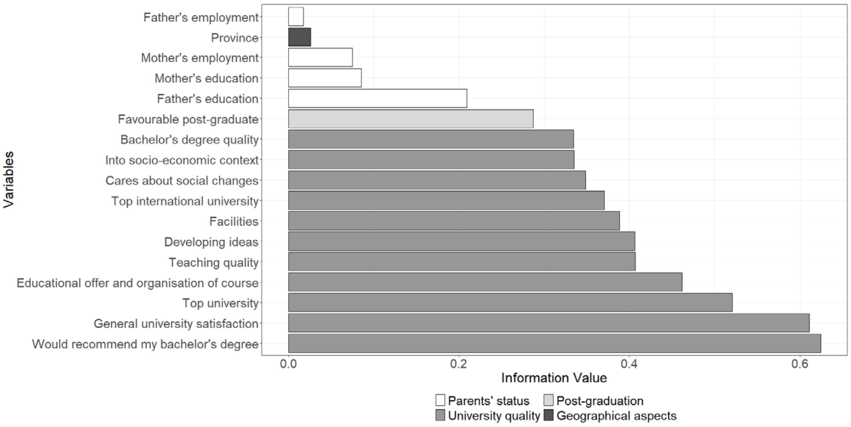


Figure 5: Variables with an *IV* value higher than 0.3 and other variables from dimensions of interest

with the highest *IV* values (more than 0.3) and other variables that are important for all four dimensions.

The six models[6] depicted in Table 6 investigate the determinants of outgoing mobility from University of Naples, using a similar approach to Section 5.1. Together with the four dimensions of interest described in Section 4, we added the ‘Gender’, ‘Abroad study periods’ and ‘Scientific area’ as control variables. The first column (*Model 1*) refers to the complete model, while the second (*Model 2*) describes the selection of the important variables from *Model 1* concerning the four dimensions. The remaining four columns regard respectively the models of geographical aspects (*Model 3*), parents’ status (*Model 4*), university quality (*Model 5*), and post-graduation (*Model 6*). According to the *BIC* value, *Model 2*, including at least one variable for each of the four dimensions, is clearly the best model. We should point out that the model assesses only the movement out of University of Naples and not the entire region, but we continue to refer to outgoing students as movers because University of Naples is the largest university in the south of Italy (as well as Campania) and also because only 10% of the outgoing students stay in the region.

The family variable has a large effect on moving: the probability of moving from University of Naples is 0.74 for students with a university-graduated father and 0.66 for students with an employed mother. Obviously, the higher the level

[6] See the Appendix (Section 7.3.3) for the model estimation.

of the father's education can be an important incentive for the student to have new experiences, as well as graduate from a top university outside of the region. Furthermore, the dynamics of the mother's employment status are comparable to the ones we just described in Section 5.1, but with a weaker effect visible in *Model 4* where the parameter is not significant. The perceived quality of University of Naples is based on the students' undergraduate experience. Both in the selected model and in the full one (*Model 1*), we did not notice any important effect of this dimension, except for 'Would recommend my bachelor's degree' (movers do not consider the University of Naples to be of poor quality, but overall, they do not recommend their bachelor's degree). Students who do not recommend their bachelor's degree are more likely to be movers from University of Naples: in *Model 2*, the probability related to 'More no than yes' is 0.88, while for 'Definitely not', it is 0.8. *Model 5* shows a slightly negative effect of considering the University of Naples as 'Top international university' on the probability of being a mover (0.44). The same effect is observed for the possibility of having 'a more favourable post-graduate work environment' when studying at University of Naples. In this case, the parameter is significant in *Model 1*, *Model 2* and *Model 6*, with the probability of moving from University of Naples being around 0.42. As for the models of Ta-ble 5, we did not find any gender effect, while we observed an important effect of studying abroad during bachelor's degree that, as expected, has a positive and sig-nificant impact on the probability of moving from the University of Naples (in *Model 2*, it is 0.86). Finally, a fundamental difference was found for 'Scientific area', where the effect of the reference category ('Socio-economic sciences') on the probability to move is positive as we found negative effects for all the other areas ('Health sciences', 'Humanities' and 'STEM'). This result can be related prob-ably to the fact that, in the economic area, University of Naples is not perceived at the same high level as other universities; on the contrary, University of Naples seems to keep students in the other scientific areas (Health sciences 0.126, Humanities 0.214, STEM 0.226). To conclude, the perceived quality of the University of Naples has a minor impact on the choice of students to move from it. In contrast, the father's educational and mother's employment status, the specific propensity of students to want new experiences, the field of study and future perspectives, are very important factors. Although the perceived quality of the specific charac-teristics of the University of Naples did not significantly affect the choice to move, the students who decided to enrol in another university to continue their studies would not recommend their bachelor's. Compared to high school students, the choice of bachelor students seems to be driven by more personal and qualitative reasons due mostly to their experiences and awareness rather than contextual motivation. In this regard, the motivations of bachelor students are more complex (the second best *BIC* is for the full model); indeed, the territorial aspect ('Province') disappears, and the mother's employment status decreases in effect.

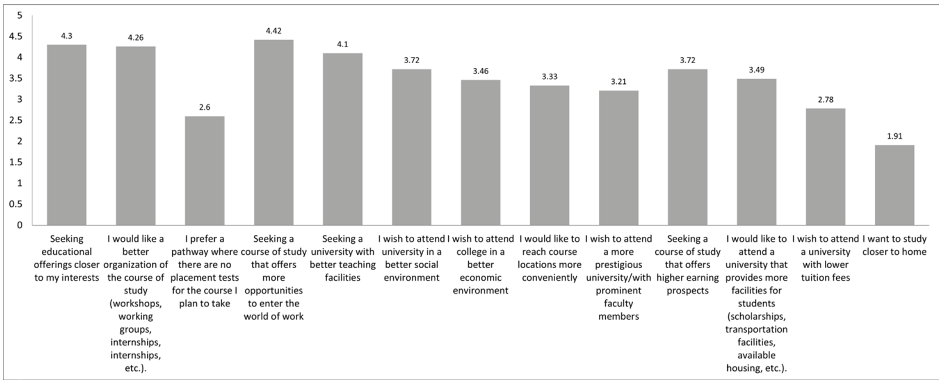
As for the high school students survey, we constructed a typical profile for a bachelor student who has a high probability of moving from the University of Naples to continue his studies. A student who has been abroad for a study period, who comes from a socio-economic scientific area, who has a highly educated father and who believes that University of Naples can not guarantee a favourable post-graduate work environment has a probability to move of almost 1 (0.99).

Comparing the entire sample and the bachelor students who decided to leave the University of Naples, the scarce importance of the geographical distance is con-firmed by the answers to the question: 'How long does it take you to get to Uni-versity of Naples from your home residence?'. Actually, we observed similar frequency distributions, that are respectively for the entire sample and the movers as follows: 'fewer than 30 minutes' (4.4% and 7.9%), 'between 30 and 59 min-utes' (39.2% and 38.1%), 'between 60 and 120 minutes' (37.6% and 34.9%), and 'more than 120 minutes' (18.8% and 19.1%).

At the same time, the question 'Would you have liked to study at a university located in a different city from your current one?' denotes a problem related to the city that may need further investigation. We observed a consistent difference between the whole sample and the movers from University of Naples: in the first case, the distribution shows a prevalence of 'No' of 69.4%, in contrast, in the second case, the distribution shows a prevalence of 'Yes' of 52.6%.

Finally, a question specifically asked to students who intend to move from University of Naples, requires special attention: 'Express your degree of agreement/disagreement (from 0 to 5) with respect to the following statements regarding your choice to continue your studies'. Since the choice of undergraduate students seems less related to the quality of University of Naples and to be more complex than that of high school students, this question may be useful at a descriptive level to detect preferences and desired characteristics for the new destination (Figure 6). Not surprisingly among the five most agreed statements are 'Seeking a course of study that offers more opportunities to enter the world of work' (the most agreed, 4.42 on average) and 'Seeking a course of study that offers higher earning prospects' (3.72), which confirm the importance of post-graduate work possibilities. At the same time, although we observed from the models that the quality of University of Naples has not significantly affected the probability of

Figure 6: Answers to the question ‘Express your degree of agreement/disagreement (from 0 to 5) with respect to the following statements regarding your choice of further education.’ The plot depicts the opinions of movers from University of Naples.



moving from it, Figure 6 shows the high expectations of students with respect to the new university. Among the five most agreed statements were ‘Seeking educational offerings closer to my interests’ (4.3), ‘I would like a better organization of the course of study’ and ‘Seeking a university with better teaching facilities’.

6. DISCUSSION AND CONCLUSIONS

The two surveys discussed in this study represent the first effort to collect data at the individual level and study the phenomenon of students’ mobility for the peculiar case of the Italian region of Campania using microdata.

Based on this initial empirical evidence, the mobility of high school and bachelor students respectively from Campania and Federico II University of Naples is driven by the motivations we indicated in Section 1 as a research hypothesis, namely that the Campania students’ mobility seems associated with the four dimensions we wanted to test. Understandably, family and geographical distance are characteristics that drive the transition from high school to university, while the choice of bachelor students seems clearer and depends very much on their previous experiences, scientific area and future perspectives.

Specifically, in the first case, high school students move from their home region for three main reasons: the quality of courses and structures, the economic possibility of the family (employed mother works like a proxy of that) and if they live far away from the closest university of Campania. In the second case, students move from the University of Naples Federico II when, based on their experience, they would not recommend their bachelor's degree, so they prefer to change university to increase the probability of being employed in the future. Usually, they have spent periods studying abroad during their first degree.

Overall, high school and bachelor students showed different behaviours in the choice to move for future studies. The mobility choices after the bachelor's degree is an additional migration with respect to the one already acted after the high school. Their motivations are then different and are driven by the university experience. While the decision-making process of high school students seems to be directed by intrinsic needs and conditions (including economic ones), the case of bachelor students, as above described, is even more complex and depends on different evaluations strictly related to the improvement of study conditions and their experiences. As evidence of this, we have seen that the choice of high school students is mostly related to the context in which they live, namely, their family and the distance to the closest university (more generally, geographical aspects), whereas, the decisions of University of Naples bachelor's students are more mature and conscious due to their prior experience and are, therefore, more complex.

Considering the movers, from the response to the specific question about the university they will attend in the future, they have a propensity to move to the northern part of the country (Rome and above). For this reason, we can also comment on the results in terms of south-north migration. In this regard, the results we have found are expected and in line with those of Santelli et al. (2022, 2019). In summary, we can highlight the following results: socioeconomic disparities (primarily influenced by family circumstances) influence the choice to move after high school; there is a migration trend among high school students that is somewhat driven by their geographical location; there is an anticipatory migration of bachelor students before entering the workforce, consistently following the south-north direction; finally, probably due to employment opportunities, students in the socio-economic sciences field are more inclined to emigrate.

In comparison to existing literature, using microdata, we can add that the quality of Federico II University of Naples and the courses offered is perceived as acceptable; indeed, it is not a significant factor in migration decisions. However, students expect very high standards from other universities, especially concerning employment prospects. These prospects seem to be influenced by factors external to the university itself and are more closely tied to the market context in which the university is situated.

In conclusion, as usually happens for surveys on specific topics, this work has its limitations. Questionnaire response rates could be improved by covering the limitation in collecting data, as described in Section 3.1; this could facilitate the estimation and interpretation of results. Second-level variables could be considered, and their potential effects on student migration could be tested. The analysis of south-north movements requires further exploration in the future by collecting more specific data on the migration destination.

References

Affuso, S. and Vecchione, G. (2012). *Migrazioni intellettuali e Mezzogiorno d'Italia: il caso della Scuola di alta formazione IPE.* McGraw-Hill.

Agresti, A. (2012). *Categorical data analysis*, vol. 792. John Wiley & Sons.

Attanasio, M. and Enea, M. (2019). La mobilità degli studenti universitari nell'ultimo decennio in italia. In *"Rapporto sulla popolazione. L'istruzione in Italia"*, 43–58. il Mulino.

Attanasio, M. and Priulla, A. (2020). Chi rimane e chi se ne va? un'analisi statistica della mobilità universitaria dal mezzogiorno d'Italia. In *"Verso Nord. Le nuove e vecchie rotte delle migrazioni universitarie"*. IT.

Attanasio, M., Ragozini, G., Porcu, M., and Giambalvo, O. (2020). Verso nord: Le nuove e vecchie rotte delle migrazioni universitarie. In *Verso Nord*, 1–208.

Bacci, S. and Bertaccini, B. (2021). Assessment of the university reputation through the analysis of the student mobility. In *Social Indicators Research*, 156: 363–388.

Beine, M., Docquier, F., and Rapoport, H. (2008). Brain drain and human capital formation in developing countries: winners and losers. In *The Economic Journal*, 118 (528): 631–652.

Biancardi, D. and Bratti, M. (2019). The effect of introducing a research evaluation exercise on student enrolment: Evidence from Italy. In *Economics of Education Review*, 69: 73–93.

Bonifazi, C. (2015). Le migrazioni tra sud e centro-nord: persistenze e novità. In *La nuova migrazione italiana. Cause, mete e figure sociali*, 57–69.

Bratti, M. and Verzillo, S. (2019). The 'gravity' of quality: research quality and the attractiveness of universities in italy. In *Regional Studies*, 53 (10): 1385–1396.

Ciriaci, D. (2014). Does university quality influence the interregional mobility of students and graduates? the case of italy. In *Regional Studies*, 48 (10): 1592–1608.

Columbu, S., Porcu, M., Primerano, I., Sulis, I., and Vitale, M.P. (2021a). Geography of Italian student mobility: A network analysis approach. In *Socio-Economic Planning Sciences*, 73: 100918.

Columbu, S., Porcu, M., and Sulis, I. (2021b). University choice and the attractiveness of the study area: insights on the differences amongst degree programmes in Italy based on generalised mixed-effect models. In *Socio-Economic Planning Sciences*, 74: 100926.

Dal Bianco, A., Ricciari, V., and Spairani, A. (2010). La mobilità degli studenti in Italia: un'analisi empirica. In *La mobilità degli studenti in Italia*, 1000–1021.

Dotti, N.F., Fratesi, U., Lenzi, C., and Percoco, M. (2013). Local labour markets and the interregional mobility of Italian university students. In *Spatial Economic Analysis*, 8 (4): 443–468.

Genova, V.G., Tumminello, M., Enea, M., Aiello, F., and Attanasio, M. (2019). Student mobility in higher education: Sicilian outflow network and chain migrations. In *Electronic Journal of Applied Statistical Analysis*, 12 (4): 774–800.

Giambona, F., Porcu, M., and Sulis, I. (2017). Students mobility: Assessing the determinants of attractiveness across competing territorial areas. In *Social Indicators Research*, 133: 1105–1132.

Impicciatore, R. and Tosi, F. (2019). Student mobility in Italy: The increasing role of family background during the expansion of higher education supply. In *Research in Social Stratification and Mobility*, 62: 100409.

Lombardi, G. and Ghellini, G. (2019). The effect of grading policies on Italian universities' attractiveness: A conditional multinomial logit approach. In *Electronic Journal of Applied Statistical Analysis*, 12 (04): 801–825.

Minerva, T., De Santis, A., Bellini, C., and Sannicandro, K. (2022). A time series analysis of students enrolled in Italian universities from 2000 to 2021. In *Italian Journal of Educational Research*, (29): 009–022.

Pugliese, E. (2002). *L'Italia tra migrazioni internazionali e migrazioni interne*, vol. 1. il Mulino.

Ragozini, G., Scolorato, C., and Santelli, F. (2016). Le determinanti della mobilità degli studenti universitari campani. In *Il sistema universitario campano tra miti e realtà. Aspetti metodologici, analisi e risultati*, 230–241. Franco Angeli.

Santelli, F., Ragozini, G., and Vitale, M.P. (2022). Assessing the effects of local contexts on the mobility choices of university students in Campania region in Italy. In *Genus*, 78 (1): 5.

Santelli, F., Scolorato, C., and Ragozini, G. (2019). On the determinants of student mobility in an interregional perspective: A focus on Campania region. In *Statistica Applicata-Italian Journal of Applied Statistics*, (1): 119–142.

Shannon, C.E. (1948). A mathematical theory of communication. In *The Bell System Technical Journal*, 27 (3): 379–423.

Silvia, C., Mariano, P., Ilaria, P., Isabella, S., and Maria Prosperina, V. (2021). Analysing the determinants of Italian university student mobility pathways. In *Genus*, 77: 1–20.

Tosi, F., Impicciatore, R., and Rettaroli, R. (2019). Individual skills and student mobility in Italy: A regional perspective. In *Regional Studies*, 53 (8): 1099–1111.

Usala, C., Porcu, M., and Sulis, I. (2023). The high school effect on students' mobility choices. In *Statistical Methods & Applications*, 1–35.

Vittorietti, M., Giambalvo, O., Genova, V.G., and Aiello, F. (2023). A new measure for the attitude to mobility of Italian students and graduates: A topological data analysis approach. In *Statistical Methods & Applications*, 32 (2): 509–543.

7. APPENDIX

7.1. HIGH SCHOOL STUDENTS: QUOTA SAMPLING, LIST OF VARIABLES AND MODELS

Following the computation of quota for the sampling strategy (Section 7.1.1), the list of the variables (Section 7.1.2) used in the logistic models, their distributions (Section 7.1.3), and the table with the estimated logistic models (Section 7.2).

7.1.1. QUOTA COMPUTATION

Table 1: Relative distribution of high school students, considering only the schools with more than 20 students enrolled at the university and more than 5 students that moved to another region (25000 students).

	Lyceum	Vocational	Technical	Total
Avellino	0.053	0.000	0.006	0.059
Benevento	0.035	0.001	0.005	0.041
Caserta	0.139	0.002	0.023	0.164
Napoli	0.455	0.004	0.078	0.537
Salerno	0.174	0.003	0.022	0.199
Total	0.855	0.010	0.134	1

Table 2: Number of students to sample in each stratum to reach a total sample of 2500.

	Lyceum	Vocational	Technical	Total
Avellino	134	0	15	149
Benevento	87	2	14	103
Caserta	347	5	58	410
Napoli	1137	11	195	1342
Salerno	434	8	55	497
Total	2138	26	336	2500

Table 3: Average number of enrolled students in one institute per stratum

	Lyceum	Vocational	Technical	Total
Avellino	67.6	0.0	37.8	62.6
Benevento	73.6	22.0	34.8	61.4
Caserta	121.0	27.0	44.9	94.3
Napoli	100.9	26.8	43.7	83.3
Salerno	95.4	38.0	42.9	82.3
Total	97.9	28.8	43.0	81.8

Table 4: Quota sampling of schools. The ratio between the number of students (Table 2) and the average number of students per stratum (Table 3).

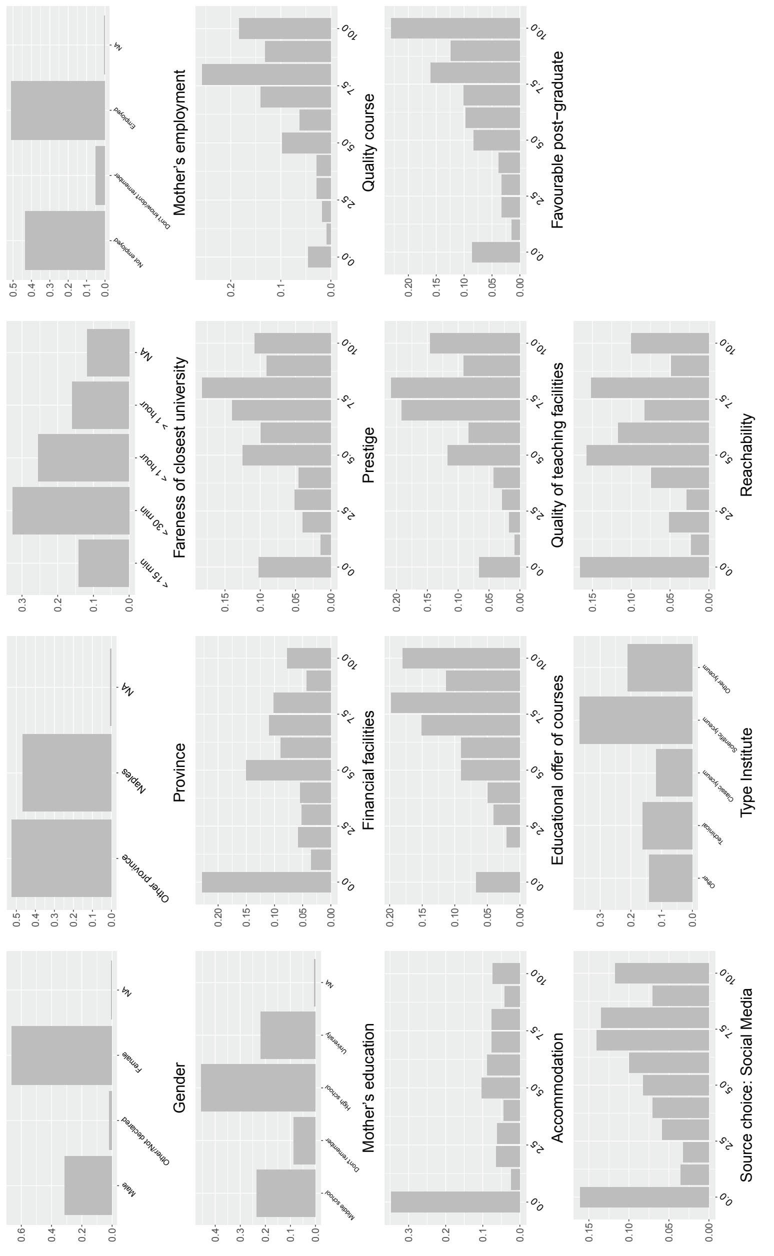
	Lyceum	Vocational	Technical	Total
Avellino	2	0	**1**	3
Benevento	1	**1**	**1**	3
Caserta	5	**1**	1	7
Napoli	17	**1**	3	21
Salerno	6	**1**	1	8
Total	31	**4**	7	42

7.1.2. LIST OF VARIABLES

The following is the list of variables depicted in Figure 2 and that are used in the logistic regression models of Table 5, where we include the additional control variable ‘Gender’.

- Gender (**Gender**)
- Province of residence (**Province**)
- How far is the university in your region most easily accessible from your home? (**Distance of closest university**)
- Mother’s employment status (**Mother’s employment**)
- Mother’s educational qualification (**Mother’s education**)
- Prestige of the university (**Prestige**)
- Which factors influenced your choice of the university you intend to enrol in? (consider that 0=not at all and 10=very much)
  - The financial support (scholarships, student accommodation, etc.) (**Financial support**)
  - The quality of the degree course you will choose (**Quality course**)
  - The presence of available university accommodation (**Accommodation**)
  - The educational offer of available courses (**Educational offer of courses**)
  - The quality of the locations and teaching facilities (**Quality of teaching facilities**)
  - A more favourable post-graduate work environment (**Favourable post-graduate**)
  - The convenience of reaching the university (**Reachability**)
- Can you indicate how important the following sources of information were for your decision to enrol at the university you indicated? (consider that 0=not at all and 10=very much): Social media (**Source choice: Social media**)
- Attended Institute (**Institute type**)

7.1.3. DESCRIPTIVE PLOTS OF THE VARIABLES



7.2. LOGISTIC REGRESSION MODELS OF THE PROBABILITY TO MOVE TOWARD ANOTHER REGION TO ATTEND UNIVERSITY

Table 5: Logistic regression models of the probability of moving toward another region to attend university.

	Dependent variable:						
	Mover from Campania Region						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Constant	-6.715^{***}	-5.202^{***}	-0.671^{**}	-2.692^{***}	-4.955^{***}	-3.144^{***}	-2.104^{***}
	(1.485)	(1.039)	(0.315)	(0.400)	(0.787)	(0.491)	(0.525)
Gender	0.664	1.531					
Other/Not declared	(1.793)	(1.736)					
Gender	0.184	−0.070					
Female	(0.574)	(0.456)					
Province	-0.924^{*}	-1.389^{***}	-1.683^{***}				
Naples	(0.545)	(0.434)	(0.367)				
Distance of closest university	−0.895	-1.190^{**}	-0.945^{**}				
<30 min	(0.635)	(0.546)	(0.455)				
Distance of closest university	-1.076^{*}	-1.066^{*}	-0.961^{**}				
<1 hour	(0.653)	(0.592)	(0.479)				
Distance of closest university	1.614^{**}	1.647^{***}	1.637^{***}				
>1 hour	(0.642)	(0.552)	(0.436)				
Mother’s employment	1.418	0.981		−0.327			
Don’t know/don’t remember	(1.169)	(1.073)		(0.812)			
Mother’s employment	1.270^{**}	1.501^{***}		1.047^{***}			
Employed	(0.514)	(0.451)		(0.316)			
Mother’s education	0.962			0.795			
Don’t remember	(0.930)			(0.609)			
Mother’s education	0.995			0.796^{*}			
High school	(0.697)			(0.425)			
Mother’s education	1.052			1.263^{***}			
University	(0.758)			(0.464)			
Prestige	0.046	0.011			0.047		
	(0.099)	(0.096)			(0.073)		
Financial support	0.043	−0.026			−0.055		
	(0.088)	(0.077)			(0.058)		
Quality course	−0.158	−0.098			−0.058		

Table 5: Logistic regression models of the probability of moving toward another region to attend university.

	Dependent variable:						
	Mover from Campania Region						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	(0.135)	(0.117)			(0.096)		
Accommodation	0.226*** (0.079)	0.171** (0.069)			0.165*** (0.054)		
Educational offer of courses	0.153 (0.115)	0.165* (0.097)			0.224*** (0.081)		
Quality of teaching facilities	0.299** (0.135)	0.250* (0.127)			0.221** (0.100)		
Favourable post-graduate	0.082 (0.086)	0.087 (0.079)				0.252*** (0.059)	
Reachability	−0.297*** (0.077)						−0.138*** (0.045)
Source choice: Social Media	0.102 (0.075)						0.181*** (0.049)
Institute type ***Technical***	−0.057 (1.089)						−1.683** (0.829)
Institute type ***Classic lyceum***	1.265 (0.927)						1.576*** (0.534)
Institute type ***Scientific lyceum***	0.741 (0.834)						0.844* (0.454)
Institute type ***Other lyceum***	0.575 (0.807)						0.006 (0.509)
Observations	314	318	341	386	334	338	340
Log Likelihood	−80.496	−93.006	−129.499	−184.177	−144.961	−164.705	−151.978
Akaike Inf. Crit.	210.991	218.013	268.999	380.355	303.922	333.410	317.955
Bayesian Inf. Crit.	304.726	278.2054	288.1584	404.0899	330.5996	341.0564	344.7579

Note: *p<0.1; **p<0.05; ***p<0.01

Model (1)	Complete model
Model (2)	Selection of important variables from complete model
Model (3)	Geographical aspects mode
Model (4)	Parents’ status mode
Model (5)	University quality model
Model (6)	Post-graduation model
Model (7)	Other varialbes model

Table 5: Logistic regression models of the probability of moving toward another region to attend university.

	Dependent variable:						
	Mover from Campania Region						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Reference categories:	Gender (***Male***) Province (***Other province***) Distance of closest university (***< 15 min***) Mother's employment (***Not employed***) Mother's education (***Middle school***) Institute type (***Other***)						

7.3. BACHELOR STUDENTS: LIST OF VARIABLES AND MODELS

7.3.1. LIST OF VARIABLES

The following is the list of variables depicted in Figure 5 and that are used in the logistic regression models of Table 6, where we include the additional control variables 'Gender', 'Abroad study periods' and 'Scientific area'.

- Gender (**Gender**)
- Did you study abroad during your studies? (**Abroad study periods**)
- What is the scientific area of the course you are attending at University of Naples Federico II? (**Scientific area**)
- Province of residence (**Province**)
- Father's educational qualification (**Father's education**)
- Mother's educational qualification (**Mother's education**)
- Father's employment status (**Father's employment**)
- Mother's employment status (**Mother's employment**)
- How satisfied are you with the following features of University of Naples Federico II?
  - The educational offer and organisation of course (**Educational offer and organisation of course**)

  - The quality of teaching (**Teaching quality**)
  - About the services offered (**Facilities**)

- How much did the following factors influence your choice of being enrolled at Federico II University of Naples? (consider that 0=not at all and 5=very much): The quality of the degree course (**Bachelor's degree quality**)

- Thinking about Federico II University of Naples , how much do you agree with the following statements?

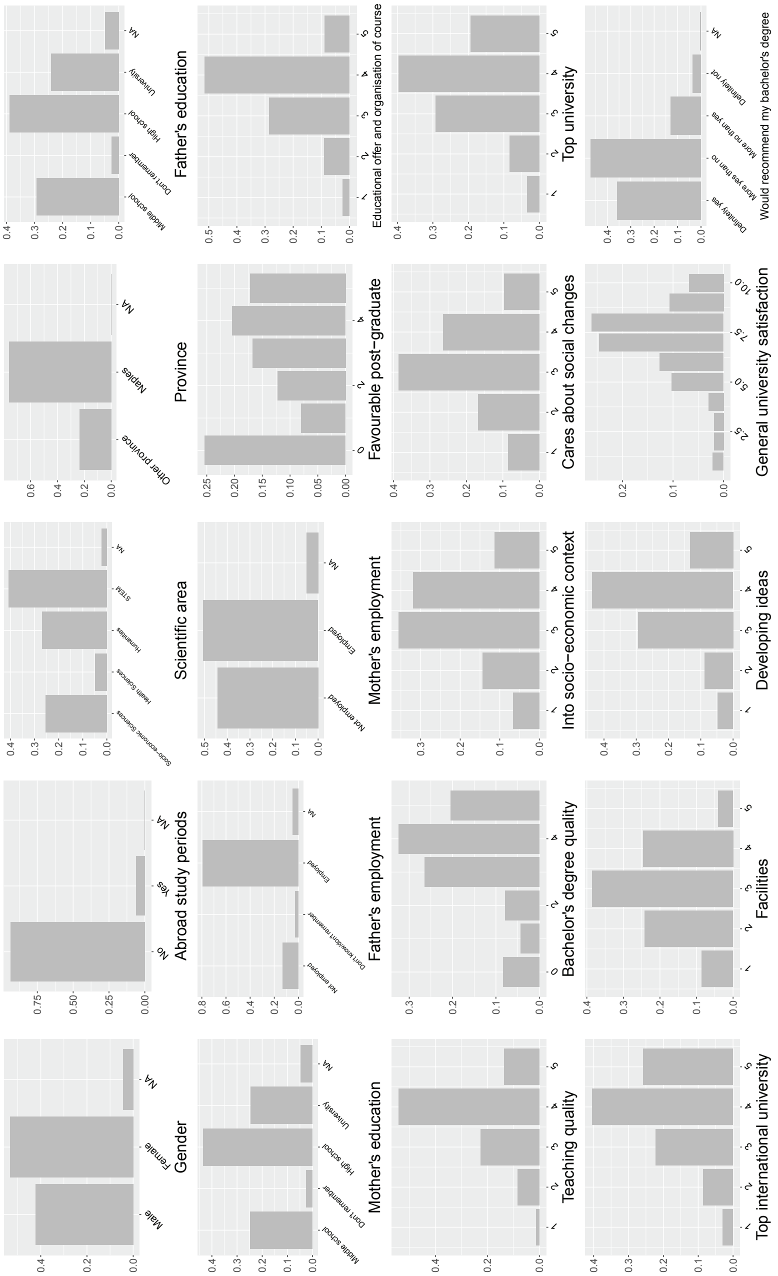
  - It is inside a socio-economic context. (**Into socio-economic context**)
  - It is attentive to changes in society (**Cares about social changes**)
  - It is a first-class university (**Top university**)
  - It is a university whose prestige is recognised internationally (**Top international university**)
  - It is a place for the development of new ideas (**Developing ideas**)

- Express your overall satisfaction with Federico II University of Naples by assigning a number from 1 (not at all satisfied) to 10 (fully satisfied) (**General university satisfaction**)

- Based on your experience, would you recommend your degree course at Federico II University of Naples ? (**Would recommend my bachelor's degree**).

- How much did the following factors influence your choice of being enrolled at Federico II University of Naples ? (consider that 0=not at all and 5=very much): A more favourable post-graduate work environment (**Favourable post-graduate**)

7.3.2. DESCRIPTIVE PLOTS OF THE VARIABLES



7.3.3. LOGISTIC REGRESSION MODELS OF PROBABILITY TO MOVE FROM UNIVERSITY OF NAPLES FEDERICO II AFTER BACHELOR DEGREE

Table 6: Logistic regression models of the probability of moving from Federico II University of Naples after bachelor's degree.

	Dependent variable:					
	Mover from University of Naples Federico II					
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.378 (1.237)	−0.116 (1.047)	−1.185*** (0.224)	−1.669*** (0.380)	1.284 (0.921)	−0.241 (0.162)
Gender ***Female***	0.287 (0.290)	0.243 (0.274)				
Abroad study periods ***Yes***	1.841*** (0.543)	1.779*** (0.520)				
Scientific area ***Health Sciences***	−1.872** (0.753)	−1.934*** (0.716)				
Scientific area ***Humanities***	−1.275*** (0.388)	−1.300*** (0.368)				
Scientific area ***STEM***	−1.152*** (0.333)	−1.230*** (0.325)				
Province ***Naples***	0.306 (0.353)	0.348 (0.335)	0.389 (0.252)			
Father's education ***Don't remember***	1.394 (0.980)	0.868 (0.840)		1.369* (0.798)		
Father's education ***High school***	0.342 (0.358)	0.338 (0.337)		0.455 (0.292)		
Father's education ***University***	1.275*** (0.431)	1.058*** (0.366)		1.286*** (0.348)		
Mother's education ***Don't remember***	−1.321 (1.407)			−1.869 (1.201)		
Mother's education ***High school***	−0.255 (0.365)			−0.254 (0.293)		
Mother's education ***University***	−0.580 (0.463)			−0.433 (0.377)		
Father's employment ***Don't know/don't remember***	0.805 (0.985)	0.607 (0.937)		0.589 (0.748)		

Table 6: Logistic regression models of the probability of moving from Federico II University of Naples after bachelor's degree.

	Dependent variable:					
	Mover from University of Naples Federico II					
	(1)	(2)	(3)	(4)	(5)	(6)
Father's employment	0.743*	0.518		0.259		
Employed	(0.431)	(0.402)		(0.337)		
Mother's employment	0.817**	0.653**		0.383		
Employed	(0.326)	(0.294)		(0.244)		
Educational offer and organisation	−0.051	−0.194			−0.185	
of course	(0.259)	(0.192)			(0.220)	
Teaching quality	0.036				0.159	
	(0.222)				(0.186)	
Facilities	−0.137				−0.223	
	(0.190)				(0.155)	
Bachelor's degree quality	−0.117				−0.133	
	(0.119)				(0.092)	
Into socio-economic context	0.135				−0.008	
	(0.185)				(0.149)	
Cares about social changes	0.177				0.014	
	(0.189)				(0.156)	
Top university	−0.021				−0.109	
	(0.215)				(0.181)	
Top international university	−0.151	−0.203			−0.232*	
	(0.171)	(0.139)			(0.141)	
Developing ideas	−0.194				−0.062	
	(0.206)				(0.165)	
General university satisfaction	−0.119				0.007	
	(0.130)				(0.109)	
Would recommend my bachelor's	0.075	0.226			0.145	
degree ***More yes than no***	(0.360)	(0.329)			(0.300)	
Would recommend my bachelor's	1.800***	2.033***			1.461***	
degree ***More no than yes***	(0.552)	(0.498)			(0.443)	
Would recommend my bachelor's	0.671	1.421*			0.887	
degree ***Definitely not***	(0.932)	(0.766)			(0.747)	
Favourable post-graduate	−0.306***	−0.331***				−0.283***

Table 6: Logistic regression models of the probability of moving from Federico II University of Naples after bachelor's degree.

	Dependent variable:					
	Mover from University of Naples Federico II					
	(1)	(2)	(3)	(4)	(5)	(6)
	(0.086)	(0.079)				(0.058)
Observations	423	426	468	444	460	465
Log Likelihood	−182.157	−187.039	−281.704	−253.936	−238.572	−267.703
Akaike Inf. Crit.	424.313	412.078	567.408	527.872	505.145	539.407
Bayesian Inf. Crit.	545.7343	489.1121	575.7054	568.8303	562.9822	547.691
Note:						*p<0.1; **p<0.05; ***p<0.01
Model (1)	Complete model					
Model (2)	Selection of important variables from complete model					
Model (3)	Geographical aspects mode					
Model (4)	Parents' status mode					
Model (5)	University quality model					
Model (6)	Post-graduation model					
Reference categories:	Gender (***Male***)					
	Abroad study periods (***No***)					
	Scientific area (***Socio-economic Sciences***)					
	Province (***Other province***)					
	Father's education (***Middle school***)					
	Mother's education (***Middle school***)					
	Father's employment (***Not employed***)					
	Mother's employment (***Not employed***)					
	Would recommend my bachelor's degree (***Definitely yes***)					

doi.org/10.26398/IJAS.0036-002

MULTIVARIATE TECHNIQUES FOR ANALYZING AND PRESENTING OFFICIAL STATISTICS INDICATORS

Ron S. Kenett
The KPA Group and the Samuel Neaman Institute, Technion, Israel

Abstract: *The production of official statistics is experiencing significant challenges including the handling of massive data sets, the application of computer intensive methods and the integration of data from different sources. Official statistics indicators provide a multivariate perspective, both in form and in content. This perspective requires an implementation of multivariate techniques for data analysis and presentation of findings. The information quality framework is a methodological approach that has been applied to many domain areas including the production of official statistics. Bayesian networks are graphical models that permit decision makers to evaluate alternative scenarios using official statistics. The article presents the information quality framework and discusses a Bayesian network application to Eurostat data. It begins with a background on current official statistics evolutionary changes and concludes with a discussion section that maps some of the challenges of official statistics.*

Keywords: *Official statistics, information quality, indicators, multivariate analysis, Bayesian networks.*

1. Background

Official statistics need to be used to be useful. Quoting from Forbes and Brown, 2012: "An issue that can lead to misconception is that many of the concepts used in official statistics often have specific meanings which are based on, but not identical to, their everyday usage meaning. All staff producing statistics must understand that … their work translate the real world into models that interpret reality and make it measurable for statistical purposes. The first step … is to define the issue or question(s) that statistical information is needed to inform. That is, to define the objectives for the framework, and then work through those to create its structure and definitions. An important element … is understanding the relationship between the issues and questions to be informed and the definitions themselves." The challenge posed by this quote is a transformation of official statistics from a producer of numbers to a generator of information. This perspective significantly expands the traditional role of official statistics. To fulfill this role, several education programs provide qualified training to producers and users of official statistics. For example, the mission of the European Master in Official Statistics (EMOS) is to enhance the ability of students to understand and analyze European official data at different levels: quality, production process, dissemination, and analysis in a national, European and international context, see EMOS, 2021. As mentioned, a key task of modern official statistics is the generation of information. A general framework for designing and assessing information quality is proposed in Kenett and Shmueli, 2014, 2016a. Kenett and Shmueli, 2016b, provide an example where administrative data, collected for operational purposes, is combined with survey-based data to enhance the information quality of official statistics. The information quality framework consists of eight dimensions and requires an explicit determination of the goals of the analysis and a clarification of the available data, the methods of analysis used and the related utility function. Of impact on information quality of official statistics is the Generic Statistical Business Process Model (GSBPM) which describes statistical production in a general and process-oriented way. It is used both within and between national statistical offices as a common basis for work with statistics production to ensure quality, efficiency, standardization, and process-orientation and is used for all types of surveys, see GSBPM, 2021.

In general, modern statistics, machine learning, data science and in general, data analytics, are having a ubiquitous impact on industry, governments, business, and services (Kenett et al, 2022, 2023a). For an example of how these impact official statistics see Barcaroli, 2017, and Bhandari et al, 2022. For a general treatment of data science and the role of data scientists see Kenett and Redman, 2019. The next section is a high-level introduction to indicators, such as those published by national bureaus of statistics of national statistics organizations.

2. Indicators

Indicators come from the Latin word "indicator" that means "who or what indicates". They represent direct and indirect data driven measures. "… an indicator is not simple crude statistical information but represents a measure organically connected to a conceptual model aimed at describing different aspects of reality" (Maggino, 2018a). The construction of indicators involves the process of synthesizing indicators through aggregative–compensative and non-aggregative approaches. These methods apply a synthesis of units with reference to one or more indicators aiming at aggregating individual values at a microlevel. This synthesis allows a comparison of macro units with references of interest. In addition to these "numerical" approaches it is common to use graphical instruments such as dashboards (Maggino, 2018b). In any case, indicators need to be validated. This is sometimes called construct validation. We expand on this when, in the net section, we introduce the information quality dimensions.

An example of indicators is provided by the United Nations Sustainable Development Goals (SDG). This initiative aims at reaching 17 goals that are defined in a list of 169 SDG Targets. Progress towards these Targets is tracked by 232 Indicators. Official statistics indicators are used in a large variety of types and applications. The most popular and politically important indicators are macro-economic statistics, such as GDP, Current Account Balance, Public Deficit, Consumer Price Index, Productivity etc. These indicators typically come from administrative accounts. In areas not covered by macro-economic accounts, such as social statistics, environment, transport, agriculture, education, etc., indicators are mostly collected through surveys (Kenett and Salini, 2012, Eurostat, 2023). Macro-economic indicators often link to scientific theory derived from economic sciences. Non macro-economic indicators typically do not rely on established theory, which raises issues with their interpretation. In general, the aim of an indicator is to provide supporting evidence to decision makers. The quality of indicators is derived from their ability to provide answers to questions posed by decision makers, with the required accuracy, timeliness, consistency, etc.

An important aspect of indicators is that they can provide a multidimensional perspective. This requires proper multivariate display and data analysis (see Kenett and Maggino, 2021). In general, the challenge of transforming numbers to information is significant. In particular, the ability to evaluate alternative scenarios based on official statistics requires methods to analysis counterfactual thought experiments. The next sections introduce a framework for planning and assessing information quality using official statistics followed by an example of a multivariate analysis using Bayesian networks to assess alternative scenarios.

3. Information quality

Information quality (InfoQ) is defined as “the potential of a data set to achieve a specific (scientific or practical) goal by using a given empirical analysis method” (Kenett and Shmueli 2014). InfoQ is determined by the data (X), the data analysis method (f) and the analysis goal (g), as well as by the relationships between them. Utility is measured using specific metric(s) (U). Setting a study goals is typically an iterative process (see Kenett et al, 2023b). By examining each of these components, and their relationships, we can learn about the contribution of a given study as a source of knowledge and insight. A mathematical formulation of information quality is: $\text{InfoQ} = U(f(X|g))$. The components of InfoQ have been mapped to eight dimensions that represent a deconstruction of the concept. Here, we present the eight InfoQ dimensions and provide some guiding questions that can be used in planning, designing and evaluating reports based on official statistics.

i) Data Resolution

Data resolution refers to the measurement scale and aggregation level of the data. The data’s measurement scale should be carefully evaluated in terms of its suitability to the goal, the analysis methods used, and the required resolution of the utility U. Questions one could ask to figure out the strength of this dimension include:

- Is the data scale used aligned with the stated goal of the study?
- How reliable and precise are the data sources and data-collection instruments used in the study?
- Is the data analysis suitable for the data aggregation level?

A low rating of data resolution is indicative of low trust in the usefulness of the study’s findings. An example of data resolution is provided by Google’s ability to predict the prevalence of flu based on the type and extent of Internet search queries. These predictions match quite well the official figures published by the Centers for Disease Control and Prevention (CDC). The point is that Google's tracking has only a day's delay, compared to the week or more it takes for the CDC to assemble a picture based on reports from doctors' clinics. Google is faster because it is tracking the outbreak by finding a correlation between what people search for online and whether they have flu symptoms, see Kenett and Shmueli, 2016a.

ii) Data Structure

Data structure relates to the type(s) of data and data characteristics such as corrupted and missing values due to the study design or data-collection mechanism. Data types include structured numerical data in different forms (e.g., cross-sectional, time series, network data) as well as unstructured, non-numerical data (e.g., text, text with hyperlinks, audio, video, and semantic data). The InfoQ level of a certain data type depends on the goal at hand. Questions to ask to figure out the strength of this dimension include:

- Is the type of data used aligned with the stated goal of the study?
- Are data-integrity details (corrupted/missing values) described and handled appropriately?
- Are the analysis methods suitable for the data structure?

A low rating of data structure reflects poor data coverage in terms of the project goals. For example, using a cross-sectional analysis method to analyze a time series warrants special attention when the goal is parameter inference, but is of less concern if the goal is forecasting future values.

iii) Data Integration

With the variety of data sources and data types available today, studies often integrate data from multiple sources and/or types to create new knowledge regarding the goal at hand (Dalla Valle and Kenett, 2015). Such integration can increase InfoQ, but, it can also reduce InfoQ. Data integration is particularly vulnerable to creation of privacy breaches. Questions to ask to figure out the strength of this dimension include:

- If the data integrated from multiple sources, what is the credibility of each source?

- How is the integration performed? Are there linkage issues that lead to dropping crucial information?
- Does the data integration add value in terms of the stated goal?
- Does the data integration cause privacy or confidentiality exposure concerns?

A low rating on data integration is indicative of missed potential in data analysis.

iv) Temporal Relevance

The process of deriving knowledge from data can be placed on a timeline that includes the periods of data collection, data analysis, and usage of results as well as the temporal gaps between these three stages. Such gaps can be due to the employment of independent contractors or internal organizational poor coordination. The different durations and gaps can each affect InfoQ. The data-collection duration can increase or decrease InfoQ, depending on the study goal, for example studying longitudinal effects versus a cross-sectional goal. Similarly, if the collection period includes uncontrollable transitions, this can be useful or disruptive, depending on the study goal. Questions to ask to figure out the strength of this dimension include:

- Considering the data collection, data analysis and deployment stages, are any of them time-sensitive?
- Does the time gap between data collection and analysis cause any concern?
- Is the time gap between the data collection and analysis and the intended use of the model (e.g., in terms of policy recommendations) of any concern?

A low rating on temporal relevance indicates an analysis with low relevance to decision makers due to data collected in a different contextual condition. This can happen in economic studies with policy implications that are based on old data.

v) Chronology of Data and Goal

The choice of variables to collect, the temporal relationship between them, and their meaning in the context of the goal at hand affects InfoQ. Questions to ask to figure out the strength of this dimension include:

- If the stated goal is predictive, are all the predictor variables expected to be available at the time of prediction?
- If the stated goal is causal, do the causal variables precede the effects?
- In a causal study, are there issues of reverse-causation?

A low rating on chronology of data and goal can be indicative of low relevance of a specific data analysis due to misaligned timing. A customer-satisfaction survey, that was designed to be used as input to the annual budget planning cycle, becomes irrelevant if its results are communicated after the annual budget is finalized (Kenett and Salini, 2012).

vi) Generalizability

The utility of *f(X|g)* is dependent on the ability to generalize *f* to the appropriate target population. Two types of generalizability are considered: statistical generalizability and scientific generalizability. Statistical generalizability refers to inferring from a sample to a target population. Scientific generalizability refers to applying a model based on a particular target population to other populations. This can mean either generalizing an estimated population pattern/model *f* to other populations, or applying *f* parameters estimated from one population, to predict individual observations in other populations. Determining the level of generalizability requires careful characterization of *g*. Generalizability is related to the concepts of reproducibility, repeatability, and replicability. Reproducibility represents insights that are replicable (but not necessarily identical), while repeatability is about achieving the same results in a repeated experiment. Replicability is used most often in genome wide association studies where a follow up experiment is conducted to identify a subset of genes as active, after following a large study investigating thousands of genes (Kenett and Shmueli, 2015). Repeatability relates to data quality and analysis quality, while reproducibility relates to InfoQ. Questions to ask to figure out the strength of this dimension include:

- Is the stated goal statistical or scientific generalizability?
- For statistical generalizability in the case of inference, does the paper answer the question “What population does the sample represent?”
- For generalizability in the case of a stated predictive goal (predicting the values of new observations; forecasting future values), are the results generalizable to the data to be predicted?

For more on Generalizability see Kenett and Shmueli, 2016b.

vii) Operationalization

Two types of operationalization are considered: construct operationalization and action operationalization of the analysis results. Constructs are abstractions that describe a phenomenon of theoretical interest. Measurable data are an operationalization of underlying constructs. The relationship between the underlying construct and its operationalization can vary, and its level relative to the goal is another important aspect of InfoQ. The role of construct operationalization is dependent on the goal, and especially on whether the goal is explanatory, predictive, or descriptive. In explanatory models, based on underlying causal theories, multiple operationalizations might be acceptable for representing the construct of interest. As long as the data are assumed to measure the construct, the variable is considered adequate. In contrast, in a predictive task, where the goal is to create sufficiently accurate predictions of a certain measurable variable, the choice of operationalized variable is critical. Action operationalizing results refers to three questions originally posed by Edwards Deming (Kenett and Redman, 2019):

- What do you want to accomplish?
- By what method will you accomplish it?
- How will you know when you have accomplished it?

Questions to ask to figure out the strength of construct operationalization include:

- Are the measured variables themselves of interest to the study goal, or is their underlying construct of interest?
- What are the justifications for the choice of variables?

Questions to ask to figure out the strength of operationalizing results include:

- Who can be affected (positively or negatively) by the research findings?
- What can he or she do about it?
- Who else?

A low rating on operationalization indicates that the study might have academic value but has little practical impact.

viii) Communication

Effective communication of the analysis and its utility directly impacts InfoQ. There are plenty of examples where the miscommunication of valid results has led to problematic outcomes. For a study of how to make more understandable National Assessment of Educational Progress (NAEP) and state test score reporting scales and reports, see Hambleton, 2002. Questions that a reviewer should ask to figure out the strength of this dimension include:

- Is the exposition of the goal, data and analysis clear?
- Is the exposition level appropriate for the readership of this report?

A low rating on communication indicates that poor communication might cover the true value of the analysis and, thereby, reduce the value of the information provided by the analysis.

Following this review of the information quality framework we now introduce Bayesian networks with an example. For more examples of applications of information quality to official statistics see Kenett and Shmueli, 2016a, 2016b.

4. Bayesian networks

Eurostat, 2023, provides survey-based information on health indicators in 36 countries over the past 20 years. It consists of data on various aspects of people's health status, which enables the analysis of public health issues, demographic patterns, socio-economic trends, and disparities in health statuses. Data on the following aspects are available:

- healthy life years

- self-perceived health and well-being
- functional and activity limitations
- self-reported chronic morbidity
- injuries from accidents
- absence from work due to health problems

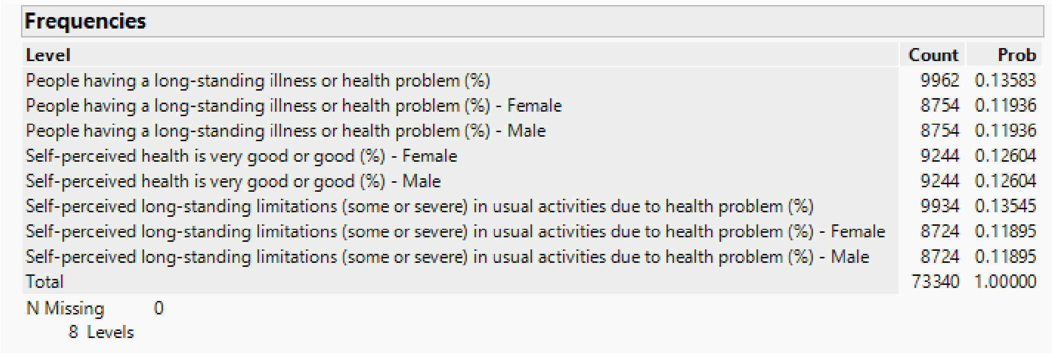
Kenett and Salini (2009) showed how Bayesian networks can be used to analyze such survey data and enable the assessment of alternative scenarios. We present here this capability in the context of Eurostat data. This approach has been implemented in a wide range of application domains such as socio-ecological system resilience, see Cai et al (2018) and Adams et al (2022) and education surveys (Pietro et al, 2015). We begin by introducing Bayesian networks.

Bayesian networks (BN) apply a graphical model structure known as a directed acyclic graph (DAG) that is popular in Statistics, Machine Learning and Artificial Intelligence. BNs are both mathematically rigorous and intuitively understandable. They enable an effective representation and computation of a joint probability distribution over a set of random variables (Pearl, 1985, Kenett et al. 2022). The structure of a DAG is defined by two sets: the set of nodes and the set of directed edges. The nodes represent random variables and are drawn as circles labelled by the variable names. The edges represent links among the variables and are represented by arrows between nodes. An edge from node *Xi* to node *Xj* represents a relation between the corresponding variables. Thus, an arrow indicates that a value taken by variable *Xj* depends on the value taken by variable *Xi*. This property is used to reduce, sometimes significantly, the number of parameters that are required to characterize the joint probability distribution (JPD) of the variables. This reduction provides an efficient way to compute posterior probabilities, given the evidence present in the data. In addition to the DAG structure, which is often considered as the "qualitative" part of the model, one needs to specify the "quantitative" parameters of the model. These parameters are described by applying the Markov property, where the conditional probability distribution (CPD) at each node depends only on its parents. For discrete random variables, this conditional probability is represented by a table, listing the local probability that a child node takes on each of the feasible values – for each combination of values of its parents. The joint distribution of a collection of variables is determined uniquely by these local conditional probability tables (CPT). The Eurostat case study presented here is based on discretized variables.

In learning the network structure, one can apply different network learning algorithms like the ones mentioned below in analyzing the Eurostat data. One can also manually include white lists of forced links imposed by expert opinion and black lists, of links that are not to be included in the network, even if the learning algorithm specifies it. In order to learn a BN that fully represents the joint probability distribution it represents, it is necessary to specify, for each node X, the probability distribution for X conditional upon X's parents. The distribution of X, conditional upon its parents, may have any form. Sometimes only constraints on a distribution are known. One can then use the principle of maximum entropy to determine a single distribution, i.e. the one with the greatest entropy given the constraints (Kenett and Salini, 2012). Often these conditional distributions include parameters which are unknown and must be estimated from data, for example using the maximum likelihood approach. When there are unobserved variables direct maximization of the likelihood (or of the posterior probability) is often complex. A classical approach to address this problem is the expectation-maximization (E-M) algorithm which alternates computing expected values of the unobserved variables conditional on observed data, with maximizing the complete likelihood assuming that previously computed expected values are correct. Under mild regularity conditions, this process converges to maximum likelihood (or maximum posterior) values of parameters (Heckerman, 1995).

Causal Bayesian networks are BNs where the effect of an intervention is defined by a ‘do’ operator that separates intervention from conditioning (Pearl, 2009). The basic idea is that an intervention breaks the influence of a confounder so that one can make a true causal assessment. The established counterfactual definitions of direct and indirect effects depend on the ability to manipulate mediators. A BN like graphical representation, based on local independence graphs and dynamic path analysis, can be used to provide an overview of dynamic relations. On the other hand, the econometric approach develops explicit models of outcomes, where the causes of effects are investigated and the mechanisms governing the choice of treatment are analyzed. In such investigations, counterfactuals are studied (Counterfactuals are possible outcomes in different hypothetical states of the world). In general, the study of causality involves: (a) defining interventions or counterfactuals,

(b) identifying causal models from idealized data of population distributions or empirical experiments and (c) identifying causal effects from actual data, where sampling variability is accounted for (Heckman, 2008). We focus here on a BN of 8 indicators from 36 countries with data from 2003 to 2022 (Figure 1). The range in the total number of surveys covered is 8724-9962. Overall, we have 73340 data points derived from the Eurostat surveys. We analyze this data with a BN.



Frequencies

Level	Count	Prob
People having a long-standing illness or health problem (%)	9962	0.13583
People having a long-standing illness or health problem (%) - Female	8754	0.11936
People having a long-standing illness or health problem (%) - Male	8754	0.11936
Self-perceived health is very good or good (%) - Female	9244	0.12604
Self-perceived health is very good or good (%) - Male	9244	0.12604
Self-perceived long-standing limitations (some or severe) in usual activities due to health problem (%)	9934	0.13545
Self-perceived long-standing limitations (some or severe) in usual activities due to health problem (%) - Female	8724	0.11895
Self-perceived long-standing limitations (some or severe) in usual activities due to health problem (%) - Male	8724	0.11895
Total	73340	1.00000

N Missing 0
8 Levels

Figure 1: The 8 indicators used in the Bayesian network analysis (JMP version 17.0)

The original data has missing values and we carried out a multivariate imputation preprocessing step to handle this. Following that, all 8 indicators were discretized into 5 categories defined by equal width bins. An algorithmic BN structure analysis with Greedy Thick Thinning produced the DAG shown in Figure 3. An alternative Bayes search gave similar results.

The root of the DAG in Figure 2 consists of percentages of "Male having a long-standing illness or health problem" (top left). The bottom right variable in the DAG is "Female having a long-standing illness or health problem". This is affected by the Male percentage directly and indirectly by "Female Self-perceived long-standing limitations (some or severe) in usual activities due to health problem". The BN indicates that 36% of Males and 10% of Females are in the lowest category. In Figure 3 and Figure 4 we condition "Male having a long-standing illness or health problem" to be 100% in the lowest category and 100% in the highest category, respectively. This ability to study the impact of such conditioning is mirroring a mental process conducted informally by decision makers. This is sometimes labeled a "what if" analysis. BNs provide the means to conduct such an analysis in a systematic and reproducible way. Here, we show what would happen if 100% of the "Male having a long-standing illness or health problem" is in the lowest category or if 100% is in the highest category, respectively. These scenarios can represent specific initiatives designed to change the current situation shown in Figure 2.

With the conditioning in Figures 3 and 4, the percent of "Female having a long standing illness or health problem" in the lowest category, increases from 19% to 29%. On the other hand, the lowest category in "Female Self-perceived long-standing limitations (some or severe) in usual activities due to health problem" dropped from 81% to 8%.

These scenarios provide an estimate of the impact of focused interventions, based on past observed data. The example indicates what would be the impact on "Female Self-perceived long-standing limitations (some or severe) in usual activities due to health problem" of changes in the conditions of "Male having a long-standing illness or health problem".

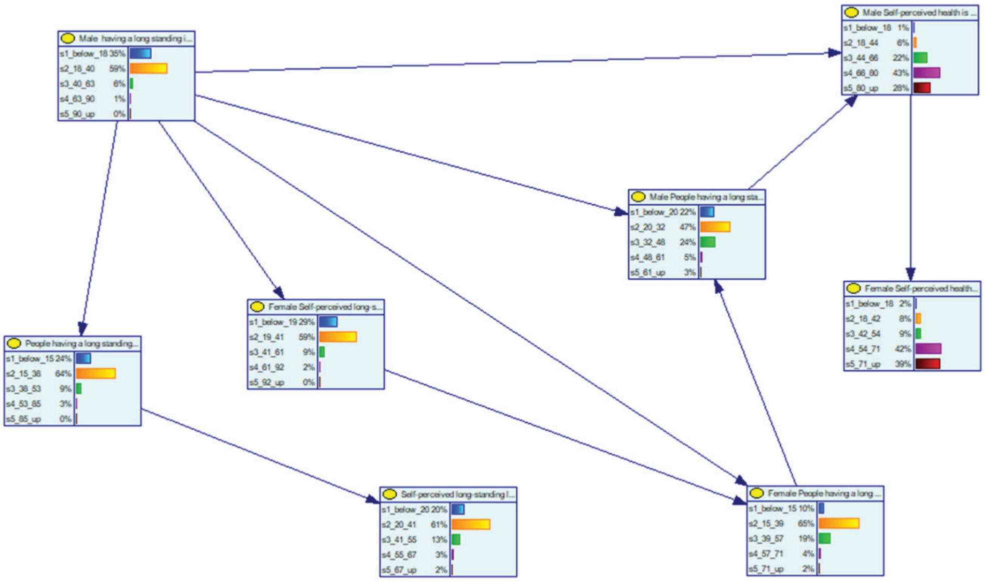


Figure 2: Bayesian network analysis of the Eurostat data (GeNie version 2.0)

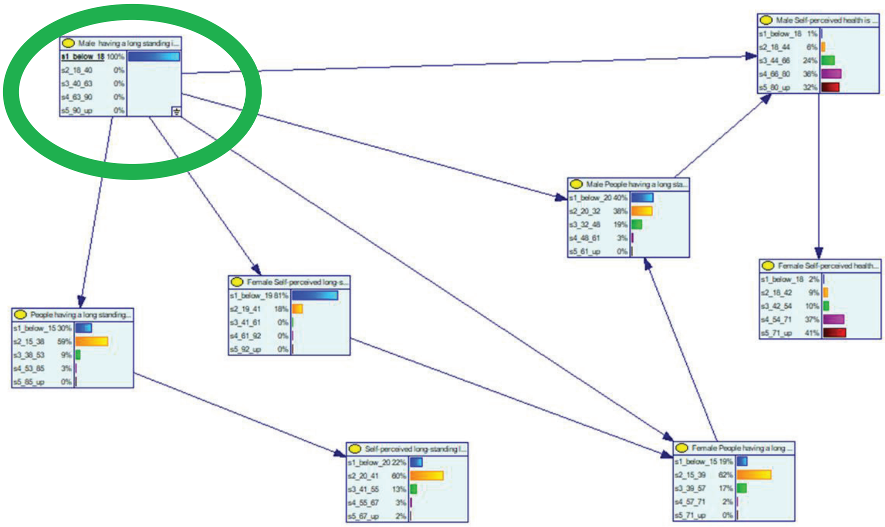


Figure 3: Low level conditioned Bayesian network analysis of the Eurostat data (GeNie version 2.0)

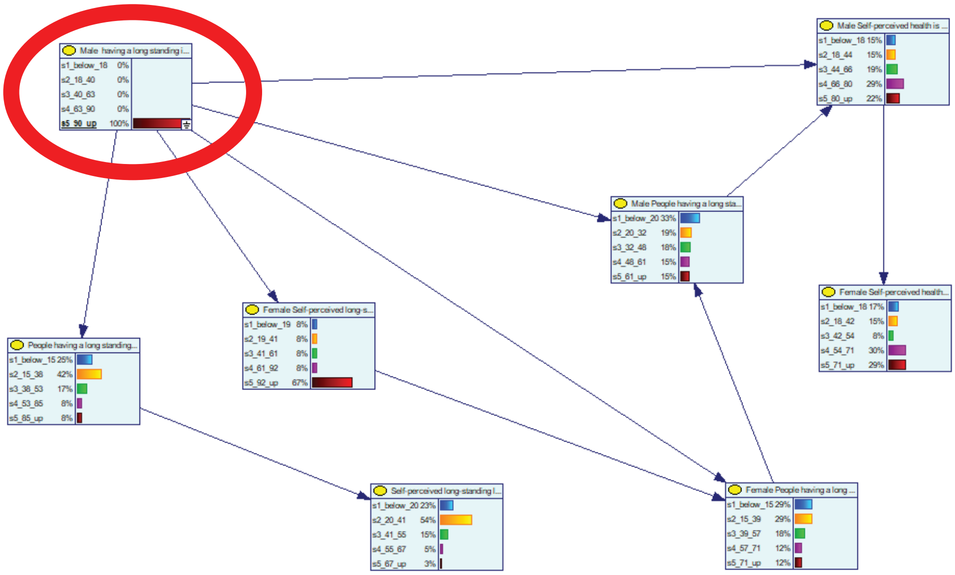


Figure 4: High level conditioned Bayesian network analysis of the Eurostat data (GeNie version 2.0)

5. Discussion

Official statistics is under strong evolutionary pressures. From a central and unique center of data production national statistics offices meet, on the one hand, alternative data providers and, on the other hand, changing expectations of users and customers. In this paper we touch on several aspects of this transformation and propose possible solutions. We list below some points that deserve more consideration in future work. We expand below on five such directions

i) A central challenge in data rich environments is data integration, the third InfoQ dimension. An example where this is needed in official statistics is in addressing survey mode effects. Surveys are typically conducted, simultaneously, on different platforms. Some surveys are conducted over the phone, some are face to face and some are based on omnibus panels. Integrating data from such sources is a much-needed competency. Dalla Valle and Kenett, 2015, propose a multivariate BN based method to calibrate such assembled data, in order to account for such mode effects.

ii) Official statistics indicators tend to be evaluated using univariate perspectives. This limits the quality of the information that can be provided. We provide an example of a multivariate analysis of survey data using Bayesian networks. Kenett and Salini, 2009, originally proposed it in the context of customer satisfaction surveys, but this also applies to official statistics.

iii) A similar transformation is occurring in the healthcare sector, see Bhandari and Kenett (2022). Both official statistics and healthcare services would benefit from coordinated initiatives, with mutual benchmarking targets.

iv) An area of research that deserves special attention is the development of methodologies for impact studies. The studies can be prospective (ex-ante) or retrospective (ex-post). They can combine observational data with randomized control interventions and case control analysis. Evaluations of ongoing interventions are called formative. Evaluations of past interventions are called summative. More knowledge is needed in conducting such studies.

v) Finally, modern statistics offers an expanded range of analysis methods for inference and predictive analytics, see Kenett et al. 2022, 2023a, 2023b. National bureaus of statistics are typically not involved in analysis and focus on data production. There is however an iterative looping cycle between data collection and data analysis so that both activities cannot be disassociated. This emphasizes the role of national bureaus of statistics as educators of decision makers and the public at large. The more sophisticated the users and producers of official statistics, the better the information generation process.

The paper is designed to map current challenges of national statistics organizations and propose possible approaches to handle them. The information quality framework is presented as a way to address a wide-angle perspective of statistical analysis and Bayesian networks as a multivariate analysis approach that enables an assessment of alternative scenarios. These are only options and, undoubtedly, more such techniques will be offered in the future. The challenge of transforming producers of numbers to generators of information used by decision makers requires both methodological and practical advances.

ACKNOWLEDGEMENT

The comments and suggestions of Dr. Smadar Lev from the Israel Central Bureau of Statistics, to an early version of this paper, are gratefully acknowledged.

REFERENCES

Adams, K. J., Macleod, C. A., Metzger, M. J., Melville, N., Helliwell, R. C., Pritchard, J., and Glendell, M. (2022) Developing a Bayesian network model for understanding river catchment resilience under future change scenarios. EGUsphere, 2022, 1-35.

Barcaroli, G. (2017) Improving the quality of official statistics by using alternative data sources: the Istat experience, Big Data in Business and Industry ECAS – ENBIS summer course, September 9-10, “Conservatorio delle Orfane”, Terra Murata - Isle of Procida – Italy.

Bhandari, D.R. and Kenett, R.S. (2022) Paradigm shift of statistical big data in healthcare: Management, analysis and future prospects, *Pravaha*, 28(1), 35–44. https://doi.org/10.3126/pravaha.v28i1.57969

Bhandari, D. R., Kenett, R.S., and Ravishanker, N. (2022) Data science for emerging application domains in Nepal. *Official Statistics for Nepal*: Issues and practices, 91, https://nepalindata.com/media/resources/items/20/bOFFICIAL_STATISTICS_OF_NEPAL_ISSUES_AND_PRACTICES.pdf#page=103

Cai, H., Lam, N. S., Zou, L. and Qiang, Y. (2018) Modeling the dynamics of community resilience to coastal hazards using a Bayesian network. *Annals of the American Association of Geographers*, 108(5), 1260-1279.

Dalla Valle, L. and Kenett, R.S. (2015) Official statistics data integration to enhance information quality, *Quality and Reliability Engineering International*, 31 (7), pp. 1281-1300.

EMOS (2021) https://ec.europa.eu/eurostat/cros/content/learning-outcomes-emos-programmes_en, retrieved June 21st, 2021.

Eurostat (2023) Health – Information on data, https://ec.europa.eu/eurostat/web/health/information-data#Health%20status

Forbes, S. and Brown, D. (2012) Conceptual thinking in national statistics offices, *Statistical Journal of the IAOS*, 28, p 89–98.

GSBPM (2021) https://ec.europa.eu/eurostat/cros/content/gsbpm-generic-statistical-business-process-model-theme_en (retrieved 21/6/2021).

Hambleton, R.K. (2002) How can we make NAEP and state test score reporting scales and reports more understandable? In *Assessment in Educational Reform*, edited by R.W. Lissitz and W.D. Schafer. 192–205. Boston, MA: Allyn & Bacon.

Heckerman, D. (1995) A tutorial on learning with Bayesian networks. Microsoft Research tech. report MSR-TR-95-06.

Heckman, J. (2008) *Econometric Causality*. International Statistical Review, 76, 1-27.

Kenett, R.S. and Salini, S. (2009) New frontiers: Bayesian networks give insight into survey-data analysis, *Quality Progress*, 42 (8), pp. 31-36.

Kenett, R.S. and Salini, S. (2012) *Modern Analysis of Customer Satisfaction Surveys: With Applications Using R*. Chichester, UK: John Wiley and Sons.

Kenett, R.S., and Shmueli, G. (2015). Clarifying the terminology that describes scientific reproducibility. *Nature methods*, 12(8), 699-699.

Kenett R.S. and Shmueli G. (2016a*) Information Quality: The Potential of Data and Analytics to Generate Knowledge*. John Wiley and Sons.

Kenett R.S., Shmueli G. (2016b) From quality to information quality in official statistics. *Journal of Official Statistics*, 32:1-22.

Kenett, R.S., and Redman, T. C. (2019) *The Real Work of Data Science: Turning Data into information, Better Decisions, and stronger Organizations.* John Wiley & Sons.

Kenett, R.S. and Maggino, F. (2021) Official Statistics Indicators: challenges in analysis and operationalization, *Journal of Official Statistics*, 37 (2), pp. 541-552.

Kenett, R.S., S. Zacks, S. and Gedeck P. (2022) *Modern Statistics: A Computer-Based Approach with Python*. Switzerland, AG: Springer Nature.

Kenett, R.S., S. Zacks, S. and Gedeck P. (2023a) *Industrial Statistics: A Computer-Based Approach with Python*. Switzerland, AG: Springer Nature.

Kenett R.S., Gotwalt C. and Poggi J.M. (2023b) An analytic journey in an industrial classification problem: How to use models to sharpen your questions. *Quality Reliability Engineering International*, 1-16. https://doi.org/10.1002/qre.3449

Maggino, F. (2018a) Indicators definition, Wiley *StatsRef: Statistics Reference Online*, DOI: 10.1002/9781118445112.stat08095, John Wiley and Sons,

Maggino, F. (2018b) Synthesis of indicators, *Wiley StatsRef: Statistics Reference Online*, DOI: 10.1002/9781118445112.stat08195, John Wiley and Sons.

Pearl, J. (1985). Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning" (UCLA Technical Report CSD-850017). Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine, CA, 329–334

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd ed., Cambridge University Press, UK.

Pietro, L.D., Mugion, R.G., Musella, F., Renzi, M.F., Vicard, P. (2015). Reconciling internal and external performance in a holistic approach: a Bayesian network model in higher education, *Expert Systems with Applications*, 42(5), 2691–2702.

doi.org/10.26398/IJAS.0036-003

TAX EVASION AND CASH PAYMENT CAP. DOES REALLY EXIST A RELATIONSHIP?

Fabrizio Antolini[1]

Department of Business Communication, University of Teramo, Campus Aurelio Saliceti, Via Renato Balzarini, 1- Teramo, Italy. ORCID: 0000-0002-3112-524X

Samuele Cesarini

Department of Business Communication, University of Teramo, Campus Aurelio Saliceti, Via Renato Balzarini, 1- Teramo, Italy. ORCID: 0000-0001-7062-1580

Abstract. *The shadow economy, which falls under the broader definition of the unobserved economy, has not found a univocal interpretation of the causes of its origin and evolution over time. The analysis becomes more difficult when extended to European countries, which differ in terms of the culture and structure of their tax systems. Despite this, to squelch a phenomenon related to the shadow economy, such as tax evasion, the European Commission has repeatedly stressed that the introduction of a cap on cash payments could be a possible tool for reducing tax evasion. Over time, different methodologies have been used to estimate both the unobserved economy and tax evasion, although the results have nonetheless converged. This does not happen in the formulation of country tax gap rankings, which change depending on whether tax evasion is used in relation to Gross Domestic Product or population. The purpose of this paper is to investigate the relationship between the levels of tax evasion and the introduction of the cash cap limits in the European countries. The existing tax regulations are different across the countries and not all have placed limits on cash payments. From the econometric estimation, the relationship between the existence of cash payment limits and the reduction in evasion was confirmed only for a threshold exceeding five thousand euros. The other variables considered – such as the tax burden on enterprises and families and the efficiency of the tax system – produce, instead, effects of a very different magnitude.*

Keywords: *Shadow economy; digital payments; tax gap; tax burden*

[1] Corresponding author: Fabrizio Antolini, e-mail: fantolini@unite.it.

1. INTRODUCTION

The shadow economy, or grey economy, is that part of the economy that avoids taxation because it frequently goes undetected (Schneider, 2011; International Monetary Fund [IMF], 2017; 2020). Initially, this term was used (Lewis, 1955) to describe unstructured forms of employment in developing countries, mainly concentrated in the service and agricultural sectors. Over time, different theories and definitions have been used to explain the causes of the shadow economy (Andrews et al., 2011; Deléchat and Medina, 2021; Dell'Anno, 2021; Elgin, 2020; Marinescu, 2019; Morales, 1997). For example, the *dualists* argue that people working in the informal sector are those who have been unable to enter the workforce due to a lack of the required technical skills. The *legalists* and *structuralists*, on the other hand, believe that it is the complex functioning of the legal system (e.g., excessive bureaucracy, high tax burden, and high labour costs) that induces workers to enter the underground economy, while finally, the *voluntarists* base it on an economic assessment of the costs and benefits associated with carrying out shadow economic operations.

The measurement of the shadow economy was addressed systematically by the European National Statistical Institutes (INs) in 1987, although it was not until 1989 that the first directive on *exhaustiveness* (89/130 EEC) was issued. The intention of this directive was that the shadow economy should be included in the calculation of gross domestic product (GDP) by countries adopting the European System of Accounts (ESA). In that period, the definition of shadow economy was a fuzzy definition, since:

> “[…] Also called the underground, informal, or parallel economy, the shadow economy includes not only illegal activities but also unreported income from the production of legal goods and services, either from monetary or barter transactions. Hence, the shadow economy comprises all economic activities that would generally be taxable were they reported to the tax authorities […].” (IMF, 2017/20).

During this same period, the concept of the *not directly observed economy* (NOE) was introduced by the United Nations in the System of National Accounts (SNA), considering NOE as the sum of underground economy (business and/or labor activities not known to the public administration, such as tax or contribution evasion), the informal economy (production activities that are legal but characterized by strong precariousness) and finally, the illegal economy (economic activities prohibited by law, or legal but carried out by unauthorized persons).The official definition was analytically reported a few years later, by OECD (2002) that classified not observed economy as in the SNA, articulating in underground economy, illegal economy and informal economy (the equivalent of shadow economy for IFM)

Although in its 1993 formulation the SNA (United Nations, 1993) had already included the illegal economy in the calculation of GDP, it was only in 2014 that national accounting systems officially included it. The concept of final domestic production (GDP) ignores any moralistic consideration and requires, as its only constraint, the existence of supply and demand. This explains why cigarette and alcohol smuggling, drug trafficking, and prostitution services are included in the GDP, while the same is not true of kidnapping. The measure of NOE is important for the credibility of GDP and more in general of national accounts estimates (OECD, 2002). Several problems are met in the NOE measurement, the first is a problem of the definition of what is to be measured. This lack of precision about the measurement target is evident looking at the range of different terms in common use – hidden economy, shadow economy, parallel economy, subterranean economy, informal economy, cash economy, black market. There is no common understanding whether they all mean the same thing and what type of relation exists with tax evasion, shuttle trade, or illegal activities. The second problem is that the estimation methods and the data sources utilized by the national statistical institutes are not always the same. Consequently, the macro-models used by the INs are not always sufficiently explained. For instance, several approaches have been used to estimate the *informal economy* (Elgin et al., 2021), defined as the market-based and legal production of goods and services that are hidden from public authorities for monetary, regulatory, or institutional reasons (Schneider et al., 2010). Among indirect methods for estimation, Schneider (2010) suggests the Multiple Indicators Multiple Causes (MIMIC) model, a structural equations model that can be used to estimate the relative size of informal economic activity. In contrast, the Dynamic General Equilibrium (DGE) model considers how optimizing households will allocate labor between formal and informal economies in each period and how the allocation will change over time (Elgin and Oztunali, 2014; Ihrig and Moe, 2004). However, both methodologies have their limitations (Elgin and Oztunali, 2014). Table 1 presents the values of the informal economy as estimated by applying the two different methodologies to the 27 states of the European Union in 2018 (the latest available period). In the official sources, these figures are reported only as a percentage of GDP. The absolute values were therefore obtained by considering the GDP recorded for the various countries in 2018. In Italy, for example, the GDP was 1,771 billion euro. Table 1 shows that when the MIMIC and DGE models are applied, the derived estimates do not necessarily coincide with those of the respective NSIs. For example, if again we consider the Italian National Institute of Statistics (ISTAT), the unobserved economy officially recorded in 2018 was 211 bn, of which 191.7 bn was the shadow economy, 91 percent of which was due to tax evasion (ISTAT, 2020). The same estimates obtained from MIMIC (28.41%) and DGE (26.08%) in relation to Italy's 2018 GDP lead to values of 503 bn and 462 bn, respectively.

Making a cross-country comparison, estimates expressed as a percentage of GDP (Table 1) indicate that Italy (28.41% and 26.08%), Lithuania (28.54% and 27.95%), Estonia (29.19% and 27.48%), Greece (29.35% and 26.17%), Croatia (29.96% and 28.66%), Romania (30.27% and 26.55%) and Bulgaria (31.75% and 27.84%) are the countries with the highest shadow economy rates. If, however, we look at estimates of the shadow economy in absolute value (billions of euro), the countries with the highest values are Germany (509.54 and 506.14), Italy (503.33 and 461.99), France (349.49 and 332.17) and Spain (269.13 and 249.49). Finally, when the estimates of absolute values are related to the resident population (per capita), the ranking of the countries changes again; now it is Luxembourg (with 9,776 and 8,858 euro per capita, as calculated by the two models), Ireland (9,593 and 9,194 euro), Denmark (8,889 and 8,554 euro), Belgium (8,701 and 8,205 euro), Sweden (8,621 and 8,038 euro) and Italy (8,415 and 7,724 euro) that have the highest share of the undeclared per capita economy.

Table 1 - MIMIC and DGE shadow (informal) economies estimates in EU countries (year 2018)

Country	MIMIC			DGE		
	% GDP	Total	Per capita	% GDP	Total	Per capita
Austria	9.51	36.66	4,138	9.24	35.61	4,020
Belgium	21.67	99.67	8,701	20.43	94.00	8,205
Bulgaria	31.75	17.85	2,550	27.85	15.66	2,237
Croatia	29.96	15.80	3,877	28.66	15.12	3,709
Cyprus	27.67	6.00	6,847	25.23	5.47	6,243
Czech Republic	16.96	35.79	3,360	16.57	34.96	3,283
Denmark	17.07	51.62	8,890	16.43	49.67	8,554
Estonia	29.19	7.57	5.713	27.48	7.13	5,379
Finland	17.79	41.52	7,525	16.10	37.58	6,810
France	14.79	349.49	5,194	14.06	332.17	4,936
Germany	15.14	509.55	6,138	15.04	506.15	6,097
Greece	29.35	52.70	4,914	26.18	47.00	4,382
Hungary	22.78	30.99	3,171	23.22	31.60	3,233
Ireland	14.40	47.05	9,593	13.81	45.09	9,194
Italy	28.41	503.33	8,415	26.08	461.99	7,724
Latvia	26.37	7.69	4,003	26.04	7.59	3,954
Lithuania	28.54	12.99	4,650	27.92	12.71	4,547
Luxembourg	9.98	6.00	9,776	9.05	5.44	8,858
Malta	23.96	3.10	6,287	24.94	3.23	6,545
Netherlands	13.05	101.00	5,844	12.57	97.32	5,631
Poland	24.66	123.03	3,240	23.28	116.15	3,059
Portugal	21.04	43.18	4,202	22.58	46.33	4,509
Romania	30.27	62.38	3,213	26.56	54.73	2,819
Slovak Republic	16.61	14.93	2,738	16.14	14.50	2,661
Slovenia	24.43	11.21	5,385	24.26	11.13	5,347
Spain	22.36	269.13	5,734	20.72	249.49	5,315
Sweden	18.74	88.19	8,621	17.47	82.23	8,038

Source: Data elaboration on CERP discussion paper "Understanding Informality" (Elgin et al. 2021.)

Whichever the methodology is utilized, there is no doubt that the existence of the unobserved economy creates a loss of tax revenue for the economic system, a distortion in the functioning of markets and, therefore, in their productivity levels. However, without a careful preliminary analysis of the causes that incentivize the shadow economy, the introduction of limits and rules in the payment system could not be a solution to catch the hidden economy.

The paper is structured as follows. In Section 2, we delve deeper into the economic rationale behind the European Commission's endorsement of cash payment limits. This section includes a comparative analysis of tax structures across different European countries and discusses the balance between tax burden and social benefits. In Section 3, a comprehensive analysis of the relationship between cash payment limits and the tax compliance gap in Europe is presented. An econometric estimation to taste the relationship between tax evasion and cash payment caps, has been applied considering two different models. Finally, the paper concludes in Section 4 with a discussion on the broader implications of our findings.

2. ECONOMIC ENVIRONMENT AND THE EUROPEAN POSITION ON THE CASH PAYMENT CAP

The European Commission has repeatedly stressed the appropriateness of introducing limits on cash payments in countries that are part of the European Union, considering it a useful approach for thwarting money laundering. The introduction of regulations on how transactions are settled, favoring card transactions, is intended to act as a deterrent. In February 2016, the Commission released a communication to the Council and Parliament on an action plan to intensify the struggle against terrorist financing (European Commission, 2016). This action plan states that ‘cash payments are widely used for the financing of terrorist activities. In this context, the relevance of potential upper limits for cash payments could also be examined’. Subsequently, on February 12, 2016, the Economic and Financial Affairs Council endorsed this and invited the Commission to assess the need to introduce appropriate restrictions on cash payments exceeding specified thresholds. Several member states, taking up these suggestions, introduced prohibitions on cash payments above a specific threshold to inhibit the anonymity of cash transactions that facilitate and incentivize the underground economy and more in general, the

NOE. This position, however, has not been supported by the evidence and, in fact, very different positions are held on the issue. For example, the IMF believes:

"*Hence, the higher the tax burden and labor costs, the more incentives individuals have to avoid these costs by working in the shadow economy*" (IMF, 2017).

The tax burden (tax and social security contributions) existing in European countries is shown in Figures 1 and 2, which show the level of the tax wedge and tax burden, respectively. Concerning the tax wedge expressed as a percentage of labor costs, in the highest quartile, we find Italy (46.5%), France (47.0%), Austria (47.8%), Germany (48.1%) and Belgium (52.6%). However, this indicator cannot be used as a proxy for the effectiveness and efficiency of public economic and welfare systems since this type of analysis should also consider the level and quality of the social benefits provided.

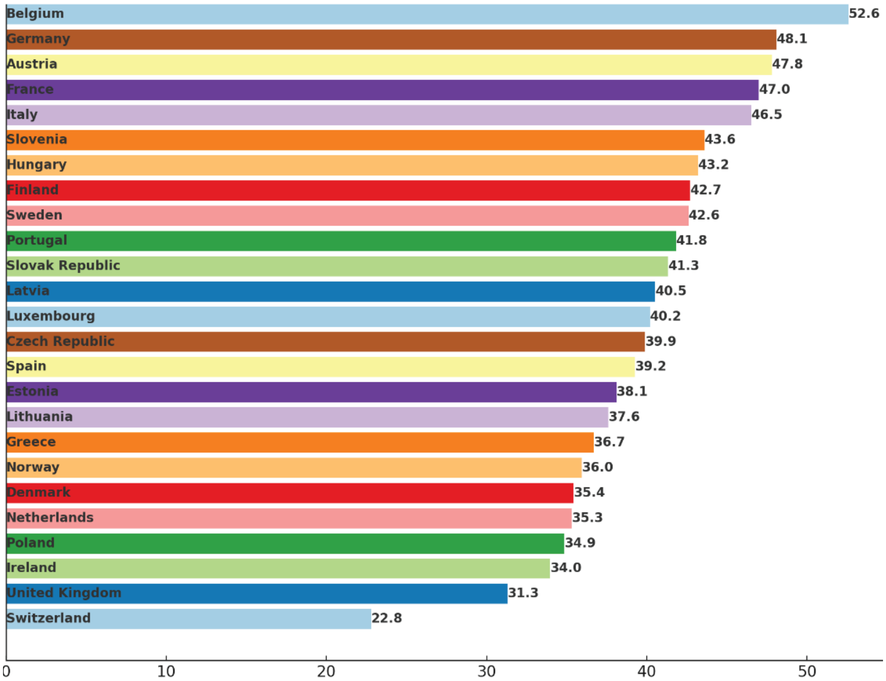


Figure 1 - Tax wedge (%) of labour costs (year 2021)

Source: Organisation for Economic Co-operation and Development (OECD) data elaboration

With regard to the tax burden (Figure 2), and considering, in particular, the tax rates on personal income in the major European countries, these exceed the 50% threshold in Denmark, France, Austria, Greece, Spain, Belgium and Portugal. Italy ranks fourteenth among the European countries considered, with a personal tax rate of 47.2%. The Czech Republic, Estonia and Hungary close the ranking with values below 25%.

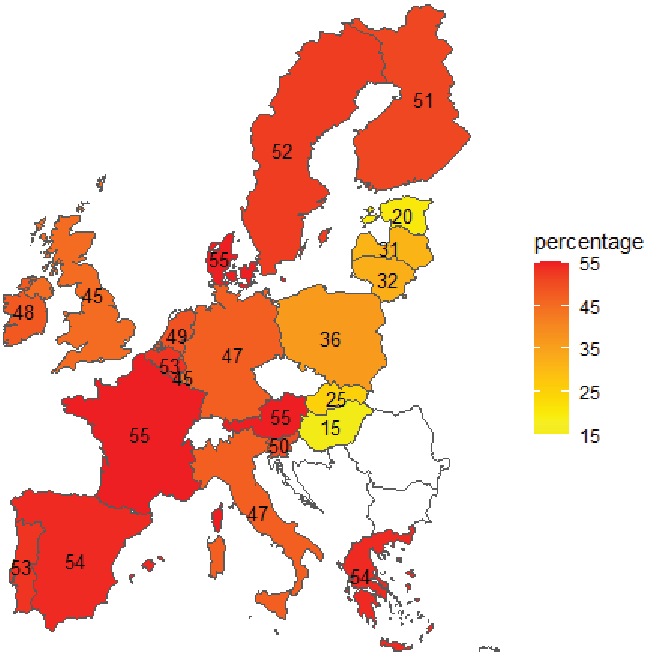


Figure 2 - Tax rates on personal income (%) in some European Countries (year 2021) *Source: OECD data elaboration*

Extending the analysis to the taxation of corporate profits, Table 2 shows the corporate tax rate, following the prescribed statutory rates at various levels of government.

Table 2 - Corporate tax rate (year 2022)

Country	Central	Sub	Total	Country	Central	Sub	Total
Austria	25.00	-	25.00	Latvia	20.00	-	20.00
Belgium	25.00	-	25.00	Luxembourg	18.19	6.75	24.94
Denmark	22.00	-	22.00	Netherlands	25.80	-	25.80
Finland	20.00	-	20.00	Norway	22.00	-	22.00
France	25.83	-	25.83	Poland	19.00	-	19.00
Germany	15.83	14.01	29.83	Portugal	30.00	1.50	31.50
Greece	22.00	-	22.00	Slovak Republic	21.00	-	21.00
Hungary	9.00	-	9.00	Spain	25.00	-	25.00
Ireland	12.50	-	12.50	Sweden	2.60	20.60	20.60
Italy	24.00	3.90	27.81	Switzerland	8.50	12.87	19.70

Source: OECD data

Not all European countries impose regional or territorial taxes on companies; however, by combining the two tax levels, it is possible to get a comprehensive overview. Italy again ranks in the highest quartile (27.8%), just behind Portugal (31.5%) and Germany (29.8%). Ireland (12.5%) and Hungary (9.0%) show the lowest values, thus influencing the flow of foreign direct investment. Overall, the tax burden in some countries, such as Italy, Luxembourg, Finland, and Australia (Fig. 3), is characterized by an unbalanced composition of the state's tax revenue. If we look at Italy, a country which, as we will have the opportunity to analyze, has a high tax gap, individual taxation contributes 26.9% of total revenue (24.0% is the average figure for OECD countries), taxation of corporate profits accounts for 4.9% (cf. 9.2% in OECD countries), social contributions account for 31.5% (cf. 26.4% in OECD countries), wealth tax accounts for 5.7% (cf. 5.6% in OECD countries) and finally, consumption tax accounts for 26.9% (cf. 32.1% in OECD countries).

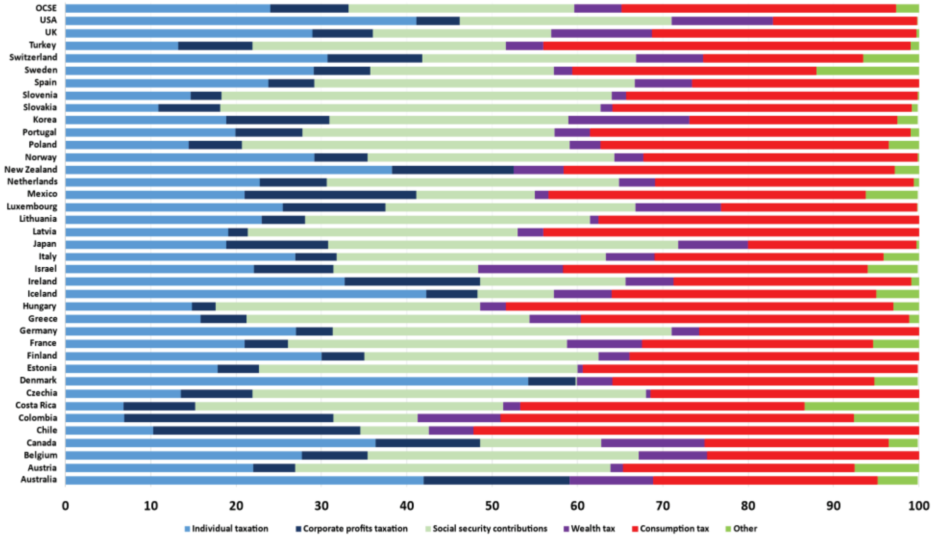


Figure 3 - Typologies of taxation in some OECD countries (year 2020 – percentage values)

Source: OECD data elaboration

Despite the European Commission's contention that they would act as a measure to curb money laundering operations, restrictions on cash payments are not applied uniformly in EU countries. Some countries (Table 3) prohibit cash payments above a certain threshold, while others do not. In September 2022, the European Consumer Centres Network (ECC), a body set up by the European Commission and the Member States to provide assistance to consumers, reported information on 30 European states, of which 27 are part of the EU, and including Iceland and Norway (ECC, 2022). Of the 30 states, only 17, which include Italy, have cash limits in force. These are mostly the southern European countries, except for Belgium and some Eastern European states.

At the top of the list of European countries with a higher limit on the use of cash is Croatia, with 15,000 euro, followed by the Czech Republic and Malta, with 10,000 euro each. A high use of cash is also authorized in Latvia (7,200 euro), Bulgaria, Slovakia, and Slovenia (5,000 euro). Italy (2,000 euro), Romania, France, and Spain (1,000 euro) hold the lowest positions, other than Greece (500 euro). Austria, Cyprus, Estonia, Finland, Germany, Hungary, Iceland, Ireland, Luxembourg, the Netherlands, Sweden, and the United Kingdom have no limits on cash circulation.

Table 3 - Limits on cash payments in Europe (year 2022 – values in euro)

Country	Private	Trade	Country	Private	Trade
Austria	-	-	Italy	2,000	2,000
Belgium	-	3,000	Latvia	7,200	7,200
Bulgaria	5,108	5,108	Lithuania	3,000	3,000
Croatia	15,000	15,000	Luxembourg	-	-
Cyprus	-	-	Malta	10,000	10,000
Czech Republic	10,509	10,509	Netherlands	-	-
Denmark	-	2,689	Norway	-	3,841
Estonia	-	-	Poland	-	3,267
Finland	-	-	Portugal	3,000	3,000
France	1,000	1,000	Romania	10,165	1,016
Germany	-	-	Slovakia	15,000	5,000
Greece	500	500	Slovenia	-	5,000
Hungary	-	-	Spain	10,000	1,000
Iceland	-	-	Sweden	-	-
Ireland	-	-	United Kingdom	-	-

Source: European Consumer Centres Network

It is widely believed that electronic payments are more likely to be used in Nordic countries, but the available information (Table 4) does not support this conclusion (European Central Bank, 2021). The value of other payment services is not reported.

Table 4 - Main payment instruments (%) in EU countries (year 2021)

Country	**AT**	**BE**	**BG**	**HR**	**CY**	**CZ**	**DK**	**EE**	**FI**	**FR**
Card	51.5	52.0	-	-	69.5	-	-	63.9	61.6	59.3
Credit transfer	27.0	36.0	-	-	18.3	-	-	29.8	38.4	17.9
Direct debits	19.4	10.2	-	-	-	-	-	-	-	18.5
E-money	0.2	0.9	-	-	3.5	-	-	-	-	0.2
Cheques	0.0	0.0	-	-	3.1	-	-	-	0.0	4.1
Country	**DE**	**GR**	**HU**	**IS**	**IE**	**IT**	**LV**	**LT**	**LU**	**MT**
Card	30.3	69.5	-	-	62.4	52.5	62.8	68.4	4.9	54.4
Credit transfer	26.1	22.4	-	-	17.2	17.4	34.0	17.4	1.4	19.3
Direct debits	43.1	1.2	-	-	6	11.5	0.0	1.0	0.4	-
E-money	0.1	2.2	-	-	11.7	15.7	-	10.2	93.3	11.6
Cheques	0.0	0.2	-	-	0.8	0.9	0.0	-	0.0	4.7
Country	**NL**	**NO**	**PL**	**PT**	**RO**	**SK**	**SI**	**ES**	**SE**	**UK**
Card	49.2	-	-	72.2	-	56.5	51.6	66.4	-	-
Credit transfer	34.2	-	-	13.4	-	37.2	32.3	13.8	-	-
Direct debits	16.6	-	-	8.8	-	3.3	8.9	18.0	-	-
E-money	0,0	-	-	2.6	-	-	0.9	0.4	-	-
Cheques	0,0	-	-	1.0	-	0.0	0.0	0.3	-	-
Average (UE)										
Card	55.7									
Credit transfer	23.9									
Direct debits	11.1									
E-money	10.2									
Cheques	0.9									

Source: European Central Bank|Eurosystem (2021)

Table 4 shows the main payment instruments in each EU country, expressed as a percentage of the total transactions in the country. The interpretation of these data can be supplemented by that presented in Table 5, which shows the average use of cards as a payment instrument over the period 2015 to 2019 (European Central Bank, 2021).

Table 5 – Average (AV) number of card transactions metered per EU country (EUC) (years 2015–2019)

EUC	**AT**	**BE**	**BG**	**HR**	**CY**
AV	765,648	1,877,696	131,948	304,243	51,143
EUC	**CZ**	**DK**	**EE**	**FI**	**FR**
AV	912,442	1.995,77	314,731	1,674,945	12,238,172
EUC	**DE**	**GR**	**HU**	**IE**	**IT**
AV	4,763,357	474,847	692,446	941,53	2,939,932
EUC	**LV**	**LT**	**LU**	**MT**	**NL**
AV	278,626	287,69	126,079	25,422	4,311,381
EUC	**PL**	**PT**	**RO**	**SK**	**SI**
AV	4,011,703	1,587,680	513,686	425,808	185.028
EUC	**ES**	**SE**	**-**	**-**	**-**
AV	4,155,485	3,322,688	-	-	-

Source: European Central Bank | Eurosystem (2021)

A final, socially oriented comment on the use of digital payments is warranted. The reality is that the use of digital instruments, including electronic payments, is often considered a proxy for a country's technological progress. However, countries experiencing an aging population do not necessarily gain the social benefits resulting from their wide distribution, since this population is characterized by insufficient literacy in information and communication technology (ICT). The digital divide may indeed produce new forms of inequality, although the digital payment system is moving increasingly towards user-friendly technologies (Forbis, 2019).

3. TAX COMPLIANCE GAP AND CASH LIMITS IN EUROPE. WHAT RELATIONSHIP REALLY EXISTS?

The annual tax compliance gap (Table 6) in Europe is estimated to be 824 billion euro. This figure was given in the report 'The European Tax Gap', which was approved by the European Parliament on 26 March 2020 (Murphy, 2019). The data contained in this report make a comparative analysis of EU countries possible. This tax gap information can be used for different purposes. It can be used to assess the effectiveness and efficiency of a tax authority, to measure inequality resulting from the failure to enforce the tax law fairly, or to measure the effectiveness, or otherwise of tax policy implementation in a particular jurisdiction. The estimated tax compliance gap shown in Table 6 is defined as the difference between the amount of taxes that could be collected under the current legislation by the tax system and the amount of taxes actually collected. For this reason, it is considered an estimate of tax evasion.

Table 6 - Tax gap estimates in Europe (year 2019 – "Total" values in billion euro – "Per capita" values in euro)

Country	Total	Per capita	Country	Total	Per capita
Italy	190.9	3,191.4	Hungary	9.1	931.2
Germany	125.1	1,506.9	Czech Republic	8.8	826.3
France	117.9	1,752.1	Ireland	6.9	1,406.9
United Kingdom	87.5	1,312.9	Slovakia	5.4	990.7
Spain	60.0	1,278.3	Bulgaria	3.8	542.9
Poland	34.6	911.2	Croatia	3.5	858.6
Belgium	30.4	2,653.7	Lithuania	3.1	1,109.4
Netherlands	22.2	1,284.6	Slovenia	2.6	1.249,5
Greece	19.9	1,855.5	Latvia	1.7	885.4
Denmark	17.5	3,014.1	Cyprus	1.6	1,826.7
Sweden	16.9	1,652.0	Luxembourg	1.6	2,606.3
Romania	16.2	834.4	Estonia	1.4	1,056.7
Austria	12.9	1.456,2	Malta	0.9	1,823.5
Portugal	11.0	1.070,4	Iceland	-	-
Finland	10.7	1.939,1	Norway	-	-
Total	**824.1**	**1,493.8 (AV)**			

Source: European Parliament "European Tax Gap" report

In absolute terms, the European country with the highest tax gap is Italy, where missed payments due to tax authorities were estimated at 190.9 billion euro (Murphy, 2019). Germany followed with 125.1 billion and France with 117.9 billion. In some countries, the NSIs produced a slight revision of the estimates included in the report. For example, in Italy, as a result of a revision by the National Economic Accounts, the estimated tax gap for 2019 was adjusted to 183.4 billion with 'unobserved' at 203.3 billion euro. Illegal activities accounted for 19.4 billion.

Evasion in relation to population size (evasion per capita) is an indicator that improves the social representativeness of the phenomenon, since taxes are mainly used to finance public services, matching changing demographic needs and the relative composition of the active and non-active populations. Using this indicator, Italy (with an average evasion of 3,191 euro per capita), together with Denmark (3,014 euro) ranked first, followed by Belgium (with 2,653 of euro evasion per capita). France and Germany, third and second, respectively, when using absolute total evasion values, drop to the ninth (1,752 euro) and eleventh (1,507 euro) positions in the per capita tax evasion ranking. The European average was 1,494 euro, and below that we find Romania (834 euro), the Czech Republic (826 euro) and Bulgaria (543 euro).

Data from the European Commission published in the EU VAT Gap Report in 2018 confirms that among EU countries, value added tax (VAT) is the tax most evaded, representing a revenue loss of 140 billion euro. The data in Table 6 are interesting, but they do not allow us to understand the reasons for the amounts of tax evasion, even if it seems to be partially correlated – but not for Belgium – with the tax wedge of labor cost described in Table 1. Table 7 shows the data for 2017 contained in the report.

Table 7 - Evaded VAT EU countries (year 2017) – value in billions of euro

Country	%	Total	Country	%	Total
Austria	9.0%	2.91	Italy	24.5%	35.44
Belgium	10.4%	3.62	Latvia	9.5%	2.56
Bulgaria	10.8%	0.61	Lithuania	25.9%	1.23
Croatia	3.5%	0.25	Luxembourg	5.1%	1.99
Cyprus	3,8%	0.77	Malta	15.1%	0.16
Czechia	12.0%	2.19	Netherlands	4.2%	2.28
Denmark	7.2%	2.25	Poland	9.9%	4.45
Estonia	5.2%	0.13	Portugal	9.6%	1.89
Finland	3.6%	0.81	Romania	33.8%	6.60
France	7.1%	12.79	Slovakia	20.0%	1.58
Germany	8.6%	22.08	Slovenia	3.8%	0.15
Greece	30.1%	0.66	Spain	6.0%	4.91
Hungary	8.4%	1.19	Sweden	0.7%	0.31
Ireland	10.6%	1.68	United Kingdom	12.2%	23.45
Total		**138.92**			

Source: European Commission (2018)

The highest VAT gap was observed in Romania, where 33.8% of the estimated VAT revenue was lost, followed by Greece (30.1%) and Lithuania

(25.9%). Sweden (0.7%), Croatia (3.5%) and Finland (3.6%) recorded the smallest percentage gaps. The largest VAT gaps, measured in absolute terms, occurred in Germany (22 billion euro), the United Kingdom (23.5 billion euro) and Italy (35.4 billion euro). The VAT gap is considered the form of tax evasion that can most effectively be prevented by the use of card payments.

Figures 4 and 5, while only considering countries with a limit on cash payments made between individuals (13 countries) and in trade (17 countries), show the potential relationship between the recorded levels of the tax gap, both total and per capita (Table 6), for each country and the limit on cash (for both individuals and in trade).

A graphical inspection of the scatter plots shows the existence of a weak inverse relationship between the per capita and total tax evasion and the existence of a cash limit, both among private individuals and in trade. From a theoretical point of view, the descriptive results presented enhance the argument (IMF, 2017) of those who state that it is the level of the existing tax and the contribution burden that is one of the causes influencing the attitude of operators who evade tax payments and not only regulations on the use of cash. This reinforces the theory that a higher level of taxation is not inevitably matched by a higher level of tax revenue (Brill and Hassett, 2007). On the other hand, it cannot be argued that a higher tax rate is always associated with a greater tax evasion. The results must, therefore, be assessed according to the institutional context of each country. For example, another factor that may represent an incentive for tax evasion is the complexity of the functioning of the tax system since this has been identified as one of the determinants of increased tax and contribution evasion (Kelmanson et al., 2019). For this reason, we introduced the Easy Pay Taxes Score (Table 8) into our analysis. This records the taxes and mandatory contributions that a medium-sized company must pay each year as well as measures of the administrative burden of paying taxes and contributions, including the time needed to comply with the major taxes and to comply with post-filing procedures (World Bank Data and PwC [PricewaterhouseCoopers], 2020). A higher score corresponds to a lower burden from the tax system as a whole and, thus to a higher level of efficiency.

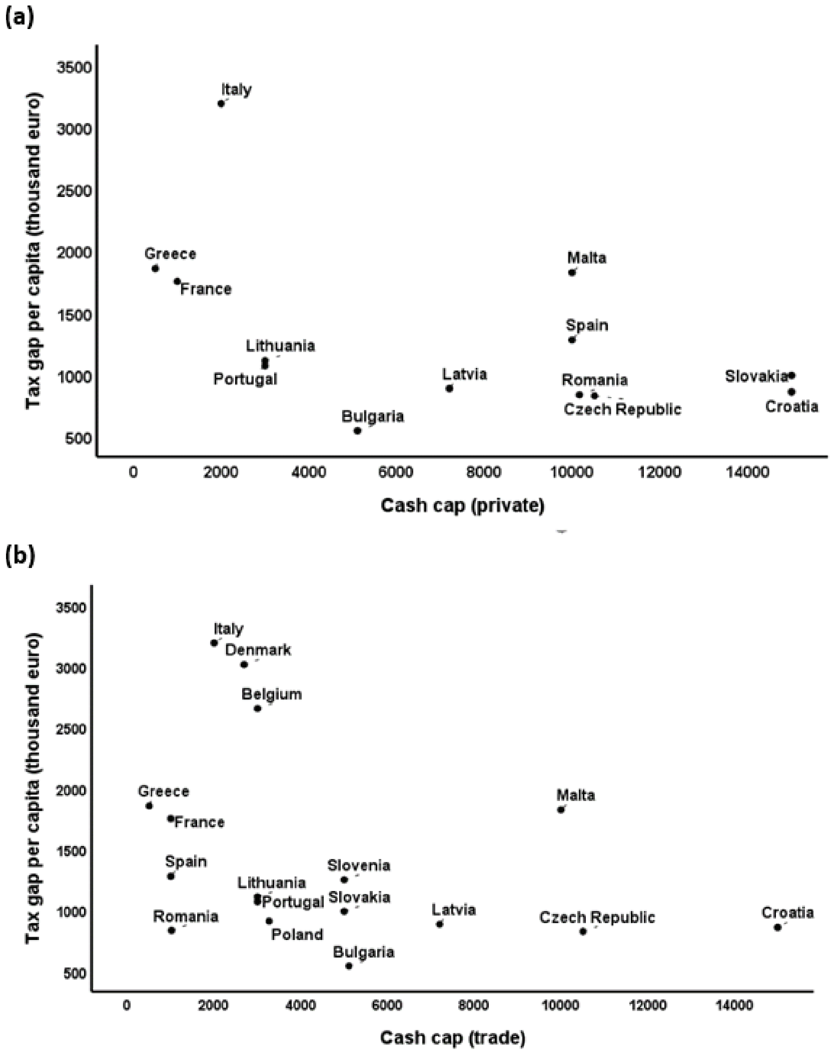


Figure 4 - Relation per capita tax gap and cash limit among private individuals (a) and in trade (b) in EU countries

Source: Data elaboration European Parliament report "European Tax Gap"- European Consumer Centres Network

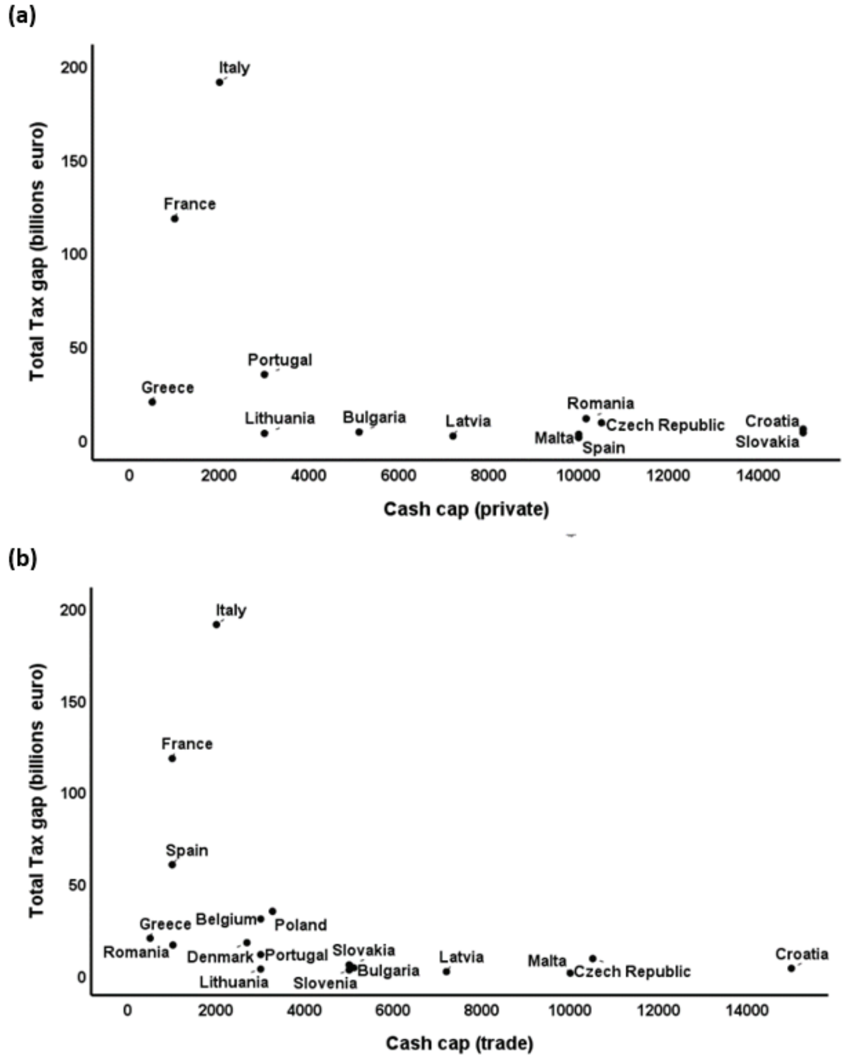


Figure 5 - Relation total tax gap and cash limit among private individuals (a) and in trade (b) in EU countries

Source: Data elaboration European Parliament report "European Tax Gap"- European Consumer Centres Network

Table 8 - Easy Pay Taxes Score for each EU country (year 2020)

Country	Easy Pay Taxes Score	Country	Easy Pay Taxes Score
Austria	83.5	Italy	64.0
Belgium	78.4	Latvia	89.0
Bulgaria	72.3	Lithuania	88.8
Croatia	81.8	Luxembourg	87.4
Cyprus	85.5	Malta	76.2
Czech Republic	81,4	Netherlands	87.4
Denmark	91.1	Poland	76.4
Estonia	89.9	Portugal	83.7
Finland	90.9	Romania	85.2
France	79.2	Slovakia	80.6
Germany	82.2	Slovenia	83.3
Greece	77.1	Spain	84.7
Hungary	80.6	Sweden	85.3
Ireland	94.6	-	-

Source: World Bank/PwC report data

Italy, with a score of 64 is in last place in this ranking, together with Malta (76.2), Poland (76.4), Greece (77.1) and France (79.2). The best performances are recorded, instead, in Ireland (94.6), Denmark (91.1) and Finland (90.9).

3.1 Tax evasion and the cash cap: an econometric estimate

To estimate the relationship between tax evasion (or the tax gap) and the existence of the cash cap, we used a regression model with dummy variables (James et al., 2020) and applied it to the 27 EU countries (Equation 1). For these countries, the data available for measuring the tax gap (Table 6) were entered as the dependent variable, while the predictors were the cash limits, divided as shown below.

The necessity of dividing cash limits into thresholds was due to the heterogeneity of levels across countries, as specified above.

$$\mathbf{y} = \beta_0 + \beta_1\mathbf{d1} + \beta_2\mathbf{d2} + \beta_3\mathbf{d3} + \varepsilon_i \qquad (1)$$

y = tax gap per capita (in thousands of euro)
d1 = 1 if country *i* has a cap on cash between 500 and 2999; otherwise, 0.
d2 = 1 if country *i* has a cap on cash between 3000 and 5000; otherwise, 0.
d3 = 1 if country *i* has a cap on cash > 5001, otherwise 0.

Model (1) is specified below:

Model 1: OLS, using observations 1-27
Dependent variable: Tax gap per capita

	Coefficient	*Error Std.*	*ratio t*	*p-value*
β_0	1566.66	204.33	7.66	<0.0001 ***
d1	420.989	333.67	1.26	0.2197
d2	−235.832	333.67	−0.70	0.4868
d3	−579.317	353.91	−1.63	0.1153

Average dependent variable	1500.52	SQM dependent variable	695.03
Sum residual squared	9602805	Error standard	646.15
R-squared	0.23	Adjusted R-squared	0.13
F(3, 23)	2.36	P-value(F)	0.09
Log-likelihood	−210.86	Akaike criterion	429.72
Schwarz criterion	434.27	Hannan-Quinn	431.27

Breusch-Pagan test for heteroscedasticity -
Null hypothesis: heteroscedasticity not present.
Test statistic: LM = 4.81
with p-value = P(Chi-square(3) > 4.81) = 0.18

Test for normality of residuals -
Null hypothesis: Error is normally distributed.
Test statistic: Chi-square(2) = 3.73
with p-value = 0.15

This model (James et al., 2020) has the advantage of interpreting the intercept β_0 – the model without dummy variables – as the reference category (baseline), which describes the situation of countries with no cash payment limits. The estimation method identifies statistical significance for the intercept, but not for the dummy explanatory variables representing the cash payment thresholds. Therefore, the introduction of cash limits would not be related to the reduction of the tax gap. In this kind of model, R^2 is usually characterized by low values, while the F-test, which is not statistically significant, is more important for confirming the hypothesis that there is no relationship between the cash limit and the reduction of the tax gap (James et al., 2020).

Economic theory, as explained above, requires the model to be integrated by including additional explanatory variables that are considered to affect the tax gap in different countries. For this reason, the Easy Pay Taxes Score (**EPTS**), which takes into account the tax burden on businesses and the level of efficiency of the existing tax system in each Member State; the tax burden on individuals in relation to GDP (**PFP/GDP**) and, finally, the average number of card payments (**card**) between 2015 and 2019, as well as its interaction effect with the cash cap above 5000 euro (**card*d3**), were included as covariates.

The identification of the model required the logarithmic transformation of the dependent variable Total tax gap to ensure the assumption of normality of the residuals. Below is the formulation (2):

$$In(\mathrm{y}) = \beta_0 + \beta_1\mathbf{d1} + \beta_2\mathbf{d2} + \beta_3\mathbf{d3} + \beta_4\mathbf{EPTS} + \beta_5\mathbf{PFP\_GDP} + \beta_6\mathbf{card} + \beta_7\mathbf{card} *\mathbf{d3} + \varepsilon_i \qquad (2)$$

The regression results are shown in Model 2 and provide further insight.

Model 2: OLS, using observations 1-27
Dependent variable: log Total tax gap

	Coefficient	*Error Std.*	*ratio t*	*p-value*
β0	10.97	2.39	4.59	0.0002 ***
d1	0.32	0.42	0.75	0.4567
d2	−0.29	0.38	−0.74	0.4626
d3	−2.00	0.57	−3.46	0.0026 ***
EPTS	−0.11	0.02	−4.03	0.0007 ***
PFP_GDP	0.08	0.03	2.56	0.0191 **
card	2.36e-07	6.32e-08	3.73	0.0014 ***
card*d3	2.42e-06	1.04e-06	2.31	0.0321 **

Average dependent variable	2.30	SQM dependent variable	1.44
Sum residual squared	9.59	Error standard	0.71
R-squared	0.82	Adjusted R-squared	0.75
F(3. 23)	12.68	P-value(F)	5.69e-06
Log-likelihood	−24.34	Akaike criterion	64.69
Schwarz criterion	75.06	Hannan-Quinn	67.78

Breusch-Pagan test for heteroscedasticity -
Null hypothesis: heteroscedasticity not present.
Test statistic: LM = 9.34
with p-value = P(Chi-square(7) > 9.34) = 0.23

Test for normality of residuals -
Null hypothesis: Error is normally distributed.
Test statistic: Chi-square(2) = 0.25
with p-value = 0.88

In Model 2, a cash cap variable greater than 5,000 euro is the only threshold to be statistically significant. The estimated results for the other variables are consistent with economic theory since they have an effect on the tax gap trend. In fact, the variables corporate tax burden and the efficiency of a country’s tax system (**ETPS**) show an inverse relation (-0.11), while the tax burden on individuals (0.08) has a positive sign. Also, positive, and statistically significant are the coefficients recorded for the number of card payments (**card**) and its interaction with the cash threshold above €5.000 (**card*d3**). For the latter two variables, however, the net effect on the tax gap is lower, although these effects should be assessed by considering the log transformation of y.

In this type of analysis, the main methodological problem will be any omitted variables and the consequent use of a misspecification model (Ahsan et al., 1992; Clarke, 2005; Lütkepohl, 1982; Pace and LeSage, 2010; Wooldridge, 2003). The verification of the lack of correlation between residuals, and the lack of systematic relationships between them and the regressors (Stock and Watson, 2019) (Figure 6a and b) is therefore necessary. For simplicity, the related test statistics were omitted.

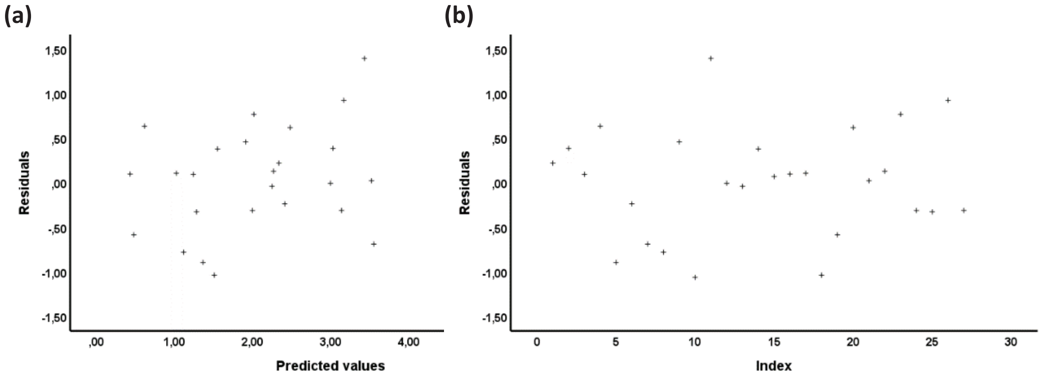


Figure 6 - Relation residuals and estimated values (a) and residuals and country index (b) for Model 2

4. CONCLUSIONS

The situation of countries with regard to their taxation systems appears quite differentiated, both in terms of types and levels of the taxes applied and the efficiency of their respective tax systems. Finding a cause or a theory that can provide a comprehensive explanation for the existing heterogeneity appears difficult. However, the unobserved economy, divided internally into underground, informal and illegal, is included in a country's GDP calculation, even if the methods of estimation used vary. If tax evasion is analyzed as an aspect of the shadow economy, it is possible to grasp the relevance of the phenomenon that can, among other things, affect the competitiveness of economic systems and their productivity levels. It is necessary, however, to understand whether the existence of possible limits to cash payments can contribute to reducing the tax gap or whether they represent an irrelevant element in comparison to a model of taxation that sees the taxpayer as a person to whom services are offered, rather than merely one from whom emoluments are requested. Taxation must be perceived as non-hostile and fair by citizens. It needs to be seen as an instrument of support for those who produce income.

The cash cap, a measure that, as demonstrated, is not in place in all countries, may find a limitation in its application in that an aging population is often characterized by insufficient ICT literacy. However, it is important to note that the preference for cash transactions in some regions may be rooted in a stronger historical connection to traditional payment methods as compared to practices in other countries. The descriptive analysis of those countries that have introduced a cash limit, conducted by means of graphical representation, showed a weak inverse relationship between the tax gap estimate and the cash limit. The econometric analysis carried out on the 27 countries of the EU provides a useful reflection for identifying a policy for tackling the tax gap. Considered analytically, in Model 1 no threshold for the use of cash payment was significant, whereas in Model 2, in which other variables were also included, a threshold above 5,000 euro became statistically significant. The results estimated for the other variables are consistent with economic theory since they influence the tax gap trend. The variables corporate tax burden and the efficiency of a country's tax system (**ETPS**) showed an inverse link (-0.11), while the tax burden on individuals (0.08) had a positive sign. Also, positive, and statistically significant were the coefficients recorded for the number of card payments (**card**) and its interaction with the cash threshold above 5,000 euro (**card*d3**). For these last two variables, however, their net effect on the tax gap was lower, although these effects should be assessed by considering the log transformation of the y performed.

There is no doubt that tax evasion depends on several factors, not least on the level of education and the importance people attribute to the public good, but it also reflects the trust that citizens place in institutions. The danger of misspecification of the model due to omitted variables has been verified, however, and therefore the model is correctly specified. The introduction of the cash cap can only be justified when it is above 5,000 euro. One of the most useful approaches in an attempt to reduce tax evasion could be to set up a tax policy that is as homogeneous as possible across countries. The existence of different rates, particularly on the taxation of corporate profits, can introduce significant distortions into the functioning of economic systems.

References

Andrews, D., Sánchez, A. C., and Johansson, Å. (2011). *Towards a Better Understanding of the Informal Economy*. OECD Publishing, Paris.

Ahsan, S. M., Kwan, A. C., and Sahni, B. S. (1992). Public expenditure and national income causality: Further evidence on the role of omitted variables. *Southern Economic Journal*, 58(3): 623–634.

Brill, A., and Hassett, K. A. (2007). *Revenue-Maximizing Corporate Income Taxes: The Laffer Curve in OECD countries*. https://ssrn.com/abstract=2235697. Last access: 29/01/2024

Clarke, K. A. (2005). The phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science*, 22(4): 341–352.

Deléchat, C., and Medina, L. (2021). *What Do We Know about the Informal Economy? The Global Informal Workforce*, 1. IMF Library.

Dell'Anno, R. (2021). Theories and definitions of the informal economy: A survey. *Journal of Economic Surveys*. *36*(5): 1610-1643.

Elgin, C. (2020). *The Informal Economy: Measures, Causes, and Consequences*. Routledge.

Elgin, C., Kose, M. A., Ohnsorge, F., and Yu, S. (2021). *Understanding Informality.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3916568. Last access: 29/01/2024

Elgin, C., and Oztunali, O. (2014). Institutions, informal economy, and economic development. *Emerging Markets Finance and Trade*. 50(4): 145–162.

European Central Bank (2021a). *Study on the Payment Attitudes of Consumers in the Euro Area (SPACE)*. https://www.ecb.europa.eu/pub/pdf/other/ecb.spacereport202012~bb2038bbb6.en.pdf. Last access: 29/01/2024.

European Central Bank (2021b). *Payments Statistics: 2021*. https://www.ecb.europa.eu/press/pr/stats/paysec/html/ecb.pis2021~956efe1ee6.en.html. Last access: 29/01/2024.

European Commission (2016). *Communication from the Commission to the European Parliament and the Council on an Action Plan for strengthening the Fight Against Terrorist Financing*. https://eur-lex.europa.eu/resource.html?uri=cellar:e6e0de37-ca7c-11e5-a4b5 01aa75ed71a1.0002.02/DOC_1&format=PDF. Last access: 29/01/2024.

European Commission (2018). *EU VAT Gap Report*. https://taxation-customs.ec.europa.eu/sites/default/files/vat-gap-full-report-2020_en.pdf. Last access: 29/01/2024.

European Consumer Centres Network (2022). *Cash Payment Limitations*. https://www.europe-consommateurs.eu/en/shopping-internet/cash-payment-limitations.html. Last access: 29/01/2024.

Forbis, S. M. (2019). *Examining and Protecting Senior Citizens from Elder Financial Exploitation within the Digital World*. Doctoral dissertation, Utica College.

Ihrig, J., and Moe, K. S. (2004). Lurking in the shadows: The informal sector and government policy. *Journal of Development Economics*. 73(2): 541–557.

ISTAT (Istituto Nazionale di Statistica) (2020). *The unobserved Economy in the National Accounts. Years 2015-2018 Report*. https://www.istat.it/it/files//2020/10/Economia-non-osservata-nei-conti-nazionali.pdf. Last access: 29/01/2024.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2020). *An introduction to statistical learning with R*. Piccin-Nuova Libraria.

Kelmanson, B., Kirabaeva, K., Medina, L., Mircheva B., and Weiss J. (2019). *Explaining the shadow economy in Europe: Size, causes and policy options*. International Monetary Fund.

Lewis, A. W. (1955). *Theory of economic growth* (1st ed.). Routledge, London.

Lütkepohl, H. (1982). Non-causality due to omitted variables. *Journal of Econometrics*. 19(2-3): 367–378.

Marinescu, C., and Valimareanu, I. (2019). The mai current (schools) of thoughts about the informal economy. *Review of International Comparative Management*. 17(5): 310–316.

Medina, L., and Schneider, F. (2017). *Shadow economies around the world: What did we learn over the last 20 years?*. International Monetary Fund.

Morales, A. (1997). Epistemic reflections on the informal economy. International *Journal of Sociology and Social Policy*. 17(3/4): 1–17.

Murphy, R. (2019). *The European tax gap*. *A report for the Socialists and Democrats Group in the European Parliament*. Global Policy.

OECD et al. (2002). *Measuring the Non-Observed Economy: A Handbook*. OECD Publishing, Paris.

Pace, R. K., and LeSage, J. P. (2010). Omitted variable biases of OLS and spatial lag models. In *Progress in spatial analysis: Methods and applications*. Springer, Berlin, Heidelberg.

Schneider, F. (2011). *Handbook on the Shadow Economy*. Edward Elgar, Northampton USA.

Schneider, F., Buehn, A., and Montenegro, C. E. (2010). New estimates for the shadow economies all over the world. *International Economic Journal*. 24(4): 443–461.

Stock, J. H., and Watson, M. W. (2019). *Introduction to econometrics* (4th Edition). Pearson Education Limited.

Wooldridge, J. M. (2003). *Diagnostic testing. A companion to theoretical econometrics*. Blackwell Publishing Ltd.

World Bank Data and PwC (PricewaterhouseCoopers) (2020). Paying Taxes 2020. Report.https://archive.doingbusiness.org/content/dam/doingBusiness/pdf/db2020/PayingTaxes2020.pdf. Last access: 29/01/2024.

United Nations. (1993). *System of national accounts*. UN.

doi.org/10.26398/IJAS.0036-04

A COMPARATIVE STUDY ON UNIVARIATE OUTLIER WINSORIZATION METHODS IN DATA SCIENCE CONTEXT

Ali Abuzaid[1]
Department of Mathematics, Al Azhar University - Gaza, Gaza, Palestine.

Iyad Alkrunz
Department of Information Technology, Al Azhar University - Gaza, Gaza, Palestine.

Abstract *Handling outliers is an important step in data analysis, and it can be approached through three different ways, namely; accommodation, omission, or winsorization. This article investigates the impact of four winsorization statistics (mean, median, mode, and quantiles) on parameter estimation through an extensive simulation study. Three probability distributions (normal, negative binomial, and exponential) are considered, each with varying degrees of contamination. The simulation results suggest that winsorization is effective for small contamination levels and large sample sizes. Furthermore, it is recommended to winsorize outliers in symmetric distributions using any of the location parameters. However, for asymmetric distributions, the median should be employed. To illustrate these findings, a real dataset on internet usage session durations for 4,500 users, comprising over 2 million records, are fitted to the exponential distribution. The identified outliers were winsorized using the aforementioned statistics.*

Keywords: *Capping; flooring; outlier; quantile-based.*

1. Introduction

Outliers refer to data values that significantly deviate from the majority of the data. The presence of outliers can have a detrimental impact on the effectiveness and accuracy of a predictive model, as they have the potential to skew estimations. Outliers can arise due to various factors, such as incorrect measurements, data entry errors, or sampling from a different population (Frost, 2020). Consequently, the issue of outlier-detection has garnered considerable attention from statisticians and data scientists.

The methods of outlier-detection are broadly classified into different classes, namely distribution-based methods, depth-based methods, and density-based methods (Preparata and Shamos, 1988, Dominguesa, et al 2018).

[1] Email: a.abuzaid@alazhar.edu.ps

The argument on the handling of outliers is continued between the belief of Tukey (1960) that rejecting outliers indiscriminately is inappropriate, and other various trimming and winsorization techniques. Thus, after detection, outliers can be handled in one of three ways: accommodation, omission, or winsorization.

Accommodation is employed by robust statistical methods to mitigate the impact of outliers on parameter estimates (Ekezie and Ogu, 2013). Outliers have the potential to undermine the conclusions of a study (Hubert et al., 2008; Farcomeni and Ventura, 2010), and thus, accommodation techniques are utilized to indirectly counteract their influence. The trimming of outliers has been extensively stud-ied, and researchers such as Lix and Keselman (1998) and Yusof et al. (2013) have proved its benefits in terms of improving robustness. Additionally, the topic of trimming, including discussions on the type (symmetric or asymmetric) and percentage of trimming, has been addressed by Babu et al. (1999) and Wilcox (2003).

In winsorization, extreme values are substituted with suitable values to mitigate the impact of outliers on estimators and modeling power (Frey, 2018). These substitute values can be any of the central tendency measures as outlined in Section 2. However, determining the appropriate winsorization percentage cut-off point and the winsorization statistic can pose challenges.

A poor choice of winsorization percentage will inflate the mean squared errors (MSE) of desired estimators. Thus, it is recommended to choose the cut-off point that minimizes the MSE compared to the classical estimator. Winsorization is recommended to avoid the loss of power (Leys, et al, 2019). Moreover, Liao et al (2017) highlighted the effectiveness of winsorization in controlling Type I error inflation and outlier impact on power based on a simulation study.

In the context of data science, practitioners used different statistics for winsorization, such as mean, median and quantiles. To the best of our knowledge, no published study has specifically examined the impact of different winsorization statistics on estimators. This article investigates the impact of four winsorization statistics viz mean, median, mode and quantile-based flooring and capping technique on the estimates of parameters of three distributions, namely normal, negative binomial and exponential distribution.

2. Outliers and Winsorization

2.1. Outliers Detection

Various methods exist for identifying outliers, such as square root transformation, median absolute deviation, Grubb’s test, and Ueda’s method, as recently discussed by Shimizu (2022). However, in this article, we use Tukey's method boxplot (1977) due to its popularity and less sensitivity of outliers' existence compared to other tests.

Boxplot is a well-known simple graphical tool to display information about continuous univariate data based on five summaries, namely, median, lower quartile Q_1, upper quartile Q_3, lower extreme, and upper extreme of a data set. Any value smaller than the lower fence $L_F = Q_1 - \nu * IQR$ or larger than the upper fence $U_F = Q_3 + \nu * IQR$ is an outlier candidate, where ν is the resistance factor and $IQR = Q_3 - Q_1$ is the interquartile range. Different values of ν can be considered, but the nominal value is $\nu = 1.5$ (Hoaglin et al, 1986). Various versions of the boxplot were also proposed (see Abuzaid et al; 2012, Saeger et al; 2016).

The following subsection discusses the treatment of outliers via winsorization.

2.2. Winsorization of outliers

The winsorization method involves replacing outlier values with a suitable statistic such as mean, median, mode or quantile-based technique as follows:

1. $Replacing\ outliers\ by\ mean$: In this technique, outliers are replaced with the arithmetic mean of the remaining observations after removing outliers.

2. $Replacing\ outliers\ by\ median$: The median value, which is the middle value of an ordered remaining observations, is used to replace the detected outliers.

3. $Replacing\ outliers\ by\ mode$: Outliers are replaced with the mode value of the remaining observations.

4. $Quantile-based\ Flooring\ and\ Capping$: in this quantile-based technique, the maximum outliers are replaced with the upper fence, U_F (capped), and the minimum outliers are replaced with the lower fence, L_F (floored).

The following section investigates the effect of the previous four considered winsorization statistics on the performance of parameter estimates for different probability distributions via a Monte Carlo simulation study.

3. Simulation

An R code has been developed to generate random datasets from three different probability distributions, namely, normal, negative binomial and exponential distribution.

3.1. Settings of Data Generation

Data were generated with four different sample sizes, n = 20, 50, 100 and 200, in such a way that $(1-\varepsilon)$ of data are generated from the original distribution (P) and the rest ε of data are generated from the contamination distribution (C). Thus, the contaminated data structure can be formulated as $P_\varepsilon = (1-\varepsilon)P+\varepsilon C$, where ε is the contamination level and ε = 0.05, 0.10, 0.15 and 0.20. The following three probability distributions are considered:

3.1.1. Normal distribution

Let X be a random variable having the normal distribution, with mean $-\infty < \mu < \infty$ and standard deviation $\sigma > 0$, $X \sim N(\mu,\sigma^2)$. The datasets were generated from the standard normal distribution with $\mu = 0$ and $\sigma = 1$. For contamination procedure, the contaminated data were generated from another normal distribution with $\mu = 4$ and $\sigma = 2$.

The maximum likelihood estimator (MLE) of the mean and standard deviation are obtained as the sample mean $\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and $\hat{\sigma} = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$, respectively. Moreover, the least squares estimation method is equivalent to the MLE, where both are sensitive to the presence of outliers.

3.1.2. Negative binomial distribution

Let X be a random variable having the negative binomial distribution, $X \sim NB(k,p)$ with mean, $\mu = \frac{k}{p}$ and variance $\sigma^2 = \frac{k(1-p)}{p^2}$, where X is the count of independent Bernoulli trials are required to achieve the k^{th} successful trials when the probability of success is a constant p, and $p \in [0,1]$. The probability of $f(X=x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}$, for $x = k, k+1, k+2, \ldots$ and $k = 0,1,2,\ldots$ The MLE of p is given by $\hat{p} = \frac{k}{x+k}$.

For the negative binomial random variable, data are generated with parameters $k = 2$ and $p = 0.2$, while the contaminated data are generated from a Poisson distribution with $\lambda = 32$, where the probability of k successes is $P(X=k) = \frac{(e^{\lambda}\lambda^k)}{k!}$.

3.1.3. Exponential distribution

The exponential distribution is the most commonly used model in reliability and life-testing analysis. The probability density function of a random variable X, having the exponential distribution is given by $f(x) = \theta e^{-\theta x}$ for $x \geq 0$ and $\theta > 0$.

The *MLE* of θ is given by $\hat{\theta} = \frac{1}{\bar{x}}$, where $\bar{x}$ is the sample mean.
Data were generated from the exponential distribution with parameter $\theta = 0.5$, and the contaminated data were generated from exponential distribution with $\theta = 0.05$.

For each combination of probability distributions, sample sizes, contamination levels and winsorization statistics, the generation procedure is repeated 1000 iterations to ensure the convergence.

3.2. Results

The impact of the four outliers winsorization statistics on the parameter estimators is measured by three common indicators as follows:

1. *Bias*, it is the difference between the estimator's expected value and the true value of the parameter being estimated.

2. $Mean\ Square\ Error$, $MSE = \frac{1}{1000}\sum_{i=1}^{1000}(\beta - \hat{\beta}_i)^2$, where β and $\hat{\beta}_i$ are the true and estimated values of the considered parameters.

3. $Goodness - of - fit\ tests$, are statistical tests aiming to determine whether a set of observed values match those expected under the applicable distribution. There are different goodness-of-fit tests, in this article the Shapiro-Wilk test is used in the case of normal distribution and exponential distri-bution, while the Kolmogorov-Smirnov test is used in the case of negative binomial distribution.

The simulation results are summarized in Tables (1-5). Regardless of the distribution, contamination level, or winsorization statistics employed, the simulation study reveals that the performance of parameter estimators improves with larger sample sizes. Specifically, the mean squared error (MSE) and bias exhibit an inverse relationship with the sample size ($n < 100$), while they stabilize as a constant function for $n \geq 100$. This relationship is partially illustrated in Figure 1.

The performance has a relatively inverse relationship with the contamination level (ε).

For the normal distribution, due to its symmetric nature, the mean, median and mode winsorization statistics have an almost similar effect on the estimators of the parameters (i.e., μ and σ^2), while they outperform the quantile-based winsorization statistic as given in Tables (1-2).

For the negative binomial distribution, the mode winsorization statistic slightly outperforms the other winsorization statistics for higher levels of contamination

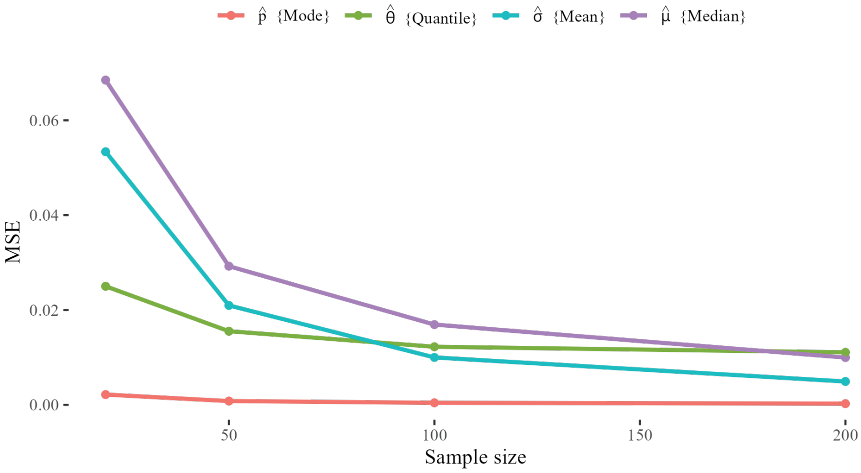


Figure 1: MSE of different parameters' estimators after using winsorization methods for different sample sizes

$(\varepsilon \geq 0.15)$, while the mean winsorization statistic performs better than other winsorization statistics for smaller levels of contamination $(\varepsilon < 0.15)$ as presented in Table 3.

For the exponential distribution (Table 4), the mean winsorization statistic has the best performance, followed by the median, mode and then the quantile-based method. This behavior may be referred to the properties of the *MLE* estimator of the parameter (θ), which is mainly the sample mean.

Table 5 presents the proportion of fitted samples by the associated distributions at 0.05 level of significance before winsorization $(\varepsilon = 0)$, where the proportions are close to 0.95 for the considered sample sizes and probability distributions. The proportions of fitted samples have an inverse relationship with the contamination level. The quantile-based winsorization statistic has the worst performance compared to the other three considered statistics because it accumulates the winsorizated values at the edges of the distribution and malforms the nature of the distribution. Thus, the mean winsorization statistic is recommended for most of the cases, especially for smaller levels of contamination $(\varepsilon \leq 0.1)$.

For normal distribution, mean, median and mode winsorization statistics have consistent performance with respect to the contamination level and sample size,

Table 1: Bias (MSE) of the normal distribution's mean estimator for different winsorization methods

		Winsorization methods			
n	ε	Quantile-based	Mean	Median	Mode
20	0	0.005 (0.053)	0.005 (0.059)	0.005 (0.059)	0.005 (0.059)
50	0	0 (0.021)	0.001 (0.023)	0.001 (0.023)	0.002 (0.023)
100	0	0.003 (0.01)	0.004 (0.01)	0.004 (0.01)	0.004 (0.01)
200	0	0.001 (0.005)	0.001 (0.005)	0.001 (0.005)	0.001 (0.005)
20	5	0.111 (0.06)	0.021 (0.058)	0.02 (0.058)	0.02 (0.058)
50	5	0.092 (0.027)	0.019 (0.021)	0.019 (0.021)	0.019 (0.021)
100	5	0.122 (0.025)	0.029 (0.012)	0.029 (0.012)	0.028 (0.012)
200	5	0.12 (0.02)	0.027 (0.007)	0.027 (0.007)	0.026 (0.007)
20	10	0.24 (0.111)	0.074 (0.068)	0.074 (0.068)	0.074 (0.069)
50	10	0.245 (0.081)	0.068 (0.029)	0.067 (0.029)	0.066 (0.029)
100	10	0.244 (0.07)	0.068 (0.017)	0.066 (0.017)	0.065 (0.017)
200	10	0.245 (0.065)	0.064 (0.01)	0.062 (0.01)	0.061 (0.01)
20	15	0.353 (0.18)	0.113 (0.08)	0.111 (0.08)	0.108 (0.082)
50	15	0.399 (0.182)	0.14 (0.049)	0.136 (0.048)	0.132 (0.048)
100	15	0.378 (0.154)	0.121 (0.028)	0.117 (0.027)	0.113 (0.026)
200	15	0.378 (0.149)	0.119 (0.022)	0.115 (0.021)	0.111 (0.02)
20	20	0.489 (0.293)	0.209 (0.118)	0.204 (0.115)	0.204 (0.115)
50	20	0.497 (0.271)	0.189 (0.067)	0.183 (0.064)	0.179 (0.062)
100	20	0.509 (0.271)	0.193 (0.054)	0.186 (0.051)	0.179 (0.049)
200	20	0.515 (0.271)	0.197 (0.047)	0.19 (0.044)	0.184 (0.042)

where the proportions of the fitted samples by normal distribution are close to 1 when the contamination level is $(\varepsilon = 0.05)$. In the case of an exponential distribution, all considered winsorization statistics perform approximately the same, where the proportions of fitted samples by exponential distribution are close to 1 regardless the sample size or contamination level.

The proportions of fitted samples by negative binomial distribution are less than the other two distributions.

4. Application

A dataset on internet usage was obtained from the Ministry of Telecom and Information Technology in Palestine. The dataset comprises more than 2 mil-

Table 2: Bias (MSE) of the normal distribution's standard deviation estimator for different winsorization methods

		Winsorization methods			
n	*ε*	Quantile-based	Mean	Median	Mode
20	0	0.037 (0.027)	0.074 (0.042)	0.074 (0.042)	0.073 (0.042)
50	0	0.02 (0.011)	0.054 (0.017)	0.054 (0.017)	0.053 (0.017)
100	0	0.011 (0.005)	0.039 (0.009)	0.039 (0.009)	0.039 (0.008)
200	0	0.009 (0.003)	0.037 (0.005)	0.037 (0.005)	0.037 (0.005)
20	5	0.095 (0.047)	0.078 (0.05)	0.077 (0.05)	0.072 (0.049)
50	5	0.094 (0.022)	0.032 (0.017)	0.032 (0.017)	0.029 (0.016)
100	5	0.125 (0.023)	0.026 (0.01)	0.026 (0.01)	0.024 (0.01)
200	5	0.123 (0.019)	0.023 (0.005)	0.023 (0.005)	0.022 (0.005)
20	10	0.226 (0.101)	0.014 (0.053)	0.013 (0.053)	0.007 (0.053)
50	10	0.25 (0.081)	0.005 (0.021)	0.005 (0.021)	0.001 (0.021)
100	10	0.252 (0.073)	0.006 (0.01)	0.006 (0.01)	0.009 (0.01)
200	10	0.255 (0.07)	0.005 (0.005)	0.005 (0.005)	0.007 (0.005)
20	15	0.35 (0.183)	0.004 (0.064)	0.005 (0.064)	0.015 (0.065)
50	15	0.401 (0.188)	0.062 (0.036)	0.063 (0.036)	0.069 (0.036)
100	15	0.387 (0.162)	0.048 (0.016)	0.049 (0.016)	0.053 (0.016)
200	15	0.392 (0.159)	0.055 (0.01)	0.055 (0.01)	0.059 (0.01)
20	20	0.487 (0.307)	0.131 (0.109)	0.132 (0.11)	0.142 (0.111)
50	20	0.514 (0.295)	0.125 (0.054)	0.126 (0.054)	0.132 (0.056)
100	20	0.526 (0.291)	0.129 (0.035)	0.13 (0.035)	0.135 (0.037)
200	20	0.532 (0.29)	0.137 (0.028)	0.137 (0.028)	0.141 (0.029)

Table 3: Bias (MSE) of the negative binomial distribution probability of success estimator for different winsorization methods

		Winsorization methods			
n	ε	Quantile-based	Mean	Median	Mode
20	0	0.006 (0.001)	0.019 (0.002)	0.02 (0.002)	0.021 (0.002)
50	0	0.004 (0.000)	0.016 (0.001)	0.017 (0.001)	0.018 (0.001)
100	0	0.003 (0.000)	0.015 (0.000)	0.016 (0.001)	0.017 (0.001)
200	0	0.002 (0.000)	0.013 (0.000)	0.014 (0.000)	0.015 (0.000)
20	5	0.013 (0.001)	0.017 (0.002)	0.018 (0.002)	0.019 (0.002)
50	5	0.014 (0.000)	0.011 (0.001)	0.012 (0.001)	0.014 (0.001)
100	5	0.017 (0.000)	0.01 (0.000)	0.011 (0.000)	0.014 (0.001)
200	5	0.018 (0.000)	0.008 (0.000)	0.01 (0.000)	0.013 (0.000)
20	10	0.031 (0.001)	0.005 (0.002)	0.007 (0.002)	0.009 (0.002)
50	10	0.034 (0.001)	0.001 (0.001)	0.003 (0.001)	0.006 (0.001)
100	10	0.034 (0.001)	0.002 (0.000)	0.000 (0.000)	0.004 (0.000)
200	10	0.034 (0.001)	0.001 (0.000)	0.001 (0.000)	0.006 (0.000)
20	15	0.045 (0.002)	0.006 (0.002)	0.004 (0.002)	0.001 (0.002)
50	15	0.049 (0.002)	0.019 (0.001)	0.017 (0.001)	0.015 (0.001)
100	15	0.047 (0.002)	0.015 (0.001)	0.013 (0.001)	0.009 (0.001)
200	15	0.046 (0.002)	0.016 (0.000)	0.014 (0.000)	0.01 (0.000)
20	20	0.056 (0.003)	0.03 (0.002)	0.029 (0.002)	0.027 (0.002)
50	20	0.056 (0.003)	0.034 (0.002)	0.032 (0.002)	0.03 (0.002)
100	20	0.056 (0.003)	0.037 (0.002)	0.035 (0.002)	0.032 (0.002)
200	20	0.056 (0.003)	0.04 (0.002)	0.038 (0.002)	0.036 (0.002)

Table 4: Bias (MSE) of the exponential distribution's rate estimator for different winsorization methods

		Winsorization Methods			
n	ε	Quantile-based	Mean	Median	Mode
20	0	0.004 (0.016)	0.116 (0.049)	0.127 (0.054)	0.147 (0.064)
50	0	0.013 (0.006)	0.101 (0.022)	0.11 (0.025)	0.127 (0.031)
100	0	0.018 (0.003)	0.094 (0.015)	0.102 (0.017)	0.118 (0.021)
200	0	0.023 (0.002)	0.092 (0.012)	0.1 (0.013)	0.116 (0.017)
20	5	0.069 (0.017)	0.077 (0.033)	0.092 (0.037)	0.115 (0.045)
50	5	0.039 (0.007)	0.08 (0.018)	0.093 (0.021)	0.116 (0.028)
100	5	0.045 (0.004)	0.07 (0.01)	0.082 (0.012)	0.107 (0.018)
200	5	0.043 (0.003)	0.066 (0.007)	0.078 (0.009)	0.104 (0.014)
20	10	0.13 (0.025)	0.055 (0.024)	0.076 (0.029)	0.106 (0.038)
50	10	0.108 (0.016)	0.052 (0.012)	0.07 (0.015)	0.103 (0.022)
100	10	0.102 (0.012)	0.051 (0.007)	0.069 (0.009)	0.103 (0.016)
200	10	0.101 (0.011)	0.047 (0.004)	0.063 (0.006)	0.098 (0.012)
20	15	0.179 (0.038)	0.038 (0.018)	0.064 (0.022)	0.1 (0.032)
50	15	0.151 (0.026)	0.042 (0.01)	0.065 (0.013)	0.103 (0.022)
100	15	0.159 (0.026)	0.03 (0.004)	0.053 (0.007)	0.096 (0.014)
200	15	0.156 (0.025)	0.029 (0.003)	0.051 (0.005)	0.095 (0.012)
20	20	0.227 (0.057)	0.035 (0.019)	0.066 (0.024)	0.11 (0.036)
50	20	0.213 (0.048)	0.029 (0.008)	0.058 (0.011)	0.11 (0.022)
100	20	0.211 (0.045)	0.018 (0.004)	0.046 (0.006)	0.097 (0.014)
200	20	0.209 (0.044)	0.016 (0.002)	0.044 (0.004)	0.097 (0.012)

Table 5: The proportion of fitted samples by associated distributions at 0.05 level of significance.

Distribution		Normal distribution				Exponential distribution				Negative binomial distribution			
n	ε	Qun.	Mean	Med	Mode	Qun.	Mean	Med.	Mode	Qun.	Mean	Med.	Mode
20	0	(0.970)				(0.964)				(0.922)			
50	0	(0.964)				(0.935)				(0.930)			
100	0	(0.944)				(0.939)				(0.924)			
200	0	(0.951)				(0.941)				(0.938)			
20	5	0.961	0.970	0.949	0.924	0.999	1.000	0.996	0.984	0.748	0.912	0.878	0.848
50	5	0.926	0.970	0.965	0.958	1.000	1.000	1.000	0.997	0.442	0.682	0.606	0.538
100	5	0.614	0.968	0.960	0.934	1.000	1.000	1.000	1.000	0.294	0.450	0.418	0.356
200	5	0.137	0.962	0.953	0.916	1.000	1.000	1.000	1.000	0.152	0.254	0.200	0.198
20	10	0.873	0.951	0.939	0.910	0.997	0.998	0.986	0.964	0.582	0.846	0.792	0.782
50	10	0.474	0.938	0.918	0.879	0.998	0.999	0.996	0.984	0.256	0.542	0.488	0.380
100	10	0.060	0.890	0.873	0.805	0.998	1.000	1.000	0.998	0.090	0.352	0.314	0.234
200	10	0.000	0.777	0.752	0.664	1.000	1.000	1.000	0.997	0.008	0.258	0.192	0.110
20	15	0.743	0.937	0.908	0.868	0.994	0.994	0.968	0.920	0.514	0.770	0.720	0.700
50	15	0.108	0.785	0.737	0.674	0.992	1.000	0.990	0.951	0.128	0.428	0.346	0.296
100	15	0.006	0.666	0.617	0.540	0.988	0.999	0.995	0.945	0.034	0.218	0.168	0.150
200	15	0.000	0.251	0.217	0.146	0.982	1.000	0.999	0.944	0.000	0.146	0.116	0.064
20	20	0.551	0.839	0.802	0.740	0.956	0.989	0.951	0.874	0.414	0.646	0.580	0.592
50	20	0.043	0.631	0.576	0.518	0.931	0.998	0.988	0.866	0.134	0.316	0.250	0.196
100	20	0.000	0.256	0.226	0.173	0.914	1.000	0.994	0.831	0.014	0.134	0.112	0.072
200	20	0.000	0.021	0.017	0.010	0.817	1.000	0.998	0.738	0.002	0.032	0.026	0.008

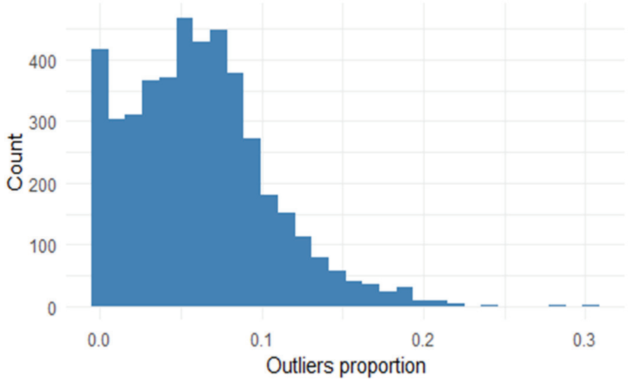


Figure 2: Histogram of detected outliers proportion

lion session records for 4,500 randomly selected users from an internet service provider company in Palestine. Each session in the dataset includes various features such as start-time, end-time, traffic, and duration.

In this example, we are interested only in sessions' durations, which are commonly hypothesized to be exponentially distributed (see Akmeroth and Ammaram, 1996, Sripanidkulchai, et al, 2004, Chetlapalli, et al, 2020). Consonance with that, we assume that sessions' duration are exponentially distributed; therefore, the sessions rows are aggregated for each user. A total of 1,416 (31.467%) of user sessions' duration have been fitted by exponential distribution at 0.05 level of significance according to Shapiro-Wilk goodness-of-fit test.

The outliers of sessions' duration for each user have been detected. Figure 2 presents the proportions of detected outliers for each user, it ranges between 0% and 30%, with mean of 6% and it is an obvious positively skewed distribution.

Three winsorization methods are applied to users' sessions duration data, which are identified as outliers. The summary of fitted users before and after winsorization is presented in Table 6. The results show that the proportion of fitted users data after winsorization is increased significantly, where the mean has the highest proportion, followed by the median and then the quantile-based method which are consistent with the findings of the simulation study. The Chi-square test of independence shows that there are significant associations between the status of users' sessions duration data (i.e fitted by exponential distribution) before and after winsorization at 0.05 level of significance. These associations reveal that an insignificant number of the exponentially fitted users data before winsorization

Table 6: Summary of fitted users data before and after winsorization by exponential distribution

		After Outliers Winsorization		
Statistics	Before	Mean	Median	Quantile-based
Proportion of fit	0.31	0.67	0.61	0.56
Chi-square test	-	1011.42	1311.84	1632.54
p-value	-	0.00	0.00	0.00

has been alternated to be not fitted by exponential after winsorization has been conducted.

5. Conclusions

The winsorization techniques to handle outliers in univariate data have been evaluated via a simulation study. The findings revealed that the nature of the data, including its distribution shape, sample size and contamination level, are the key factors. Thus, it is recommended to use winsorization techniques for large samples $(n \geq 100)$ with a small level of contamination $(\varepsilon \leq 0.1)$. In the case of symmetric distributions, any of the central tendency measures can be used, while for the asymmetric distributions, the use of the median is recommended. This article has focused on three commonly used probability distributions: normal, negative binomial, and exponential. However, further studies could explore other univariate and multivariate distributions to gain a more comprehensive understanding. Additionally, robust statistics remain a viable alternative to winsorization, and it would be valuable to compare their performance in outlier handling techniques.

References

Abuzaid, AH., Mohamed, IB. and Hussin, A.G. (2012). Boxplot for circular variables. *Computational Statistic*s. 27 (3), 381-392.

Almeroth, KC., and Ammarm, MH. (1996). Collecting and modeling the join/leave behavior of multicast group members in the MBone. In *Proceedings of International Symposium on High Performance Distributed Computing* (HPDC).

Babu, G.J., Padmanabhan, A.R. and Puri ML (1999). Robust one-way ANOVA under possibly non regular conditions. *Biometrical Journal*, 41: 321-339.

Chetlapalli, V., Iyer, K.S.S. and Agrawal, H. (2020) Modelling time-dependent aggregate traffic in 5G networks. *Telecommun Syst* 73, 557-575. https://doi.org/10.1007/s11235-019-00629-w

Dominguesa R, Filipponea M, Michiardia P and Zouaouib J. (2018) A Comparative Evaluation of Outlier Detection Algorithms: *Experiments and Analyses*, Pattern Recognition 74: 406-421.

Ekezie, D.D., and Ogu, A.I. (2013) Statistical Analysis/Methods of Detecting Outliers in Univariate Data in A Regression Analysis Model. *International Journal of Education and Research*, 1(5): 1-24.

Frey B (2018). *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation* (Vols. 1-4). Thousand Oaks,, CA: SAGE Publications, Inc. doi: 10.4135/9781506326139

Frost, J. (2020). Hypothesis testing: An intuitive guide for making data drives decisions. Statistics by Jim Publishing State College, Pennsylvania, U.S.A.

Hoaglin, D.C., Iglewicz B. and Tukey, J.W. (1986) Performance of some resistant rules for outlier labeling. *J Am Stat Assoc* 81(396):991-999

Hubert, M., Rousseeuw, P.J. and Van Aelst, S. (2008). High-breakdown Robust Multivariate Methods. *Statistical Science*, 23(1): 92-119.

Kwak, S.K. and Kim, J.H. (2017) Statistical data preparation: management of missing values and outliers. *Korean Journal of Anesthesiology* 70(4): 407-411.

Leys, C., Delacre, M., Mora, Y.L., Lakens, D. and Ley, C. (2019). How to classify, detect, and manage univariate and multivariate outliers, with emphasis on preregistration. *International Review of Social Psychology*, 32(1): 5, 1-10.

Liao, H, Yanju, Li and Brooks, GP (2017). Outlier impact and accommodation on power. *Journal of Modern Applied Statistical Methods*, 16(1): 261-278.

Lix, L.M. and Keselman, H.J. (1998). To trim or not to trim: Tests of location equality under heteroscedasticity and non-normality. *Educational and Psychological Measurement*, 115: 335-363.

Nyitrai T and Miklos M (2019) The effects of handling outliers on the performance of bankruptcy prediction models. *Socio-Economic Planning Sciences*, 67, 34-42.

Preparata, F., Shamos, M. (1988) *Computational Geometry: An Introduction*, Springer-Verlag, Berlin.

Saeger, T., Kleven, B., Otero, I., Wallace, M. and Ziglar, R. (2016) Outlier labeling method for univariate data for module test and die sort. *IEEE transactions on semiconductor manufacturing*, 29(4): 330-335.

Shimizu, Y. (2022) Multiple desirable methods in outlier detection of univariate data with R source codes. *Front. Psychol*. 12:819854.

Sripanidkulchai, K., Maggs, B. and Zhang, H. (2004). An analysis of live streaming workloads on the internet. In *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement* (IMC'04). Association for Computing Machinery, New York, NY, USA, 41-54. DOI:https://doi.org/10.1145/1028788.1028795

Tukey, J.W. (1977) *Exploratory Data Analysis*. Addison-Wesley, Reading.

Tukey, J.W. (1960). *A Survey of Sampling from Contaminated Distributions.* Princeton, New Jersey: Princeton University.

Wilcox, R.R. (2003). *Applying Contemporary Statistical Techniques*. Academic Press: San Diego, CA.

Yusof, Z.M., Othman, A.R. and Syed Yahaya, S.S. (2013). Robustness of Trimmed F statistics when handling nonnormal data. *Malaysian Journal of Science*, 32(1): 73-77.

doi.org/10.26398/IJAS.0036-005

DETERMINANTS OF STUDENTS' ATTITUDE TOWARDS ONLINE LEARNING IN HIGHER EDUCATION DURING COVID-19

Emiliano del Gobbo[1]
Department of Economics, Management and Territory, University of Foggia, Foggia, Italy

Alfonso Guarino[2]
Department of Humanities, University of Foggia, Foggia, Italy

Barbara Cafarelli[3]
Department of Economics, Management and Territory, University of Foggia, Foggia, Italy

Lara Fontanella[4]
Department of Legal and Social Sciences, University of Chieti-Pescara, Pescara, Italy

Abstract *The emergence of online environments due to the COVID-19 outbreak has changed the landscape of educational learning. The pandemic emotionally affected students around the world: some benefited from online/blended learning, others lost motivation. In this paper, exploiting a structural equation model approach, we report on observed characteristics and latent factors impacting students' attitude toward online/blended learning during the pandemic emergence. The research was carried out at the University of Foggia, Italy, and encompassed 2,420 students after two years of COVID-19 emergency education. The study results have shown how the overall attitude of students is positive towards online learning, but several features have an essential role in influencing this attitude. The main findings are that students with higher motivation and engagement with professors are more prone to favor online learning, while students with higher engagement with classmates and worse pandemic emotional impact exhibit a lower level of satisfaction. Furthermore, considering contingency factors, commuter and working students display a more positive attitude, but the ones enrolled in scientific-technological courses show a lower level of satisfaction with online learning. Considerations and implications of these results are provided.*

Keywords: *online learning, attitude, covid-19, structural equation model*

[1] emiliano.delgobbo@unifg.it (**Corresponding author**)

[2] alfonso.guarino@unifg.it

[3] barbara.cafarelli@unifg.it

[4] lara.fontanella@unich.it

1. Introduction

The COVID-19 pandemic has been representing a complex health challenge for the entire world. To reduce the transmission of the coronavirus disease, sev-eral countries established infection prevention and control measures by limiting contact between people. Governments suggested or ordered physical distancing and movement restrictions. As a result of the COVID-19 outbreak, the educational system has moved to deliver courses online during Spring 2020 (Ali, 2020; Daniel, 2020; Hodges et al., 2020; Murphy, 2020), thus online learning rushed pervasively in students' daily life (Ali, 2020; Huang et al., 2020), introducing new organizational challenges for educational institutions (Abdullah and Kauser, 2022; Musella et al., 2022). Some universities were offering *asynchronous* classes where instructors prepare assignments or record lectures, and students can complete them at their own pace (Hodges et al., 2020). Some institutions used *synchronous* learning that occurs at a specific time via a specific medium. In Italy, online learning was expanded from Fall 2020 to Spring 2021 (Appolloni et al., 2021; Favale et al., 2020). Some Italian campuses have extended the period to Fall-Winter 2021 and Spring 2022 as well. Although, during the pandemic, online learning has represented a valid alternative to traditional learning, students have been more exposed to stress and difficulties compared to the traditional face to face teaching approach (Farooqui, 2020). Therefore, universities are interested in understanding how online learning impacts on their students. In general terms, by understanding students' attitudes, challenges, and preferences, universities can develop aiding strategies in case there are further waves of COVID-19 or any other disaster requiring an emergency and sudden transition to remote learning.

The shift towards online education during COVID-19 pandemic has triggered the proliferation of many studies focused on perceived learning outcomes and students' satisfaction in this new learning environment. In this context, the present paper explores Italian University students' perceptions about online learning after COVID-19 government measures ("*stay-at-home*" and/or "*physical distance*"). In particular, we focus on the case of the University of Foggia (south of Italy), a young University (about 20 years since its foundation) with about 12 thousand students.

The goal of our research is to investigate students' attitude towards online learning and the features that impact on such an attitude. Online learning attitude pertains to an individual's inclination toward exerting effort in online learning. Students' attitudes toward educational technology directly impact their learning process (Aguilera-Hermida, 2020). Botero et al. (2018) studied the factors that affect behavioral intentions and the use of mobile-assisted learning in the context of language learning. The research has shown that students' attitude significantly impacts their intention to adopt mobile technology learning. Investigating this kind of attitude is of utmost importance to elaborate and allow decision makers to make the best choices to support students' online education, lowering the barrier to education and building inclusive learning environments.

In this research, we try to answer to the following research questions:

- **RQ1.** Does pandemic emotional impact positively affect student attitude towards online learning?

- **RQ2.** Does engagement affect student attitude towards online learning?

- **RQ3.** Does student motivation affect student attitude toward online learning?

- **RQ4.** Is shyness a factor related to student attitude towards online learning?

- **RQ5.** Do exogenous variables, such as gender, age, enrollment year, student worker and commuter status affect student attitude towards online learning?

In our analysis, we assume that the unobserved constructs (i.e., latent variables) influence ordered polytomous observed variables (i.e., observed indicators) measured using Likert scales. Therefore, to address the previous research questions, we employ a generalized structural equation modeling (GSEM) framework (Muthèn, 1984). Structural equation models (SEMs; Bollen, 1989) allow researchers to model complex relationships, account for measurement error, test theoretical frameworks, and visually communicate their findings. Unlike conventional SEMs, which primarily concentrate on continuous and normally distributed data, GSEM expands its applicability to encompass a wider array of data types, including categorical, count, and even mixed types.

The paper is organized as follows. Section 2 provides an overview of the literature. Section 3 presents and offers details concerning the materials and methodology used for this research. Section 4 highlights our results and provides insights for University to improve the learning environment fitting the educational approach to the students. Lastly, Section 5 concludes the paper with final remarks and future works.

2. Literature review

This section provides an overview of previous papers in the literature that have surveyed students to understand their attitudes towards e-learning. After briefly discussing key or seminal papers on students' attitudes towards online learning, we focus on those published after the COVID-19 outbreak.

Lee et al. (2005) developed one of the first studies investigating students' acceptance of an Internet-based learning medium. The model proposed captures both extrinsic (perceived usefulness and ease of use) and intrinsic (perceived enjoyment) motivators for explaining students' intention to use the new learning medium. The results show that both perceived usefulness and enjoyment significantly and directly impact intention to use, while perceived ease of use did not significantly influence students' acceptance of Internet-based learning.

Mehra and Faranak (2012) developed an 83-item attitude towards e-learning scale on six domains, namely "Perceived usefulness", "Intention to adopt e-learning", "Ease of e-learning use", "Technical and pedagogical support", "E-learning stressors" and "Pressure to use e-learning".

Returning to our primary literature focus, we will now provide a chronological overview of studies published after the pandemic outbreak.

Hussein et al. (2020) performed a qualitative study to investigate undergraduate students' attitudes toward their experience with emergency online learning during the first few weeks of the mandatory shift to online learning caused by COVID-19. Cost and time effectiveness, safety, convenience and improved participation were the most frequently cited positive aspects. At the same time, distraction and reduced focus, heavy workload, problems with technology and the internet, and insufficient support from instructors and colleagues were the most recurrent negative aspects. Serhan (2020) investigated students' attitude towards the use of Zoom in remote learning and their perceptions of its effects on their learning and engagement in comparison to traditional face-to-face teaching. The results suggest that students held a negative attitude toward using Zoom and perceived it as having a detrimental impact on their learning experience and motivation. Flexibility was listed as the main advantage. Rafiq (2020) analyzed students' attitudes toward e-learning in higher education in Pakistan using a Technology Acceptance Model (TAM). The findings show that male students and students at higher levels of education have a more positive attitude towards e-learning. Aguilera-Hermida (2020) found out that students preferred face-to-face learning over emergency online learning. Their results show how attitude, motivation, self-efficacy, and use of technology play a significant role in students' cognitive engagement and academic performance.

Another set of papers considered a more extended period of online/remote learning. Tzafilkou et al. (2021) developed and validated a multidimensional scale to measure remote learning attitude of students. The scale was tested on students from a Greek university. The study highlights how the major impact is due to students' prior experience in distance learning and field of study, while gender and age do not show a significant influence. Law (2021) showed that most students have a positive attitude and are satisfied with online learning delivery concerning learning materials, assessments, communication, technological tools, and technical support. Dikaya et al. (2021) analyzed the relation between students' psychological traits and attitudes toward forced remote learning. The findings indicate that students with a more positive attitude toward forced re-mote learning have a higher percentage of assimilated learning materials during the lockdown. Moreover, the authors found statistically significant associations between interpersonal communicative skills (self-regulation, shyness, alienation, manipulative and cooperative communication styles) and thinking styles (right-hemispheric and integrated), on the one hand, and attitude to remote learning, on the other. Gonzalez-Frey et al. (2021) evaluated students' attitude toward remote learning through qualitative analysis. The gathered data revealed that remote ed-ucation was somewhat worse than regular education, and four themes emerged to be of strict impact on online learning: *(i)* communication between students and faculty, *(ii)* flexibility with assignments, *(iii)* increased virtual interaction, and *(iv)* support. Afroz et al. (2021) investigated students' and teachers' attitudes toward online learning during the COVID-19 situation in Bangladeshi government colleges. Findings reveal that cost and time-effectiveness, safety, convenience, and improved participation were the most frequently cited positive aspects. At the same time, distraction and reduced focus, heavy workload, problems with technology and the internet, lack of ICT knowledge and poor network infrastructure, limited availability of educational resources, low attendance of learners, uncooperative learners and insufficient support from instructors and colleagues were the most recurrent negative aspects. Çelik and Uzunboylu (2022) developed a scale to measure the students' attitude towards distance learning including factors such as usefulness, communication, preference for distance learning, and preference for face-to-face learning. The analysis showed how usefulness and preference for distance learning indicated a positive attitude, while social presence and preference for face-to-face learning indicated a negative one. Gender had no impact on the scale dimensions. Chen et al. (2022) investigated differences in students' attitudes toward remote learning by comparing two cohorts of students: those based in Australia and those in China. Both cohorts were studying the same units with the same group of teaching staff. Australian students preferred remote learning due to its convenience and the availability of video recordings. In contrast, students in China preferred face-to-face learning, possibly due to their lack of prior experience in an English-speaking learning environment and hesitance to engage with lecturers and learning activities. These results show how students accept remote learning in a familiar language and learning environment. In contrast, if the teaching is delivered in a second language using unfamiliar teaching methods, remote learning will require additional scaffolding to enhance the learning experience. Assaf and Nehmeh (2022) evaluated the attitude towards remote learning in Lebanon. The study evidenced that the students felt isolated due to remote learning; in addition, online communication was not helpful in improving learning. Zagkos et al. (2022) measured attitudes of students of five Greek universities towards the distance learning process. The data reflects the substantial agreement of the students that face-to-face teaching cannot be replaced by distance learning, especially when it comes to laboratory training. The consensus is also that remote learning has abased pedagogical relationships between professors and classmates and among the latter as well. Findings indicate that students come to a meeting of minds about the educational inequalities, which are worsened by the lack of digital equipment and undeveloped technological infrastructure. Furthermore, this study reveals a correlation between the responses of the sample and their demographic and social characteristics, something that offers possibilities for additional research. Huang and Wang (2022) examined the effects of student motivation and engagement on students' academic achievement in online learning from the perspective of self-determination theory. The study evidenced how online emergency learning environments satisfying students' psychological needs of autonomy and competence promote optimal motivation, positive engagement and academic achievement. This study also contributed to revealing the sophisticated nature of relatedness satisfaction in the case wherein its specific effects depend on the cultural configuration of the contexts and the specific types of engagement. Moreover, the research evidenced how students' engagement acted as partial or full mediators between the satisfaction of the psychological needs and academic achievement. Specifically, the effects of autonomy and competence satisfaction on students' academic achievement were partially mediated by the extent to which they cognitively, emotionally and behaviorally engaged in online learning activities. Radovan and Makovec (2022) studied students of a Slovenian university, evidencing how home setting/environment affected students' attitudes towards distance learning, their assessment of competence for distance learning, as well as their motivation to study and their sense of being overwhelmed. Thus, more study difficulties, negative attitudes and motivation problems were observed among students who were not provided with adequate study conditions. The study indicates that distance learning has also potential, but this potential can only be realized if all those involved in the process are provided with the right conditions.

From this overview, we deduce that students all over the world acknowledge the utility of online/remote learning, especially as an answer to the pandemic emergence, and that they like its peculiar features such as lessons' video recordings. At the same time, they feel the downsides of remote learning affecting specific categories of students. More importantly, students state that for the majority of learning activities face-to-face modality cannot be replaced entirely.

For the sake of clarity, the key features of the papers included in our overview are described in Table 1.

3. Materials and methods

In this section, we first present the research design offering details on the questionnaire administered. Next, we display and discuss the structural equation model used to analyze the collected data.

3.1. The questionnaire

This research was conducted on a sample of students enrolled at the University of Foggia. The first part of the questionnaire is dedicated to gathering general socio-demographic information of respondents. From the questions in this section, we derived some exogenous variables to be included in our model: *Gender*, *Age*, *Average Mark*, *Working student* (if the students works full time or part time), *Commuter Student* (if the student has daily commute from a city outside Foggia to reach the University facilities). We also derived some artificial variables: *Enrollment Year*, *Disciplinary Area* and *Progress Score*. In the current Italian educational system, several types of graduate programs exist; they also have diverse durations and degree. This makes it difficult to compare the student enrollment year across different programs. To solve this problem, we created a variable, *Enrollment Year*, where the students enrolled in the first three years of the higher degree are indicated by their actual enrollment year, while all the others (students in supplementary years or enrolled in Master's degrees) are indicated with 4. For example, a student enrolled in the second year of study, is indicated with 2. A

Table 1: Studies on attitude of students towards online learning published after the COVID-19 outbreak.

Study	Number of partecipants	Main remarks
Aguilera-Hermida (2020)	270 (89% from USA University, 11% otherwise)	Analysis of acceptance of emergency online learning.
Hussein et al. (2020)	45 from University of Abu Dhabi	Qualitative investigations of strength and weakness of emergency remote learning.
Rafiq (2020)	2160 from Pakistan Universities	Application of Technology Acceptance Model to analyze attitude towards e-leaning in higher education.
Serhan (2020)	31 from a USA University	Investigating student attitude toward remote learning via Zoom.
Afroz et al. (2021)	100 from Bangladeshi Government Colleges	Analysis of students' and teachers' attitudes towards online learning.
Dikaya et al. (2021)	280 students from a Russia University	Analysis of students psychological traits (self-regulation, shyness, alienation, manipulative and cooperative communication styles) and attitude toward emergency remote learning.
Gonzalez-Frey et al.(2021)	93 from a college (unspecified institution)	Qualitative analysis of students' attitude.
Law (2021)	97 from Kuching (Malaysia)	Relation between attitude and learning materials, assessments, communication, technological tools and technical support.
Tzafilkou et al. (2021)	142 from a Greece University	Tested a scale to measure remote learning attitude.
Ferrer et al. (2022)	574 (usable)	Relation of Attitude with Engagement, Motivation and Study Mode.
Huang and Wang (2022)	14,935 across 39 universities	Measured participants' perceptions of needs satisfaction, engagement and academic achievement, respectively, during the emergency online learning.
Assaf and Nehmeh (2022)	928 Grade 9 and Third Secondary learners a of formal academic learning in the Lebanese educational system	Evaluation of learners attitude toward remote learning in Lebanon.
Çelik and Uzunboylu (2022)	384 for Exploratory Factor Analysis + 305 for Confirmatory Factory Analysis	Development of a scale to measure attitude towards remote learning.
Chen et al. (2022)	368 from Australia Universities + 40 from China Universities	Comparison of attitude towards online learning between students in Australia and China.
Radovan and Makovec (2022)	1,827 from Faculty of Arts at the University of Ljubljana, Slovenia	Analysis of correlation of attitudes and experience about distance education with variables such as living conditions, study conditions, gender, etc.
Zagkos et al. (2022)	807 from 5 Greece-based universities	Measure of students' attitude toward online learning.

student in his first supplementary year is indicated with 4, as well as a student enrolled in his second year of a Master's degree. The main reason behind this coding is to evaluate the difference between novel students in the university system and experienced students who are confident with the university educational system. *Disciplinary Area* has been derived from the program the students were enrolled in. We grouped their courses by disciplinary area according to the official directives of the Italian Ministry of Education. The variable *Progress Score* is computed as the ratio of the number of credits acquired to those required. It ranges from 0 to 1, with a mean value equal to 0.50 and standard deviation of

0.21. Students in supplementary years are penalized by increasing of the ratio's denominator: the penalty is equal to the number of credits required for an additional year of study.

To measure respondents' traits, we considered several scales proposed in the literature. Building a questionnaire of this type is challenging because it should "optimize" a series of criteria. On the one hand, the questionnaire must pursue the research aims; on the other hand, it must ensure that respondents do not spend too much time and effort in taking it. This maximizes the number of participants that join the research and minimizes the number of responses and incomplete questionnaires. Moreover, some scales needed to be adapted to the context of Italian university and educational environment. To pursue this aim and maximize students' involvement, we picked a subset of questions from each scale corresponding to the most relevant questions for each factor.

Student attitude toward online learning is assessed through a Likert type scale (see Table 2) derived from the work of Serhan (2020).

An aspect considered relevant to explain attitude toward online learning is shyness. Shyness as a personality trait may be defined as excessive self-focus characterized by negative self-evaluation that creates discomfort or inhibition in social situations and interferes with pursuing one's interpersonal or professional goals (Henderson et al., 2010). The nature of online learning impacts the possibility of interaction with others, therefore shy students might have different attitudes towards online learning. Shyness is measured by means of a 6-item Likert scale (see Table 3) based on McCroskey and Richmond (1982).

Another important aspect of our analysis is engagement with professors and with other students. These latent traits are measured through a scale (see Table 4) derived from the work of Freda et al. (2021). Engaged students are not just students who simply attend and participate in lessons, but they are able to sustain efforts, commitments, self-regulate behaviors and choices, negotiate and share their goals with others (colleagues, peers, teachers, families, etc.), accept the challenge of their limits in learning processes (Freda et al., 2021). Students' engagement is generally associated with a positive view of their own study activity, not illusory optimistic, but capable of showing and developing resources in terms of industriousness, activity and initiative (Freda et al., 2021). The engagement with other peers is a relevant aspect of engagement in educational context, and in this study we consider this could have a role in students' attitude towards online learning. Engagement with professors is another aspect of educational engagement, and involves the relationship with the teachers.

Emotional impact of the pandemic is measured by means of the scale provided by Ballou et al. (2020). This scale has been adapted considering the different moments we asked participants to answer, so they were asked to compare their current status to the one previous the COVID-19 outbreak. The scale is composed of 8 items (see Table 5). Global pandemic introduced unprecedented fear/worry about one's own safety as well as the health and safety of all; fears about local and global economic/political instability; frustration/disappointment regarding the complete disruption of daily activities (Ballou et al., 2020). The scale measures the emotional impact of the present pandemic over individuals. More negatively affected people could have more concerns in attending crowded places such as a classroom.

All the items in the scales above are measured on a 5-point rating scale.

The last scale, based on Vallerand et al. (1992), measures students' motivation to attend university. This scale has been validated in Italian by Alivernini and Lucidi (2008). From the original scale we picked 4 out of 6 factors (i.e., Amotivation, Extrinsic motivation – external regulation, Extrinsic motivation – Introjected, Extrinsic Motivation - Identified), with 3 items each. Students were asked if the items matched the reasons they enrolled at university, providing a score between 1 (Matches perfectly) and 7 (It doesn't match at all) (see Table 6). Individuals are amotivated when they do not perceive contingencies between outcomes and their own actions. Amotivated individuals believe that external causes beyond their own control are to blame for their actions. They start doubting themselves and wonder why attending university in the first place.They might quit taking part in academic activities (Deci and Ryan, 1985; Vallerand et al., 1992). Extrinsic motivation pertains to a wide variety of behaviors which are involved as a means to an end and not for their own sake (Deci, 1975; Deci and Ryan, 1985; Vallerand et al., 1992). Extrinsic motivation can be distinguished in three types, ordered from lower to higher level of self-determination: *External Regulation*, *Introjection*, and *Identification*. *External Regulation* occurs when the behavior is regulated through external means, such as rewards and constraints (Deci and Ryan, 1985, 1990; Vallerand et al., 1992). *Introjection* occurs when students internalize the reasons of their actions and act to behave as good students are supposed to do (Vallerand et al., 1992). *Identification* occurs when the behavior becomes valued and judged important for the individual, especially that is perceived as chosen by oneself (Vallerand et al., 1992).

In our questionnaire, we also included an open-ended question, asking the students their opinion on their online learning experience.

The choice of the described latent variables aims to better understand what are the factors that affect the most the attitude of students towards online learning. Motivation is an important factor to have success at higher school (Nur'aini et al., 2020). Personal traits, such as shyness, could have a relevant role in making the students more comfortable at home – typical learning setting for online learning – as they do not perceive the reduction of social opportunities as relevant as the others. Engagement is another driving factor for students' attitude (Ferrer et al., 2022; Sanders et al., 2016), and we aimed to analyze how the engagement with professors and peers interacts with this factor. Finally, we wanted to understand what relation is present between students with negative emotional impact and their attitude towards online learning.

The questionnaire has been submitted through academic emails to all the students: 2,420 of them participated between 24 May 2022 and 14 June 2022. For the University of Foggia, such a time range matches the end of the term, just before the exam session.

Before participating in the survey, all participants provided informed consent with respect to the research scope, and the management and gathering of anonymous data. No compensation was provided for participating in the study.

3.2. Structural Equation Modeling

In our study, we employ a GSEM (Muthèn, 1984) to examine the connection between the endogenous latent variable "attitude towards online learning" and the exogenous latent and observed variables detailed in Section 3.1.

A classical SEM consists of two components: a measurement model that clarifies the relationship between continuous latent variables and their continuous observed indicators, facilitating the estimation of latent factors, and a structural model that illustrates the interactions between endogenous and exogenous variables. When dealing with ordered categorical items, such those obtained through Likert scales, the conventional measurement model, which relies on continuous observed indicators, necessitates the incorporation of a threshold model. This threshold model links each observed categorical indicator to an underlying continuous variable defining specific cut-off points on the continuous underlying variable, which correspond to different response categories on the observed categorical item. The threshold model essentially serves as a bridge that quantifies how the responses on a categorical scale are associated with the unobserved latent construct. This adaptation leads to a GSEM.

In our specification, we assume that there is one endogenous latent variable (namely, η: attitude toward online learning) and eight exogenous latent variables $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_Q)'$.

Denoting by $\mathbf{Y} = (Y_1, \ldots, Y_K)'$ the observed categorical variables measuring the endogenous latent variable η, and with $\mathbf{Z}^{(y)} = \left(Z_1^{(y)}, \ldots, Z_K^{(y)}\right)'$ the corresponding underlying continuous variables, the reflective measurement model for the endogenous latent variable can be expressed as

$$\mathbf{Z}^{(y)} = \boldsymbol{\lambda}^{(y)}\eta + \boldsymbol{\epsilon}^{(y)}, \quad \text{with} \quad Y_k = c \quad \text{if } \gamma_{k,c-1}^{(y)} \leq Z_k^{(y)} \leq \gamma_{k,c}^{(y)},\ k = 1, \ldots, K \quad (1)$$

where the K-dimensional vector $\boldsymbol{\lambda}^{(y)} = \left(\lambda_1^{(y)}, \ldots \lambda_K^{(y)}\right)'$ contains the factor loadings.

Along the same lines, denoting by $\mathbf{X} = (X_1, \ldots, X_L)'$ the observed categorical variables measuring the $Q = 8$ exogenous latent variable $\boldsymbol{\theta}$, and with $\mathbf{Z}^{(x)} = \left(Z_1^{(x)}, \ldots, Z_L^{(x)}\right)'$ the corresponding underlying continuous variables, the reflective measurement model for the exogenous latent variables can be expressed as

$$\mathbf{Z}^{(x)} = \boldsymbol{\Lambda}^{(x)}\boldsymbol{\theta} + \boldsymbol{\epsilon}^{(x)}, \quad \text{with} \quad X_l = c \quad \text{if } \gamma_{l,c-1}^{(x)} \leq Z_l^{(x)} \leq \gamma_{l,c}^{(x)},\ l = 1, \ldots, L. \quad (2)$$

Here, in a confirmatory perspective, the pattern of fixed and free loadings in the $L \times Q$ factor loadings matrix $\boldsymbol{\Lambda}^{(x)}$ is specified according to a multi-unidimensional schema (Sheng and Wikle, 2007), also known as independent cluster structure (McDonald, 2000), where each item loads only on a specific latent variable. Within our model, we assume correlations between the exogenous latent variable.

In equations (1) and (2), $\gamma_{k,0}^{(y)} = -\infty \leq \gamma_{k,1}^{(y)} \leq \ldots \leq \gamma_{k,C}^{(y)} = \infty$ and $\gamma_{l,0}^{(x)} = -\infty \leq \gamma_{l,1}^{(l)} \leq \ldots \leq \gamma_{l,C_l}^{(x)} = \infty$ are the ordered thresholds for item Y_k and X_l respectively, and $\boldsymbol{\epsilon}^{(y)}$ and $\boldsymbol{\epsilon}^{(x)}$ are normally distributed errors.

The structural component of the model can be written as:

$$\eta = \boldsymbol{\beta}'\boldsymbol{\theta} + \mathbf{g}'\mathbf{w} + u \quad (3)$$

where $\boldsymbol{\beta}$ and $\mathbf{g}$ are vectors of regression coefficients, $\mathbf{w}$ is the observed exogenous covariate vector, and u is a normally distributed error.

The structural component is represented in the path diagram provided in Figure 1, where rectangles symbolize the observed exogenous covariates, while ovals rep-resent both endogenous and exogenous latent variables.

4. Results

In this section, we show the results of our study. In more details, we first offer an overview of the participants in our study (Section 4.1); then, we show the

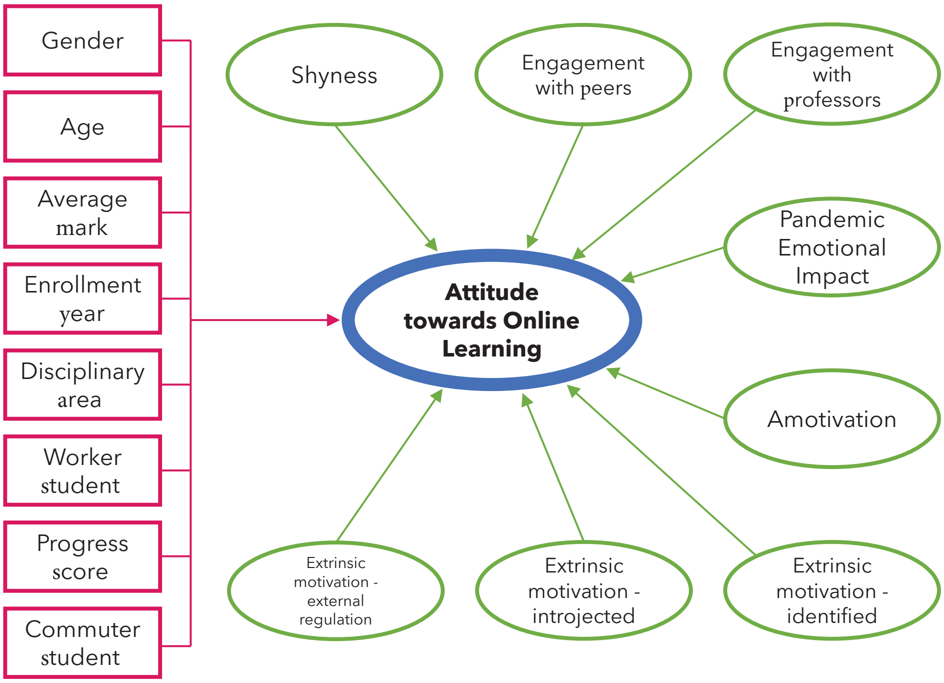

Gender
Age
Average mark
Enrollment year
Disciplinary area
Worker student
Progress score
Commuter student
Shyness
Engagement with peers
Engagement with professors
Pandemic Emotional Impact
Attitude towards Online Learning
Amotivation
Extrinsic motivation - external regulation
Extrinsic motivation - introjected
Extrinsic motivation - identified

Figure 1: Path diagram for the structural component in the proposed GSEM. ***Rectangles symbolize the observed exogenous covariates, while ovals represent both endogenous and exogenous latent variables.***

results of the measurements model, i.e., how the items impact on the estimates of latent factors (Section 4.2), and the structural relation between the exogenous latent variables and attitude towards online learning (Section 4.3). Next, Section 4.4 is devoted to presenting the findings of this research with regards to the research questions proposed in the introduction. Lastly, in Section 4.5, we report the analysis made on the students' opinions expressed in natural language.

The GSEM estimation here reported has been performed through the R package *lavaan* (Rosseel, 2012), a well-tested library that provides the essential tools for SEM analysis. To deal with the ordinal nature of the observed indicators in the measurement model, we use the weighted least squares mean-and variance-adjusted (WLSMV) estimator, which is the most common method in the SEM literature to analyze ordered categorical variables (see Li, 2016, and references therein). WLSMV is a limited information estimation method that for ordered categorical indicators utilizes polychoric correlations. As a limited information method, WLSMV is not only a robust method but also computationally fast, especially when the sample size and the number of dimensions are large (Flora and Curran, 2004).

The model with 323 parameters has been estimated with 2,016 observations after pruning incomplete observations. We remark we have dropped from the study the students that declared “other” as *gender*, since they were just 6.

4.1. Sample composition

The sample of respondents covers 18% of the students’ population. In the Economics, Management, Territory department, the coverage peaks 35% while the lowest value is 13% for the department of Clinical and Experimental Medicine (see Figure 2). Therefore there is broad coverage of all departments.

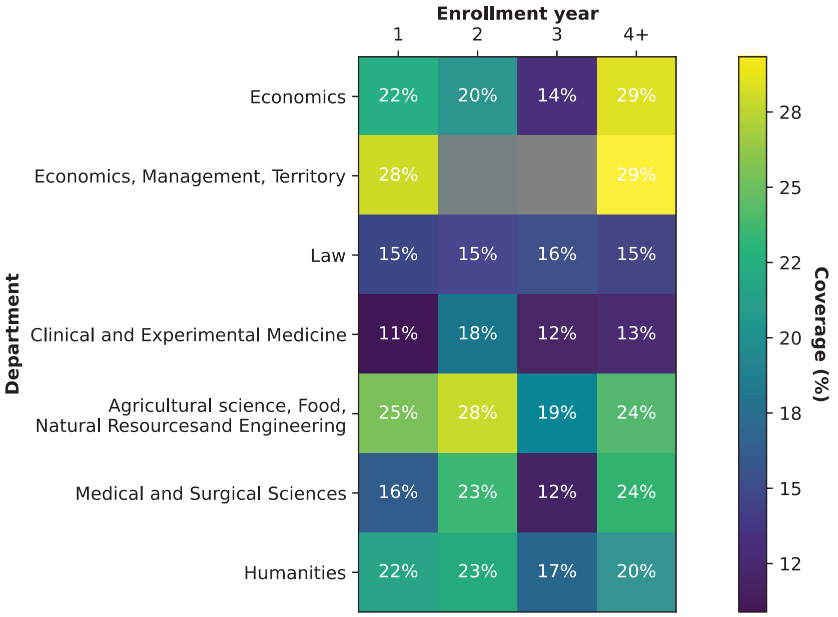
70.5% of respondents are females, 29.2% males, and 0.3% other, covering the 20% of the female population and the 15% of the male population). The mean age of participants is 25 years. 70.9% of respondents are commuters, 17.6% of them are from Foggia city, and the remaining 11.5% are students from outside Foggia which had taken accommodation in the city. 97% declared that they owned adequate devices for online learning and 93% said they had an adequate connection. 70.3% of respondents are enrolled in a Bachelor program, while 16.3% is in a Master’s program. The remaining 13.4% is enrolled in a Unique cycle master’s degree (5-6 years).

4.2. Measurement equation estimates

Tables 2 - 6 show the estimates of the factor loadings of the measurement equations, along with the ordinal alpha coefficient (Zumbo et al., 2007) for each dimension. Ordinal alpha is conceptually equivalent to Cronbach’s alpha (Cronbach, 1951). The critical difference between the two is that ordinal alpha is based on the polychoric correlation matrix, rather than the Pearson covariance matrix, and thus is more suitable to assess the reliability of scales with ordinal indicators. Since all latent variables exhibit ordinal alpha coefficients exceeding 0.75, it can be inferred that the measurement instrument demonstrates high reliability. Furthermore, all the factor loadings are significant with p-value < 0.001. The negative factor loadings are related to items that measure the latent trait in a reverse direction. The estimated threshold parameters are provided in Table A1 in the Online Resource.

Table 7 shows the correlations across the exogenous latent variables in the proposed GSEM.

Figure 2: Coverage of respondents respect the number of enrolled students by department and year of enrollment in higher education.



- Gray background indicates course year without enrolled students (due to a new department).
- Master's degree students have been offset to 4th year, as they require a 3-year bachelor's program to enroll in a master's course.

The highest correlations are observed among the latent variables associated with the *Extrinsic motivation* subscales. This highlights a clear positive relationship between behaviors driven by external rewards and constraints (*External regulation*), and the internalization (*Introjection*) and identification (*Identified*) with the importance of university studies. Moreover there is a moderate negative correlation between *Extrinsic motivation – Identified* and *Amotivation*. This shows that the higher is the self-determination level the lower the propensity to be amotivated. *Amotivation* is directly correlated with *Pandemic emotional impact*, indicating that students experiencing a more severe emotional impact tend to be more amotivated.

We also find a positive correlation between *Engagement with professors* and *Engagement with peers*. Both these forms of engagement are inversely correlated with *Amotivation*, and directly correlated with *Extrinsic motivation - Identified*. These findings emphasize how engagement levels vary with the degree of

Table 2: Measurement model for the latent variable *Attitude toward online learning*: factor loading estimates. The question prompts the students to indicate their agreement with the items.

Id.	**Item**	**Factor Loading Estimate**
Factor: ATTITUDE TOWARDS ONLINE LEARNING		
A1	I enjoyed using the online learning platform during the class.	1.000
A2	I would like to use online learning platform in other classes.	1.038
A3	The use of online learning platform allowed flexibility in my learning schedule.	1.007
A4	Overall, I enjoyed using online learning platform in the class	1.029
A5	The use of online learning platform helped me learn the class content.	1.029
A6	The use of online learning platform helped me develop confidence in the subject.	1.030
A7	The use of online learning platform helped me participate in the class in ways that enhanced my learning.	1.031
A8	The use of online learning platform motivated me to actively participate in class activities.	1.039
A9	The use of online learning platform made it easier for me to be more engaged in the class discussions.	1.044
A10	The use of online learning platform increased my interaction with my instructor.	1.020
A11	The use of online learning platform increased my interaction with my classmates.	0.964
A12	The use of online learning platform motivated me to seek help from tutors, classmates, and the instructor.	0.959
A13	The activities during online learning platform sessions motivated me to learn the class content more than the ones in the face-to-face traditional class meetings.	1.012
A14	I participated more in the online learning platform sessions in comparison to the traditional face-to-face class meetings.	0.991
A15	My attention to the class tasks during the online learning platform sessions was greater in comparison to the traditional face-to-face class meetings.	0.999
A16	It was easier to participate in group activities in the online learning platform sessions in comparison to the traditional face-to-face class meetings.	0.986
Ordinal alpha= 0.986		

All the free items estimates are significant with p-value < 0.001.

self-determination. Students who are less inclined to attribute their outcomes to external factors and are more motivated by future career prospects and job market-relevant skills display higher levels of engagement. *Engagement with peers* is inversely correlated with *Shyness*, and this relation is expected, as more shy persons are less likely to engage with others.

4.3. Structural component estimates

Table 8 shows the estimated regression coefficients for the structural component.

Shyness, *Engagement with professors*, *Extrinsic motivation - Introjected* positively affect *Attitude towards online learning*. *Engagement with peers* shows a negative impact: students with higher engagement with their colleagues are less satisfied with online learning.

Table 3: Measurement model for the latent variable *Shyness*: factor loading estimates. The question prompts the students to indicate how much the sentences apply to themselves

Id.	Item	Factor Loading Estimate
Factor: Shyness		
B1	I am a shy person.	1.000
B2	Other people think I talk a lot.	-0.346
B3	I tend to be very quiet in class.	1.159
B4	I am a quiet person.	1.266
B5	I talk more in a small group (3-6) than others do.	0.729
B6	I talk more in class than most people do.	-0.597
Ordinal alpha= 0.768		

All the free items estimates are significant with p-value < 0.001.

Table 4: Measurement model for the latent variables related to the *Engagement*: factor loading estimates. The question prompts the students to indicate how much the sentences apply to themselves.

Id.	Item	Factor Loading Estimate
Factor: Engagement with University Professors		
C1	My teachers are interested in my opinions and what I say.	1.000
C2	Teachers are usually available to discuss my work.	1.095
C3	Teachers clarify what they expect of us students.	0.994
Factor: Engagement with University Peers		
D1	I feel like I'm part of a group of friends at University.	1.000
D2	I like to meet friends at university.	0.987
D3	I've made meaningful friends with some college colleagues.	1.077
D4	I have good relationships with my University colleagues.	1.113
D5	Studying with other students is useful to me.	0.671
Engagement with University Professors: **Ordinal alpha**= 0.837		
Engagement with University Peers: **Ordinal alpha**= 0.896		

All the free items estimates are significant with p-value < 0.001.

Extrinsic motivation – Identified has a negative impact on the attitude. Instead, *Amotivation* does not seem to influence attitude toward online learning.

Pandemic impact negatively affects *Attitude towards online learning*. We initially expected students significantly impacted by the pandemic to hold a more positive attitude towards online learning. This anticipation was based on the fact that online classes are typically taken from the safety of one's home, potentially reducing exposure to COVID-19 by avoiding crowded places. However, the coun-

Table 5: Measurement model for the latent variable *Pandemic Emotional Impact*: factor loading estimates. The question prompts the students to indicate how much the sentences apply to themselves, comparing the current status to the one previous the COVID-19 outbreak.

Id.	Item	Factor Loading Estimate
Factor: PANDEMIC EMOTIONAL IMPACT		
E1	Feeling more frustrated about not being able to do what you usually enjoy doing	1.000
E2	Having more difficulty concentrating	1.359
E3	Feeling more grief or sense of loss	1.417
E4	Being less productive	1.415
E5	Feeling more angry or irritated	1.362
E6	More difficulty sleeping	1.166
E7	Feeling more lonely or isolated	1.269
E8	More anxious	1.210
Ordinal alpha= 0.941		

All the free items estimates are significant with p-value < 0.001.

terintuitive results obtained from our analysis might be attributed to the general low mood of students. It's also possible that a bidirectional effect, not accounted for by the model, could exist, where a negative attitude towards online learning worsens the impact of the pandemic.

Concerning the exogenous variables, male students show a more negative *Attitude towards online learning*. Being an older student has positive impact. As expected, Working Students (both Full-Time and Part-Time) have a more positive attitude. The impact of *Average mark*, *Enrollment year* and *Progress score* is not significant. In the disciplinary area, only scientific technological courses have a significant negative impact on student attitude.

4.4. Findings related to Research Questions

In the following, we provide answers for the research questions discussed in Section 1. We remark that the SEM method adopted has proved to be effective in answering in a comprehensive way to the research questions.

RQ1. ***Does pandemic emotional impact positively affect student attitude towards online learning?*** *Pandemic emotional impact* has a significant relevance in explaining attitude towards online learning with an estimated negative effect. A bidirectional effect could also be possible, given that a negative experience in

Table 6: Measurement model for the latent variables related to the *Motivation*: factor loading estimates. The question prompts the students to indicate how much the sentences match the reasons they enrolled in University.

Id.	**Item**	**Factor Loading Estimate**
Factor: AMOTIVATION		
F1	Honestly, I don't know; I really feel that I am wasting my time in school.	1.000
F2	I once had good reasons for going to college; however, now I wonder whether I should continue.	1.064
F3	I can't see why I go to college and frankly, I couldn't care less.	1.009
Factor: EXTRINSIC MOTIVATION - EXTERNAL REGULATION		
F4	Because with only a high-school degree I would not find a high-paying job later on.	1.000
F5	In order to obtain a more prestigious job later on.	1.990
F6	Because I want to have "the good life" later on.	2.127
Factor: EXTRINSIC MOTIVATION - INTROJECTED		
F7	To prove to myself that I am capable of completing my college degree.	1.000
F8	Because of the fact that when I succeed in college I feel important.	0.913
F9	To show myself that I am an intelligent person.	1.020
Factor: EXTRINSIC MOTIVATION - IDENTIFIED		
F10	Because I think that a college education will help me better prepare for the career I have chosen.	1.000
F11	Because eventually it will enable me to enter the job market in a field that I like.	0.994
F12	Because I believe that a few additional years of education will improve my competence as a worker.	0.966
AMOTIVATION: **Ordinal alpha**= 0.926		
EXTRINSIC MOTIVATION - EXTERNAL REGULATION: **Ordinal alpha**= 0.831		
EXTRINSIC MOTIVATION - INTROJECTED: **Ordinal alpha**= 0.911		
EXTRINSIC MOTIVATION - IDENTIFIED: **Ordinal alpha**= 0.892		

All the free items estimates are significant with p-value < 0.001.

an online learning environment might increase the emotional negative impact of the pandemic.

RQ2. ***Does engagement affect students attitude towards online learning?*** Both *Engagement with professors* and *Engagement with peers* significantly impact the attitude. *Engagement with Professors* has a positive effect; therefore, students experiencing higher engagement with professors are more prone to evaluate positively the online learning experience. This can indicate that the relationship between students and professors is essential and affects the appreciation of the deployed learning approach. Instead, *Engagement with peers* has a negative effect; therefore, students that have a better relationship with other students feel it more difficult to adapt to an online learning context. This result is expected because

Table 7: Correlations between the latent variables.

Latent Variable	1.	2.	3.	4.	5.	6.	7.
1. Shyness							
2. Engagement with Professors	-0.061^{*}						
3. Engagement with Peers	-0.297^{**}	0.372^{**}					
4. Pandemic Emotional Impact	0.184^{**}	-0.182^{**}	-0.053^{*}				
5. Amotivation	0.231^{**}	-0.362^{**}	-0.301^{**}	0.418^{**}			
6. Extrinsic Motivation - External Regulation	0.037	0.064^{*}	0.029	0.062	0.166^{**}		
7. Extrinsic motivation – Introjected	0.005	0.117^{**}	0.128^{**}	0.068^{**}	-0.028	0.522^{**}	
8. Extrinsic Motivation - Identified	-0.054^{*}	0.248^{**}	0.206^{**}	-0.067^{**}	-0.319^{**}	0.666^{**}	0.639^{**}

$^{\circ}$ p-value < 0.10, * p-value < 0.05, ** p-value < 0.01

Table 8: Results of the GSEM Structural component: estimated impact on *Attitude towards Online Learning*.

Regressor	Regression coefficient estimate	Std.Err	z-value	P(>|z|)
Shyness	0.086	0.033	2.614	0.009
Engagement with Professors	0.177	0.035	5.064	0.000
Engagement with Peers	-0.112	0.030	-3.764	0.000
Pandemic Emotional Impact	-0.270	0.040	-6.729	0.000
Amotivation	-0.008	0.059	-0.133	0.894
Extrinsic motivation – External Regulation	0.257	0.114	2.258	0.024
Extrinsic motivation – Introjected	0.164	0.035	4.617	0.000
Extrinsic motivation – Identified	-0.130	0.068	-1.906	0.057
Gender	-0.104	0.048	-2.173	0.030
Age	0.012	0.003	3.914	0.000
Worker Full Time	0.266	0.074	3.577	0.000
Worker Part Time	0.223	0.050	4.463	0.000
Commuter Student	0.204	0.047	4.305	0.000
Average Mark	0.012	0.010	1.157	0.247
Enrolled at first year	0.020	0.060	0.325	0.745
Enrolled at second year	0.011	0.061	0.176	0.860
Enrolled at third year	-0.042	0.068	-0.619	0.536
Disciplinary Area Medical	-0.049	0.078	-0.627	0.531
Disciplinary Area Sanitary	-0.094	0.085	-1.108	0.268
Disciplinary Area Scientific-Technological	-0.121	0.055	-2.195	0.028
Progress Score	0.107	0.120	0.893	0.372

For *Gender*, the baseline is female.
For *Worker Full Time* and *Worker Full Time*, the baseline is the student is not a worker.
For *Worker* status, the baseline is not a worker.
For the year of *Enrollment* the baseline is enrollment in a year over the third.
For the *Disciplinary Area* the baseline is enrollment in a Humanities Area course.

online methodologies provide fewer opportunities to establish relationships with colleagues than traditional learning in classrooms (Kaufmann and Vallade, 2022).

Such a finding could indicate that efforts are needed to compensate for this factor. The literature provides a handful of tools for facing this issue: from the community of inquiry framework – a well-known approach – for building collaborative online learning environments (Fiock, 2020), to the limitless possibilities offered by the so-called "metaverse", where learners will move away from online learning as we know today to more immersive content based on virtual reality tools, gamified learning with an increased engagement (Phakamach et al., 2022).

RQ3. ***Does student motivation affect students attitude toward online learning?*** The impact of motivation on attitude is not equal among all different aspects of motivation evaluated through the scale exploited in our research. Amotivation, which is a lack of perception of contingencies between outcomes and one's actions, does not affect the evaluation of online learning. Students with higher extrinsic motivation for studying tend to have a more positive attitude towards online learning, particularly when their motivation is associated with external factors such as the belief that a university degree will lead to a job. This motivation is prevalent among students who primarily view university as a means of obtaining an academic title. It is noteworthy to examine the result for *Extrinsic motivation - Identified*. Unlike other extrinsic motivations, this type shows an opposite trend—it has a negative effect on attitudes toward online learning. We suspect this is because highly motivated students view the university as a place to acquire knowledge, skills, and competencies that enhance their well-being and career prospects. This perspective suggests that highly motivated students may not perceive online learning outcomes as effectively as traditional methods.

RQ4. ***Is shyness a factor related to student attitude towards online learning?*** As shown in Section 4, *Shyness* has a positive impact on how online learning is experienced: students with a higher level of shyness are more prone to give a favorable evaluation of online learning.

RQ5. ***Do exogenous factors, as gender, age, enrollment year, student worker, and commuter status affect attitude towards online learning?*** The results evidence how the enlisted exogenous features significantly impact the attitude towards online learning. In particular, male students, older students, working students and commuter students are more favorable towards online learning. These results are reasonable: workers students and commuters are the ones experiencing the most significant difficulties in attending face-to-face classes. The online learning, jointly with the recorded lessons, provided them all the flexibility to attend the lessons. The average mark does not result to be significant, while it is interesting how students enrolled in disciplinary courses in the scientific-technological Area seem to be less favorable towards online learning. This is in line with the findings of other studies (Newsome et al., 2022; Ngah et al., 2022; Owston et al., 2020), that discovered how STEM students' perception was worse than non-STEM courses. The University of Foggia was already experimenting with forms of online learning before the COVID-19 outbreak in humanities courses, preparing teachers and adapting courses to this scope. Instead, other programs suffered a rush emergency transition, which may be a possible reason for the different of impact. The *Average mark*, the *Progress score*, and the *Year of enrollment* do not seem to be significant in explaining the attitude toward online learning.

4.5. Textual analysis

Textual questions are intriguing because they allow respondents to freely express their opinions, but they are more challenging to analyze than conventional structured inquiries because they cannot be processed quantitatively. Text mining has been used to approach this type of data. Text mining is a toolbox of methods that enables the extraction of knowledge from text-based data. In this case, we used a bi-gram method to extract information from students' comments. In such an instance, the individual responses are seen as an ordered list of words. A bi-gram consists of two sequential words in a document. From the document, we can extract a set of bi-grams, transforming it into an unordered set for quantitative analysis. In the scope of this research, we created a wordcloud. Wordcloud is a widely-used data visualization technique in textual analysis, presenting a list of terms with varying sizes based on an associated factor.

Figure 3 shows the wordcloud of the most frequent bi-grams in the corpus of textual comments gathered from the open-ended question asking participants to provide a general comment about their online learning experience. The most frequent bi-grams are “dual mode”, “positive experience”, “off site”, “online learning”, “take exams”, “study management”, “great experience”, “working student”, “work study” and “reach university”. Figure 3 indicates how students particularly appreciated online learning. Integrating the results shown in the wordcloud, and the manual reading of a sample of students' answers, we deduced a positive appreciation for the opportunity of better managing own time mainly for students that are commuters and/or workers. The overall impression is a broad appreciation and the wish to continue with remote learning.

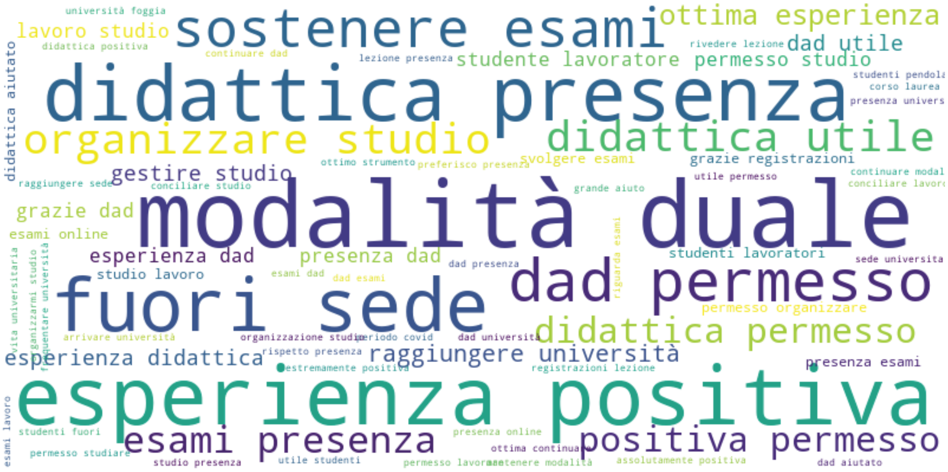


Figure 3: Wordcloud of the most frequent bi-grams from the answers to the open ended questions to provide a general comment about the online learning experience.

5. Conclusion

Online learning represents an excellent opportunity for the educational sys-tem, and the pandemic outbreak rushed out the introduction of new technologies in educational institutions. The end of the emergency phase opens up a new phase where more attention can be paid to studying the factors influencing the success of online learning platforms to improve students' outcomes. The timing of our survey offered a great opportunity to grasp the long experience of students with online learning approaches. Nonetheless, the broad participation (18% of the total number of students enrolled at the University of Foggia) offered a great opportunity that other papers lacked.

The results of our research show how attitude towards online learning has an evident connection with personal needs of students. This is highlighted by the significant impact of being commuter or working students on their attitude: students with more time constraints due to work, or because they have to come from a city outside Foggia to attend lessons, have more positive attitudes towards online learning. Comments in the open-ended question explicitly confirms the "better time management" of online learning.

Engagement is another factor influencing student attitude. Students more engaged with professors have a more favorable attitude toward online learning. This confirms the critical role of engagement in learning, which should induce universities to try to improve the connection and the relation, preparing the teachers to create more enjoyable environments in online learning. Instead, engagement with peers is negatively related to attitude. This point leverages the primary flaw of online learning systems that, in many implementations, cannot let students establish strong relationships. Therefore, a more significant investment should be made by universities to reduce the gap between online and face-to-face in establishing engagement with peers. Although there exist different collaborative online learning environments for facing this issue, their full adoption still has a long way to go. More efforts should be put to implement environments that are more collaborative like the communities of inquiry (Fiock, 2020), designed to increase student engagement and to encourage their motivation in online courses. Already since the first decade of 2000, researchers who were studying the upcoming online learning era were also advocating that videoconferencing for holding classes was not enough, and that the future of online learning was tightly stitched with the design of new forms of learning and new ways to promote collaboration among students as it was done in face-to-face learning (Garrison and Akyol, 2012; Lambert and Fisher, 2013). In this regard, the pandemic could provide the adequate boost for the adoption of adequate solutions, such as collaborative online learning environments (Di Cerbo et al., 2008). Another flourishing field, in this sense, is that of the upcoming metaverse, where researchers are already designing immersive collaborative environments for learning purposes (Jovanović and Milosavljević, 2022).

Overall, this research has been carried out after a prolonged emergency time due to the pandemic, and the evaluation of the relation between the pandemic's emotional impact and attitude is of absolute interest. The results showed that people with a higher emotional impact also had the worst attitude toward online learning. Of course, this result has to be taken carefully; a bidirectional relation could be present between the two variables. However, this result suggests it is essential for university to take care of students' emotional situation, e.g., by offering support services. The University of Foggia was already introducing online teaching, making the transition smoother. Therefore, students widely appreciated online learning and expressed the wish to continue. This is particularly evident from the textual analysis but also from digging into other questionnaire answers. However, such a result shows that online learning is promising, but not all students have the same attitude, and several factors can influence it. This is relevant for universities to consider and continue to adapt the learning systems to fulfil the requirements and the natural attitude of all students. At the same time, they should support them in the most challenging steps. This kind of help could also change students' attitudes.

Finally, further investigation is needed to understand STEM students' worst attitude toward online learning and the reasons behind it. The limitations of this study could be helpful in guiding further research. Some of the relations are not casually explainable, and further investigation is needed. Some of the scales have been used with reduced items so to lighten the questionnaire and achieve broad participation. Future similar studies in other universities and countries could address our research findings. Lastly, we expect that more complex and explainable models could be proposed to explain students' attitude.

Acknowledgments

Publication realized within of MUR funding, FSE – REACT EU resources of the PON Research and Innovation 2014-2020.

References

Abdullah, F. and Kauser, S. (2022). Students' perspective on online learning during pandemic in higher education. In *Quality & Quantity*. doi:10.1007/s11135- 022-01470-1.

Afroz, R., Islam, N., Rahman, S., and Anny, N.Z. (2021). Students' and teachers' attitude towards online classes during Covid-19 pandemic. In *International Journal of Research in Business and Social Science*, 10: 462–476. doi:10.20525/IJRBS.V10I3.1155.

Aguilera-Hermida, A.P. (2020). College students' use and acceptance of emergency online learning due to COVID-19. In *International Journal of Educational Research Open*, 1: 100011. doi:10.1016/J.IJEDRO.2020.100011.

Ali, W. (2020). Online and remote learning in higher education institutes: A necessity in light of COVID-19 pandemic. In *Higher Education Studies*, 10: 16–25. doi:10.5539/hes.v10n3p16.

Alivernini, F. and Lucidi, F. (2008). The Academic Motivation Scale (AMS): Factorial structure, invariance, and validity in the Italian context. In *TPM-Testing, Psychometrics, Methodology in Applied Psychology*, 15: 211–220.

Appolloni, A., Colasanti, N., Fantauzzi, C., Fiorani, G., and Frondizi, R. (2021). Distance learning as a resilience strategy during Covid-19: An analysis of the Italian context. In *Sustainability*, 13 (3). doi:10.3390/su13031388.

Assaf, J. and Nehmeh, L. (2022). The remote learning experience in Lebanon: Learners' attitudes and practices. In *Pedagogical Research*, 7 (1): em0115. doi:10.29333/pr/11551.

Ballou, S., Gray, S. and Palsson, O.S. (2020). Validation of the pandemic emotional impact scale. In *Brain, Behavior, & Immunity - Health*, 9: 100161. doi:10.1016/j.bbih.2020.100161.

Bollen, K.A. (1989). *Structural Equations with Latent Variables.* Wiley series in probability and mathematical statistics. Applied probability and statistics section. John Wiley & Sons, Oxford, England. doi:10.1002/9781118619179.

Botero, G.G., Questier, F., Cincinnato, S., He, T. and Zhu, C. (2018). Acceptance and usage of mobile assisted language learning by higher education students. In *Journal of Computing in Higher Education*, 30: 426–451. doi:10.1007/S12528-018-9177-1.

Chen, H., van Reyk, D., Reyna, J. and Oliver, B.G. (2022). A comparison of attitudes toward remote learning during the COVID-19 pandemic between students attending a Chinese and an Australian campus. In *Advances in Physiology Education*, 46: 297–308. doi:10.1152/ADVAN.00141.2021.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. In *Psychometrika*, 16 (3): 297–334. doi:10.1007/BF02310555.

Daniel, J. (2020). Education and the COVID-19 pandemic. In *Prospects*, 49 (1): 91–96. doi:10.1007/s11125-020-09464-3.

Deci, E.L. (1975). *Intrinsic Motivation*. Springer US. doi:10.1007/978-1-4613-4446-9.

Deci, E.L. and Ryan, R.M. (1985). Intrinsic motivation and self-determination in human behavior. In *Intrinsic Motivation and Self-Determination in Human Behavior*. doi:10.1007/978-1-4899-2271-7.

Deci, E.L. and Ryan, R.M. (1990). A motivational approach to self: integration in personality". In *Nebraska Symposium on Motivation*, 38: 237–288.

Di Cerbo, F., Dodero, G. and Succi, G. (2008). Extending moodle for collaborative learning. In *Proceedings of the 13th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education, ITiCSE 2008*, vol. 40, 324. doi:10.1145/1597849.1384367.

Dikaya, L.A., Avanesian, G., Dikiy, I.S., Kirik, V.A. and Egorova, V.A. (2021). How personality traits are related to the attitudes toward forced remote learning during COVID-19: Predictive analysis using generalized additive modeling. In *Frontiers in Education*, 6: 108. doi:10.3389/FEDUC.2021.629213.

Farooqui, S. (2020). Education in the time of Covid-19: How institutions and students are coping. In *Business Standard*. URL https://www.business-standard.com/article/education/education-in-the-time-of-covid-19-how-institutions-and-students-are-coping-120043001575_1.html.

Favale, T., Soro, F., Trevisan, M., Drago, I. and Mellia, M. (2020). Campus traffic and e-learning during COVID-19 pandemic. In *Computer Networks*, 176: 107290. doi:10.1016/j.comnet.2020.107290.

Ferrer, J., Ringer, A., Saville, K., Parris, M.A. and Kashi, K. (2022). Students' motivation and engagement in higher education: The importance of attitude to online learning. In *Higher Education*, 83: 317–338. doi:10.1007/S10734-020- 00657-5.

Fiock, H. (2020). Designing a community of inquiry in online courses. In *The International Review of Research in Open and Distributed Learning*, 21 (1): 135–153. doi:10.19173/irrodl.v20i5.3985.

Flora, D.B. and Curran, .J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. In *Psychological Methods*, 9 (4): 466 – 491. doi:10.1037/1082-989X.9.4.466.

Freda, M.F., Raffaele, D.L.P., Esposito, G., Ragozini, G. and Testa, I. (2021). A new measure for the assessment of the university engagement: The SInAPSi academic engagement scale (SAES). In *Current Psychology*. doi:10.1007/s12144-021-02189-2.

Garrison, D.R. and Akyol, Z. (2012). *The Community of Inquiry Theoretical Framework*, chap. 7. Routledge. doi:10.4324/9780203803738.ch7.

Gonzalez-Frey, S.M., Garas-York, K., Kindzierski, C.M. and Henry, J.J. (2021). College students' attitudes towards remote instruction during the coronavirus pandemic: Future directions. In *Excelsior: Leadership in Teaching and Learning*, 13: 96–112. doi:10.14305/jn.19440413.2021.13.2.02.

Henderson, L., Zimbardo, P. and Carducci, B. (2010). *Shyness*, 1–3. John Wiley & Sons, Ltd. doi:10.1002/9780470479216.corpsy0870.

Hodges, C.B., Moore, S., Lockee, B.B., Trust, T. and Bond, M.A. (2020). The difference between emergency remote teaching and online learning. In *EDUCAUSE Review*. URLhttps://er.educause.edu/articles/2020/3/the-difference-between-emergency-remote-teaching-and-online-learning.

Huang, R., Liu, D., Guo, J., Yang, J., Wei, X., Knyazeva, S., Li, M., Zhuang, R., Looi, C. and Chang, T. (2020). Guidance on Flexible Learning during Campus Closures: Ensuring course quality of higher education in COVID-19 outbreak. Beijing: Smart Learning Institute of Beijing Normal University.

Huang, Y. and Wang, S. (2022). How to motivate student engagement in emergency online learning? Evidence from the COVID-19 situation. In *Higher Education*. doi:10.1007/S10734-022-00880-2.

Hussein, E., Daoud, S., Alrabaiah, H. and Badawi, R. (2020). Exploring undergraduate students' attitudes towards emergency online learning during COVID-19: A case from the UAE. In *Children and Youth Services Review*, 119: 105699. doi:10.1016/J.CHILDYOUTH.2020.105699.

Jovanović, A. and Milosavljević, A. (2022). VoRtex metaverse platform for gamified collaborative learning. In *Electronics*, 11 (3): 317. doi:10.3390/electronics11030317.

Kaufmann, R. and Vallade, J.I. (2022). Exploring connections in the on-line learning environment: Student perceptions of rapport, climate, and loneliness. In *Interactive Learning Environments*, 30 (10): 1794–1808. doi:10.1080/10494820.2020.1749670.

Lambert, J.L. and Fisher, J.L. (2013). Community of inquiry framework: Establishing community in an online course. In *Journal of Interactive Online Learning*, 12 (1): 1–16.

Law, M.Y. (2021). Student's attitude and satisfaction towards transformative learning: a research study on emergency remote learning in tertiary education. In *Creative Education*, 12: 494–528. doi:10.4236/ce.2021.123035.

Lee, M.K., Cheung, C.M. and Chen, Z. (2005). Acceptance of Internet-based learning medium: The role of extrinsic and intrinsic motivation. In *Information & Management*, 42: 1095–1104. doi:10.1016/J.IM.2003.10.007.

Li, C.H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. In *Behavior Research Methods*, 48 (3): 936 – 949. doi:10.3758/s13428-015-0619-7.

McCroskey, J.C. and Richmond, V.P. (1982). Communication apprehension and shyness: Conceptual and operational distinctions. In *Central States Speech Journal*, 33 (3): 458–468. doi:10.1080/10510978209388452.

McDonald, R.P. (2000). A basis for multidimensional item response theory. In *Applied Psychological Measurement*, 24 (2): 99–114. doi:10.1177/01466210022031552.

Mehra, V. and Faranak, O. (2012). NOTE FOR EDITOR: Development an instrument to measure university students' attitude towards e-learning. In *Turkish Online Journal of Distance Education*, 13: 34–51.

Murphy, M.P.A. (2020). COVID-19 and emergency eLearning: Consequences of the securitization of higher education for post-pandemic pedagogy. In *Contemporary Security Policy*, 41 (3): 492–505. doi:10.1080/13523260.2020.1761749.

Musella, F., Vicard, P. and De Angelis, M.C. (2022). A bayesian network model for supporting school managers decisions in the pandemic era. In *Social Indicators Research*, 163 (3): 1445–1465. doi:10.1007/s11205-022-02952-3.

Muthèn, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. In *Psychometrika*, 49 (1): 115 – 132. doi:10.1007/BF02294210.

Newsome, M.L., Pina, A.A., Mollazehi, M., Al-Ali, K. and Al-Shaboul, Y. (2022). The effect of learners' sex and STEM/non-STEM majors on remote learning: A national study of undergraduates in Qatar. In *Electronic Journal of e-Learning*, 20: 360–373. doi:10.34190/EJEL.20.4.2262.

Ngah, A.H., Kamalrulzaman, N.I., Mohamad, M.F.H., Abdul Rashid, R., Harun, N.O., Ariffin, N.A. and Abu Osman, N.A. (2022). Do Science and social science differ? Multi-group analysis (MGA) of the willingness to continue online learning. In *Quality & Quantity*. doi:10.1007/s11135-022-01465-y.

Nur'aini, K.D., Werang, B.R. and Suryani, D.R. (2020). Student's learning motivation and learning outcomes in higher education. In *3rd International Conference on Social Sciences (ICSS 2020)*, 463–466. Atlantis Press. doi:10.2991/assehr.k.201014.101.

Owston, R., York, D.N., Malhotra, T. and Sitthiworachart, J. (2020). Blended learning in STEM and non-STEM courses: How do student performance and perceptions compare?. In *Online Learning*, 24: 203–221. doi:10.24059/olj.v24i3.2151.

Phakamach, P., Senarith, P. and Wachirawongpaisarn, S. (2022). The metaverse in education: The future of immersive teaching & learning. In *RICE Journal of Creative Entrepreneurship and Management*, 3 (2): 75–88. doi:10.14456/rjcm.2022.12.

Radovan, M. and Makovec, D. (2022). This is not (the New) normal. students' attitudes towards studying during the COVID-19 pandemic and the determinants of academic overload. In *Electronic Journal of e-Learning*, 20: 257– 269. doi:10.34190/EJEL.20.3.2366.

Rafiq, F. (2020). Analyzing students' attitude towards e-learning: A case study in higher education in Pakistan. In *Pakistan Social Sciences Review*, 4: 367–380. doi:10.35484/PSSR.2020(4-I)29.

Rosseel, Y. (2012). lavaan: An r package for structural equation modeling. In *Journal of Statistical Software*, 48 (2): 1–36. doi:10.18637/jss.v048.i02.

Sanders, L.D., Daly, A.P. and Fitzgerald, K. (2016). Predicting reten-tion, understanding attrition: A prospective study of foundation year students. In *Widening Participation and Lifelong Learning*, 18 (2): 50–83. doi:10.5456/WPLL.18.2.50.

Serhan, D. (2020). Transitioning from Face-to-face to remote learning: Students' attitudes and perceptions of using Zoom during COVID-19 pandemic. In *International Journal of Technology in Education and Science*, 4: 335–342. doi:10.46328/IJTES.V4I4.148.

Sheng, Y. and Wikle, C.K. (2007). Comparing multiunidimensional and unidimensional item response theory models. In *Educational and Psychological Measurement*, 67 (6): 899 – 919. doi:10.1177/0013164406296977.

Tzafilkou, K., Perifanou, M. and Economides, A.A. (2021). Development and validation of a students' remote learning attitude scale (RLAS) in higher education. In *Education and Information Technologies*, 26: 7279–7305. doi:10.1007/S10639-021-10586-0.

Vallerand, R.J., Pelletier, L.G., Blais, M.R., Briere, N.M., Senecal, C. and Vallieres, E.F. (1992). The academic motivation scale: A measure of intrinsic, extrinsic, and amotivation in education. In *Educational and Psychological Measurement*, 52 (4): 1003–1017. doi:10.1177/0013164492052004025.

Zagkos, C., Kyridis, A., Kamarianos, I., Dragouni, K.E., Katsanou, A., Kouroumichaki, E., Papastergiou, N. and Stergianopoulos, E. (2022). Emergency remote teaching and learning in Greek Universities during the COVID-19 pandemic: The attitudes of university students. In *European Journal of Interactive Multimedia and Education*, 3. doi:10.30935/EJIMED/11494.

Zumbo, B.D., Gadermann, A.M. and Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. In *Journal of Modern Applied Statistical Methods*, 6 (1): 21–29. doi:10.22237/jmasm/1177992180.

Çelik, B. and Uzunboylu, H. (2022). Developing an attitude scale towards distance learning. In *Behaviour & Information Technology*, 41: 731–739. doi:10.1080/0144929X.2020.1832576.