doi: 10.15503/emet2017.95.113

Challenges, pitfalls and opportunities of researching shadow education through the internet

Vít Šťastný
Faculty of Education, University of South Bohemia,
Jeronýmova 10, České Budějovice, Czech Republic
E-mail address: vstastny@pf.jcu.cz

Abstract

The private supplementary tutoring phenomenon (in the literature often described as the shadow education) has expanded worldwide, and thus captured the interest of researchers and scholars in various fields. The usage of the internet as a source of potential information on the phenomenon is rather scarce as only a few studies have benefited from its potential so far. The aim of the paper is to thoroughly describe the methodological challenges and pitfalls encountered in the author's quantitative content analysis of online advertising of private tutoring lessons. The paper deals with problems and issues in various stages of research proces (e.g., definition of population, sampling strategy, ways of acquiring and coding data and assessing its validity) and suggests ways to cope with these challenges. It addresses the limitations of the research and mentions the opportunities for employing the methodology in international comparative studies or as a secondary method for data triangulation.

Key words: shadow education, private supplementary tutoring, online advertising, methodology

Introduction

In the recent decades, it is possible to observe the worldwide growth and expansion of what the scholars termed a "shadow education" (Stevenson, & Baker, 1992; Bray, 1999; Bray, 2009; Bray, & Kwo, 2014 etc.). The term shadow education encompasses various forms and types of private supplementary tutoring, which can be defined as "tutoring in academic subjects which is provided by the tutors for financial gain and which is additional to the provision by mainstream schooling" (Bray, & Kwok, 2003, p. 612). As this kind of educational provision in forms of extra lessons (additional instruction) exists parallel to the mainstream education system, the metaphor of the shadow is appropriate for several reasons: the private supplementary tutoring exists only because of the existence of mainstream education system, changes in the shape and nature of the mainstream education system imply changes in the shadow education system, much more attention of the public focuses on the mainstream rather than the shadow education system, and the features of the shadow education system are much less distinct than the ones of the mainstream education (Bray, 1999, p. 17).

Private supplementary tutoring helps pupils to learn, to extend their human capital, which in turn contributes to economic development; it may have a valuable social functions such as provision of constructive activities and opportunities for children (Bray, 2009, p. 13). Besides these positive features, private supplementary tutoring can also have negative implications. Due to its paid character it contributes to maintaining and exacerbating the social inequalities as the better-off families can afford higher qualities and quantities of private tutoring (Bray, 2011); it may also have a negative impact and a backwash on mainstream schooling by affecting the dynamics of teaching and learning in mainstream classes (e.g., when all students receive supplementary tutoring, teachers are not obliged to work so hard, or it can generate greater disparities in students' performance), it can cause a lack of tutored pupils' interest in regular classes, mainstream teachers who also work as tutors might prefer tutoring and thus might neglect their duties in school etc. In some settings the mainstream school teachers might not be so supportive of or concerned with slow learners, whose parents find themselves forced to arrange a paid supplementary tutoring for their children to keep up with the minimum level of competence required by the school. In some settings, teachers may abuse their power and teach only a part of a required syllabus during regular lessons and require students to attend their after-school tutoring classes (Bray, 1999; Bray, 2009; Silova, Būdiene, & Bray, 2006).

Methodological issues related to shadow education research

As the phenomenon got more visible in many countries, the amount of research in the field has increased. Vít Šťastný (2016a) analysed 50 research studies which focused exclusively on private supplementary tutoring in the European context. The study found that 38 papers in the selected sample were published after 2005, almost half (24 studies) in 2012 or later. However, despite the fact that the shadow education is better documented thanks to the researchers in universities and comparable bodies, governments, international agencies, journalists, and others (Bray, 2010, p. 3), the research on private supplementary tutoring has lagged behind the expansion and diversification of the phenomenon. Although the methodology of the research of out-of-school education, to some extent, ressembles the research in other domains of education, respectively social science, the researchers in the field have encountered many issues and methodological problems linked to the specific object of research, partly because it is less structured (Bray, Kwo, & Jokić, 2015a, p. 3). As the research of shadow education intensified in the last decades, the methodological questions arised, but it took some time for them to be addressed in a more systematic way in the scholarly literature. Arguably the first methodological debates on shadow education research became most visible with the publication of professor Mark Bray's (2010) paper[9]. The methodological discussions were fur-

[9] M. Bray (2010) identified several issues related to the research of shadow education, namely the definition of the focus of investigation (how to define private tutoring, what forms to include etc.), securing the data (which might be problematic because of inability or unwillingness to provide data by various agents related to private tutoring; M. Bray pointed out also problems related to instruments for data collection) or interpreting the data.

ther intensified recently with the methodological criticism of big international surveys PISA and TIMSS for shadow education research (Bray, & Kobakhidze, 2014) and culminated with the publication of a methodologically oriented monograph (Bray, Kwo, & Jokić, 2015b) analyzing and thoroughly reviewing the experience of a dozen of shadow education researchers with their research form various contexts, and bringing up several lessons to be learned and shared with researchers in the field in the future. Their contributions were mostly focused on a field work using questionnaires, interviews or other 'conventional' methods. However, no attention was paid to the research of shadow education by the method of content analysis (Šťastný, 2017a).

The aim of this paper is to describe the methodological challenges and pitfalls encountered in the author's quantitative content analysis of online advertising of private tutoring lessons supply (Šťastný, 2017b)[10] and to suggest possible ways to cope with these challenges. By this paper, the author would like to contribute to the methodological discussion on shadow education research, and to critically review his own work.

This paper will first review the previous studies of shadow education, which have used the internet as a data source. In the similar manner as in the methodological book of M. Bray, Ora Kwo and Boris Jokić (2015b), the paper then turns to author's own methodological experience made during the research of private tutoring lessons internet supply focusing on the employed research methods and overall strategy, sampling frame, legal and ethical issues of the research, data acquisition and processing, coding the open data and questioning the data validity. The author then discusses the possibilities of using the content analysis[11] in further studies and sums up the lessons learned from the research.

Previous use of internet as a source of data in the study of shadow education

Whilst in other areas of inquiry the analyses of web content are quite common, the use of internet as a source of data is quite rare in the scholarly literature on the shadow education phenomenon (Šťastný, 2016a). This might be surprising, because the internet, respectively websites, social media posts and other virtual documents are the potential source of data in their own right (Bryman, 2015, p. 299) and represent a cheap and affordable way to get data on the studied phenomenon. In the following paragraphs, the author comments on the papers which

[10] This paper further develops the methodological issues mentionned only briefly in the original study (Šťastný, 2017b), and some remarks were previously published in author's dissertation (Šťastný, 2016b).

[11] Content analysis is the technique which enables researchers to analyze the human communications, thus enabling to study their behavior indirectly (Fraenkel, & Wallen, 2009, p. 472) and has long and traditionally been used in educational research. Since the 1950s, when the content analysis as a research method was developed, the types of analyzed texts and the media in which this text is stored have evolved dramatically, so the websites too became an object of inquiry in scholarly studies (Gavora, 2015, p. 345).

have previously used internet websites as a source of data on the private supplementary tutoring phenomenon and its further analysis.

British research team led by Emily Tanner (Tanner, et al., 2009) has developed a methodology for mapping the supply of private tuition providers in the United Kingdom and further investigation of the costs, arrangements and other characteristics of private tutoring in the country. The researchers constructed two databases. The first one contained a list and characteristics of private tuition agencies operating in England that maintained a presence on the internet (that is they were indexed and returned by the search engine Google) and consisted of 506 entries, and the second one contained individual tutors working within three local areas. They stored the data in Microsoft Access database (p. 8). In their research report, they provided a detailed list of the database fields (like contact details, tutoring agency size, specialism, target age groups, ways of tutor recruitment, fees and costs etc.) and categories (possible values) used with each database fields. They also thoroughly described the way of internet search, including employed search engines (the authors did not rely only on Google, but used also Metacrawler and Zapmeta as secondary search engines which combine the results of multiple search engines) and keywords, business directories, specialist websites or local area databases (Appendix B of the report). The authors had to deal with difficult issue, what kind of provider or website to include in the database, as this was not always easy to determine (e.g., different levels and standards of information was provided on the pages, thus making it difficult to ensure consistency across the dataset, most sites did not provide any timestamp by which to date the information they contained, or they encountered inconsistent or contradictory information within some sites (pp. 9–10). In the data analysis, the authors remained rather on a descriptive level (showing for example, percentages of various private tuition provider types) and did not use any inferential statistics.

Another paper which brought relevant methodological contribution to the shadow education analysis using the internet was authored by Armand Faganel and Anita Trnavcević (2013). These Slovenian authors sought to understand the current discourse about private tutoring, that is how private tutoring is negotiated and constructed. Thus, they designed an exploratory study of discourses and representations of private tutoring in online chat rooms. The choice of chat rooms as a data source was justified by their popularity among Slovenian children and youth, illustrated by relevant research findings. Authors performed an internet search (through Google) to locate relevant websites, out of 30 websites found, and they chose one with most concentrated, organised and geographically spread information about tutoring; authors also included the data on average number of visitors of the site. The selected website contained 536 advertisements to which 81 postings, opinions and commentaries were submitted (p. 170). Because the texts on chat rooms were not lengthy or rich in content, the authors chose a summative approach to the content analysis, which, according to Hsiu-Fang Hsieh and Sarah Shannon (2005), involves counting and comparisons, usually of keywords, followed by the interpretation of the underlying context. This approach then reflected also in the presentation of findings, which were partly of quantitative character (e.g., authors state, how many comments were positive, negative or neutral), partly of qualitative character (e.g., authors mention particular stories shared by tutors and tutees).

Olga Kozar's (2013, 2015) studies also used the internet websites to investigate the private supplementary tutoring phenomenon. The first paper (Kozar, 2013) aimed to find out, which are the most popular subjects for private tutoring, and to obtain information on the background of sought-after private tutors in Moscow. She used tutor-listing websites as a source of data for her analysis. To identify relevant websites, O. Kozar used the Russian popular search engine Yandex to identify tutor-listing websites, and consequently chose the two largest ones for the analysis (the size was determined according to the respective number of tutors that each of the websites listed). To analyze the data present on these websites, the content analysis used a combination of quantitative approach (e.g., to analyze total number of registered tutors, proportions of different subjects for tutoring) and qualitative approach (the detailed analysis of 32 top-ranking tutor profiles and their age, educational background or profession).

In the second paper, O. Kozar (2015) paper chose a slightly different approach. First, she had chosen 17 websites of private online language schools. To analyse their content, she framed her investigation in a critical discourse analysis (CDA) using the method of thematic analysis of the selected websites' content in order to find out the rhetorical strategies and discursive practices, which might reflect a wider ideology underlying the website content. Thus, rather than focusing on the quantitative aspects of the websites, O. Kozar undertook an analysis which required more elaborate interpretation and included more subjective views. She screened all websites of the genre (belonging to "online schools category") in the top 50 search results (using the search engine Yandex) and finally chose 17 relevant websites appropriate for analysis. As an output of her linguistic analysis of the website content, she was able to identify common themes (using inductive analysis), types of sentences used in the communiqués (declarative, interrogative, imperative and exclamatory), rhetorical devices (use of pronouns, nominalisation, colloquialisms, hedging, ellipses, jokes, etc.) as well as appraisal elements (appreciation, judgement, mono/heterogloss and graduation).

V. Šťastný's (2017b) research built upon the previously mentioned studies, the study aimed to analyze the socio-demographic background of individual private tutors advertising online and their distribution within a country and to assess the macro and micro factors underlying the advertised price (fees) for a tutoring lesson. To accomplish the task, the author chose a quantitative approach, and located the private tutor advertisements on various types of webpages (individual tutor website, notice boards, mediated notice boards etc.). Mediated notice boards were chosen as a source of data on private tutors. From eight mediated noticeboards identified through websearch using Google and Seznam search engines, the author chose one. The single tutors' advertisements appearing on the webpage were then submitted to content analysis and data on age, qualifications, gender etc. were processed (for results, see: Šťastný, 2017b; Šťastný, 2016b).

The above-mentioned studies varied in the degree of details provided about the methodology (which could relate to the form of the paper, as research report has generally less lenght limitations than the standard journal papers or a book chapter). The studies recognize that numerous tutors do not have an online presence, and thus the findings relate only to a part of the tutoring market and also recognized the limited validity of the data obtained on the internet (Tanner et al., 2009, p. 10; Faganel, & Trnavcević, 2013, p. 175; Šťastný, 2017b, p. 7; etc.). To find the relevant material (websites with data on the phenomenon), search engine Google was mainly used (Tanner, et al., 2009, p. 9; Faganel, & Trnavcević, 2013, p. 169). The specific situation is in Russia, where the most popular search engine is Yandex (Kozar, 2013) and in the Czech Republic with search engine Seznam.cz (Šťastný, 2017b, p. 6). The data on private tutoring found on internet was analyzed using a content analysis, which, however, differed in approach. Whilst the mapping of online private tutoring in UK (Tanner et al., 2009) was mainly quantitative, the studies of A. Faganel and A. Trnavcević (2013) and O. Kozar (2013 and 2015) used a mixed approach (that is, qualitative and quantitative) towards the dana analysis. None of the above-mentioned studies mentioned how the process of obtaining data from websites was actually managed and how long did it last. The overview of the studies is provided in the table 1 in Appendix.

Research process of private tutoring lessons online supply in detail

Content analysis is a traditional method of research focused on analysis and interpretation of texts (Krippendorf, 2004; Gavora, 2015). The digital age has fundamentally changed the ways in which content analysis can be conducted (Neuendorf, 2002, p. 79) as well as the types of texts which can be analyzed expanded significantly and includes e.g., web pages, email discussions and email correspondence etc. As mentioned by Susan Herring (2010), there are many issues related to the analysis of these new electronic forms of communications, and in some cases, the 'non-traditional' approach has to be adopted[12]. Bearing this in mind, the author of the present paper had to deal with several challenges when researching the internet supply of private tutoring lessons (for the original study see Šťastný, 2017b). These are described in the following paragraphs within the steps of the analysis which were adopted from Louis Cohen, Lawrence Manion and Keith Morrisson (2011). These authors recommend to conduct the content analysis in eleven steps[13] (which,

[12] S. Herring (2010) distinguished a traditional approach when using this method and non-traditional approach. Sally McMillan (2000) enumerated the traditional approach in five basic steps in the process of content analysis while considering their application to the web: (1) the researchers formulate research questions and/or hypotheses, (2) they select the sample of content to be analyzed, (3) categories (units of analysis) are defined, (4) the selected content is coded, and (5) collected data are analyzed (pp. 81–82). However, S. Herring (2010) noted a number of issues related to the traditional approach even in analyses of old media (e.g., using non-random samples, coding categories emerging from the data, standard statistical tests applied to non-random samples) and suggested that strict adeherence to it might under certain circumstances contraproductive in the analysis of new media.

[13] 1) Define the research questions to be addressed by the content analysis; 2) Define the population from which units of text are to be sampled; 3) Define the sample to be included; 4) Define the

for the sake of simplicity, were merged into eight), according to which this chapter is structured.

Define the research questions to be addressed by the content analysis

In the early stages of empirical inquiry, research questions are usually formulated, and these questions imply an appropriate methodology for their answering (Fraenkel, & Wallen, 2009, p. 27). In the traditional approach, the research questions derive from a theoretical framework and researchers who apply content analysis to the Web should find a context for them either in existing or emerging theory (McMillan, 2000, p. 81). The analysis was based on hitherto empirical evidence on private tutoring supply, which is, however, quite limited compared to other domains of shadow education research (Šťastný, 2016a). In this way, there are many gaps in the knowedge on private tutoing providers, their motivations, characteristics, ways of operation, marketing strategies etc., which opened floor to many possible research questions. When devising the research questions, the pragmatic view was considerd and questions that are "…empirically answerable from the available data" (Herring, 2004, p. 7) were chosen. In the initial internet tutor advertising webpages pre-screening, the type and character of the data available online was monitored, and bearing the availability of data in mind, this step helped to formulate and precise the research questions[14]. To some extent, the questions underwent an evolution, as they were confronted with the available data and before any data was processed or downloaded; it was considered ahead by the author, whether the data could satisfactorily answer the questions and vice-versa. Research questions[15] were, therefore, reformulated several times to fulfill all characteristics of good research questions as defined by Klaus Krippendorf (2004, pp. 32–33)[16] and, at the same time, to reflect the data available for analysis.

context of the generation of the document; 5) Define the units of analysis; 6) Decide the codes to be used in the analysis; 7) Construct the categories for analysis; 8) Conduct the coding and categorizing of the data; 9) Conduct the data analysis; 10) Summarize; 11) Make speculative inferences (Cohen, Manion, & Morrisson, 2011, p. 476–483).

[14] Of course, these questions also had to relate to the overall project objectives, which were defined as "diagnosis" of the shadow education phenomenon (Šťastný, 2016b) and in the specific part of the research design, the aims of the sub-study were to analyse the socio-demographic background (age, gender, qualifications, professional profile) of individual private tutors advertising online and their distribution within a country; and to assess the macro and micro factors underlying the advertised price (fees) for a tutoring lesson.

[15] 1) How does the supply of private tutoring lessons differ in terms of number of providers? 2) How does the supply of private tutoring lessons differ in terms of qualifications, gender and age? 3) How does the supply of private tutoring lessons differ in Prague and Moravian-Silesian Region in terms of academic subjects for tutoring? 4) How does the average price differ in Czech regions and which factors underlie it?

[16] 1) They are presumably answerable by examination of body of texts, 2) They delineate possible answers, 3) They are concerned with at the moment inaccessible phenomena, 4) They allow to be (in)validated by undertaking the inquiry by other means or methods (Krippendorf, 2004, 32–33).

Define the population from which units of text are to be sampled

The basic limitation incorporated in content analysis is that the researcher observes the phenomenon indirectly by drawing inferences from the bodies of text (Krippendorf, 2004, pp. 24–25). In case of shadow education, the object of the study (private tutoring lessons supply) is reflected in the texts who are presumably posted by the providers. However, as already noted by O. Kozar (2013, p. 79), "one cannot state with confidence that the number of online profiles corresponds to the actual number of real-life individuals providing fee-based instruction, the findings from the tutor-listing websites are nevertheless significant as they point, at the very least, to the popularity of private tutoring services."

The consideration of definition of population had to start from this point. The first idea was to define the population of text as "all private tutors' advertisements on Czech webpages," however this would be problematic, because not only that no sampling frame was available for such defined population, but also the validity of data and findings could be threatened by a lot of missing information (e.g., in some adverts, tutors did not state very clearly all the information needed to answer research questions), information that could be duplicate (e.g., some tutors post their advert on many sites to increase their chance of being spotted by their customers) and the population of texts would presumably be very large as the text of individual tutor advertisements could be located practically anywhere on the internet. In cases, when the population cannot be easily identified or the population is extremely large, John Creswell (2012, p. 145, see also Weare, & Lin, 2000, p. 280) recommends to use multistage (cluster) sampling in which getting a complete list of members of population within subgroups (or clusters) in the population might be possible.

In case of shadow education online supply, these subgroups could be identified based on the types of websites in which the individual private tutors advertise their services. After the initial websearch, it turned out that these agents advertise on three types of webpages (typology adapted from Tanner et al., 2009): 1) single webpages of individual tutors; 2) notice boards (website notice board for individual tutor advertisements with contact details allowing clients to negotiate directly with the tutors); 3) mediated notice boards (list of registered tutors, from which clients select, but no individual tutor contact details are provided, so contact with the tutor is mediated by the provider).

Webpages of "individual tutors" advertising their lessons exist, however, they did not seem to be a standard way of advertising for private lesson (compared to number of tutors registered in other types of webpages, e.g., notice boards, their number was, at least according to internet search negligeable) and for most of the tutors, it seems easier to post their advertisements on different kinds of webpages with already prepared structure (notice board or mediated notice board type). The reasons may be different, e.g. they do not want to pay for separate webhosting, do not know how to or want to create personal page, or do not have sufficient content to put on the page (e.g., tutors who begin their tutoring career might not have enough references of past experience etc.). Thus, analysing single tutor webpages was not considered an appropriate for the research purpose.

Notice boards seemed a better option, however, the content of the advert usually does not have a fixed structure and not all adverts contained the information neccessary to answer the research questions, which presented a thread to data validity. There was as well a problem with data duplication, as many tutors were active in re-posting their advertisements after several days in order to get higher in the ranked order of displayed adverts.

The most appropriate option then seemed the mediated notice boards, which usually had a standardized presentation (profile) of single tutors with (obligatory) information about the qualification, tutored subjects, places of operation, advertised price etc.

The appropriate websites could be located by using search engines, using collector sites (in which individuals and organizations is to collect and post lists of links to related sites) or by identifying sites according to their popularity based on some special engines measuring the number of visits (Weare, & Lin, 2000, pp. 276–279). Second option was decided for, that is by using search engines to locate the appropriate websites. Christopher Weare and Wan-Ying Lin (2000, p. 278) also have a legitimate objection that not all webpages are properly indexed and catalogued and frames constructed in such a way could be biased towards more heavily trafficked parts of the Web. As the provision of shadow education is a private business and tutors are seeking profits, it was assumed that they will tend to make their online-presence as visible as possible, thus ensuring to advertise on sites which are indexed by the search engines.

The choice of search engines was another step taken by the author. The issue with some search engines is that they might purposefully omit some pages or sites and do not index them. Such problem was detected by for example, Dina Borzekowski, Summer Schenk, Jenny Wilson and Rebecka Peebles (2010, p. 1532), who researched pro-eating disorder websites. In the Czech context, the private supplementary tutoring is definitely not such a delicate topic as pro-eating disorder topic might elsewhere be, and thus the purposeful deindexing of such sites referring to private tutoring from search results is unlikely. However, in some contexts or countries (e.g., with rather restrictive educational policies on private tutoring like Korea in the past), the risk of such practice from the search engines might exist (e.g., in accordance with the government policies, which in some cases try to reduce the private tutoring, see e.g. Bray, & Kwo, 2014). As mentionned in the literature review, previous studies usually employed Google search engine (Tanner et al., 2009; Faganel, & Trnavcević, 2013), or a nationally traditional engine such as Yandex in Russia (Kozar, 2013; Kozar, 2015). Therefore both Google and Seznam, which, according to Webcertain report (2013), share 71 and 26 percent of the market, were decided to be used. Seznam is the traditionnal search engine which has been extensively used since the 1990s with the introduction of internet in the Czech Republic (e.g., in 1998, Seznam became the most visited Czech website). Alan Bryman (2015, p. 299) notes that a researcher has to be very patient to try as many keywords and their combinations as possible. In both search engines, dozens of keywords have been tried which define or characterise the phenomenon in scope: private tutor, private lessons, supplementary tutoring, help with academic subjects, remedial teaching to schoolchildren, extra lessons, private instruction... For each keyword, every result until the 7th page of search results was checked, then the pages 7 to 25, only the links which seemed relevant were checked. In the first round, various internet pages were screened in search for information on commercial offers of private tutoring lessons. The results in both engines did not differ too much.

By internet search (using both search engines Google and Seznam.cz), eight mediated notice boards with individual private tutor profiles were identified. The population was then defined as "texts of individual tutors' advertisements located on these eight websites" (Šťastný, 2017b).

Define the sample to be included

Another issue of consideration was the sampling of texts of private tutor profiles present on these eight mediated notice boards. The first idea was to include all tutor profiles present at all eight portals. This option turned out problematic, because the initial screening showed that some private tutors may advertise on two or more mediated notice boards (under different nicknames), and inclusion of all texts in the analysis would create a duplicate records in the database, and would bias the results. Therefore, analysing only one of these eight web-portals seemed more appropriate (usually, mediated notice boards do not allow the tutors to register more than once), and lead to another decision about the choice of the appropriate mediated notice board. Unfortunately, no previous data were available on the popularity of these webportals among users (pupils or parents), so the choice was to be made according to some externally-set criteria. The main chosen criteria (the number of registered private tutors in relevant academic subjects for tutoring, the amount of information the individual tutors provided, and countrywide coverage in all regions) were believed to reflect the popularity and relevance of the web-portal and at the same time this information was also easily accessible on all eight portals). Another way to identify the appropriate website for analysis could rely on some previous studies which identified the most popular websites (e.g., a quantitative suvey of representative sample of pupils or parents could identify, which ways of finding tutor is most widespread and which websites are used the most).

After assessing pre-set criteria on all eight portals, website www.naucim.cz fulfilled the aforementioned criteria to the highest degree (Šťastný, 2017b). As the study was a part of a wider research design (see: Šťastný, 2016b), there was also a possibility to verify the choice of the portal ex-post by interviewing several tutors advertising there. Most of them claimed they have on average two or three new requests for tutoring lessons per week, and compared to other ways of advertising, they considered advertising on the selected website rather effective.

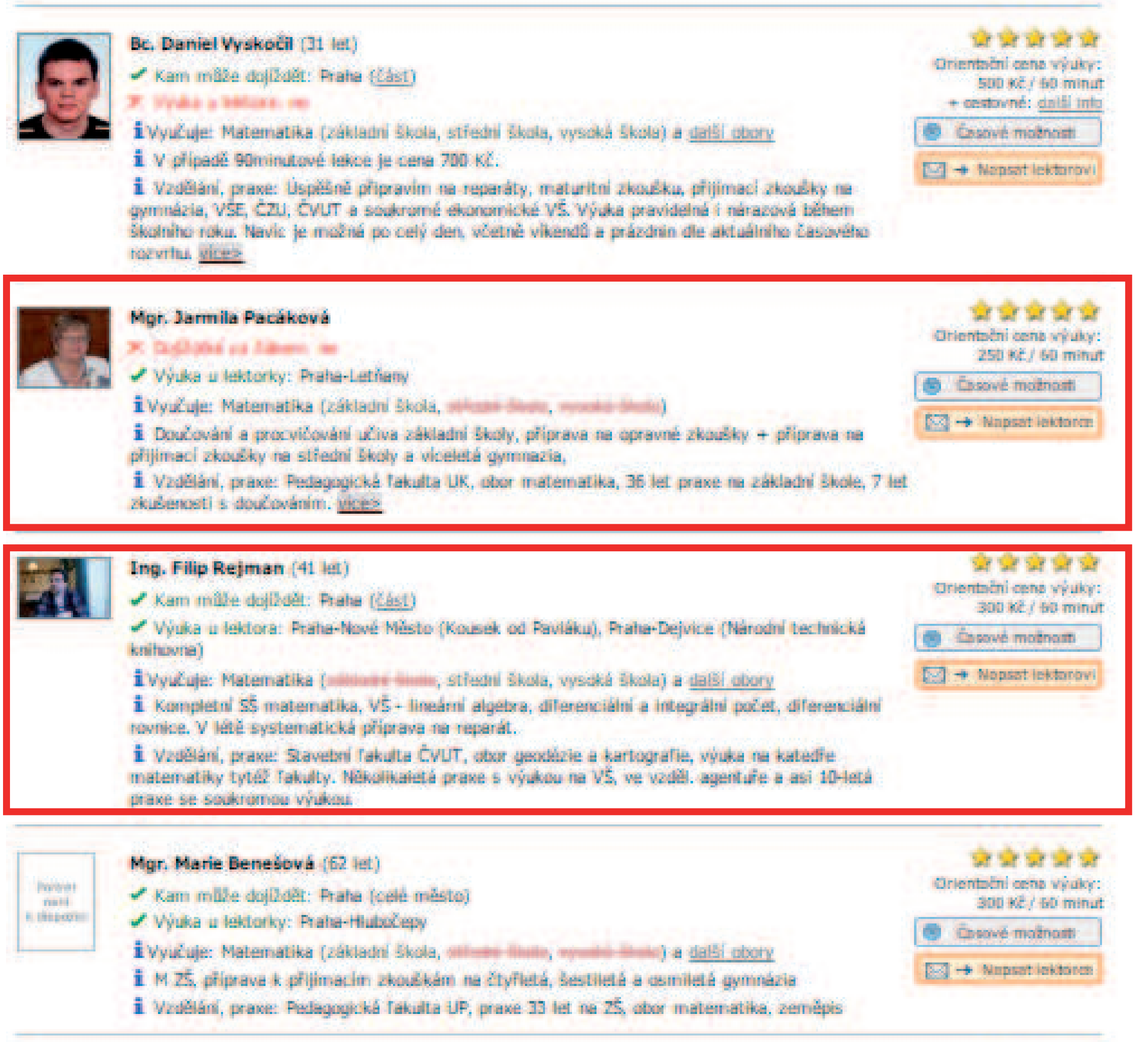
Define the context of the generation of the document

The internet is ever changing and problems also occur when the websites disappear and the researcher might find in the middle of the analysis that the data are gone (Bryman, 2015, p. 299). There are various ways and possibilities of automation that the researcher can use. For example, Christian Bauer and Arno Scharl (2000) employed the software *Webanalyser*, or Inhwa Kim and Jasma Kuljis (2010) used *LocalWebsite Archive* to download and archive all versions of the websites they analysed. Such software could be efficient, but also costly. Another disadvantage in using such kinds of software to download website content for quantitative analyses is the fact that the data has to be transferred into a specific format which enables further quantitative statistic processing (e.g., Excel spreadsheets, SPSS format etc.) and using software which downloads website content as text would anyway require a further data extraction and editing. Thus, the standard and widespread software of Microsoft's Office package Excel 2013 was employed. The versions of Microsoft Excel 2007 and newer enable the user to run so called web query, which downloads the content of the website (or only part of it) to a spreadsheet, and when neccessary, Excel refreshes the query. The website URL structure was very convenient for this purpose, as the standard format of the address was www.mainsite.cz/tutoring-subject/-/region (e.g., www.naucim.cz/doucovani-biologie/-/praha/). The VBA (Visual Basic for Applications) programming language is inherently included in the above-mentioned software and is used for automation of standard tasks performed in the environment of Microsoft Office applications. One of its functionalities is to loop through a list of items: two separate lists have been created (with subjects and regions) and then a makro was programmed (Excel programme written in VBA) to combine the two lists, open the pages and download them into the spreadsheet. The makro programming took two hours including the testing, and compared to manual downloading or editing the data, it spared a lot of time.

Define the units of analysis

In the classical handbook on content analysis, K. Krippendorf (2004, p. 349) distinguishes three types of units in the analysis: 1) sampling units, which are mutually exclusive units of text that are selectively included in an analysis; 2) recording units, which are either equal to or contained in the sampling units, but separately described, coded, or recorded in the terms of a data language; and 3) context units, which set limits on the amount of text to be consulted in determining what a recording unit means. As noted by S. McMillan (2000, p. 93), earlier content analyses of traditional media have developed traditional context units and measures, such as word count for newspapers, column-inch, seconds of broadcast etc. Due to the multiple media combination present on the web, the content analyses do not follow an established standard. In web based inquiries, the majority of studies define their sampling unit as a single website, but usually do not then distinguish web pages belonging to the website as recording units, which C. Weare and

W. Lin (2000, p. 281-282) recommend. The advertisements of private lessons are usually structured in pieces of text, which have apparent borders and are dintiguishable from each other on the particular webpage, thus fulfilling the requirements of units of analysis as defined above by K. Krippendorf (2004, p. 349). Such was the case of most of the tutoring websites, which were considered for analysis (see above) and also of the chosen mediated notice board webportal (see Figure 1). The text of the private tutor profile was chosen as the unit of analysis.



Bc. Daniel Vyskočil (31 let)
✓ Kam může dojíždět: Praha (část)
✗ Výuka u lektora: ne
ℹ Vyučuje: Matematika (základní škola, střední škola, vysoká škola) a další obory
ℹ V případě 90minutové lekce je cena 700 Kč.
ℹ Vzdělání, praxe: Úspěšně připravím na reparáty, maturitní zkoušku, přijímací zkoušky na gymnázia, VŠE, ČZU, ČVUT a soukromé ekonomické VŠ. Výuka pravidelná i nárazová během školního roku. Navíc je možné po celý den, včetně víkendů a prázdnin dle aktuálního časového rozvrhu. více>
Orientační cena výuky: 500 Kč / 60 minut
+ cestovné: další info

Mgr. Jarmila Pacáková
✗ Dojíždí za žákem: ne
✓ Výuka u lektorky: Praha-Letňany
ℹ Vyučuje: Matematika (základní škola, ~~střední škola~~, ~~vysoká škola~~)
ℹ Doučování a procvičování učiva základní školy, příprava na opravné zkoušky + příprava na přijímací zkoušky na střední školy a víceletá gymnazia,
ℹ Vzdělání, praxe: Pedagogická fakulta UK, obor matematika, 36 let praxe na základní škole, 7 let zkušenosti s doučováním. více>
Orientační cena výuky: 250 Kč / 60 minut

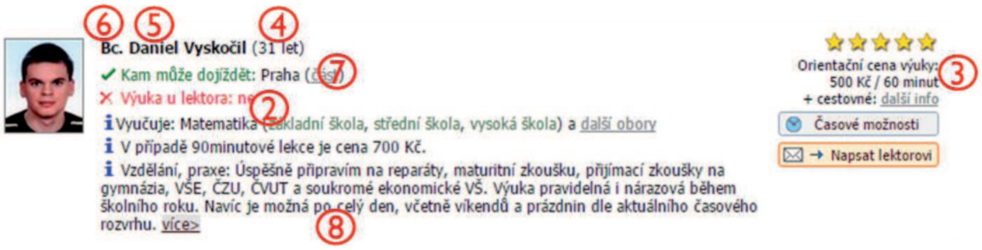
Ing. Filip Rejman (41 let)
✓ Kam může dojíždět: Praha (část)
✓ Výuka u lektora: Praha-Nové Město (Kousek od Paviáku), Praha-Dejvice (Národní technická knihovna)
ℹ Vyučuje: Matematika (~~základní škola~~, střední škola, vysoká škola) a další obory
ℹ Kompletní SŠ matematika, VŠ - lineární algebra, diferenciální a integrální počet, diferenciální rovnice. V létě systematická příprava na reparát.
ℹ Vzdělání, praxe: Stavební fakulta ČVUT, obor geodézie a kartografie, výuka na katedře matematiky téže fakulty. Několikaletá praxe s výukou na VŠ, ve vzděl. agentuře a asi 10-letá praxe se soukromou výukou.
Orientační cena výuky: 300 Kč / 60 minut

Mgr. Marie Benešová (62 let)
✓ Kam může dojíždět: Praha (celé město)
✓ Výuka u lektorky: Praha-Hlubočepy
ℹ Vyučuje: Matematika (základní škola, ~~střední škola~~, ~~vysoká škola~~) a další obory
ℹ M ZŠ, příprava k přijímacím zkouškám na čtyřletá, šestiletá a osmiletá gymnázia
ℹ Vzdělání, praxe: Pedagogická fakulta UP, praxe 33 let na ZŠ, obor matematika, zeměpis
Orientační cena výuky: 300 Kč / 60 minut

Fig. 1. Example of private tutors' advertisements on the page.
Source: website www.naucim.cz.

Decide the codes to be used in the analysis, construct the categories for the analysis

The advantage of selecting texts of tutor profiles as units of analysis in this case was that the texts were structured and the required information was stored on a specific position in the advertisement (see Figure 2).



Bc. Daniel Vyskočil (31 let)

✓ Kam může dojíždět: Praha (části)

× Výuka u lektora: ne

Vyučuje: Matematika (základní škola, střední škola, vysoká škola) a další obory

V případě 90minutové lekce je cena 700 Kč.

Vzdělání, praxe: Úspěšně připravím na reparáty, maturitní zkoušku, přijímací zkoušky na gymnázia, VŠE, ČZU, ČVUT a soukromé ekonomické VŠ. Výuka pravidelná i nárazová během školního roku. Navíc je možná po celý den, včetně víkendů a prázdnin dle aktuálního časového rozvrhu. více>

Orientační cena výuky: 500 Kč / 60 minut + cestovné: další info

Fig. 2. Example of private tutor profile advertisement.
Source: website www.naucim.cz.

Each tutor profile had a standardized structure, which made the definition of codes easier. There were in total 8 codes: (1) region(s) of operation, (2) academic subject(s) offered for private tutoring lessons, (3) offered price per lesson, (4) age, (5) gender, (6) highest attained education level (coded according to ISCED 2011 framework),[17] (7) willingness to commute to the tutee's place of abode, (8) closer description of the tutor's offer (including e.g., occupation status).

Most pieces of the above mentionned information were explicitely mentionned by the tutors in the text of their advertisement (2-7), and thus their coding was quite reliable and straightforward (the region of operation was defined in the page URL). The author tried to register as a tutor on the portal in order to get more insight about the process and kind of information which has to be provided by the tutors. For finding out the possible values of the above mentionned observed variables, it turned out beneficial to mock register as a tutor on the selected portal to see the process, and registration requirements for the tutors. By undertaking the registration process the researcher found out that some fields had predefined values (e.g., the gender – male; female or commuting to student's place of abode – yes; no; yes, but with costs reimbursement). The only information which required a coding (in a strict sense) was the tutor's occupation (8). Each record had to be evaluated separately, finally, seven categories of occupation emerged from the qualitative evaluation of the tutor advertisement texts (upper-secondary student; university student; university teacher or PhD candidate; primary or secondary school teacher; other pedagogical worker; other profession; not stated or unable to identify).

Conduct the coding and categorising of the data

To assure the reliability of the coding, the commonly recommended step in the process is to engage two or more researchers, who code the content independentlly (e.g., Krippendorf, 2004). Thus, the fact the author of the analysis was the sole author can be considered a drawback of the study. On the other hand, most of the material for coding was manifest (except the identification of the private tutor's occupation), and thus quite unambiguous for reliable coding.

[17] Some tutors may have not stated their academic title despite the fact they possess one. However, presumably for the marketing purposes, the tutors arguably tend to publish every information that could attract their clients, including their academic achievements; at the same time, they are asked to fill in their titles upon registration (though the field was not mandatory).

Data analysis

In total, the database included data on 2058 tutors, but contained 6911 records (rows), because the tutors could provide private lessons in more subjects and also in more regions, some of them were included in the database in more records (e.g., the tutor identified according to a name and surname occupied three records in the database if he or she offered service mathematics in Prague and Central Bohemian region and English language in Prague). Thus, in the created database, it was important to create a "helper columns," which would identify the duplicate records according to the tutor's residence and subject of tutoring, because various analyses required different filters to be applied. For example, when analysing uniquely tutor's age in general, it would be redundant to include all records, as some tutors would be included more times. Or, when analyzing the regional differences, it was neccessary to include the offers of all tutors operating in the region (the tutors could have had a different price policy in the two regions), so only records unique according to tutor's name and region of operation could have been included.

Discussion

In the previous text, several challenges encountered during the content analysis of webportal advertising private tutoring lessons were shared. Using internet as a research tool and source of data in the field of shadow education is advantageous either as a fully-fledged research approach, or could be used as a supplement for data triangulation (Šťastný, 2017b, p. 15). Jack Fraenkel and Norman Wallen (2009, p. 474) observe that content analysis may be helpful in validating the findings of a study or studies using other research methodologies. The same applies also in the shadow education domain, where the validity of findings from content analysis of online texts concerning shadow education can be further increased by combination with other methods of research. For instance, the selection of analyzed web portals could be preceded by a representative questionnaire survey of target group (pupils), which could find out the most common ways of how the pupils (or parents) find their tutors (to see the position of internet among other options) and in case of internet search of tutors, the questionnaire survey may find out the most popular websites for tutor search. Consequently, such information could serve as a basis for alternative sampling procedures/strategies like purposive sampling based on previous empirical evidence.

Another use of such studies is in comparing the supply of private tutoring in multiple countries. In this case, the shadow education researcher will face the problem of standardization and data availability, which may vary accross various web-portals based in various countries. More appropriate approach could be the analysis of internationally operating private tutoring webportals. An example of such Czech webportal could be e.g., www.domelie.cz, which has equally a Slovakian mutation www.domelia.sk. Both sites operate under the same provider and the private tutor profiles advertising on these sites have the same structure, thus making them relatively easy to compare. In case of using more sites and comparing the private lessons supply internationally, the legal requirements may differ in the both countries, thus before using and downloading the data from the internationally operating websites, the researcher should be aware of these requirements, and check which jurisdiction the analyzed webportal belongs to and if the operations with the data he or she intends to do is legal in every country of interest.[18]

M. Bray (2009, p. 17; p. 6) notes that obtaining data on shadow education from tutors is difficult, because they usually try to avoid attention as many of them provide tutoring as an informal activity generating untaxed income, and in some cases, tutors may also avoid sharing information on qualifications, premises, curricula, teaching methods etc. However, when studying the internet supply of private tutoring lessons, it was remarked that the tutors' interest was (usually) to provide enough information on their services and (professional) background (in the text of their advertisement) in order to attract the attention of their potential clients. On the other hand, the limitation of the data authenticity (as mentioned already by O. Kozar, 2013) has to be acknowledged, because online profiles should be viewed rather as a self-presentation of individuals offering their services on the market and might not always correspond to real-life facts (on the other hand, by stating false data, the tutors would risk a loss of credibility in case their customers find it out). Also, the study has similar limitations as the one of E. Tanner et al. (2009, p. 10), that is that the research took in only a section of the tutoring market, which may not reflect the full extent of tutoring in the particular country.

Conclusion

The data available online for research on the shadow education are relatively rich, less difficult to collect compared to field work (e.g., collecting them via questionnaires) and currently underutilized by shadow education analysts, despite the fact that its research can bring interesting insights and findings. This paper tried to point out some challenges and pitfalls the author encountered during his research, and the ways he addressed them. These shared insights will hopefully prove useful for further research in this domain.

References

[1] Bauer, C., & Scharl, A. (2000). Quantitive evaluation of Web site content and structure. *Internet Research, 10(1),* 31–44.

[2] Borzekowski, D. L. G., Schenk, S., Wilson, J. L., & Peebles, R. (2010). e-Ana and e-Mia: A Content Analysis of Pro–Eating Disorder Web Sites. *American Journal of Public Health, 100(8),* 1526–1534.

[3] Bray, M. (1999). *The shadow education system: Private tutoring and its implications for planners*. Paris: International Institute for Educational Planning.

[4] Bray, M. (2009). *Confronting the shadow education system: What government policies for what private tutoring?* Paris: United Nations Educational, Scientific and Cultural Organization.

[18] In the Czech Republic, exploitation of such online databases is not permitted, unless it is performed for education or scientific purposes (Copyright Act No. 121/2000).

[5] Bray, M. (2010). Researching shadow education: Methodological challenges and directions*. Asia Pacific Education Review, 11(1),* 3–13.

[6] Bray, M. (2011). *The challenge of shadow education: Private tutoring and its implications for policy makers in the European Union*. Brussels: European Commission.

[7] Bray, M., & Kobakhidze, N. (2014). Measurement issues in research on shadow education: Challenges and pitfalls encountered in TIMSS and PISA. *Comparative Education Review*, *58(4),* 590–620.

[8] Bray, M., & Kwo, O. (2014). *Regulating private tutoring for public good: Policy options for supplementary education in Asia*. Paris: UNESCO a Hong Kong: Comparative Education Research Centre, HKU.

[9] Bray, M., & Kwok, P. (2003). Demand for private supplementary tutoring: Conceptual considerations, and socio-economic patterns in Hong Kong. *Economics of Education Review*, *22(6),* 611–620.

[10] Bray, M., Kwo, O., & Jokić, B. (2015a). Introduction. In: M. Bray, O. Kwo, & B. Jokić (Eds.) *Researching Private Supplementary Tutoring: Methodological Lessons from Diverse Cultures* (pp. 3–19), Hong Kong: Comparative Education Research Centre (CERC), The University of Hong Kong, and Dordrecht: Springer.

[11] Bray, M., Kwo, O., & Jokić, B. (Eds.). (2015b). *Researching private supplementary tutoring: Methodological lessons from diverse cultures*. Hong Kong: Comparative Education Research Centre.

[12] Bryman, A. (2015). *Social research methods.* New York: Oxford University Press Inc.

[13] Cohen, L, Manion, L., & Morrison, K. (2011). *Research methods in education.* New York: Routledge.

[14] Copyright Act (n. 121/2000) on copyright and rights related to copyright and on amendment to certain acts. https://www.mkcr.cz/doc/cms_library/12-az_2006_v_aj-2005.pdf

[15] Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research.* New York: Pearson Education Inc.

[16] Faganel, A., & Trnavčevič, A. (2013). Constructions of private tutoring in Slovenian online chatrooms. In M. Bray, A. Mazawi, & R. Sultana (Eds.), *Private tutoring across the mediterranean* (s. 167–176). Rotterdam: Sense Publishers.

[17] Fraenkel, J. R., & Wallen, N. E. (2009). *How to design and evaluate research in education*. New York: McGraw-Hill.

[18] Gavora, P. (2015). Obsahová analýza v pedagogickom výskume: Pohľad na jej súčasné podoby. *Pedagogická orientace, 25(3),* 345–371.

[19] Herring, S. C. (2004). Computer-mediated discourse analysis: An approach to researching online behavior. In: S. A. Barab, R. Kling, & J. H. Gray (Eds.), *Designing for virtual communities in the service of learning* (pp. 338-376). New York: Cambridge University Press.

[20] Herring, S. C. (2010). Web Content Analysis: Expanding the Paradigm. In J. Hunsinger, L. Klastrup, & L. Allen (Eds.), *International Handbook of internet Research* (pp. 233–249), Springer: Dordrecht.

[21] Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative health research*, *15(9),* 1277–1288.

[22] Kim, I., & Kuljis, J. (2010). Applying content analysis to web-based content. *CIT. Journal of Computing and Information Technology, 18(4),* 369–375.

[23] Kozar, O. (2013). The face of private tutoring in Russia: Evidence from online marketing by private tutors. *Research in Comparative and International Education*, *8(1)*, 74–86.

[24] Kozar, O. (2015). Discursive practices of private online tutoring websites in Russia. *Discourse: Studies in the Cultural Politics of Education, 36(3),* 1–15.

[25] Krippendorff, K. (2004). *Content analysis: An introduction to its methodology.* Sage Publications: Thousands Oaks.

[26] McMillan, S. J. (2000). The microscope and the moving target: The challenge of applying content analysis to the World Wide Web. *Journalism & Mass Communication Quarterly*, *77(1),* 80–98.

[27] McMillan, S. J. (2000). The microscope and the moving target: The challenge of applying content analysis to the World Wide Web. *Journalism & Mass Communication Quarterly*, *77(1),* 80–98.

[28] Neuendorf, K. (2002). *The content analysis guidebook.* Sage Publications: Thousands Oaks.

[29] Pace, L. A. & Livingston, M. (2005). Protecting human subjects in internet research. *EJBO-Electronic Journal of Business Ethics and Organization Studies, 10(1)*, 35–41.

[30] Silova, I., Būdiene, B., & Bray, M. (Eds.) (2006). *Education in a hidden marketplace: Monitoring of private tutoring.* New York: Open Society Institute.

[31] Šťastný, V. (2016a). Klíčová témata a metody ve výzkumu soukromého doučování [Key Topics and Methods in Private Supplementary Tutoring Research]. *Orbis Scholae*, *10(1),* 35–62.

[32] Šťastný, V. (2016b). Private supplementary tutoring phenomenon as a shadow education system

in the Czech Republic (Dissertation thesis, Charles University). Retrieved from https://is.cuni.cz/webapps/zzp/download/140052563/?lang=en

[33] Šťastný, V. (2017a). Book review: Researching private supplementary tutoring: Methodological lessons from diverse cultures by Mark Bray, Ora Kwo and Boris Jokic (Eds.) [collective monograph review]. *Asia Pacific Journal of Education.* Retrieved from http://www.tandfonline.com/doi/full/10.1080/02188791.2017.1319897

[34] Šťastný, V. (2017b). Private tutoring lessons supply – Insights from online advertising in the Czech Republic. Compare: A Journal of Comparative and International Education. Retrieved from http://www.tandfonline.com/doi/full/10.1080/03057925.2016.1259064

[35] Stevenson, D. L., & Baker, D. P. (1992). Shadow education and allocation in formal schooling: Transition to university in Japan. *American Journal of Sociology*, *97(6)*, 1639–1657.

[36] Tanner, E., Ireson, J. M., Rushforth, K., Smith, K., Day, N., Tennant, R., & Turczuk, O. (2009). *Private tuition in England.* Research report DCSF-RR081. London: Department for Schools and Families. Retrieved from http://www.dcsf.gov.uk/research/data/uploadfiles/DCSF-RR081.pdf.

[37] Weare, C. & Lin, W.Y. (2000). Content analysis of the World Wide Web: Opportunities and challenges. *Social Science Computer Review, 18(3),* 272–292.

[38] Webcertain (2013). *Global search and social report (Q4 2013)*. Retrieved from http://internationaldigitalhub.com/en/publications/the-webcertain-global-search-and-social-report-2013/download

APPENDIX

Table 1 Previous studies of shadow education using internet as a source of data

	E. Tanner et al. (2009)	A. Faganel & A. Trnavcević (2013)	O. Kozar (2013)	O. Kozar (2015)	V. Šťastný (2017b)
Objective	to provide a national profile of private tuition providers … and to offer more detailed information on the characteristics of private tuition transactions, (costs, location, frequency and length of sessions…)	to explore the content and nature of discourses on different forums and websites reflecting both demand and supply of tutoring	to find out, which are the most popular subjects for private tutoring, and to obtain information on the background of sought-after private tutors in Moscow	to investigate whether the private tutoring related websites exhibit similar discourse and ideology and whether they might belong to the same 'genre prototype'	to analyse the socio-demographic background (age, gender, qualifications, professional profile) of individual private tutors advertising online and their distribution within a country; to assess the macro and micro factors underlying the advertised price (fees) for a tutoring lesson
Methodological approach	Quantitative[1]	Mixed	Mixed	Qualitative	Quantitative
Search engine used	Google.uk, Metacrawler, Zapmeta	Google	Yandex	Yandex	Google, Seznam.cz

Method	Construction of database of tuition agencies with web presence and further quantitative analysis of data	Content analysis of communications about private tutoring in chatrooms	Quantitative analysis of all tutor profiles; qualitative analysis of private tutor profiles	Critical discourse analysis of thematic websites	Quantitative analysis of private tutor profiles
Sample	504 records in the database (each record representing one agency with online presence)	One of 30 thematic websites chosen; of 536 advertisements on the site 81 were analyzed	Two largest tutor-listing websites were selected for the analysis (size determined as number of registered tutors), further analysis of tutor advertisements on site http://repetitors.info/ (n=21 497) and http://repetitor-baza.ru/ (n=8 074); and of 32 private tutor profiles	17 websites belonging to “online schools category” identified through internet search	One of eight mediated notice board websites chosen, in total 2 058 tutor profiles analyzed