

# Interpretative Quantitative Methods in Science Education: Approaches to Multivariate Data Analysis

Matheus Monteiro Nascimento  Brazil  
Estevão Antunes Júnior  Brazil  
Cláudio José de Holanda Cavalcanti  Brazil  
Fernanda Ostermann  Brazil

The field of research in Science Education still lives under the shadow of positivism, which hides in both quantitative and qualitative methodologies. Traditional quantitative methods are shown to be strong agents of the positivist perspective of research, mainly due to the naive interpretation that the numerical data represent the reality of the facts. In order to move away from this view, an advance in the direction of the development of the so-called mixed methods has been observed, especially in regards to the quantitative interpretive analysis, which encourages a greater variety of analytical methods in order to better understand the object of study. Thus, the objective of this paper is to discuss Multidimensional Scaling (MDS) and Correspondence Analysis (CA), which are configured as possibilities of interpretative quantitative methods that are routinely used in the multivariate analysis of data, as well as examples applied in research in the area of Science Education. The result of our presentation is that duly substantiated and reflected interpretative quantitative methods can move away from the positivist paradigm that still permeates the area of Science Education.

**Keywords:** Correspondence Analysis; Multivariate Analysis; Multidimensional Scaling.

## 1. Introduction

Many fields of knowledge rely on scientism to legitimize themselves. This view became hegemonic in Brazil in the 50's and 60's of the last century, when educational policies were strongly directed to the perception of the North American philosopher John Dewey. The objective was to embody education with a scientific method that would provide a sequence of prescribed steps, that is, a science of education (Cunha, 2001).

History shows that the natural sciences have gained great relevance at various times as they represent a supposed pure and absolute truth. Thus arises the so-called scientific method to explain phenomena in science. Such a movement is currently seen as a vision of outdated knowledge production, including in more traditional areas such as physics itself. Although not admittedly present in these discourses, especially scientific ones, this view is implicitly present in theoretical and methodological frameworks.

Methodologically, the idea of seeking the absolute truth has been consistently associated over time with quantitative data analysis processes (Bryman, 1984).

Philosophically, this identification assumes that quantitative research, in this perspective, attempts to find a supposed objective reality in the data through the aid of statistical representations. This would be done by simulating the influence of independent variables on dependent variables, seeking to find causal relationships. Furthermore, when quantitative research is analyzed under this perspective, as if the data actually represented what the researcher needs to analyze, the main purpose of the analysis is subverted and non-numerical factors that may affect the results are disregarded, as well as the researcher's own perception and influence. For example, in large-scale exams such as PISA there is a very strong correlation between student performance and average household income. A light interpretation could conclude that rich students are smarter than low-income students. This simplistic view disregards all elements of children's social context that may influence their school performance.

Still, remnants of this belief in objective reality survive far beyond the tradition of quantitative research, as qualitative research methodologies can incorporate this philosophy in order to give greater credibility to non-numerical data. On qualitative research methodologies that underlie aspects of this idea, there is the Grounded Theory, which admits that raw qualitative data can be fractionated through categories and subcategories in order to extract theories (theories emerge from the data) to reach reliable conclusions (Age, 2011; Cho; Lee, 2014).

It is to no one's surprise that methodological research questions concerns researchers in the field of science education. At the National Meeting of Science Education Research (ENPEC) in 2001, an event organized by the Brazilian Association for Research in Science Education (ABRAPEC), a round-table debate was held to discuss this topic. Some points discussed at the event contribute to the justification of this paper and deserve some highlight. Medeiros (2002) points out the hegemony of the qualitative paradigm in research studies on the field, reversing the quantitative trend observed in the educational field since the mid-1960s. This reversal, however, took the situation to the extreme, to the point that journals would not accept quantitative research (Medeiros, 2002). The author further expresses his disagreement with the Manichean separation of research on Education into two stagnant poles: the qualitative and the quantitative. In this sense, Greca (2002) recommends the integration of qualitative and quantitative approaches in order to minimize the intrinsic limitations of each method. Greca (2002) also pointed out that research studies on science education presents some methodological issues, either because they do not offer an explicit discussion about the adopted methodology or because they do not make an approximation between the theoretical framework used and the research methodology. Regarding investigations supported by statistical analysis, the author points out that they seem to be limited only to survey the frequency of answers to questionnaires or to the exclusive use of statistical averages without further detail on the data.

Along the same lines, Carvalho, Oliveira and Rezende (2009) carried on an analysis of 83 articles published in the Brazilian Journal of Research in Science Education,

a journal administered by ABRAPEC, between 2004 and 2008. The objective of the research was to analyze the specific content, the subject, and the type of methodological approach present in these articles. The authors emphasize that only one work was identified as quantitative, six as qualitative-quantitative and 76 as qualitative. More recently, Schneider, Fujii and Corazza (2017) conducted a similar survey in science education journals between 2015 and 2016. The authors concluded, from the analysis of 240 articles, that none of them took a quantitative approach, and only ten percent used a mixed quali-quantitative perspective. Therefore, despite the strong recommendation of researchers in the field regarding the integration between qualitative and quantitative approaches, what is observed is the preponderance of qualitative analyzes.

In the intention of integrating these two approaches, there is an advance in the literature towards the development of the so-called mixed methods, which combine quantitative and qualitative analyzes (Brannen, 2017; Johnson; Onwuegbuzie, 2004). Therefore, the present paper aims at presenting some quantitative methods that align with the so-called quantitative interpretative analysis (Babones, 2016), which is within the methodological paradigm of mixed methods (Johnson, & Onwuegbuzie, 2004). In this view, the researcher does not limit his investigation to the causal relationship of objective variables. On the contrary, interpretative quantitative analysis attempts to depart from the traditional view of quantitative analysis, in which hypotheses are tested to determine the statistical significance of results on the relationship between variables. From this interpretative perspective, which has historically been associated with qualitative research, a greater variety of analytical methods are used and recommended in order to better understand the relationship between latent concepts (Babones, 2016). Thus, quantitative interpretative methods are potential alternatives for researchers who seek to broaden the understanding of the variables involved in the investigation, thereby becoming able to “see” beyond the raw data.

In order to illustrate this potentiality, Babones (2016) cites the most famous case of a quantitative, interpretative analysis, made way before the existence of modern statistics. The author refers to the British physician John Snow, who discovered a cholera-contaminated well in 1854 in London that was causing a significant number of deaths, judging only from the distance from residents to the well and the presence of the disease in these residents (Cerdeira, & Valdivia, 2007). Although unable to observe cholera either in the well or in the victims, the doctor was able to warn the authorities about the risks of ingestion of those contaminated waters (Babones, 2016). That is, from an interpretation of observable data – the distance from the residents to the well, and presence or absence of the disease – it was possible to infer the existence of an unobservable variable and even formulate a policy of intervention.

From the time of John Snow until today, important technological advances have been observed in the computational analysis of psychological, sociological and educational data. This is evidenced by the ease with which computers can analyze ever larger amounts of data. With this advance, researchers have been aiming to improve

theoretical models, especially those focused on multivariate analysis, prominent tools within the mixed methods paradigm (Johnson; Onwuegbuzie, 2004). Generally speaking, a multivariate analysis refers to any statistical method that simultaneously analyzes multiple measurements of some object under investigation (Hair, Black, Babin, Anderson, & Tatham, 2009).

Knowing the difficulties and complications inherent to the process of understanding these more sophisticated quantitative methods, and which are fundamental for interpretative analysis, the objective of this paper is to contribute to the minimization of these difficulties by discussing two recurrent quantitative treatments used in multivariate analysis data and its applicability in research in the field of Science Education. Multidimensional Scaling (MDS) and Correspondence Analysis (CA) are addressed. The importance of this work lies in the fact that, besides presenting these two sophisticated statistical tools, it discusses and exemplifies the potential of these methods in the context of interpretative quantitative analysis. That is, it intends to show researchers the evolution of statistics and philosophy that grounds quantitative analyzes guided by an interpretative view, which sheds light on new questions and broadens the spectrum of analytical possibilities in relation to the problems of Science Education.

In the following sections, each method is presented from the following strategy: at first, a generic problem that is unrelated to actual research is addressed to illustrate the use of these analytical tools. After detailing these examples, some recent research studies in the area of Science Education that have relied on these methods and which fit into the quantitative interpretative approach are highlighted. It is not considered possible to exhaust all possibilities of using these statistical treatments within a single paper; on the contrary, the purpose of this text is to offer an introductory view of the subject and to present examples of its use in the field of Science Education. Moreover, this text functions as an initial reading for those researchers who aim to carry out investigations based on the approach of quantitative interpretative methods.

It is utterly important to highlight that all the analysis presented in this work were performed in the R programming environment (R Core Team, 2015).

## **2. Multidimensional Scaling**

Multidimensional Scaling is one of the simplest multidimensional analysis tools, initially developed to map distances between objects (Johnson, & Wichern, 2007). This procedure allows the researcher to identify similarities between cases in a two-dimensional plane, taking into account certain variables. The tool provides functions to convert similarities between perceptions about a specific theme into distances presented in a multidimensional space. In this space, each object or event is represented by a point, where the distance between the pairs of points indicates the similarity relationship between them. Its most common applications are in the field of administration (Borg; Groenen; Mair, 2017; Young, 2013) – consumer perceptions of products or brands – and in the field of ecology (Cox; Cox, 2000; Dixon, 2003) – mapping of species in different

regions.

In the field of Science Education, especially in Brazil, MDS originated from research on cognitive mapping and similarity between physical concepts (Greca; Moreira, 2001; Santos, 1978; Santos; Moreira, 1991). More recently, however, this procedure has been used in studies on the social representations of the sciences (Hilger, 2009; Rodrigues; Borges; Pietrocola, 2018). Despite the outstanding investigations, the potential of the applicability of MDS is still little discussed in the Brazilian academic scenario. Therefore, in this section, an MDS will be performed on generic data, in order to make clear to the reader the idea of using this tool. Furthermore, will be presented investigations that relied on MDS for studies different than those described above, in order to illustrate its multiple applications.

Multidimensional Scaling is strongly related to the concept of dissimilarity, which in turn has a complementary relationship with the concept of similarity, as shown by the mathematical relationship  $d = 1 - s$ , where  $d$  is the dissimilarity and  $s$  is the similarity. Thus, if  $s = 1$  (two identical subjects or objects according to a series of parameters), we have  $d = 0$  and vice versa. As discussed at the beginning of this section, MDS is able to represent the similarity between cases on a multidimensional map. For this, it is first necessary to build the so-called dissimilarity matrix. A classic and intuitive example of a dissimilarity matrix for performing an MDS is an array of Euclidean distances between cities in a country (the names of cities are in the rows and columns). The elements of this matrix would be the distances between cities. Thus, the “dissimilarity” between city pairs would be quantified solely by the Euclidean distance between them. It is important to stress that the concept of dissimilarity is not limited to that of Euclidean distance. There are several measures of dissimilarity in statistics, and only one of them is the Euclidean distance, which is probably one of the sources of confusion between the concepts of distance and dissimilarity. The dissimilarity matrix resembles an inverse correlation matrix (correlation is closer to similarity).

As an example, Figure 1 shows (Euclidean) distances between some cities in Europe, constituting a typical matrix of dissimilarity. These data were obtained from the eurodist dataset, which is present in the R programming environment and contains distances, in kilometers, between 21 cities in Europe.

In this case, the diagonal elements of this matrix are null because they represent the distance from a city to itself (in other words, this represents complete similarity). The greater the distance between any two cities, the greater their dissimilarity. The greatest dissimilarity, therefore, is observed between Lisbon (LIS) and Athens (ATH), which means that they are the most distant cities in geographical terms. The process of MDS consists in reproducing, from this matrix of dissimilarity and specific algorithms, a pictorial representation of these dissimilarities that would redevelop the positions of these cities in the geographical map. This way, the MDS will result in a graph where each point will represent a city.

	Athens	Barcelona	Brussels	Calais	Cherbourg	Cologne	Copenhagen	Genève	Gibraltar	Hamburg	Hook of Holland	Lisbon	Lyon	Madrid	Marseille	Milan	Munich	Paris	Rome	Stockholm	Vienna
Athens	0	3313	2963	3175	3339	2762	3276	2610	4485	2977	3030	4532	2753	3949	2865	2282	2179	3000	817	3927	1991
Barcelona	3313	0	1318	1326	1294	1498	2218	803	1172	2018	1490	1305	645	636	521	1014	1365	1033	1460	2868	1802
Brussels	2963	1318	0	204	583	206	966	677	2256	597	172	2084	690	1558	1011	925	747	285	1511	1616	1175
Calais	3175	1326	204	0	460	409	1136	747	2224	714	330	2052	739	1550	1059	1077	977	280	1662	1786	1381
Cherbourg	3339	1294	583	460	0	785	1545	853	2047	1115	731	1827	789	1347	1101	1209	1160	340	1794	2196	1588
Cologne	2762	1498	206	409	785	0	760	1662	2436	460	269	2290	714	1764	1035	911	583	465	1497	1403	937
Copenhagen	3276	2218	966	1136	1545	760	0	1418	3196	460	269	2971	1458	2498	1778	1537	1104	1176	2050	650	1455
Genève	2610	803	677	747	853	1662	1418	0	1975	1118	895	1936	158	1439	425	328	591	513	995	2068	1019
Gibraltar	4485	1172	2256	2224	2047	2436	3196	1975	0	2897	2428	676	1817	698	1693	2185	2565	1971	2631	3886	2974
Hamburg	2977	2018	597	714	1115	460	460	1118	2897	0	550	2671	1159	2198	1479	1238	805	877	1751	949	1155
Hook of Holland	3030	1490	172	330	731	269	269	895	2428	550	0	2280	863	1730	1183	1098	851	457	1683	1500	1205
Lisbon	4532	1305	2084	2052	1827	2290	2971	1936	676	2671	2280	0	1178	668	1762	2250	2507	1799	2700	3231	2937
Lyon	2753	645	690	739	789	714	1458	158	1817	1159	863	1178	0	1281	320	328	724	471	1048	2108	1157
Madrid	3949	636	1558	1550	1347	1764	2498	1439	698	2198	1730	668	1281	0	1157	1724	2010	1273	2097	3188	2409
Marseille	2865	521	1011	1059	1101	1035	1778	425	1693	1479	1183	1762	320	1157	0	618	1109	792	1011	2428	1363
Milan	2282	1014	925	1077	1209	911	1537	328	2185	1238	1098	2250	328	1724	618	0	331	856	586	2187	898
Munich	2179	1365	747	977	1160	583	1104	591	2565	805	851	2507	724	2010	1109	331	0	821	946	1754	428
Paris	3000	1033	285	280	340	465	1176	513	1971	877	457	1799	471	1273	792	856	821	0	1476	1827	1249
Rome	817	1460	1511	1662	1794	1497	2050	995	2631	1751	1683	2700	1048	2097	1011	586	946	1476	0	2707	1209
Stockholm	3927	2868	1616	1786	2196	1403	650	2068	3886	949	1500	3231	2108	3188	2428	2187	1754	1827	2707	0	2105
Vienna	1991	1802	1175	1381	1588	937	1455	1019	2974	1155	1205	2937	1157	2409	1363	898	428	1249	1209	2105	0

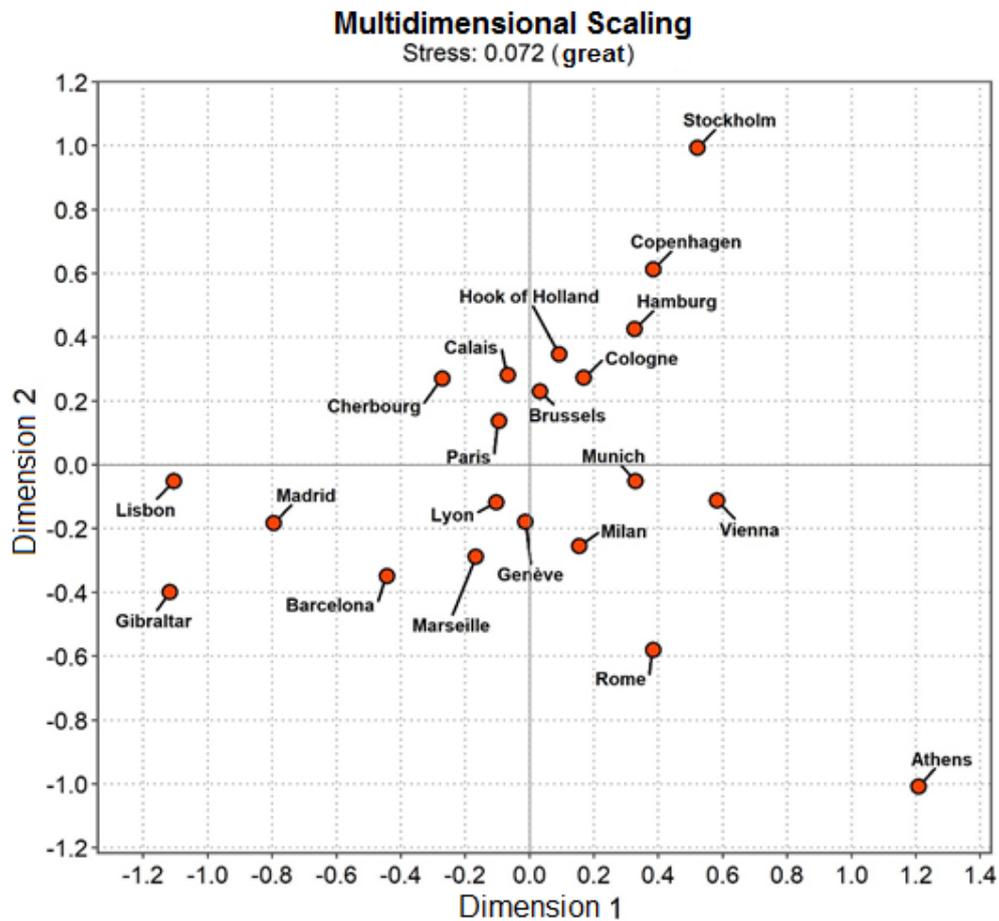
Figure 1. Dissimilarity matrix indicating Euclidean distances between cities in Europe

Source: Authors (Data: The Cambridge Encyclopedia).

Figure 2 shows the resulting MDS map from the dissimilarity matrix of Figure 1, using the R’s smacof package (De Leeuw, & Mair, p., 2009). A simple inspection shows that the position of cities in the resulting MDS map bears relative resemblance to the actual positions observed on a geographical map of the European continent. Similar to what was observed in the dissimilarity matrix, the cities of Lisbon and Athens are the furthest from the map (values marked in blue in the matrix of Figure 1). Conversely, the two closest cities on the map, which logically present the least dissimilarity, are the cities of Lyon and Genève (values marked in red in the same matrix). It should be emphasized that the amount called stress, synthetically named  $\sigma$  (Borg et al., 2017, p. 23; De Leeuw; Mair, 2009, section 3.1), gives a measure of how good the fit of points on the map was. The following criteria can be used for evaluating the quality of the representations (Levshina, 2015, p. 341):  $\sigma \geq 0,2$  (poor);  $0,1 \leq \sigma < 0,2$  (good/ok);  $0,05 \leq \sigma < 0,1$  (very good) e  $\sigma \leq 0,05$  (excellent).

This type of analysis, in which dissimilarities are the Euclidean distances, is not suitable for most research studies, including those on science education. Dissimilarity is not immediately measurable and should be estimated according to qualitative criteria, then transformed into a numerical result by some mathematical model. Thus, for these investigations, the construction of a dissimilarity matrix is no longer a trivial issue as illustrated in Figure 1. For the sake of clarity, we will construct a dissimilarity matrix from raw data such as those obtained often in educational investigative questionnaires, where respondents may opt trough a Likert-type scale (Likert, 1932). As an example, we

used data from a science and technology perception questionnaire applied to European citizens in the early 1990s (Reif; Melich, 1995). Seven questions were selected with four answer options each, ranged between strongly disagree to strongly agree. This data is found in the R's ltm package (Rizopoulos, 2006).



**Figure 2.** Map obtained from the MDS performed in the dissimilarity matrix of Figure 1. The map in general is called *similarity map* because it allows to associate similarities with the proximity between the points. Here, however, the map reproduces the approximate distance between cities and their relative position (compare it to a geographical map)

In order to simplify the analyses, the first 30 respondents of this questionnaire (denoted by R1 to R30) were selected from a total of 392. The answers of these respondents are illustrated at the top of Figure 3 – and below are the questions for each column. For exercise purposes, the following research question will be proposed: how similar are the respondents to this questionnaire? In order to answer such a question, an MDS will be performed. The scale of the answers is as follows: 1 (strongly disagree), 2 (disagree), 3 (agree) and 4 (strongly agree). As seen, for the elaboration of the resulting MDS map, it is first necessary to obtain the dissimilarity matrix from the raw data. Constructing an array in this format means calculating dissimilarity between all pairs of respondents. Although it is feasible to calculate it by hand, it would be unfeasible for such data as it would lead

to a 30×30 matrix, resulting in 450 dissimilarities to be calculated since the matrix is symmetrical. Even though the calculation is simple individually, this is not practical. From the table with the raw data, which are the answers of the different respondents – arranged in the rows of the table in Figure 3 (a) –, we calculate the dissimilarity matrix using R – here the subjects are the respondents (which we want compare to each other) and the attributes are the answers to the questionnaire items. To calculate dissimilarities, it is necessary to use a mathematical model that proposes to quantify it. There are many types of models for calculating dissimilarities, each suited to a data type. For categorical data (in this case, ordinal, since it is a Likert scale), the Euclidean distance is not the most appropriate dissimilarity measure. For this type of data, it is possible to use the Bray-Curtis dissimilarity (Quinn, & Keough, 2002, p. 414), which can be easily obtained with R, for example, through the *vegan* package (Oksanen et al., 2019).

There are better dissimilarity measures for categorical or ordinal data, but far more complex than the Bray-Curtis dissimilarity, so in this example we will adopt it for simplicity. Although R calculates all dissimilarities between pairs and builds the matrix without major problems, as the aim of the present work is to make the use of these tools as accessible as possible to the reader, the dissimilarity between the first two respondents will now be calculated. Bray-Curtis dissimilarity is given by the following equation (Johnson; Myatt, 2014):

$$d_B(i, j) = \frac{\sum_{k=1}^n |x_{ik} - x_{jk}|}{\sum_{k=1}^n x_{ik} + \sum_{k=1}^n x_{jk}} \quad (1)$$

In the above expression, we have:

- $d_B(i, j)$  → Bray-Curtis dissimilarity between respondents  $i$  and  $j$  (lines).
- $x_{ik}$  → value assigned by respondent  $i$  (line) to the question  $k$  (column).
- $x_{jk}$  → value assigned by respondent  $j$  (line) to the question  $k$  (column).

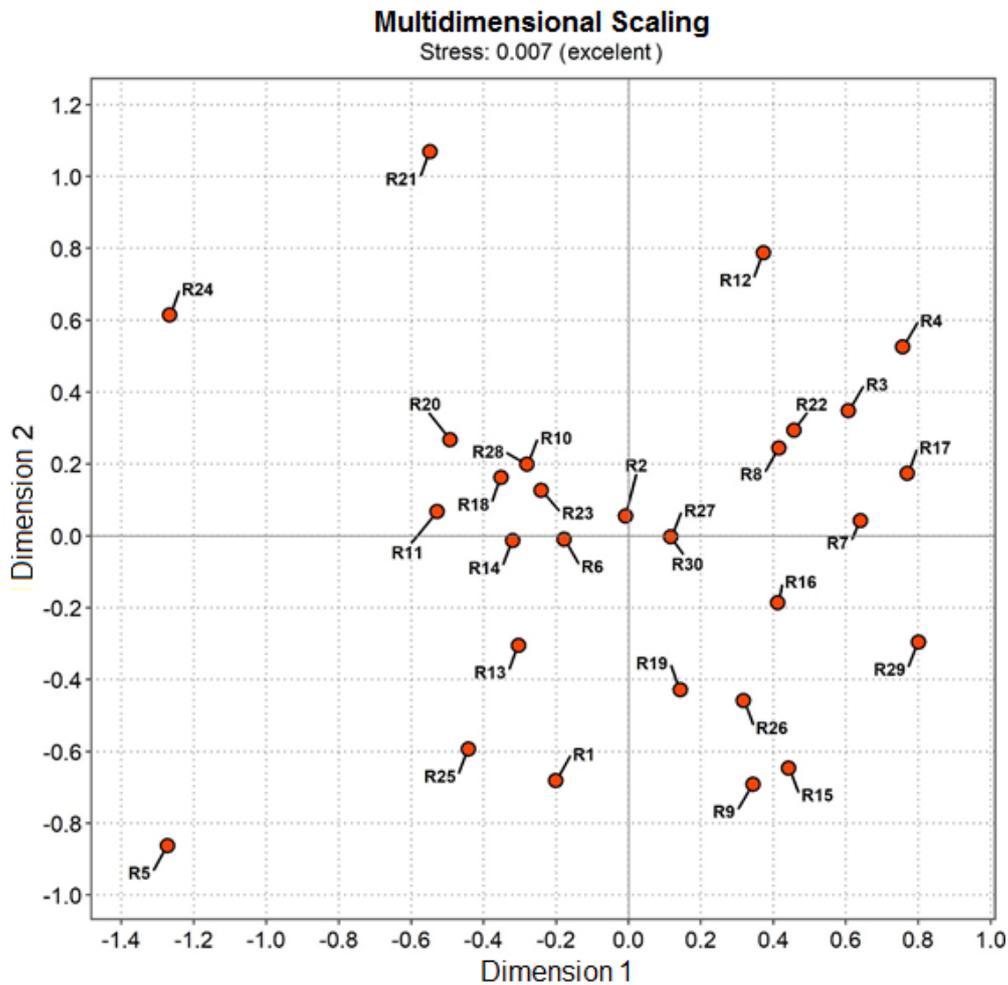
Using equation (1) and the values assigned by respondents 1 and 2 – taken from the upper table of Figure 3 – it is clear that the dissimilarity between them is 0.09. The calculation of this dissimilarity and for other respondents is represented in the lower table of Figure 3. Clearly, the dissimilarity matrix is symmetrical, and the main diagonal elements are null (dissimilarity between a subject and himself). In this case, these quantities estimate how distinct the response sequence was from one respondent to another.

After the dissimilarity matrix is built, the MDS can be run, and the dissimilarity map obtained. Remember that the hypothetical research question is to determine the similarity between the thirty respondents of the Science and Technology perception questionnaire, that is, how the respondents assemble or antagonize in terms of the answers to each item of the questionnaire. Figure 4 shows the map obtained from the MDS. The interpretation is very similar to that given on the map of Euclidean distances from European cities.

Respondent	Confort <sup>a</sup>	Environment <sup>b</sup>	Work <sup>c</sup>	Future <sup>d</sup>	Technology <sup>e</sup>	Industry <sup>f</sup>	Benefits <sup>g</sup>	
R1	4	4	4	3	4	3	2	<sup>a</sup> Science and technology are making our lives healthier, easier and more comfortable.
R2	3	4	3	3	3	3	3	
R3	3	2	2	2	4	4	3	
R4	3	3	2	2	4	4	3	
R5	3	1	4	4	2	3	1	<sup>b</sup> Scientific and technological research cannot play an important role in protecting and repairing the environment..
R6	4	3	4	3	3	4	3	
R7	3	2	2	3	4	4	4	
R8	3	2	2	3	3	4	4	
R9	3	3	3	4	4	4	2	<sup>c</sup> The application of science and new technologies will make the work more interesting.
R10	4	3	3	3	3	3	3	
R11	3	3	3	4	2	3	3	
R12	3	3	1	2	2	4	3	
R13	3	2	3	3	4	3	3	<sup>d</sup> Thanks to science and technology, there will be more opportunities for future generations.
R14	3	2	3	3	2	2	2	
R15	3	4	2	3	4	3	2	
R16	3	4	3	3	4	4	3	
R17	3	3	1	2	3	2	2	<sup>e</sup> New technology does not depend on basic scientific research.
R18	3	3	3	3	2	3	3	
R19	3	2	2	3	3	3	2	
R20	3	3	3	3	2	2	3	
R21	2	3	3	1	2	3	3	<sup>f</sup> Scientific and technological research does not play an important role in industrial development.
R22	3	3	1	3	3	4	3	
R23	3	3	3	3	3	3	3	
R24	1	2	4	2	2	3	3	
R25	3	1	3	3	3	3	2	<sup>g</sup> The benefits of science outweigh any detrimental effects it may have.
R26	3	3	2	3	3	3	2	
R27	3	4	2	3	3	3	3	
R28	4	3	3	3	3	3	3	
R29	3	4	1	3	4	4	2	
R30	3	4	2	3	3	3	3	

Respondent	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16	R17	R18	R19	R20	R21	R22	R23	R24	R25	R26	R27	R28	R29	R30
R1	0.00	0.09	0.18	0.16	0.19	0.08	0.17	0.20	0.11	0.09	0.16	0.24	0.11	0.17	0.07	0.08	0.20	0.14	0.14	0.16	0.22	0.18	0.11	0.22	0.14	0.12	0.11	0.09	0.11	0.11
R2	0.09	0.00	0.14	0.12	0.20	0.09	0.14	0.12	0.11	0.05	0.07	0.15	0.07	0.13	0.07	0.04	0.16	0.05	0.10	0.07	0.13	0.10	0.02	0.18	0.10	0.07	0.02	0.05	0.12	0.02
R3	0.18	0.14	0.00	0.02	0.26	0.14	0.05	0.07	0.12	0.14	0.17	0.11	0.07	0.19	0.12	0.09	0.17	0.15	0.11	0.18	0.19	0.10	0.12	0.19	0.16	0.13	0.12	0.14	0.12	0.12
R4	0.16	0.12	0.02	0.00	0.28	0.11	0.07	0.10	0.09	0.12	0.14	0.08	0.10	0.21	0.10	0.07	0.14	0.12	0.13	0.15	0.16	0.07	0.10	0.21	0.18	0.10	0.10	0.12	0.10	0.10
R5	0.19	0.20	0.26	0.28	0.00	0.19	0.25	0.23	0.17	0.20	0.13	0.28	0.18	0.14	0.23	0.24	0.29	0.16	0.17	0.19	0.26	0.26	0.18	0.20	0.11	0.19	0.23	0.20	0.28	0.23
R6	0.08	0.09	0.14	0.11	0.19	0.00	0.13	0.11	0.11	0.04	0.11	0.14	0.11	0.17	0.16	0.08	0.20	0.09	0.14	0.12	0.17	0.09	0.07	0.17	0.14	0.12	0.11	0.04	0.16	0.11
R7	0.17	0.14	0.05	0.07	0.25	0.13	0.00	0.02	0.11	0.14	0.16	0.15	0.07	0.18	0.12	0.09	0.21	0.14	0.10	0.17	0.23	0.10	0.12	0.23	0.15	0.12	0.12	0.14	0.12	0.12
R8	0.20	0.12	0.07	0.10	0.23	0.11	0.02	0.00	0.14	0.12	0.14	0.13	0.10	0.16	0.14	0.11	0.19	0.12	0.08	0.15	0.21	0.07	0.10	0.21	0.13	0.10	0.10	0.12	0.14	0.10
R9	0.11	0.11	0.12	0.09	0.17	0.11	0.11	0.14	0.00	0.11	0.09	0.17	0.09	0.15	0.09	0.06	0.18	0.12	0.12	0.14	0.20	0.10	0.09	0.25	0.12	0.10	0.14	0.11	0.09	0.14
R10	0.09	0.05	0.14	0.12	0.20	0.04	0.14	0.12	0.11	0.00	0.07	0.15	0.07	0.13	0.12	0.09	0.16	0.05	0.10	0.07	0.13	0.10	0.02	0.18	0.10	0.07	0.07	0.00	0.16	0.07
R11	0.16	0.07	0.17	0.14	0.13	0.11	0.16	0.14	0.09	0.07	0.00	0.13	0.10	0.11	0.14	0.11	0.19	0.02	0.13	0.05	0.11	0.12	0.05	0.16	0.13	0.10	0.10	0.07	0.19	0.10
R12	0.24	0.15	0.11	0.08	0.28	0.14	0.15	0.13	0.17	0.15	0.13	0.00	0.18	0.20	0.18	0.14	0.12	0.11	0.17	0.14	0.14	0.05	0.13	0.20	0.22	0.14	0.13	0.15	0.13	0.13
R13	0.11	0.07	0.07	0.10	0.18	0.11	0.07	0.10	0.09	0.07	0.10	0.18	0.00	0.11	0.10	0.07	0.19	0.07	0.08	0.10	0.16	0.12	0.05	0.16	0.08	0.10	0.10	0.07	0.14	0.10
R14	0.17	0.13	0.19	0.21	0.14	0.17	0.18	0.16	0.15	0.13	0.11	0.20	0.11	0.00	0.16	0.17	0.15	0.08	0.09	0.06	0.18	0.19	0.11	0.18	0.09	0.11	0.16	0.13	0.21	0.16
R15	0.07	0.07	0.12	0.10	0.23	0.16	0.12	0.14	0.09	0.12	0.14	0.18	0.10	0.16	0.00	0.07	0.14	0.12	0.08	0.15	0.21	0.12	0.10	0.26	0.13	0.05	0.05	0.12	0.05	0.05
R16	0.08	0.04	0.09	0.07	0.24	0.08	0.09	0.11	0.06	0.09	0.11	0.14	0.07	0.17	0.07	0.00	0.20	0.09	0.14	0.12	0.17	0.09	0.07	0.22	0.14	0.12	0.07	0.09	0.07	0.07
R17	0.20	0.16	0.17	0.14	0.29	0.20	0.21	0.19	0.18	0.16	0.19	0.12	0.19	0.15	0.14	0.20	0.00	0.17	0.12	0.14	0.21	0.11	0.14	0.27	0.18	0.09	0.14	0.16	0.14	0.14
R18	0.14	0.05	0.15	0.12	0.16	0.09	0.14	0.12	0.12	0.05	0.02	0.11	0.07	0.08	0.12	0.09	0.17	0.00	0.11	0.03	0.08	0.10	0.02	0.14	0.11	0.08	0.07	0.05	0.17	0.07
R19	0.14	0.10	0.11	0.13	0.17	0.14	0.10	0.08	0.12	0.10	0.13	0.17	0.08	0.09	0.08	0.14	0.12	0.11	0.00	0.14	0.20	0.11	0.08	0.20	0.06	0.03	0.08	0.10	0.13	0.08
R20	0.16	0.07	0.18	0.15	0.19	0.12	0.17	0.15	0.14	0.07	0.05	0.14	0.10	0.06	0.15	0.12	0.14	0.03	0.14	0.00	0.11	0.13	0.05	0.17	0.14	0.11	0.10	0.07	0.20	0.10
R21	0.22	0.13	0.19	0.16	0.26	0.17	0.23	0.21	0.20	0.13	0.11	0.14	0.16	0.18	0.21	0.17	0.21	0.08	0.20	0.11	0.00	0.19	0.11	0.12	0.20	0.17	0.16	0.13	0.26	0.16
R22	0.18	0.10	0.10	0.07	0.26	0.09	0.10	0.07	0.12	0.10	0.12	0.05	0.12	0.19	0.12	0.09	0.11	0.10	0.11	0.13	0.19	0.00	0.07	0.24	0.16	0.08	0.07	0.10	0.07	0.07
R23	0.11	0.02	0.12	0.10	0.18	0.07	0.12	0.10	0.09	0.02	0.05	0.13	0.05	0.11	0.10	0.07	0.14	0.02	0.08	0.05	0.11	0.07	0.00	0.16	0.08	0.05	0.05	0.02	0.14	0.05
R24	0.22	0.18	0.19	0.21	0.20	0.17	0.23	0.21	0.25	0.18	0.16	0.20	0.16	0.18	0.26	0.22	0.27	0.14	0.20	0.17	0.12	0.24	0.16	0.00	0.20	0.22	0.21	0.18	0.32	0.21
R25	0.14	0.10	0.16	0.18	0.11	0.14	0.15	0.13	0.12	0.10	0.13	0.22	0.08	0.09	0.13	0.14	0.18	0.11	0.06	0.14	0.20	0.16	0.08	0.20	0.00	0.08	0.13	0.10	0.18	0.13
R26	0.12	0.07	0.13	0.10	0.19	0.12	0.12	0.10	0.10	0.07	0.10	0.14	0.10	0.11	0.05	0.12	0.09	0.08	0.03	0.11	0.17	0.08	0.05	0.22	0.08	0.00	0.05	0.07	0.10	0.05
R27	0.11	0.02	0.12	0.10	0.23	0.11	0.12	0.10	0.14	0.07	0.10	0.13	0.10	0.16	0.05	0.07	0.14	0.07	0.08	0.10	0.16	0.07	0.05	0.21	0.13	0.05	0.00	0.07	0.10	0.00
R28	0.09	0.05	0.14	0.12	0.20	0.04	0.14	0.12	0.11	0.00	0.07	0.15	0.07	0.13	0.12	0.09	0.16	0.05	0.10	0.07	0.13	0.10	0.02	0.18	0.10	0.07	0.07	0.00	0.16	0.07
R29	0.11	0.12	0.12	0.10	0.28	0.16	0.12	0.14	0.09	0.16	0.19	0.13	0.14	0.21	0.05	0.07	0.14	0.17	0.13	0.20	0.26	0.07	0.14	0.32	0.18	0.10	0.10	0.16	0.00	0.10
R30	0.11	0.02	0.12	0.10	0.23	0.11	0.12	0.10	0.14	0.07	0.10	0.13	0.10	0.16	0.05	0.07	0.14	0.07	0.08	0.10	0.16	0.07	0.05	0.21	0.13	0.05	0.00	0.07	0.10	0.00

Figure 3. Answers to the Science and Technology Perceptions Questionnaire (top table), dissimilarity matrix calculated from the raw questionnaire data (bottom table). The quiz questions regarding the items placed in the top table columns are shown below that table



**Figure 4.** Map obtained in the MDS from the dissimilarity matrix calculated with the data from the Science and Technology perception questionnaire

However, now the distance on the map indicates dissimilarities between respondents. It is possible to notice that four respondents are totally similar to each other, as is the case of respondents R28 and R10 and respondents R27 and R30. Visually they are all overlaid on the map. This is because the answers attributed to the seven questionnaire items were strictly the same (resulting in null dissimilarity, marked with red in the lower table of Figure 3). The detailed examination of the assigned answers also allows to infer what makes these individuals similar, that is, they answered the items in a similar way. This is the case of respondents R8 and R22, R18 and R28, R18 and R10, among others.

It can also be inferred which individuals answered the questionnaire items more antagonistically, namely, disagreeing more with each other. The “most antagonistic” respondents are R24 and R29, as they have the greatest dissimilarity to each other (marked in blue in the lower table of Figure 3) - this feature is well reproduced in the map of Figure 4, since it is noted that R24 and R29 are the farthest points from each other. In addition, respondents R5 and R24 and R21 have disagreeable attitudes towards

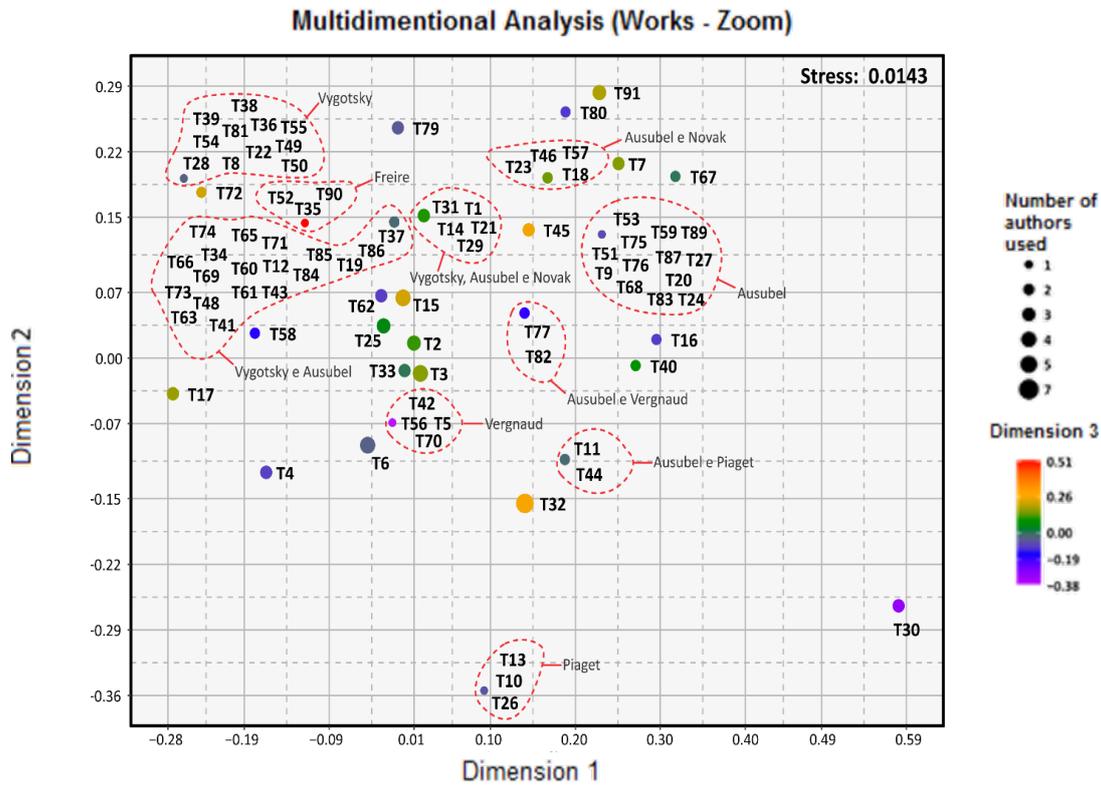
Science and Technology in comparison to the other respondents, as they appear isolated on the map, away from the rest of the group.

Despite being a simple example, we can see that the application of Multidimensional Scaling in research on the field of Science Education is a viable and interesting alternative when we intend to infer relationships between subjects (i.e. the respective questionnaire response patterns analyzed), which are difficult to derive directly from raw data. Moreover, in the area of Science education, data-based investigations whose answers are ordinal, coming from questionnaires, are very common. The analysis of these questionnaires, when performed using traditional methods such as the chi-square, only indicates the association between the categories, but says nothing about the relationship between the respondents of the questionnaire. Multidimensional Scaling makes it possible to make such an association, as we illustrated with the above examples. Evidently, the MDS alone shows similarities and differences between respondents, not allowing to obtain reasons for these differences. To go deeper into the issue of these differences and similarities, this method can be used in conjunction with research methodologies of qualitative nature, in an investigative setting based on solid theoretical frameworks.

## 2.1 The MDS in a recent research study on the field

As an example of recent research on the field of science education that relied on MDS to determine the similarity between dissertations of a professional master's degree course in Physics Education, we can mention that carried on by Nascimento (2016). The author's objective was to analyze the similarity between 90 dissertations from the choice of theoretical references. The raw data used in the analysis were obtained from a table whose lines were the dissertations (T1, T2, T3 ...), and the columns were the theoretical references adopted in the respective works (Ausubel, Vygotsky, Piaget, among others). If a particular dissertation had a Freireian approach as a theoretical framework, for example, then a number 1 was marked in the column named Freire, leaving 0 for all other columns. The investigated similarity focused only on the authors chosen and tended to increase when two works adopted at least one author in common, since some had more than one author as theoretical reference. In this particular case, considering that the data are binary, the Jaccard (Johnson, & Myatt, 2014) dissimilarity was calculated and not that of Bray-Curtis as in the previous example.

As long as they are binary and indicate presence/absence (or yes/no), these data are characterized as categorical. However, they can also be characterized as ordinal data, since there is a natural ordering: if a given author is present as a reference in a certain work (encoded with 1), it can arguably be placed in a higher hierarchy than those who are not (encoded with 0). Due to the nature of the data, the best visualization occurred for three dimensions. This kind of dissimilarity measure can also be calculated by R's *vegan* package.



**Figure 5.** Similarity analysis between dissertation works of a Professional Master's degree course in Physics Education

Source: Nascimento (2016)

Figure 5 shows the resulting MDS map performed on the dissimilarity matrix calculated from the presence/absence data of theoretical references in the dissertations from a master's program. Each point on the map indicates a dissimilarity, whereas the color indicates the position in the third dimension. In addition, the larger the point size, the more authors that work used as a theoretical framework. The red dotted line delimits the totally similar work groups, indicating the authors that constitute the theoretical references of these work groups. There is clearly the formation of a main group, from which works T30 and works of group T10-T13-T26 are removed. There are 11 works that used only Vygotsky, and 48 in total that used the same author, namely, there are 37 works that use Vygotsky among other authors and with which these 11 have some similarity. It turns out that the 11 works that use only Vygotsky have much more similarity to the others than the group of three that use Piaget – exactly as an “attraction effect” from these similarities.

The same reasoning can be used to explain why works that use only Ausubel are also closer to the main group than those that use only Piaget – 13 use only Ausubel, and 49 out of the total 62 that use the same author bring other authors too and are distributed in the main group. The works T35, T52 and T90 use only Freire, T5, T42, T56 and T70 use

only Vergnaud, and T64 and T78 have no definite theoretical framework. On the map, works that use only one author tend to position themselves more at the edges except for T5, T42, T56 and T70, which are close to T6 (that uses Vygotsky, Ausubel, Piaget, Vergnaud and Bruner) and T4 (Vygotsky, Piaget and Vergnaud). This is an example of how important the dimension 3 is because T6 looks closer to the T5-T42-T56-T70 group than T4. However, the color of T6 shows that it is closer to the dimension than group T4. The map offset between the T5-T42-T56-T70 and T6 group is 0.366, and between that same group and T4 is 0.330, slightly lower. What the MDS provided were some clues as to how the theoretical references were being viewed in the program in question, elements that could hardly be identified from a simple inspection of the data.

### 3. Correspondence Analysis

Correspondence Analysis (CA) “provides a multivariate representation of interdependence for nonmetric data that is not possible with other methods” (Hair et al., 2009, p. 20). In other words, the objective of CA is to visualize the association between categorical variables in a small map. This technique allows to determine the degree of association between rows and columns of a table. In practical terms, there is more than one type of CA, but all can be divided basically into two large groups, namely: Simple Correspondence Analysis (SCA) – used to study associations between two categorical variables; and Multiple Correspondence Analysis (MCA – a generalization of SCA to study more associations between multiple categorical variables) (Greenacre, 2017). The primary objective of MCA is to provide dimensionality reduction, namely, to visualize multidimensional data in a reduced number of dimensions (usually two or three, i.e. on a flat or three-dimensional space, so that it reproduces the table’s total variance well). Thus, MCA analyses are data mining techniques that do not necessarily aim at generalizing to a wide variety of contexts, but rather at explaining important patterns or associations that can be articulated with quantitative or qualitative analyses. This technique has its mathematical origin in the work of Hirschfeld (1935) and since then it has been used in different contexts, such as ecology and psychology. As it facilitates the analysis of the association between categorical variables, CA has become an important tool for research in the humanities, health and social sciences (Greenacre; Blasius, 1994; Ferreira, 2003), especially since the sixties, with the works of the French sociologist Pierre Bourdieu (De Nooy, 2003). In the field of science education, on the other hand, CA is also a technique that is not yet part of the methodological repertoire of most researchers, with a few exceptions (Adamuti-Trache; Andres, 2008; Andres; Adamuti-Trache, 2008; Nascimento; Cavalcanti; Ostermann, 2017; Shanka; Quintal; Taylor, 2006).

Correspondence Analyses (CA) have contingency tables (cross tables that show the frequency distribution of common occurrences among the various categories of variables) as input data. As an example, Figure 6 (a) indicates the contingency table from hypothetical data of a survey on how often certain groups of people smoke cigarettes. These data were adapted from Greenacre (1984). This table contains five rows (groups

of people) and four columns (smoking categories), where the frequencies of smoking categories appear for each employee group in a fictitious organization. For example, 25 people in Group C claim not to smoke, while 13 people in Group D claim to smoke a lot. These quantities are called observed values. Correspondence Analysis allows the visualization of the association between the categories of these two variables, which are the group of people and the smoking categories.

(a) Contingency table

	None	Light	Medium	Heavy	PO <sub>1</sub>
Group A	4	2	3	2	0.056995
Group B	4	3	7	4	0.093264
Group C	25	10	12	4	0.264249
Group D	18	24	33	13	0.455959
Group E	10	6	7	2	0.129534
PO <sub>2</sub>	0.316062	0.233161	0.321244	0.129534	

(b) Expected values

	None	Light	Medium	Heavy
Group A	3.48	2.56	3.53	1.42
Group B	5.69	4.20	5.78	2.33
Group C	16.12	11.89	16.38	6.61
Group D	27.81	20.59	28.27	11.4
Group E	7.90	5.83	8.03	3.24

(c) Standardized residuals

	None	Light	Medium	Heavy
Group A	0.28	-0.35	-0.28	0.48
Group B	-0.71	-0.58	0.51	1.09
Group C	2.21	-0.55	-1.08	-1.01
Group D	-1.86	0.77	0.89	0.47
Group E	0.75	0.07	-0.36	-0.69

**Figure 6.** (a) Contingency table, (b) expected values and (c) standardized residuals for smoker data from a fictitious company

In Figure 6 (a), the last column (PO<sub>1</sub>) and last row (PO<sub>2</sub>) indicate, respectively, the proportion of occurrence of each row (groups of people) and column (smoking categories). For example, the proportion of people who claim to never have smoked (category *None*) is given by the ratio between the sum of the frequencies in the corresponding column and the total number of cases, that is,  $(4 + 4 + 25 + 18 + 10) / 193 \approx 0.316062$ . This number can be understood as the probability of occurrence of this group within the table, that is, the probability of a person claiming that they do not smoke is approximately 31.6 percent within the total population investigated. Figure

6 (b) shows the expected values for frequencies, accurate to two decimal places. This magnitude, the expected value of each case, can easily be calculated from the observed values. To do this, simply multiply the row occurrence ratio by the column occurrence ratio and the total number of cases. For non-smokers within Group C, for example, the expected value is given by  $0.264249 \times 0.316062 \times 193 \approx 16.12$ .

Figure 6 (c) shows the most important results for performing Correspondence Analysis, which are the standardized residues. The standardized residue  $R_p$  is defined as

$$R_p = \frac{V_o - V_e}{\sqrt{V_e}}, \tag{2}$$

- $V_o$  → is the observed value;
- $V_e$  → is the expected value.

The standardized residual informs the deviation in relation to the expected value in square root units of the expected value. That is, it indicates how far the observed variable deviated from the expected value. The lower (higher) the modulus of the residue, the smaller (larger) will be the association between the categories. This matrix with the standardized residuals is a key part for the map construction with the visualization of the association between the variables from the CA. Mathematically, CA relies on the singular value decomposition of the residuals matrix to graph the rows and columns of the contingency table as points in small vector spaces (Souza, Bastos & Vieira, 2010). This decomposition is performed from standardized residuals and will be detailed below.

Singular Value Decomposition (SVD) is a linear algebra technique used to factor matrices (Anton, & Busby, 2006). SVD decomposes a matrix into a product of the factors of the other three matrices. Thus, whether  $C$  is a real or complex matrix, then there are two unit orthogonal matrices  $U$  and  $V$ ; and a diagonal matrix  $\Sigma$  as in:

$$C = U\Sigma V^T \tag{3}$$

where  $V^T$  is the transposed matrix of  $V$ . The diagonal values of  $\Sigma$  are called singular values<sup>1</sup>, expressed as a vector  $\sigma$  and correspond to the square root of the eigenvalues  $\lambda$  of matrix  $C$ .

In the case of Correspondence Analysis, matrix  $C$  is simply the ratio between the standardized residuals matrix and the square root of the total number of cases, counted from the contingency table in Figure 6 (a). The  $x_{row}$  and  $x_{column}$  coordinates of the map are obtained from matrices  $U$  and  $V$  based on the relationships (Greenacre, 2017):

$$x_{row} = \frac{U\sigma}{\sqrt{PO_1}} \tag{4}$$

---

<sup>1</sup> For further details on singular value decomposition, we suggest the following links: [http://web.mit.edu/be.400/www/SVD/Singular\\_Value\\_Decomposition.htm](http://web.mit.edu/be.400/www/SVD/Singular_Value_Decomposition.htm) and <http://web.cs.iastate.edu/~cs577/handouts/svd.pdf>

and

$$\mathbf{x}_{\text{column}} = \frac{\mathbf{V}\boldsymbol{\sigma}}{\sqrt{PO_2}}. \quad (5)$$

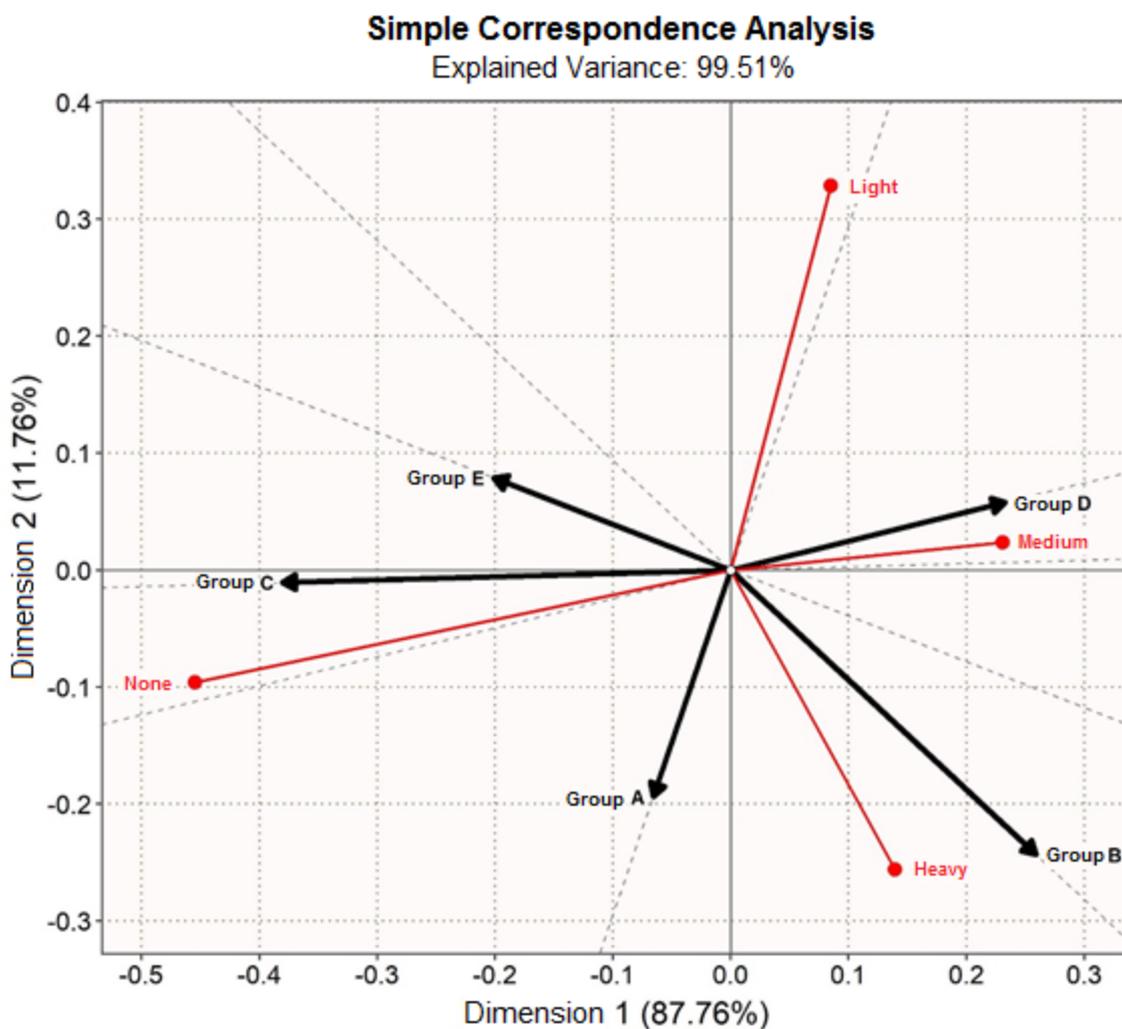
Therefore, the mathematical problem of an CA is the determination of the matrices  $\mathbf{U}$  and  $\mathbf{V}$  from the input matrix  $\mathbf{C}$ . To perform these operations, one must take into account that  $\mathbf{C}\mathbf{C}^T = \mathbf{V}\boldsymbol{\Sigma}^T \boldsymbol{\Sigma}\mathbf{V}^T$  and that  $\mathbf{C}\mathbf{V} = \mathbf{U}\boldsymbol{\Sigma}$ . It is important to highlight that the calculations for these matrices become much more complex as the number of categories involved increases. Thus, numerical methods to perform such complex operations are necessary. The programming environment R has this advantage, since it performs these operations even if the number of categories increases substantially. R's *ca* package (Nenadic, & Greenacre, 2007) is quite complete and allows the computation of CA and more than one type of MCA.

As presented, Correspondence Analysis seeks to study whether a given category relates appreciably to another one(s). The graphical representation is usually shown on a two-dimensional map by straight line segments that start at the origin (some with orientation for reference), so that their length on the map indicates how much information that category adds to this map. In addition, the larger the cosine of the angle between two segments and their lengths, the greater the association between them. Thus long segments oriented so that there is a small angle relative to each other are strongly associated. Long segments oriented at an angle greater than 90 degrees (negative cosine) are anti-associated (anti-correlated, where this anti-correlation is maximum if the angle is close to 180 degrees, i.e. maximum opposition). If the angle is close to 90 degrees, the categories have very little association with each other. If the total variance explained by the visualization is not too high (below 80 percent) and the segments have small map lengths, the interpretation of the associations involving these categories should be carefully done. If the total variance explained by the map is high (above 80 percent), short segments indicate that the category is not associated (or anti-associated) with the others. In other words, this indicates that occurrences of categories represented by short segments do not deviate much from the expected value.

Figure 7 shows the CA resulting from the contingency table data of Figure 6 (a). It indicates the association between the level of tobacco use and worker groups of a fictional company. The quality of the visualization is given by the high variance explained by its dimensions – 99.51 percent – that is, a very high fidelity map which reproduces visually very well the residuals on the table from Figure 6 (c). Tobacco consumption was used as a reference variable, represented by arrows in red. The five groups of workers were positioned on the map and highlighted by black arrows.

It is important to understand that CA measures relative associations. For example, from the visualization of Figure 7 it is possible to affirm that the members of Group C are more characterized by not smoking than the others. That is, among all groups, this is the one that most exceeds the expected value in this regard, as shown in the table in

Figure 6 (c) – the standardized residue is 2.21. Group B participants are characterized more than others by smoking a lot (standardized residue 1.91). Thus, it is expected that Group C will be anti-correlated with the Heavy consumption category, which can be seen by the angle of more than 90 degrees between its segments. The same goes for the anti-correlation between Group B and the category None. It is also worth noting that Group A is practically not strongly associated with any category of consumption, which is revealed by the angles between the None group and the Heavy categories, or by the small residues for this same intersection of attributes. Group D is the furthest from the expected value in the Medium consumption category, and is also the most anti-correlated with the None category. Thus, CA allows inferring associations between groups (entities) and some attributes (consumption category), which allows to understand the behavior of groups and relationships between them according to their consumption habits. Next, we will present a research study on the field of Science Education with this method.



**Figure 7.** Map produced by Simple Correspondence Analysis from the data in the table represented in Figure 6(a)

### 3.1 An investigation on the field supported by CA

The previous example shows how a CA can help highlight associations between variables that are hidden in contingency tables. Recently, Antunes Jr., Ostermann and Cavalcanti (2019) used Correspondence Analysis to investigate the relationships between the formation of academic advisors, the theoretical references used and the regions of the country in which the dissertations of the National Master's Degree Programme in Physics Education (MNPEF) are being published. The database for the analysis was the dissertations and / or educational products presented under the MNPEF. Issues such as student-teacher training, advisor and co-advisor training, nature of the educational product, and articulation of the theoretical framework with the educational product are some of the factors that were taken into account for data compilation. As an example, a more detailed examination will be presented only taking into account the advisor's background and research history, as well as the articulation of the theoretical framework in the educational product, as illustrated in Figure 8.

It is easy to see from the map that the association that results between the advisor attributes with a degree and publication on the field (*ocf\_p*) and well-articulated theoretical reference (*bem\_art*) is greater than the association between advisor without training but with production on the field (*osf\_p*) and *bem\_art*. The visualization is quite reliable as the variance explained in both dimensions is 100 percent. It is also clear that works in which the theoretical framework is disarticulated from the elaboration of the educational product (*desart*) are much more associated with supervisors without training and without publication on the field (*osf*). From this analysis, the author concludes that an adequate education or extensive experience on the field of teaching by the advisors results in dissertations whose educational products are much better articulated with the theoretical references. Since the visualization of Figure 8 is accurate, as it explains 100 percent of the variance, the short segments are unrepresentative compared to the others, meaning that the attribute for advisors *without training and production under development on the field* (*osf\_d*) is not relevantly associated with the levels of articulation of the theoretical framework with the educational product. Even so, the *osf\_d* category tends to be characterized more by well-articulated theoretical references and the *osf* category by totally disjointed theoretical references of the educational product. This conclusion becomes more evident by performing this joint analysis of multiple factors from CA.

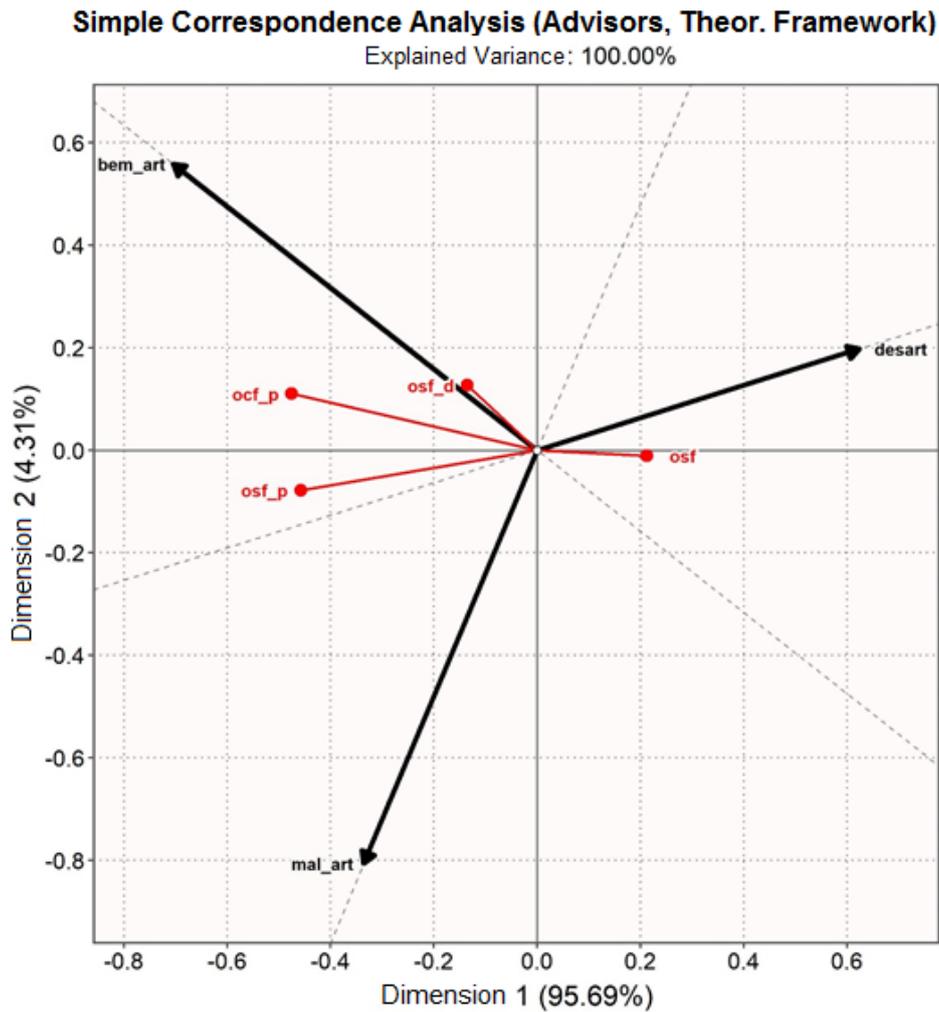


Figure 8. Simple Correspondence Analysis for the guiding variables and theoretical framework

Source: Antunes Jr.; Ostermann; Cavalcanti (2019).

#### 4. Final Considerations

In the present work some interpretative quantitative methods were discussed from an approach that departs from the traditional quantitative approaches. The strategy was to present the methods based on their effective use with the support of data obtained freely from the internet. This paper dealt especially with multivariate data analysis methods: Multidimensional Scaling and Correspondence Analysis. After the presentation of the methods, we introduced recent researches on the field of Science Education as examples of their effective use. All visualizations and statistical analyses were performed with the support of the R programming environment (R Core Team, 2015).

Generally speaking, Multidimensional Scaling (MDS) is a method used for visualizing dissimilarities – which can be understood, for purposes of comprehension, as Euclidean distances – between categorical variables (Johnson, & Wichern, 2004). An

example of MDS was based on data from a questionnaire applied in the early 1990s, which infers the perception of European residents regarding science and technology (Reif; Melich, 1995). Then, in order to illustrate the use of MDS in the field of Science Education, we summarized the results of a research that investigated the similarities among dissertations and theoretical references of a Professional Master's Degree Programme in Physics (Nascimento, 2016; Nascimento; Ostermann; Cavalcanti; 2017).

On the other hand, Correspondence Analysis (CA) is a tool that allows the visualization of the association between a set of categorical variables in a map of reduced dimensions (Hair et al., 2009). In this paper, CA was used in the interpretation of the association between workers of a fictional company and their tobacco use (Greenacre, 1984). Then the exemplification of the use of Correspondence Analysis in a recent research on the field came from the investigation by Antunes Jr., Ostermann and Cavalcanti (2019) on the formation of MNPEF advisors and the use of theoretical references in the development of educational products.

Finally, it is noteworthy that this text aimed to present possible alternatives for the use of quantitative methods, in interpretative mode, especially for research in Science Education, showing that analyses of this nature may move away from the belief that the data express an absolute reality by itself. Nonetheless, it was not our intention to delve into philosophical and epistemological questions imbricated in the appropriation of quantitative methods. The weight of this discussion has been diminished precisely to invest in the discussion on how these tools can be incorporated into the methodological repertoire of researchers on the field. Firstly, we showed how these methods are used in practice so that later we could present recent research on the field.

## References

Adamuti-Trache, M., & Andres, L. (2008). Embarking on and persisting in scientific fields of study: Cultural capital, gender, and curriculum along the science pipeline. *International journal of science education*, 30(12), 1557–1584.

Age, L.-J. (2011). Grounded theory methodology: Positivism, hermeneutics, and pragmatism. *The qualitative report*, 16(6), 1599–1615.

Andres, L., & Adamuti-Trache, M. (2008). University attainment, student loans, and adult life course activities: A fifteen-year portrait of young adults in British Columbia. In R. Finnie, R. E. Mueller, A. Sweetman, & A. Usher (Eds.). *Who goes? Who stays? What matters? Accessing and persisting in post-secondary education in Canada* (pp. 239–275). Retrieved from [http://higheredstrategy.com/mesa/pdf/MESA\\_Andres\\_Adamuti\\_Trache.pdf](http://higheredstrategy.com/mesa/pdf/MESA_Andres_Adamuti_Trache.pdf)

Anton, H., & Busby, R. C. (2006). *Álgebra linear contemporânea*. São Paulo, Brasil: Bookman Editora.

- Antunes Jr., E., Ostermann, F., & Cavalcanti, C. J. H. (2019) A subvalorização da formação continuada de professores: dos orientadores à articulação do referencial teórico no contexto do Mestrado Nacional Profissional em Ensino de Física. *Alexandria: Revista de educação em Ciências e Tecnologias*, 12(2), 267–291.
- Babones, S. (2016). Interpretive quantitative methods for the social sciences. *Sociology*, 50(3), 453–469.
- Borg, I., Groenen, P. J., & Mair, P. (2017). *Applied multidimensional scaling and unfolding*: Berlin: Springer. <https://doi.org/10.1007/978-3-642-31848-1>
- Brannen, J. (2017). *Mixing methods: Qualitative and quantitative research*: Routledge.
- Bryman, A. (1984). The debate about quantitative and qualitative research: a question of method or epistemology? *British journal of Sociology*, 75–92.
- Carvalho, R. C. D., Oliveira, I., & Rezende, F. (2009). Tendências da pesquisa na área de educação em ciências: uma análise preliminar da publicação da ABRAPEC. *VII Encontro Nacional de Pesquisa em Educação em Ciências*, 1–12.
- Cerda J., & Valdivia G. (2007). John Snow, la epidemia de cólera y el nacimiento de la epidemiología moderna. *Revista chilena de infectología*, 24(4), 331–334. <https://dx.doi.org/10.4067/S0716-10182007000400014>
- Cho, J. Y., & Lee, E.-H. (2014). Reducing confusion about grounded theory and qualitative content analysis: Similarities and differences. *The qualitative report*, 19(32), 1–20.
- Cox, T. F., & Cox, M. A. (2000). *Multidimensional scaling*. Nova Iorque, US: Chapman and hall/CRC.
- Cunha, M. V. (2001). John Dewey e o pensamento educacional brasileiro: a centralidade da noção de movimento. *Revista Brasileira de Educação*, 86–99.
- De Leeuw, J., & Mair, P. (2009). Multidimensional Scaling Using Majorization: SMACOF in R. *Journal of Statistical Software*, 31(3), 1–30. <https://doi.org/10.18637/jss.v031.i03>
- De Nooy, W. (2003). Fields and networks: correspondence analysis and social network analysis in the framework of field theory. *Poetics*, 31(5–6), 305–327.
- Dixon, P. (2003). VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science*, 14(6), 927–930.
- Ferreira, M. C. (2003). Consumo cultural e espaços sociais: os vestibulandos das universidades públicas na cidade do Rio de Janeiro, 1990. *Opinião Pública*, 9(1), 170–189.
- Greca, I. M. (2002). Discutindo aspectos metodológicos da pesquisa em ensino de ciências: algumas questões para refletir. *Revista Brasileira de Pesquisa em Educação em Ciências*, 2(1), 73–82. Recuperado de <https://periodicos.ufmg.br/index.php/rbpec/article/view/4152>.

- Greca, I. M., & Moreira, M. A. (2001). O uso da análise multidimensional na pesquisa em ensino de ciências. *Revista Brasileira de Pesquisa em Educação em Ciências*, 1(3). Recuperado de <https://periodicos.ufmg.br/index.php/rbpec/article/view/4162>.
- Greenacre, M. (2017). *Correspondence analysis in practice*. Nova Iorque, US: Chapman and Hall/CRC.
- Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. Londres, Inglaterra: Academic Press.
- Greenacre, M. J., & Blasius, J. (1994). *Correspondence analysis in the social sciences: Recent developments and applications*. Londres, Inglaterra: Academic Press.
- Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). *Análise multivariada de dados*: São Paulo, Brasil: Bookman Editora.
- Hilger, T. R. (2009). *Representações sociais da física quântica* (Dissertação de Mestrado). Universidade Federal do Rio Grande do Sul. Porto Alegre, Brasil. Recuperado de: <https://www.lume.ufrgs.br/handle/10183/17743>.
- Hirschfeld, H. O. (1935). *A connection between correlation and contingency*. Trabalho apresentado em Mathematical Proceedings of the Cambridge Philosophical Society.
- Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis*. 6th. New Jersey, US: Pearson Prentice Hall.
- Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational researcher*, 33(7), 14–26.
- Johnson, W. P., & Myatt, G. J. (2014). *Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining 2nd Edition*: John Wiley & Sons.
- Kincheloe, J. L., & Tobin, K. (2009). The much exaggerated death of positivism. *Cultural Studies of Science Education*, 4 (3), 513–528.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.
- Levshina, N. (2015). *How to do linguistics with R: data exploration and statistical analysis*. Amsterdam, Holanda: John Benjamins Publishing Company.
- Medeiros, A. (2002). Metodologia da pesquisa em educação em ciências. *Revista Brasileira de Pesquisa em educação em ciências*, 2(1). 66–72. Recuperado de: <https://periodicos.ufmg.br/index.php/rbpec/article/download/4151/2716>.
- Nascimento, M. M., Cavalcanti, C., & Ostermann, F. (2017). Análise de Correspondência aplicada à pesquisa em Ensino de Ciências. *Enseñanza de las ciencias*, (Extra), 1319-1324. Recuperado de: <https://ddd.uab.cat/record/184468>.
- Nascimento, M. M. (2016). *Análise de produtos educacionais desenvolvidos no âmbito de um mestrado profissional em ensino de física* (Dissertação de Mestrado). Universidade Federal do Rio Grande do Sul. Porto Alegre, Brasil. Recuperado de <https://www.lume.ufrgs.br/handle/10183/139386>.

- Nascimento, M. M., Ostermann, F., & Cavalcanti, C. (2017). Análises multidimensional e Bakhtiniana do discurso de trabalhos de conclusão desenvolvidos no âmbito de um mestrado profissional em ensino de Física. *Ciência & educação*, 23(1), 181–196. <https://doi.org/10.1590/1516-73132017001001>.
- Nenadic, O., Greenacre, M. (2007). Correspondence Analysis in R, with two-and three-dimensional graphics: The ca package. *Journal of Statistical Software* 20(3), 1–13. <https://doi.org/10.18637/jss.v020.i03>
- Oksanen, J., Blanchet, F. G, Friendly, M., Kindt, R., Legendre, P., McGlenn, D., Minchin, P. R., O’Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H., Szoecs, E., & Helene Wagner (2019). Vegan: Community Ecology Package. R package version 2.5–4. <https://CRAN.R-project.org/package=vegan>
- Quinn, G. P., & Keough, M. J. (2002). *Experimental design and data analysis for biologist*. Cambridge, Reino Unido: Cambridge University Press.
- R Core Team. (2015). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Recuperado de <https://www.R-project.org>
- Rizopoulos, D. (2006). ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses, *Journal of Statistical Software*, 17(5), 1–25. <https://doi.org/10.18637/jss.v017.i05>
- Reif, K., & Melich, A. (1995). *Euro-barometer 38.1: Consumer protection and perceptions of science and technology, november 1992*: Inter-university Consortium for Political and Social Research.
- Rodrigues, E. V., Borges, A. E., & Pietrocola, M. P. (2018). *História da Ciência no Ensino de Física como representação social dos licenciandos. XVII Encontro de Pesquisa em Ensino de Física*. Campos do Jordão - SP.
- Santos, C. A., & Moreira, M. A. (1991). *Escalonamento multidimensional e análise de agrupamentos hierárquicos*. Porto Alegre, Brasil: Editora da Universidade.
- Santos, C. A. d. (1978). *Aplicação da análise multidimensional e da análise de agrupamentos hierárquicos ao mapeamento cognitivo de conceitos físicos* (Dissertação de Mestrado). Universidade Federal do Rio Grande do Sul. Porto Alegre, Brasil. Recuperado de <https://www.lume.ufrgs.br/handle/10183/29922>.
- Schneider, E. M., Fujii, R. A. X., & Corazza, M. J. (2017). Pesquisas quali-quantitativas: contribuições para a pesquisa em ensino de ciências. *Revista Pesquisa Qualitativa*, 5(9), 569–584.
- Shanka, T., Quintal, V., & Taylor, R. (2006). Factors influencing international students’ choice of an education destination—A correspondence analysis. *Journal of Marketing for Higher Education*, 15(2), 31–46.

Souza, A. C., Bastos, R. R., & Vieira, M. d. T. (2010). *Análise de Correspondência Simples e Múltipla para Dados Amostrais Complexos. 19º SINAPE: São Pedro/SP.*

Young, F. W. (2013). *Multidimensional scaling: History, theory, and applications:* Psychology Press.

### **Matheus Monteiro Nascimento**

 <http://orcid.org/0000-0001-8179-5391>  
Universidade Federal do Rio Grande do Sul  
Departamento de Física  
matheus.monteiro@ufrgs.br

### **Estevão Antunes Júnior**

 <http://orcid.org/0000-0002-9989-0335>  
Universidade Federal do Rio Grande do Sul  
Programa de Pós Graduação em Ensino de Física  
eantunesjr@gmail.com

### **Cláudio José de Holanda Cavalcanti**

 <http://orcid.org/0000-0002-2477-3150>  
Universidade Federal do Rio Grande do Sul  
Departamento de Física  
claudio.cavalcanti@ufrgs.br

### **Fernanda Ostermann**

 <http://orcid.org/0000-0002-0594-2174>  
Universidade Federal do Rio Grande do Sul  
Departamento de Física  
fernanda.ostermann@ufrgs.br

**Submitted on 04th july 2019**

**Accepted on 27th november 2019**

**Published on 26th december de 2019**