UNIVERSITY OF TURKU

CONDITIONS AND EFFECTS OF AN INTELLIGENT TUTORING SYSTEM USAGE FOR RUSSIAN HIGH-STAKES EXAM IN ENGLISH

Faculty of Education

Department of Teacher Education

Master's thesis

Author(s):

Alexey Tarasov

30.05.2023

Turku

Master's thesis

Subject: Education
Author(s): Alexey Tarasov
Title: Conditions and the effects of an intelligent tutoring system usage for Russian high-stakes exam in English
Supervisor(s): Koen Veermans
Number of pages: 67 pages
Date: 30.05.2023

Abstract

The aim of the proposed study was to dwell on the field of intelligent tutoring systems as applied to high-stakes exam settings in foreign languages. The main research hypothesis of this paper was the following: Does the study attempt frequency within the suggested intelligent tutoring system affect the overall students' learning performance in preparation for the Speaking part of the Russian high-stakes exam in the English language? Addressing this research hypothesis also resulted in acquiring understanding on key stakeholders' perception of preparation for the Russian high-stakes exam in English. Research literature was thoroughly analyzed and the suggested intelligent system was described in detail. Data was collected through a computer-based automated procedure with further randomization and sampling. As a result of the study, three cohorts of users of the intelligent tutoring system were defined. Each cohort maintained a positive study dynamics experienced through the use of the intelligent tutoring system. Also, continuous aspiration for implementing online self-training environments was identified within the majority of a foreign language teachers' community. The framework developed for the research can be used in future research as a foundation for investigating self-regulated learning environments created for the Speaking part preparation of high-stakes exam in foreign languages.

Key words: computer-based learning, computer-assisted learning, high-stakes exam, intelligent tutoring system, self-regulated learning, second language learning, second language acquisition, speaking tasks, speaking tests.

Table of contents

Chapter 1. Background and Significance

1.1 Introduction

In recent years, Russia has experienced, undoubtedly, the biggest technological switch in the state exam testing settings. Despite the attempts to ‘humanize’ the language testing procedure throughout the course of demo periods, educational authorities finally have agreed on incorporating a computer-based testing system within the framework of high school Russian State Exam (abbreviated as EGE and alternatively called in research works as Unified Russian State exam or USE for short) in foreign languages.

Previously in 2009, State Exam was implemented into the high school curriculum dismissing the in-school final exam framework alongside the university exam entrance procedure. However, this change wasn’t anyhow linked with the technology and was referred as a way to fight the corruption in the university enrollment (Denisova-Shmidt & Leontyeva, 2014). Foreign languages state exams (English, French, German, and Spanish) have been a part of the EGE since its introduction and was literally a written examination having no oral (Speaking) part.

Such a controversial mode for a language examination without any oral communicative task has paved the way for both professional and public discussions on the issue. And since 2013 a number of oral part examinations have been suggested, and one even has been tested in the OGE (Russian state exam for secondary school leavers) settings. However, none of them have been accepted as a working solution. Eventually, a computer-based oral part procedure was first tested and then approved by the authorities. No changes have been made to the written exam including Grammar, Vocabulary, Listening, Reading, and Writing parts.

The implementation of the computer-based oral testing system (further Speaking part) took place in 2015 when the oral part of the EGE was first introduced as a component of the exam. As the language exams weren't an obligatory part of EGE, the oral exam became an option which can add up to 20 points to the overall maximum score of 100. Nearly all test takers participate making it virtually a compulsory part which by default is trained by all the potential test takers.

Different features have been suggested as the reasons for implementing the computer-based form for the Speaking part, including variable examiners’ language mastery and high

expenses of the paper-based procedure. Yet none of them have revealed a big ideological switch: from a dialogue-based scenario (initially tested in a human-based mode) to monologue-oriented tasks. Initially, the Speaking part framework consisted of a few tasks incorporating real-life discussion with a human examiner. Eventually, this framework has been replaced by a monologue format that is fully conducted in a computer-based format.

Although it is a monologue in the formal sense, the existing 4-task structure includes one task which can be considered dialogical to some extent. In this a task test taker has to ask a grammatically and semantically correct question following the guidelines on the screen (in the task the typical situation includes a real-life scenario in which a test-taker has to ask questions requesting a standard information from service providers, for instance, travel agent or a swimming pool administrator). All the other tasks (reading a passage, describing a picture and comparing pictures) were to become spontaneous monologue presentations excluding the reading task which by nature is a warming up task. The introduction of the exam criteria has unfolded a strict-structure scenario which should be mastered and demonstrated by test takers.

Bearing this in mind, well-established publishing houses have issued a variety of course books aiming at preparing for the Speaking part. Nearly all the coursebooks were backed up with a CD or an online application which was called ‘self-training systems’. All the systems provide the exam-like interface with only one yet important feature – to record the voice and the playback option. Though this provides some support, this falls short for fulfilling the goal of autonomous or semi-autonomous study. All the systems provide the exam-like interface with only one yet important feature – to record the voice and the playback option.

The observation of students’ in-class preparation has shown that in addition to the support provided by the materials from publishers, teachers provide assistance to the computer-based exam with paper-based materials. If this combination would together result in satisfactory outcomes, that could be the end for the discussion. Yet the first statistics on the Speaking part which was altogether considered as ‘the easiest thing possible’ has shown that students experience difficulties in managing the Speaking tasks in comparison to other exam parts.

Evidently, the computer-based exam has become a troublesome issue not because of the language complications but rather because of the unusual non-human environment to which the students need to get accustomed to literally during the simulation and real exams. And the distribution of the Speaking part results in comparison to the other exam components casts doubt on whether the currently used non-self-regulated (‘non-computerized’ in other words)

methods of preparation are a good match to the computer-based exam in question. This can be illustrated with results from the statistics of FIPM (Federal Institute of Pedagogical Measurements – a government body which has suggested, implemented the Speaking part and further provided follow-ups) is presented in Figure 1.

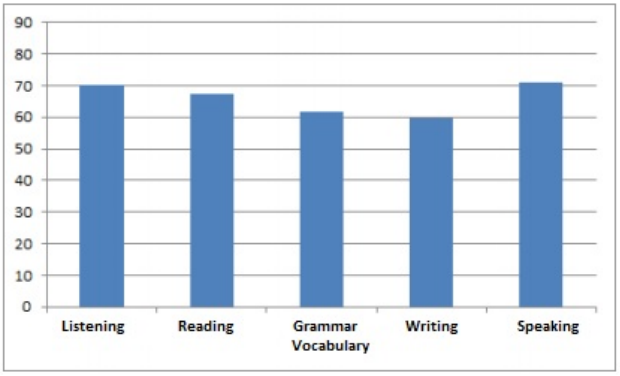

Figure 1. Distribution of EGE results (all Russia) in 2015

In 2015, the year of test implementation, test takers performed better in Speaking than in all the other parts of the exam scoring 71% on average. Listening section was 2nd on the list with 70% of the tasks completion.

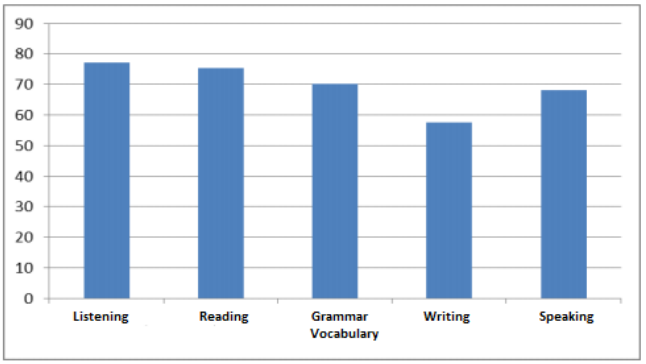

Figure 2. Distribution of EGE results (all Russia) in 2017

In 2017 the picture wasn't upside down, but the change is vivid (see Figure 2). In all the reproductive skills tested (Listening, Reading) the performance was better than in productive skills (Speaking and Writing). In Speaking test takers have scored 68% lowering the average figure by 3%. It seems important to note that both productive language skills have shown divergent dynamics although for the Speaking part the downturn is not quite significant.

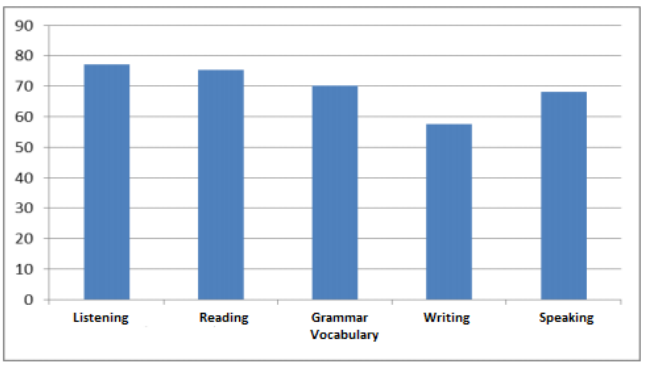

Figure 3. Distribution of EGE results (all Russia) in 2018

The 2018 statistics has demonstrated an identical pattern when the reproductive skill performance is on an upward trend while productive skills results are on a downward path (see Figure 2). For Speaking the drop was 2 % showcasing a 66% success rate.

2.1 Significance survey

These results on the Speaking part after it was implemented as a computer-based addition to the existing exam framework, could be expected at the time of introduction, but one would not expect this to be lasting or even become larger. This illustrates that the preparation for the Speaking part by the publisher material and teachers tried to tutor students both aiming at preparing students.. In case of teachers this was mostly done with paper based materials and given the dropping scores of the Speaking part, it might be reasonable to look for alternative methods, preferably first introduced to the teacher - a decision-maker for the study materials usage. A survey being conducted within the population might reveal overall acceptance of the paper-based material malfunction thus potentially securing a computer-based study path as ‘making-the-difference’ trigger. The population awareness of the necessity of such a switch is to secure the significance of the issue, which might be resolved by implementing an intelligent tutoring system, sufficient enough for EGE Speaking part preparation.

The significance survey was conducted within the following framework: a preliminary stage, two main stages, and a data analysis stage. The framework secures the application of test-retest principle which has contributed to the research reliability which, in its turn, highlights the significance of the issues in question.

Assuming the view of the teacher as a major ‘make-a-difference trigger’ to the student learning performance (Hattie, 2003), it seemed important to study the teachers’ attitude

towards ITS as it could be defined as an essential tool for the exam preparation. This reflection is an observational one coming from putting together a new computer based form of language testing (speaking part) and open stats available at https://fipi.ru/ege/analiticheskie-i-metodicheskie-materialy (only in Russian) covering the recent years exam outcomes.

As it was stated earlier, we assume that the teacher's role can be of great importance in such settings as the case taken under consideration could be regarded as a challenging one due to lack of certainty (new testing tool) and a high-stakes nature of the examination. Although the statistics have not shown improvement over time, it is still reasonable to assume that the teacher's leadership might be an instructive insight for students willing to reach learning goals. What was called reflection in the introductory passage might be better understood through the relevance motive to the idea in question. It seems obvious that teachers do focus on a practical side of the instructive intervention. Thus, it is plausible to claim that a would-be useful tool might be considered as such when teachers see its relevance to the existing exam situation. Any learning or testing tool, if introduced from outside, becomes an external feature. Thus, its implementation under the aforementioned circumstances targets improving the existing preparation framework. It seems reasonable to assume that teachers can be such 'external providers' for students' population.

To the naked eye, the divergent trend in EGE Speaking part exam results could be assigned to various reasons including variable assessment and exam papers. However, such controversy in the Speaking part might have something to do with the technological innovation and students' learning process: students experience more and more difficulties due to growing pressure and a number of minor changes such as task wording therefore acquiring anxiety in the course of preparation instead of getting confidence and useful skills. It seems reasonable to assume the aforementioned view as a probable one as no significant changes have been made in the testing materials over the first 4 years of the exam practice. Moreover, an improvement trend or at least minor fluctuations could be expected as a normal course of events when it comes to replicating the same exam formula (in case of EGE the reproductive skills statistics is a vivid example of the opposite trend when teachers and students adjust to the familiar exam framework). And it seems quite reasonable to assume that the lack of matching-the-task educational environments could contribute to the downward trend in the exam performance. Moreover, teachers do realize that students experience great difficulties

during EGE Speaking part preparation due to the exam's linguistic features and learning tools in use.

To test the assumption and its correlation with the teachers' reflection on the learning process, it seems critically important to look at the teachers' overview on the preparation stage to Speaking part focusing on the learning tools and materials in use. To reach the goal, three surveys have been carried out aiming at understanding general patterns holistically. Below the specifics of each stage is presented (the main stat parameter is given in the brackets as well as the type of data used). The measurement tool is 'correlation coefficient'. The nominal data (utterances of teachers' population) are processed by means of content analysis.

1) *Preliminary survey* (the mean number/qualitative and quantitative data)
2016 - The aim was to study teachers' understanding of the would-be innovations for the EGE Speaking part. For this matter one close-ended question has been taken into account (the other questions of the survey were intended for investigating the whole process of the EGE preparation).
2) *Main survey – 1st stage* (the mean/qualitative and quantitative data)
2017 - 2018 - The aim was to investigate teachers' perception of online tools implementation into the EGE Speaking part preparation. The suggested framework with open-ended questions has contributed to the stage validity as there were no such lead-in inquiries as close-ended questions.
3) *Main survey - 2nd stage with further analysis* (the mean number + correlation/qualitative and quantitative data)
2020-2021 - The aim was to observe a consistency pattern (or a lack of such) in comparison to the 1st stage. The framework is a replication of the 1st stage followed by correlation analysis.

The suggested design of the significance survey is to compensate for its obvious drawback - no matching populations throughout conducting all survey stages: teachers' cohorts were made up of different individuals. Hence, there is no solid ground for treating the suggested analysis as the inter-rater agreement. Nonetheless, applying a test-retest principle is to contribute to the overall reliability of the suggested design.

It should be mentioned here that the inter-rater processing in the current study has been done unconventionally due to the basic feature of the surveys – different teachers' cohorts were being questioned during both phases. Thus, it is assumed that teachers' cohorts bear very similar characteristics (close to identical) and could be treated as 'pairs' within the timespan of 2 surveys. The overall framework for the significance survey is aimed at detecting a pattern in support of the online tools implementation. For this matter the following design was suggested. However, it is still reasonable to treat both cohorts similar, but independent.
The first preliminary survey about EGE preparation contained 10 questions. It was conducted online via SurveyMonkey service throughout 2016 and 1 question was taken into consideration – 'What materials could facilitate preparation to the computer-based Speaking part'. The question was closed and contained 4 options including 'practice with a native speaker' and 'your choice option. Two options – 'self-training systems' (coded as 1 see Appendix 1; a copy translation from Russian into English which is a conventional name in coursebooks) and 'more practice in the exam format' (coded as 2) – seem to be corresponding to the idea of 'self-training' educational environments. Even the latter option should be assumed as a relevant match as long as the exam format represents computer-based mode which, if practised, requires a specific training environment enhancing students to interact and 'collaborate' with the machine. The total number of participants was 66. All the participants are considered, but only relevant answers (n=51) were taken for measurement. The design of the survey was aimed at revealing a percentage of people who consider computer-based preparation as a sufficient option, an additional element of the exam preparation. Thus, reaching a level of .77 could be assumed as high teachers' awareness in necessity for self-regulated learning environment implementation. Although the figures seem to serve as a confession towards innovation, there might be a high probability that the preliminary survey's framework with closed questions has driven teachers to choose a brand-new option regardless of their real intentions and claims. Therefore, the main focus of the research is placed on the following 2 iterations boasting the content analysis paradigm through open-ended questions. A relatively high level of teachers' 'innovation' intent can be partly attributed to the design of the preliminary stage with reliance on a close-ended question, which might drive the participants into picking up 'the most expected option'. That is why the preliminary stage has to be considered as a rough draft serving a benchmark function for the main stages of the survey. As said earlier, the idea behind conducting the significance survey was to see a general pattern, which can be regarded as a trendline showcasing continuous changes (or the

lack of the ones) in the teachers' attitude towards using the intelligent tutoring system in question.

The main survey had two stages and was organized as a set of entrance questions to EGEnglish.ru facebook page. This public was made up of English teachers who are involved or interested in the EGE preparation process (https://www.facebook.com/groups/284294695249462/). The first questionnaire went online in 2017-2018 and contained 3 open questions one of which has been taken into account: 'What online tools do you lack for the EGE preparation?'. The second iteration was released online in the summer of 2018 and also was made up of 3 open-ended questions. In this case, the procedure included collecting answers to the following question: 'What did you lack most for a better the EGE preparation?'. The total number of participants in two iterations of the questionnaire was 384. The next stage was to single out relevant replies using content analysis based on the binary principle: answers containing 'Speaking or Tutoring entities' were taken into account while others were discarded (n=305). By treating 'tutoring entities' reply as a ITS reference, we assume that teachers are aware of the existing online tutoring systems which are to great extent Speaking learning environments. Thus, mentioning the term in general which falls into the group is highly likely a reference to the Speaking learning environment. The sample wasn't randomized as long as the total number of relevant answers was 79 (for the 1st stage). In order to see a match with the preliminary stage of the survey, the first 51 answers from the 1st stage sample were further analysed in order to pinpoint replies containing the following keywords (originally in Russian and given in English translation):'tutoring system', 'self-training system', 'online system', 'computer-based system', 'self-training environment', 'speaking apps' (coded universally as 1 in the Survey 2 column; see Appendix A). The content analysis has shown that 33 participants stated the lack of tutoring systems for EGE Speaking preparation.

The calculated score for the 1st stage of the main survey is .65 which could be assumed as a majority's desire within teachers' populations to implement 'self-training systems' into the preparation practice to Speaking part of EGE. Moreover, the following mean figure seems to be relevant due to content analysis use.

The second stage was performed in 2020-2021 in the same facebook group with the same framework of questions offered to the target audience. The population had the same basic

features like their counterparts in the 1st stage iteration: teachers of English, preparing or interested in preparing for the EGE exam. None of 1st stage participants were allowed to take part in the 2nd stage due to the procedure implemented: only newcomers were offered a questionnaire.

The collection of the data was carried out online following the same binary principle of the intended entities extraction with only one discrepancy: «automated systems», «online learning environment» were added to the 'Speaking or Tutoring entities' as differential clues for coding relevant answers. This was done due to the fact that by the time of the survey implementation (2020-2021 study year) no other automated online systems had been developed apart from the Speaking online tutoring system in question (EGEnglish.ru). The same set of keywords (originally in Russian and given in English translation) was taken into account: 'tutoring system', 'self-training system', 'online system', 'computer-based system', 'self-training environment', 'speaking apps'. Sticking to the 'online system' entity might still seem to be questionable, but having only the Speaking part as the only exam part entitled to the computer-based procedure appears to be a reasonable foundation for assuming such a broad term as an applicable notion for the analysis.

The initial idea of the 2nd stage was to replicate the 1st page path in terms of collecting the exact number of raters to ease further processing of the data. Unfortunately, the Facebook policy implementation discarded the entrance questionnaire barely stopping the initial data collection mechanism. It was decided not to roll out a new edition in different settings as it could violate the framework significantly.

Finally, the data extracted (see Appendix 2) were processed. The figure marking support for a self-regulated mode for the 2nd stage is .57 showing a proximity to the result of the 1st stage which exercises a .07 margin (after putting the results of both stages together, there is the following calculated margin). Therefore, it is plausible to note that there is a continuous trend in which a simple majority of teachers advocate for implementation of an online tutoring system.

After processing the 2nd stage figures, it is also possible to claim that all participants of both stages literally belong to the same big cohort of teachers involved in an online community which serves as an online reference or a teachers' helpdesk. Judging by the thin margin

between the 1st and the 2nd stages it is reasonable to conclude that on both occasions teachers have shown particular interest in applying online tutoring possibilities for the EGE preparation.

Content analysis of the data was challenging due to the necessity to single out methodology (pedagogical) aspects from the exam administration entities. For instance, chunks of speech containing administrative exam features (Speaking assessment) were discarded. At the same time it is important to mention troubleshooting techniques that have been used during the content analysis. Some entities extracted from the questionnaire were vague in terms of their linguistic bias hence limiting the power of the measurement tool. For the sake of unfolding the bias the following technique was implemented: a) the linguistic biased entities were considered separately in regard to their relevance to the keywords in question, b) connection to the 2nd question of the survey (What are three difficulties of the EGE preparation?). Below there is elaboration on the nature of the detected biased entities.

For instance, the entity ‘automated equipment for recording’ was regarded as a ‘Speaking entity’ due to the following reasoning: a recording option is a basic feature of the online learning environment for the EGE preparation. Some teachers’ replies were vague due to the very general manner of vocabulary in use. For instance, a participant used the coining ‘lack of technical resources’ which can be attributed to the Speaking part only by means of studying the other question which revealed a teacher’s concern over the Speaking issue thus confirming its connection to the primary question.

The framework, allocating two stages, secures the application of test-retest principle which has contributed to the research reliability. As long as both stages were carried out online, the following ethical considerations were taken into account: open access without the need to distribute personal information and the confidentiality guarantees to the specific unintended personal data pieces that could uncover the participants’ personality. The latter was addressed to the questionnaires of the 1st and 2nd stages which were done in a Facebook public group. By aligning both stages we can detect a few patterns highlighting a continuity in terms of teachers’ outlook towards implementing close-to-life learning environment:
1) Both stages have experienced a similar participants’ ratio referring to the Speaking part of the test. Thus, for the 1st stage it was 17,5%. For the 2nd stage it was 10% meaning that both had fallen into 10-20% benchmark. The downward trend can be interpreted as a foreseen

expectation due to loss of the concept's newness which was obvious in the first edition of the questionnaire (within 2 years after Speaking part implementation).

2) The samples of those participants treating ITS systems as a valuable asset to the EGE preparation are almost of similar value.

Chapter 2. Literature review

For the last four decades many research papers have strived to highlight the efficiency of learning environments based on the assumption that learners through acquiring a certain degree of independence get higher educational outcomes. Rooted in Piaget's ideas of constructivism and Vygotsky's social-cultural approach, the collaborative principle has been repeatedly nominated as one of the major triggers stimulating students' better knowledge development. Thus, establishing interaction in teacher-students and student-student learning teams presupposes building far more complex environments taking into account continuous interactivity which is obtained by having guided activity, reflection, feedback, pacing control and pretraining (Moreno et. al, 2007). The following principles have conceptualized instructional design as a relevant framework for emerging learning environments.

The set of aforementioned principles emphasizes technological opportunities of computer-based instruction which is coined in the term of 'powerful learning environments' providing 'students with optimally supported possibilities for high-level learning, improving students' adequate self-regulation and facilitating the advancement of their conceptions of knowledge, learning, and instruction' (Lowyck et. al,. 2003). Although the given definition has no direct connection to collaborative nature of learning, it seems evident that all the mentioned features clearly explain deeper and more intelligent interactions between students and a learning agent which since then is regarded as an electronic or online system with permanent or occasional teachers' intervention into the learning process. And the collaborative human-like nature is still maintained within tutoring systems as they aim at replicating the instruction dialogue.

The idea of an intelligent tutoring system is rooted in Mastery learning (ML), the concept of (Bloom, 1971). The ML approach assumes organizing learning in a number of stages when the transition from one to another is obtained by getting formative assessment reflecting what was learned and what should be studied. The scheme is then reciprocated until the goal of each stage is reached. It is necessary to note that ML was conceptualized in the human instruction mode when the teacher is in charge of both instructional and implementation phases and there is no 'competing instructional entity'.

The appearance of e-learning tools has widened a range of educational opportunities providing nearly instant assistance which is not commonly supported by the human instructor. As a result, the concept of 'self' has gained significance as instruction and testing phases could be fully automated while teaching methods are incorporated in both. And two new similar learning strategies can be regarded as development of ML approach: self-directed and self-regulated learning. Both terms are often considered parallel and used interchangeably. However, it seems necessary to draw a line between two concepts by closely studying the context of both.

In the further review the focus is on the concepts, which are directly related to the notion of intelligent tutoring system, thus the text refrains from using the umbrella term of computer-assisted language learning (CALL) and some of its notions, which are not relevant to the research in question.

2.1 Self-directed learning and learning autonomy

Self-directed learning is seen as one's own learning independence in relation to instructional forces by choosing learning goals and methods to reach the destination (Tobin, 2000). The definition shows a broad nature of the notion referring to ideas of independence and choice. And research works clearly articulate that the learning process could be truly self-directed in a context motivating quite free learning manipulation of the material for further study. Wiklund-Engblom (2013) described the following e-learning structure which was offered to the staff in corporate training settings. The two described e-learning iterations of the educational environment did have recurrent feedback mechanisms with a very high level of independence (materials could be used practically randomly, especially in the first iteration) and thus might be considered as a self-directed educational environment.

In language learning research, there is another concept, which can be regarded as a more 'instructional' and less self-imposed method. It is learning autonomy. Although the term is treated as an elusive entity with such a distinctive feature as possessing responsibility for learning, the suggested plethora of the other descriptors – decision-making, choice, control, independence, capacity to learn, self-awareness, active learning, self-direction, strategic competence, motivation, metacognition, behaviour, reflection, goal-setting, self-assessment, time management (Garrison, 2003; Hurd, 2005; Scharle & Szabo, 2000; White, 2003) –

clearly articulate zero possibility of their acquisition through solitary process of learning discovery without teachers' instruction.

Achieving autonomy is viewed through the lens of teacher-student collaboration which can positively affect this non-solitary process (Andrade & Evans, 2013; White, 2003). However, in language learning there is another perspective on the issue claiming the students' necessity to have developed autonomy skills in various degrees prior to approaching autonomy-oriented learning models (Nunan, 1997; Scharle & Szabo, 2000).

2.2 Self-regulated learning

Pintrich (2000) defined self-regulated learning as a constructive process when learners set goals for their learning and attempt to monitor, regulate, and control their cognition, motivation and behavior. Later on, self-regulated learning was specified as a notion about mastering and monitoring one's skills in the learning process in order to succeed in the specific task that one has chosen (Brand-Gruwel et al., 2014). In that sense, self-regulation refers to our ability to adapt to the tasks and context in order to master a skill or a set of skills in order to succeed in the learning process, while self-direction concerns our independence in choice of content in relation to instruction and goals.

Although self-regulation is regarded as a set of teachable skills, it is claimed that there is another dimension for self-regulation which emerges from experience (Paris & Paris, 2001). Self-regulated learning is often associated with students being able to acquire a set of effective learning strategies and their further application for a particular task (Andrade & Evans, 2013). Students, possessing a wider array of strategies, generally gain higher learning outcomes in comparison to those individuals who have acquired a limited number of strategies (Zimmerman & Martinez-Pons, 1986).

The categories of self-regulated learning, i.e. metacognitive, motivation, cognitive, behaviour, are almost identical to Oxford's four language learning strategies: metacognitive, affective, cognitive, social-interactive (Oxford, 2008).

Classroom observations from research papers highlight the teacher's leading role in helping students to acquire a status of a self-regulated learner. In an appropriate learning environment,

facilitating instruction – exposure to complex tasks, offering study choices, providing opportunities for self and peer evaluation – altogether contribute to a profile of a confident, resourceful, and curious learner who engage in regular use of metacognitive, cognitive, motivational, and behavioral strategies (Andrade & Evans, 2013; Perry et al, 2002).

2.3 Intelligent tutoring system

The shift from behaviourism to cognitivist approach to building knowledge alongside the development of microcomputers have given a way to the appearance of a new instruction domain – computer-assisted instruction (CAI) which has provided a framework for Intelligent tutoring system (ITS) development.

The Intelligent tutoring system's learning method based on cognitive psychology and ML approach has become a further stage in developing computer-based instruction which was previously assigned to a stimulus-response behaviorist approach. The 'intelligence' of ITSs is ascribed to Artificial Intelligence (AI) application which has provided mixed-initiative instruction dialogue, personalized to the needs of the individual student (Brown & Sleeman, 1982; Wenger, 1987). ITSs are generally seen as a replica of human instruction which incorporates knowing of 'what to teach, who to teach and how to teach' (Nwana, 1990).

The key features of ITSs are concerned with differentiation of such tutoring systems from their predecessors and include adaptivity, balanced control between students and ITSs, in-built domain specific knowledge (Brown & Sleeman, 1982).

The functionality of ITSs is generally attributed to the four-folded model including domain module (knowledge on the subject, mainstream and alternative explanations), tutorial module (teaching goals and plans, provide instruction and learning activities, diagnose misconception, intervene in case a student experiences difficulties), student module (maintain information about student's cognition), and interface module (Garito, 1991; Nwana, 1990). Yet sometimes the emphasis is put on the control function enlarging the model to 5 elements by adding a control module which is in charge of treating detected errors by adapting to the students' level of advancement (Padayachee, 2002) and the teaching methodology reducing the model to 3 components (Self, 1999).

The efficiency of ITSs is raised by the research community within the paradigm of computer vs. human instruction. Thus, meta-analyses and meta-analytical reviews seem to be of greater value as long as they present more reliable evidence and by far diverse research perspectives resulting in different investigation outcomes. VanLehn (2011) has arrived at the conclusion that ITSs are comparable to human instruction mode in terms of efficiency. Also, it was pointed out that 'there is an interaction plateau rather than a steady increase in effectiveness as granularity decreases' meaning from-no-to-moderate efficiency of ITSs with subsequent dividing of learning units. Kulik and Fletcher (2016) argued that in most cases ITS students outperformed their counterparts from conventional classes with a higher performance improvement for ITSs compared to human instruction. Ma (2014) claimed that significant positive mean effect sizes were traced regardless of the ITSs' usage type (principal means of instruction, a supplement to teacher-led instruction, an integral component of teacher-led instruction, aid to homework). Also, this research work highlights that there was no significant difference between learning from ITS and learning from individualized human tutoring or small-group instruction while ITS outperformed large-group instruction, non-ITS computer-based instruction and textbooks/workbooks learning mode.

Research papers spanning the current and the preceding decade have revealed a range of hot topics related to ITSs application and its importance for the educational process. The problem of understanding learners' emotional states is clearly under the spotlight of the research community (D'Mello et. al, 2007; Vail et. al, 2016; Taub et. al, 2018) Also, dialogue-based learning scenarios (Graesser et. al, 2005), metacognition development (Ramandalahy et. al, 2010; Roll et. al, 2011; Trevors et. al, 2014), and collaborative support (Bernacki et. al, 2014; Olsen et. al, 2014) are widely investigated.

2.4 Simulation learning

Learning through simulation environments is not a new topic in the field of CALL, although speaking simulations run by virtual agents has sparked the research community in the recent decades due to widespread distribution of AI web and mobile applications.

Speaking within the framework of second language learning is often diagnosed as a skill that lacks continuous practice in the classroom context (Grobler & Smits, 2017; Sydorenko et. al, 2018). Various real-life simulations including oral practice by means of storytelling (Kim,

2014), role-plays (Martınez-Flor & Uso-Juan, 2010; Yen et al, 2015), and telecollaborative discussions via Skype (Barron & Black, 2015; LoCastro, 2011) are researched and critically analysed. However, 'these tasks are either not highly structured or do not provide practical opportunities for intrinsic feedback (especially with large numbers of students) nor the modelling needed to move language development forward' (Sydorenko et al, 2019).

In recent years, simulations in language learning have been researched on a basis of mediums which maintain real-life linguistic environments or similar to such conditions. Among these mediums two have received significant attention within the CALL research community: augmented/virtual reality tools and dialogue/conversational chatbots. The striking feature of the following works is retained in their primary focus on general understanding of the technologies (Adnan, 2020; Nghi et al, 2019; Smutny & Schreiberova, 2020; Tu, 2020) or the non-productive skills development including grammar and vocabulary skills (Kim, 2019; Liu et al., 2022) as well as actors' perception of the tools' application to the classroom settings (Pokrivcakova, 2022, Chuah & Kabilan, 2021). The following outlook does have its own relevance to the field, but it clearly may drive away from the productive track for which the systems were initially engineered. Therefore, the papers advocating for speaking skills acquisition by means of chatbots and AR/VR tools (Chien, 2020; Hakim & Rima, 2022; Yang et al, 2022) are of great value as they ideologically support the methodological grounds for the aforementioned innovations rooted in ITS and autonomy learning frameworks.

It is worth noting that chatbots generally rely on using such state-of-the-art learning assistance tools as speech-to-text analyzers, natural language processing algorithms, parts-of-speech taggers – all of them work on a probability basis thus presupposing an error rate in the provided learning feedback. On the one hand, such a real-life condition without absolute accuracy stands for the human-like response nature of chatbots and VR/AR tools. However, the issue of learning (corrective) feedback is seen as one of the most challenging strata of the automated speaking agents. According to teachers' perception, the meaningful feedback rate with minimal teacher's intervention is seen as a parameter that needs to be improved (Chuah, 2021). Subtle (non-explicit) or non-immediate feedback prevents learners from noticing (Holden & Sykes, 2013). In addition, the range of feedback mechanisms should be expanded and might take the form of generalities (Sydorenko et al, 2018).

After carrying out the significance survey, it has become clear that the teachers' population has welcomed self-regulated tools which might accompany the exam preparation practice. Also, reviewing topical literature has confirmed a possibility of implementing a kind of intelligent tutoring system in the EGE Speaking part settings as a differential element aiming at strengthening students' awareness and necessary exam skills. These inferences have been taken into account for developing an ITS system described below.

Chapter 3. Intelligent tutoring system

3.1 Overview

The EGEnglish.ru platform is an online platform possessing qualities of the intelligent tutoring system. It had been specifically developed as a Russian state exam online preparation tool aiming primarily at the Speaking part of the exam. However, later a few courses were added to the range of interactive study units, which currently employ both Speaking and Writing study kits not just for EGE test-takers, but also for the public willing to improve their productive skills in the English language. Below there is a landing page of EGEnglish.ru online platform.

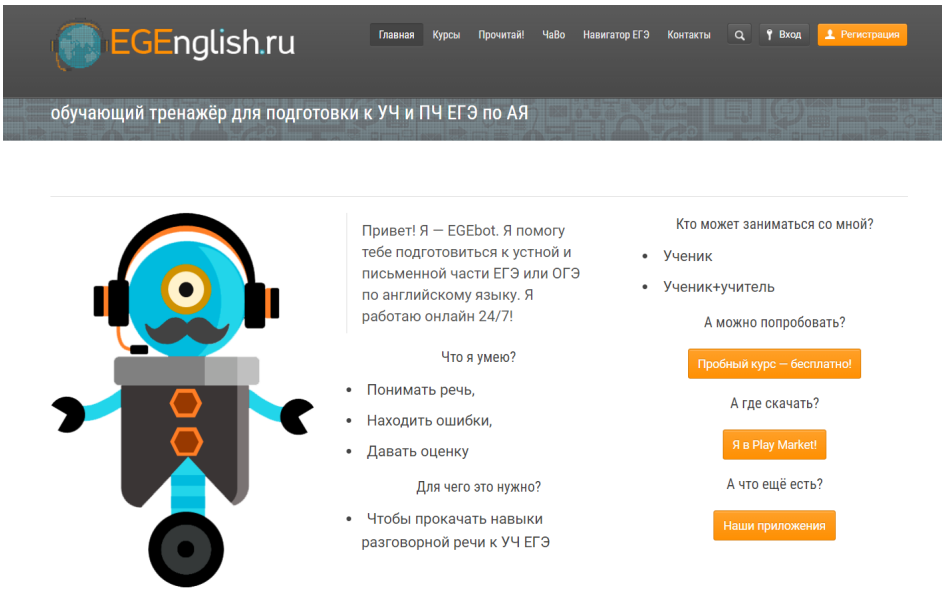

Figure 4. Landing page of EGEnglish.ru (available only in Russian)

EGEnglish.ru was developed by a team of private individuals including English language teachers and software engineers. One of the co-founders is the author of the research. The first iteration was released in April 2016 as a beta version which was tested by a limited number of test-takers and teachers. From 2017 to 2021 it was being updated on the regular basis due to the minor changes of the tasks' wording. Also, a few additional features, including a grammar checker, automated criteria-based assessment, and taboo words 'eliminator' have been added to the core technology – speech-to-text engine providing nearly instant feedback to every single attempt of a test taker. Nonetheless, the basic functionality of EGEnglish.ru was in tact throughout the research period as test-takers were

able to listen to their own learning attempts, a common feature for other existing online tutoring systems, as well as get hold of 'written imprints' of their oral performances. This feature makes it possible to define EGEnglish.ru as the only interactive platform with nearly instant feedback. The online platform is hosted at https://egenglish.ru and accessible through desktop and mobile devices.

From the very outset the core of the online environment was automated speech recognition (ASR) engine. Later on it was updated to the current technological framework comprising a natural language processing engine (content analysis of the script by the EGE criteria) and an in-built grammar analyzer. All these instructional layers are aimed at processing students' speech flow by means of a microphone and web browser tools. Below there is a standard feedback for the user of the ITS: an approximate task score (based on NLP and grammar analyzer), ASR-generated script of the utterance, a detailed criteria-based report of the approximate task score (replicating a human assessment framework).

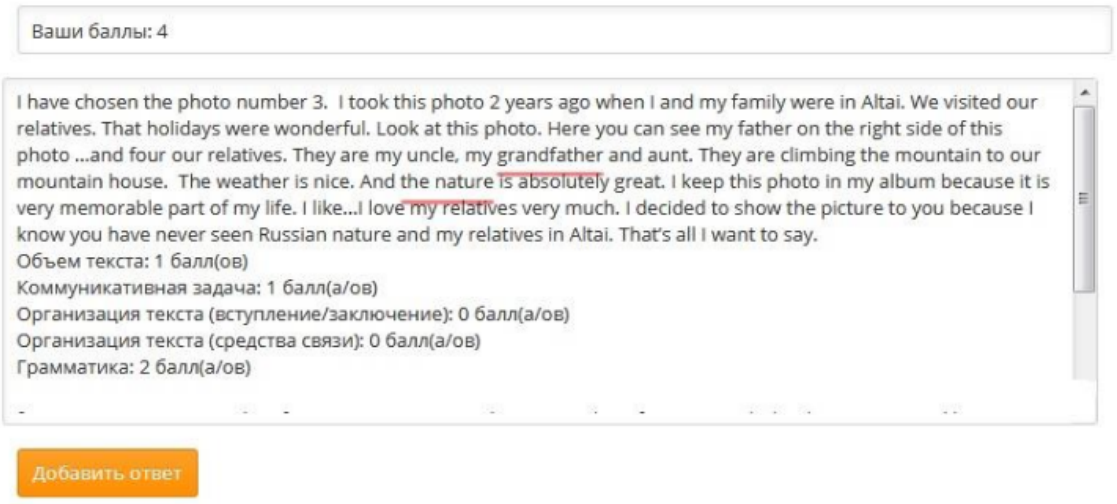

Figure 5. A standard report for the EGEnglish.ru user (partly in Russian)

3.2 Study path

The aforementioned ITS four-folded model has been implemented with a control module in order to maximise a learner's study experience. The domain module includes audio and text samples for the tasks; the tutorial module incorporates mind maps, a suggested training scheme; the student module retains study attempts information with received scores from the system (natural language processing assessment); the interface modules naturally includes all the buttons in charge of learning activities (play/stop button — for listening to samples, save

button — for saving a script in the ITS database, script form — which by default is used for presenting a script and can also serve for manipulating with a script (see the next passage). The control module, invisible to users, calculates and demonstrates the criteria-based feedback with a final score for a performed task.

Users of EGEnglish.ru are encouraged to follow such a training scheme: 1) read the given task (identical to the exam format), 2) listen to a sample response, 2) read a sample script (identical to an audio response), 4) produce an utterance using supportive tools (mind maps on the task, sample scripts), 5) study the system's feedback (an automated NLP report assessing students' performance using the current EGE criteria including content and grammar analysis) — this stage might also involve manipulation with the received script — editing it to a maximum score, 6) 2nd attempt producing an utterance without supportive tools, 7) study the system's feedback.

The automated feedback steps (5th and 7th) within the study paths show a would-be students' performance (score) in close to real exam settings. The report of the system includes a caution which states that automated feedback doesn't necessarily correlate with a human examiner's feedback.

The aforementioned cycle is viewed as a complete study path within the first attempt's framework which, however, does not fully cover the suggested learning path within a bigger cycle. In general, the learning cycle could be described with the following pattern: the 1st attempts > feedback (trigger) > 2nd and successive attempts.

All the steps are automated and do not involve any interaction with human assessors (examiners) apart from the personal account feature which makes students' scripts visible for their teachers who have previously registered and linked their students to their own personal teachers' accounts.

3.3 Learning method

As the automated EGEnglish.ru feedback provides an 'arbitrary and harsher' assessment which doesn't generally match the human scoring and basically downgrades student' performance, users by default are encouraged to make another attempt in order to reach a

maximum score which is assumed as an ultimate goal in high-stakes testing settings. Thus, such a learning chain in theory guarantees a learner's basic follow-up to the first and following attempts.

3.4 High-stakes tasks at EGEnglish.ru

The Speaking part of EGE consists of 4 tasks with each focused on specific linguo-pragmatic skill (in the brackets the maximum score):

1) Reading a text according to normative pronunciation rules (1 point)
2) Dialogue - asking 5 questions according to grammar rules (5 points)
3) Monologue - describing a picture (7 points)
4) Monologue - comparing 2 pictures (7 points)

All these tasks are accessed via EGEnglish.ru through 2 EGE-oriented study units: 'free course' (1 set of EGE tasks) and 'full course' (15 sets of EGE tasks). The full course was a paid course with no free access to the general public, but the free access promocodes were distributed regularly for test groups. Judging by the score distribution for EGE tasks, it becomes obvious that Task 3 and Task 4 account for the biggest share of the total score. Moreover, the EGEnglish.ru study methodology maintains its full capacity within complex tasks aiming at coherent speech analysis. Only Task 3 and Task 4 descriptors manifest such manipulation. Therefore, it was decided to focus on these two tasks (see Figures 6 and 7).

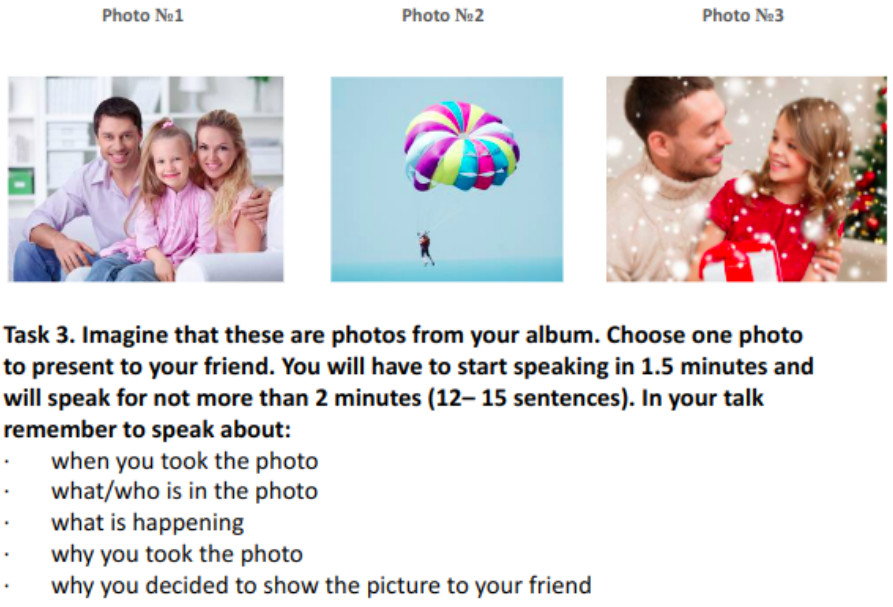

Figure 6. Task 3 from the EGEnglish.ru full course

Task 3 is considered a ‘basic level task’ by the exam developers. This coining acquires its meaning only in comparison to Task 4, which is regarded as an ‘advanced level task’. Although Task 4 description almost fully replicates the guidelines for Task 3, the comparative nature of Task 4 secures its ascribing to a more advanced level task type.

Figure 7. Task 3 from the EGEnglish.ru full course

The following link provides some insights in English for EGE Speaking part preparation: https://www.slideshare.net/MacmillanRussia/m-mann-speaking. Due to the nature of the Russian national high-stakes exam, all official materials regarding it are given only in Russian. That is why all the above stated descriptions of the EGE tasks are extracted from the FIPI official web-resource (the authority in charge of implementing the exam): https://fipi.ru/

3.5 Support

The technical support is provided by the EGEnglish.ru support team. Also, all the users can learn more about the suggested study path by watching the intro video which is presented at the home page and EGEnglish.ru YouTube channel.

No academic support is provided to test-takers by the project team. However, teachers who might have been involved in preliminary preparation stages (direct instruction to students) do get methodology training through webinars and social media communities of the project.

Chapter 4. Design and Methodology

4.1 Research objectives

The significance survey results clearly state teachers' concern towards the lack of appropriate preparation tools. The collected suggestions on the matter as well as common sense advocate for implementing such an online tutoring system which can give meaningful feedback to students.

This study aims at describing a self-regulated study path of test takers who prepare for the Speaking part of the Russian high-stakes exam in English (EGE). The benchmark is the test-taking situation which is observed by means of the intelligent tutoring system and assessed by qualified examiners. The key participants include test-takers who are predominantly 11th grade graduates and, possibly, former school leavers who, by default, are also a target population of the exam as it is open to the general public with no age limits. One another group is teachers who definitely contribute to students' learning outcomes. Although such an important group of participants cannot be excluded from the research, its design manifests that the teachers' 'interference' to the instruction might be limited due to the application of certain data collection principles. In detail, they will be described in the Data analysis section of the research.

Initially, the following setup was regarded for the research: students' self-regulated learning path might be compared to the other existing instruction modes including teachers' direct interference and a mixed type (self-regulated learning and teachers' instruction). The aforementioned framework has eventually been dismissed. It happened not to be feasible due to the lack of data from teachers about the preparation process as well as newly discovered variables in test groups such as private tutors' (teachers') support during the training within the intelligent tutoring system, and variance in foreign language mastery of study groups. This inference will be elaborated on in the successive sections of the research.

Finally, a newly developed framework took the place of the initial approach. Despite being completely different in terms of the groups distribution, the second framework bears a similar research ideology aiming at investigating various study paths of test takers. Moreover, it preserves the same ecological principle, i.e. looking into existing realms of high-stakes exam preparation, which, as we assume, varies in terms of intensity within the normally distributed

cohorts of participants. Therefore, it can be also stated that the initial plan (idea) has been fulfilled as the suggested study cohorts represent the following exposure to the intelligent tutoring system:
1) ‘high intensity group’ (5 or more study attempts within the ITS)
2) ‘moderate intensity group’ (3-4 attempts)
3) ‘low-intensity group’ (1-2 attempts)

The introduced coinings clearly indicate the learning involvement ratio for the population in question. To the naked eye, all these cohorts can be determined as a random distribution. In fact, the grouping has been made following the same ecological principle with the help of empirical observation on users’ online learning behaviour within the studied online environment. Approximately 80% of all registered users have approached it, occasionally not exceeding the 2 attempts’ limit. On the opposite side of the population there is a small cohort which has not more than 3% of test takers. Such extremes and the in-between cohort apparently define the whole population based on learning scope. The apparent shift towards collecting data only from the training system itself has contributed greatly to the study feasibility.

This study is to look at these cohorts of the population in order to get insights on students’ learning performance, and, also, highlight how the variable study paths correlate with the high-stakes exam preparation in general.

4.2 Research questions

The questions of the research are the following:
Do multiple attempts affect the students’ learning performance in preparation for the Speaking part of the Russian high-stakes exam in the English language?
Does the study attempt frequency within the proposed intelligent tutoring system affect the overall students’ learning performance in preparation for the Speaking part of the Russian high-stakes exam in the English language?

The questions have an ethical issue as it shades the teacher’s instruction mode. Virtually, it appears that the teacher’s impact has been underestimated in the research in question. Nevertheless, the author strongly advocates for treating mastery learning principle as the

essential foundation of the learning environment where a student is always supported on the metacognitive level. And the teacher is instrumental in this process, which is usually a missing feature in various tutoring systems. Therefore, the research questions do not underestimate the necessity of teachers' instructional, motivational and metacognitive support.

4.3 Methodology background

The following paper is to inform further high-stakes testing research in the field of Intelligent tutoring system (ITS) by providing theoretical frameworks and suggesting some plausible techniques for implementing the aforementioned type of research in self-regulated learning environments aimed at developing speaking proficiency. The research in question is naturally limited in the scope of potential research designs due to its non-human nature thus endorsing the data analysis mechanisms that rely on tracking students' learning behaviour through their digital imprints. In other words, the collected datasets frame the overall design of the research.

The chosen design is a derivative from the interaction-based research framework (Mackey & Gass, 2005). Despite being not quite fitting the purpose due to the markedly humanistic term, the suggested methodology, rooted in manipulating learners' interactions and the received feedback in order to determine the correlation between components of interaction and language proficiency, is in line with the research questions and the initial idea of tracing the feedback from the ITS in question. The deviation from the authentic framework is in the interpretation of the 'interaction' as in the current study it is plausible to assume that learners collaborate with the ITS in a kind of dialogue mode resulting in providing instructional feedback.

The design of the thesis follows a computer-facilitated subtype of the interaction-based research framework that was quoted earlier. The subtype is defined as computer-mediated communication (CMC) and it is 'a text-based medium that may amplify opportunities for students to pay attention to linguistic form as well as providing a less stressful environment for second language practice and production' (Mackey & Gass, 2005). All the mentioned CMC's peculiarities correlate with the EGEnglish.ru in-built features.

The rationale for using quantitative methods partly stems from the nature of the data collection proposed for the research as well as a low reliability of qualitative methods in case of their implementation for the ICT in question (EGEnglish.ru): content analysis is limited due to its' users variable usage of the computer-generated feedback responses (written scripts) which are incomparable for the whole population and acquired samples. The other, and probably the most important, reason for applying a quantitative method is the choice of the instrument. Discourse completion task (DCT) has been chosen as an appropriate tool matching the research ideology. DCTs are widely used to establish the pragmatic features of a specific interlanguage functioning by manipulating large quantities of data through which it is possible to generate a significantly large corpora of comparable, varied speech acts (Ogiermann, 2018). Although the DCT is mainly associated with the conversational mode involving learners engagement in communicative exchanges (Mackey & Gass, 2005), it seems reasonable to treat the ITS users as human agents responding to computer-generated study stimuli, i.e. the ITS tasks replicating the EGE format tasks. Although DCT responses are claimed as being different from real-life language performance, they do represent "a participant's accumulated experience within a given setting" (Golato, 2003). And, in case of investigating the ITS in question, this inference can be supported by the fact that the EGE tasks are 'artificial' by default. Therefore, the retrieving mechanism through the ITS framework almost fully replicates real-life exam settings eliminating constraints which potentially put at risk authenticity and pragmatic value of the DCTs in use. On a practical level, DCT is viewed as the task type which can be easily distributed to considerably large groups of research participants within a short period of time thus making it a sophisticated instrument for the contrastive study of speech acts (Aston, 1995; Barron, 2003).

In the research, the elicitation of students' speech acts was done through their collection in response to DCTs tasks aligned with the analysed high-stakes exam tasks №3 and №4. The tasks' completion was performed in an asynchronous way within the ITS online learning environment. The retrieval of audio files, representing students' responses to DCT tasks, was performed by the author of the research through accessing the EGEnglish.ru databases in 2021-2022.

4.4 Sampling

The initial study of the participants' population started from identifying the audio recordings representing the ITS users' attempts to perform task №3 and task №4. The starting point for the procedure was May of 2017, the month marking an open-to-public release of EGEnglish.ru after completing an aplha and beta testing periods. The end point was July of 2021, the final study period with the initial Speaking task framework operation. The following month the EGE Speaking tasks were authorized for being recalibrated eliminating further opportunities for using up-to-date study settings for retrieving comparable data. Data collection procedure was being performed via MobaXterm software in a manual mode throughout the 2021-2022 study year. For search purposes, the internal EGEnglish.ru coding system was used. The following codes were tracked down in the list of recordings: 2196, 2215, 2823, 2888, 2925, 2929. They have been automatically assigned to the corresponding Task 3 or Task 4 options by the client-server backend application. Thus, it is possible to assume with very high probability that each of these codes had highlighted the affiliation of the responses to both tasks unless research participant had deliberately performed some different tasks, which are not corresponded to the task code (on the grouping and the assessment stages it has been put under the spotlight in order to secure matching between the tasks and the title codes). Also, each title (token) of recordings contains a personal login of the user or the automatically assigned one showing the time of task execution.

The search was being conducted by the following scheme:
1) chronological detection of the recordings bearing the tracked codes of Task 3 and Task 4.
2) identifying the 'repeated entities' (recordings) by matching the parts of the file titles with users' logins.
3) storing and further downloading of the recordings database containing only the studied population of the research falling into the sampling criteria.

The 'repeated entity' notion has to be clarified from the very outset. The overall design of the research is aimed at tracking the initial contact with the ITS (pre-test) and the final iteration (post-test), although such points of intersection might not be firmly regarded as the first and the last study attempts of users. The EGEnglish.ru backend system does store all the key tokens of the user-system interaction, but it is possible to assume that the users might have

entered the ITS before creating a personal student profile. Conversely, the user could have had some study attempts after logging out of the system.

Despite acquiring a majority in the significance survey, teachers appear to be fairly reluctant in implementing cutting-edge online tools due to various reasons. Partly, this suspicious outlook was reasonably motivated by teachers' complaints on the Internet speed, which seems to be quite common for teachers approaching new-computer based environments (Kay, 2009). In addition, there was one more bias during the data collection stage. A number of recordings with different voices have been detected. As long as the voices were giving away the task's guidance, it is reasonable to assume that they belonged to teachers or private tutors. Therefore, the initial idea of building up experimental groups with the teacher's direct involvement was discarded. Also, this decision affected the sampling procedure as such recordings were not taken into account. Luckily, in the collected samples there were no such recordings with multiple speakers within performing tasks.

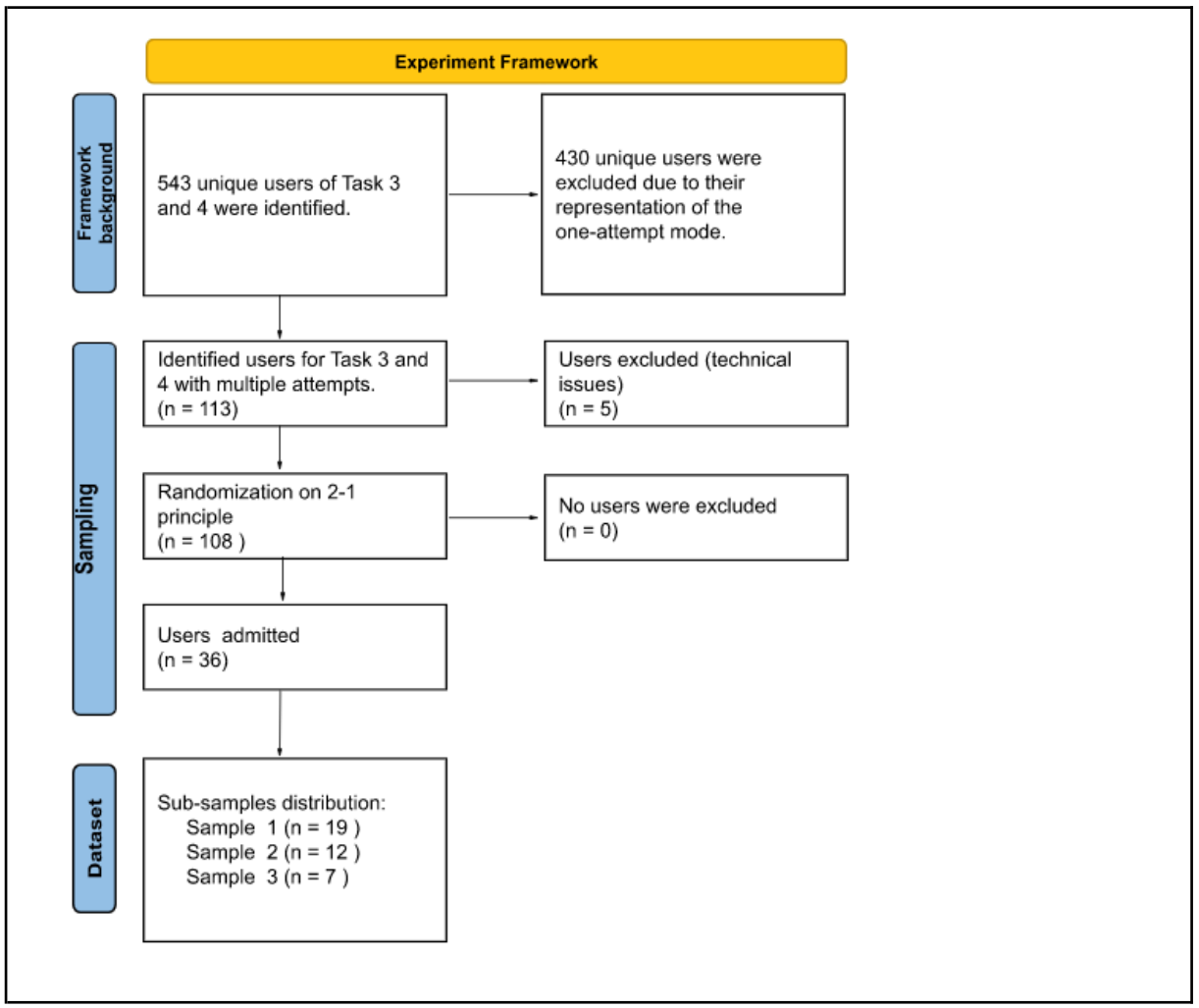

Figure 8. Diagram with stages within the sampling procedure

The above given diagram depicts the complete sampling procedure including the experiment background with sorting out raw data. As a result, 113 users were allocated, each of which producing study attempts within the range of 2 to 8. The task 3 sub-set included 68 user tokens containing a title string with the following tags: login (user name) – task code – date of recording. The task 4 sub-set with the same tag structure incorporated 45 user tokens. The accumulated data was then put under the timing analysis aiming at detecting the initial and the final attempts dates. This was done solely for marking the pairs of recordings which would be extracted for randomization and assessment.

The collected bank of binary user tokens consists of 113 user names which has undergone a simultaneous process of grouping and randomization. The latter was executed for the sake of research feasibility and, as mentioned before, for keeping the unsystematic variation to a minimum. However, before implementing randomization the author of the research has applied a grouping mechanism based on frequency parameter which, in its turn, was decomposed from the authentic distribution of the recordings. The biggest group, compiled of the ITS users who have performed 2 study attempts for Task 3 and Task 4, consists of 62 user tokens. The second group in size, representing the users with 3 or 4 performances, was made up of 37 users. The smallest group has accumulated only those who have performed the tasks 5 or more times. The total number of such people was 14 students.

The following titles have been coined for the groups: low frequency group (not more than 2 study attempts), mid frequency group (ranging from 3 to 4 attempts), high frequency group (5 and more attempts). The first 2 groups were taken for sampling without any preliminary classification apart from the above mentioned frequency benchmark. The only exception was the high frequency group which for randomization purposes can be split into 2 sub-groups. The first of which contains 10 recordings representing Task 3 study attempts. The second sub-group stands for Task 4 performances which accounted for only 4 study attempts. Therefore, the second sub-group couldn't be processed through a simple randomization procedure as with such a number there was no practical sense in singling out the only recording from the cohort.

Bearing in mind the following feature of the sample, the next step presupposed splitting of the would-be randomized entities from the exceptional cohort. Thus, the following sub-groups were formed making the 109-4 ratio (division?). Despite placing no initial requirements for

the study groups, prior to executing a randomization procedure the sample had to be carefully calibrated for eliminating bias related to the technical features of test takers' recordings. First, assessors, who would be recruited for the sample evaluation based on the EGE criteria, were asked to give an empirical rough approximation to the minimum length of a 'meaningful' recording which should, to some extent, guarantee acquiring at least 1 point for the communicative task benchmark. The suggested length of 20 seconds, in practice, results in an audio recording file of roughly 1 mb. Given the following limit, all the files from the initial sample were put under the scrutiny. As a result, 4 recordings from the randomized cohort were discarded, and only 1 was removed from the non-randomized cohort. The first 2 were removed due to breaching the minimum band limit for the recordings. One another recording has been excluded from the sample due to failing the sound test. Also, one more audio attempt has to be discarded because of the research design aiming at processing a succession of the initial performance (pre-test condition or in-test condition) and the final presentation (post-test condition). By collecting 105 (+3 for the non-randomized cohort of high-frequency Task 4 performances) recordings, the technical calibration came to its end. However, there is one more preliminary procedure that has to be conducted as the length of the recordings might be a good predictor to a recording's eligibility, but, certainly, it might be a misleading feature as long as there is always a chance of collecting a silent recording, which falls into the predetermined limitations. Therefore, before the assessment took place all the recordings from the sample had been checked for detecting 'death silence' cases. Finally, no such recordings have been detected.

The acquired cohorts of the users' population were randomized by 2-1 principle: every 3rd student was chosen for the experiment. The randomized sample with a non-randomized chunk was compiled accounting for 38 user tokens which bear the information about the user profile, task references, and timing of the study attempts. There is one important consideration that is not directly connected to the randomization process, but it does have an impact on the visualization of the data. For testing reliability of the recruited assessors proficiency the following 'trick' was implemented: 3 user tokens (6 recordings) were randomly added to the list in order to have a complete list of the pre-research and research data sets. In detail, the implemented procedure will be elaborated on in the 'Assessment' section.

Implementing randomization seems to be a crucial element of the repeated measure design of the research. Random assigning of the users to the sample mitigates potential systematic

variation of the study. In fact, the whole population was having training sessions at different parts of the study year prior to taking a real-life examination. Therefore, it is reasonable to assume that their level of familiarity with the exam tasks was varying from low awareness in the beginning of the study year to a considerably high level in the final stage of the exam preparation. Although it seemed impossible to eliminate such a research threat with the training workload surrounding mostly the final month preparation, it was still feasible to cross out 'successive' users by applying a 2-on-1 simple randomization technique. It was done on a chronological basis when each 3rd recording was being extracted for the sample. This approach has allowed to acquire a more stable sample bearing general characteristics of the population thus making the user cohort more dispersed in terms of its timing distribution within the study period.

As a result of execution of the mentioned procedures, the list of anonymized entities representing user tokens have been compiled (see Appendix C). Each user's code refers to the pre-test and post-test recordings, attempts being undertaken, and task affiliation. For delivering recordings to assessors the list was first anonymized and then regrouped by the recordings' numbers through an automated randomizing tool at random.org (https://www.random.org/sequences/min=1&max=82&col=1&format=html&rnd=new). The 82 recordings representing 38 users from the sample and 3 test users for assessment validation were coded by the suggested numerical codes. This procedure has finalized sampling and initial data representation.

4.5 Assessment procedure

As long as only Task 3 and Task 4 are included in the experiment, these tasks are the only to be desribed in terms of their assessment criteria. In Task 3 the test-taker has to describe a personally chosen picture out of a set of 3 images. The maximum score is 7 points. The following criteria are used: Coherence and Cohesion, Grammatical correctness.
In Task 4 a test-taker has to compare and describe two given pictures. The maximum score is 7 points. The criteria for Task 3 are applied to Task 4, although sub-criteria for one more criterium vary. This differential feature is elaborated on below.

Apart from the shared criteria for both tasks, there is one another, which might be considered a 'supercriterion' due to its overall value to the score of each task. The arbitrary translation

for it can be ‘Description accuracy’. The following criterion has sub-criteria reflecting the very idea of each task. In case of Task 3 emphasis in put on describing the action features of the picture (place, participants, action), while for Task 4 the most important part is a comparative feature (similarities and differences between the given pictures). In case of failing to fulfill at least 3 subcriteria out of 5 a test-taker must be given a zero score for the criterion and the overall score must be also 0. The other criteria, under such circumstances, are not taken into account.

To secure the research reliability, a stepwise assessment phase was implemented before starting the main procedure. This was done in order to test mimicking the real-life procedure at a small scale. The most important reason for this move is that even in real EGE settings an assessment team is comprised of 3 individuals, 2 main ones and 1 supervisor evaluator. The same cast was hired for the research in question. In this stepwise phase, a recruited supervisor served as a quality control tester. This person, being an experienced EGE supervisor, assessed these 6 test recordings (pre-test and post-test entities belonging to Tester 12, Tester 36, Tester 38) from the pack in order to set a probation benchmark — these 6 recordings were coded with no reference to their belonging to either a pre-test or post-test attempt thus the real-life conditions have been fully replicated. After that, the main assessors were exposed to the sample database and asked to deliver the first stage of assessment in which both research recordings and test entities were included. The following mechanism was incorporated for hiding the test phase within the research procedure. As a result, the team of main assessors could not pay special attention to the test entities thus maintaining the best possible level of assessment with no biased evaluation scenarios. Eventually, the received marks were analyzed in order to see a correlation with the evaluator’s assessment: no strong deviations (higher than 2 points) have been detected thus securing a high probability of close-to-real-life evaluation in the research.

The assessment criteria are taken without any modifications from the official EGE assessment methodology. All the students' recordings were stored on the servers of EGEnglish.ru technical support and were delivered to the chosen certified examiners who are involved in the EGE official testing period.

The summoned team of 2 assessors blindly assessed the recordings and the average score was taken into account unless the given grades (scores) were in fluctuation of more than 2 points.

In such cases a third examiner (supervisor) was invited to assess the ‘problematic’ recording and their score was considered for the experiment. The same procedure was also applied to the cases when a test taker received a zero score from either of the initial examiners.

4.6 Experiment description

It is assumed from the very outset that students have entered the EGEnglish.ru using two specific pathways: searching independently or through teachers’ instruction. Although it is obvious to note that both cohorts have peculiar characteristics and were really diverse in terms of varying preparation levels to the exam in question, the research focus rests in the intelligent tutoring system application to the exam preparation and its short-term effects on the learning process. Therefore, teachers’ intervention is not considered as a variable due to the above mentioned reason and an obvious observation from the real-life exam conditions – all high-stakes exam students receive pre-exam instruction from either school teachers or private tutors on a regular basis.

The experiment was set up in conventional educational settings without any intervention from the EGEnglish.ru team into the training process apart from providing continuous technical support to teachers and students who were contacting the team. Two stages were suggested as sophisticated framework for collecting the data:
1) Formative stage (pre-test)
2) Evaluative stage (post-test)

The formative stage was not bound to the specific timing as its main goal was to monitor the initial students’ level of competence. Thus, the formative stage covers only the very first training attempt. Conversely, the evaluative stage comprised a further period during which the participants were performing all successive students’ training attempts. On the evaluative stage the last performed attempt was extracted for assessment. The data from all the groups was collected through EGEnglish.ru recording tools and then extracted in an audio form in order to be sent for experts’ assessment.

4.7 Data storage and confidentiality

The data is stored on the EGEnglish.ru servers under encrypted passwords with no access to any individuals except for the author of the research. To ensure confidentiality, each of the

participants was assigned a code name (such as Tester1) before delivering the recordings for assessment. As long as no private information was stored in the personal accounts of the research participants, it was decided not request any permission regarding research involvement. Although during the registration process at EGEnglish.ru all the new users were notified that their recordings could be used for research purposes. The sample was summoned by the users of the ITS users who had been exposed to the aforementioned notification as well as the unregistered users (without any credentials stored in the ITS personal account space) who had been exposed to this notification presented at each web-page containing EGE tasks.

Chapter 5. Data Analysis

5.1 Normality testing

First and foremost, the assessed data collection was stocked in an appropriate xsl format for further manipulation. All the assessed recordings (n=76) are grouped in the table containing all the basic credentials as well as scores and deviations of the sample participants (see Appendix D).

After receiving assessors' scores it was reasonable to start studying the sample in order to establish its basic statistical characteristics, which are instrumental in identifying the sample's match to the population in question. By calculating the mean = 3.42 and standard deviation = 2.20 it has become manageable to see a central tendency value. Additionally, processing skewness and kurtosis methods have allowed to align the sample to normal distribution qualities. Despite being quite like bimodal in the histogram, the sample skewness, after being processed through the AI-Therapy calculator (https://www.ai-therapy.com/psychology-statistics/), is suggested as a similar to normal distribution with skewness of -0.48 with a standard error of .27. Achieving a p-value of .08 (z-value is -1.74) made it possible to assume the claim on normality to be relatively significant, although a suggested value greater than 1.96 for significance at $p < .05$ (Field, 2013) has not been maintained. Nevertheless, kurtosis benchmarks has rejected a normality inference stating that population excess kurtosis (unbiased) = -1.021, with a standard error of .545 (z-value is -3.31 and p-value < .001). Below all presented calculations are processed through AI-Therapy calculator unless the other measurement tool is mentioned.

The aforementioned binary outcome is not quite surprising given the conditions of the research design. The left-side cluster concentrated around 0 score manifests the very idea of the learning intervention: quite a few of the ITS users have entered the pre-test phase with a quite low awareness of Task 3 and Task 4 performance in a computer-assisted environment simulating the real-life exam conditions. Therefore, the close-to-normal distribution visually defines the proposed assumption and clearly summarizes scores' variance within the sample (See Figure 9).

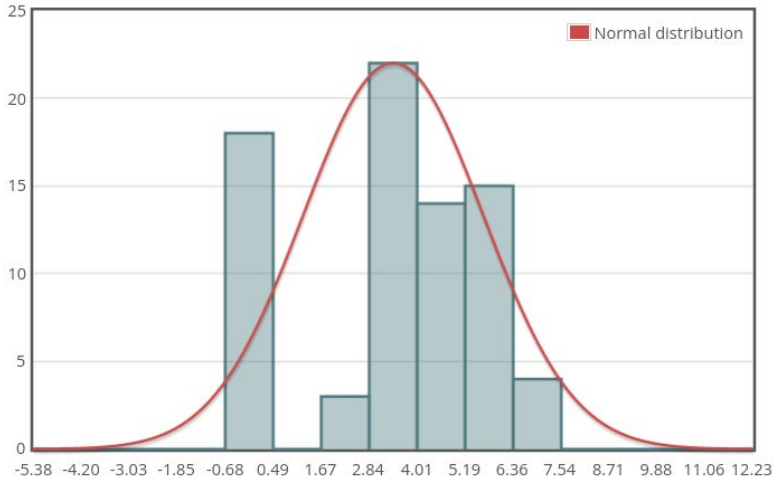

Figure 9. Normal distribution test for the whole sample

The overall visual statistics shown at the first histogram (Figure 9) precisely depicts a broad bigger picture. However, it seems to be fruitful to study a final chunk of data by applying the same central tendency approach. For the sake the post-test group is to be processed through the identical measurement tools. In this sub-test, measuring skewness and kurtosis enables us to narrow down the focus and conceptual field of the research, i.e. the post-test benchmark reflecting the scope of the ITS interference.

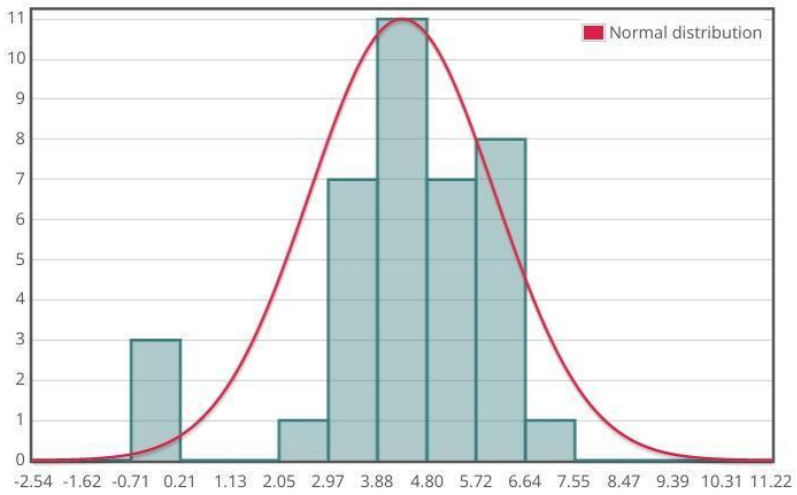

Figure 10. Normal distribution test (post-test)

The above given visual imprint of the post-test data reveals securing peakedness in the left side of the chart (see Figure 10). Although z-value of 1.52 and p-value of .12 represents a smaller probability and diminished significance in comparison to the overall dataset, population excess kurtosis (unbiased) = 1.23 with a standard error of .75 stands for a similar to normal distribution status. The figures for skewness have not claimed normality

(population skewness (unbiased) = -1.08, with a standard error of 0.38). The following bipolarity in statistical representations cannot and should not hide an important correlation, which does not seem to be evident as it requires manipulation with some retrospective data for the EGE Speaking part.

5.2.Descriptive data

In the first section of the research a few charts have been outlined and commented on in order to put the EGE Speaking downward trend under the spotlight. Apart from 2015 figures the EGE Speaking mean for the whole exam population fell in the success range of 0.6-0.7. With the mean for the post-test phase (all 3 Groups) equaling 4.39 within the 0-7 scope it is not difficult to notice that the mean in the percentage form, i.e. .63, correlates with the above mentioned band. Thus, it seems plausible to assume that the research sample can be considered a match to the whole EGE population in terms of the high-stakes exam performance. It might not be an obvious thought at first glance as exam conditions vary greatly, but a more thorough outlook at the research design tells the opposite. The idea is to be elaborated on in the Results and Implications section as well as in the Discussion section. The further investigation of the data is to take its turn into the basic sample grouping, which advocates for the research hypothesis. The first group to be analyzed ought to be the biggest one in which participants opted for 2 two study attempts. The calculated scores of the group 1 (n=19), which is titled after its core peculiarity, is given below (See Figure 11).

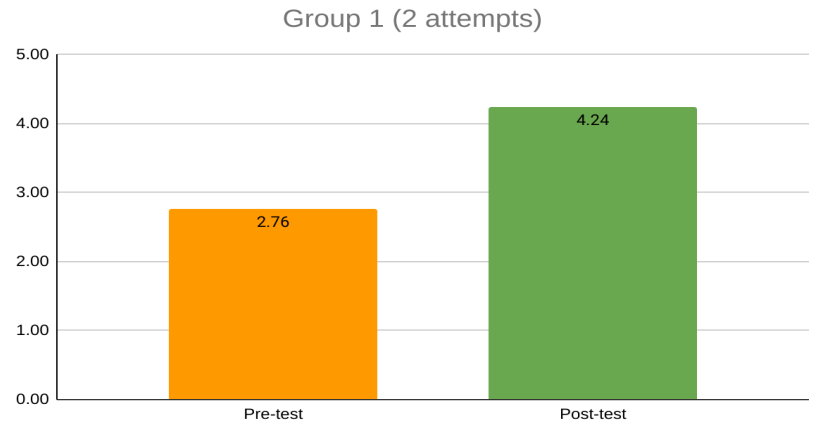

Figure 11. Comparison of Groups 1 means

The mean for the pre-test of 2.76 (out of 7 maximum) maximum) transforms into 4.24 at the post-test intervention. The mean comparison showcasing quite a moderate ascent of study performance is not the best corresponding benchmark as it narrows down would-be interpretations due to its one-dimension nature.

Instead, it seems to be essential to study the Group 1 score distribution referring to the other 2 central tendency features, median and mode, in order to have a broader look on the changes within the group (Figure 12). The band scores are grouped for pinpointing the failing mark (0) and the 'central' range (3-5).

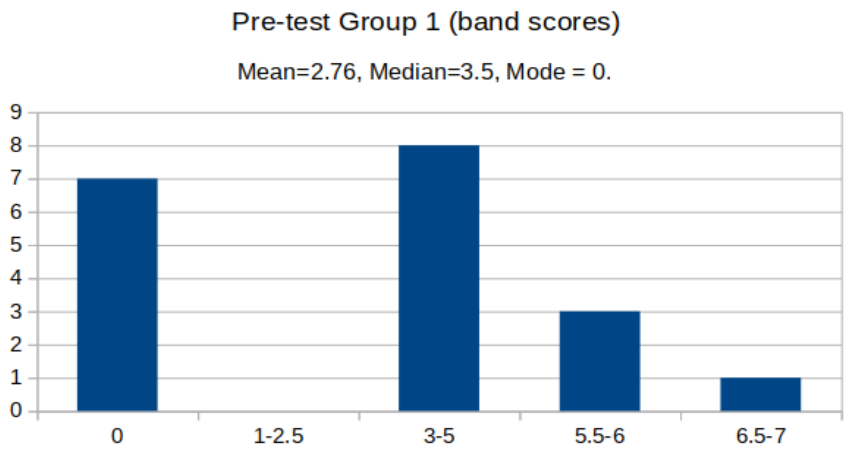

Figure 12. Central tendency parameters for Group 1 (pre-test).

The bar chart highlights both parameters which seem to be far more exposing for Group 1. The bimodal look of the descriptor contradicts with the mode that stands for 0 (count 7). Such a low mode contributes to a considerably high median parameter of 3.5 for the sample. Both benchmarks pinpoint the very nature of the pre-test for Group 1: a big part of the users have entered the ITS with low awareness of the exam in question. Nevertheless, the static position of the pre-test score can only lead up to some topical inferences.To obtain a full picture, it is necessary to study the parameters at the post-test phase (See Figure 13).

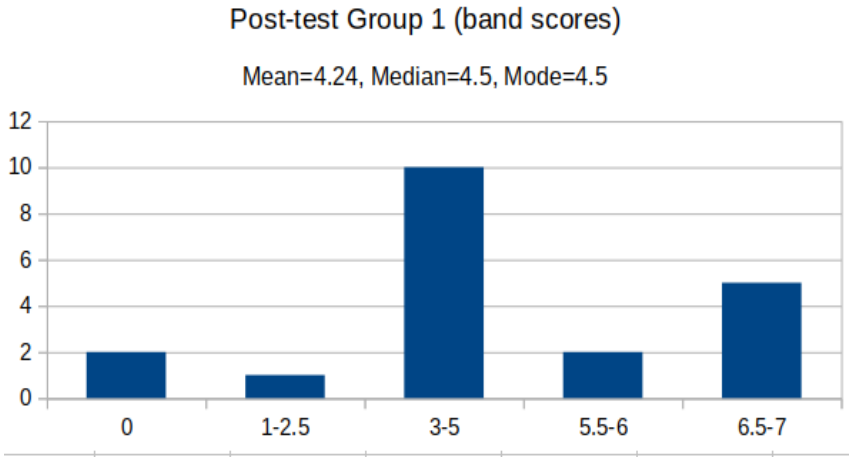

Figure 13. Central tendency parameters for Group 1 (post-test).

Undoubtedly, the power of descriptive statistics is visually maintained in the post-test chart. The corresponding to each other (count = 5) mode and median hitting 4.5 mark have clearly depicted the switch of values within this short-term stint of training. The majority of Group 1 participants have managed to fulfill the task successfully by attempting it only for the second time - the attempt that cannot be looked down upon and taken for granted. The decision of the group to have finalized the EGE speaking part training is supported by the quite good study performances within the tasks.

Thanks to the descriptive statistics and the central tendency parameters it has become feasible to see the study dynamics within Group 1. Although it seems to be of value, such a cramped field of investigation cannot fully help in generalizing on a level of the whole sample, which seems to be a more desired scope of every single research.

However, such generalizations do need more sophisticated tools, which will be applied throughout the whole data range further in the text. But before doing so, the other two groups should be represented from the same standpoint. Below there is a chart for Group 2 (See Figure 14).

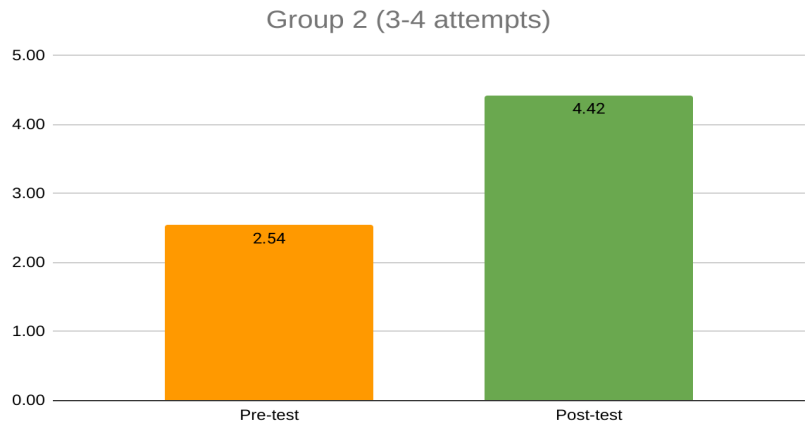

Figure 14. Comparison of Group 2 means.

Visually, the bar chart cannot be considered as a dramatic paradigm shift. However, there is a clear indication that the study performance deviation is expanding: the indicators have moved into opposite directions. In comparison to Group 1, the pre-test outcome is at

a lower point (2.76 >2.54) while the post-test outcome (4.24 > 4.42) has increased.

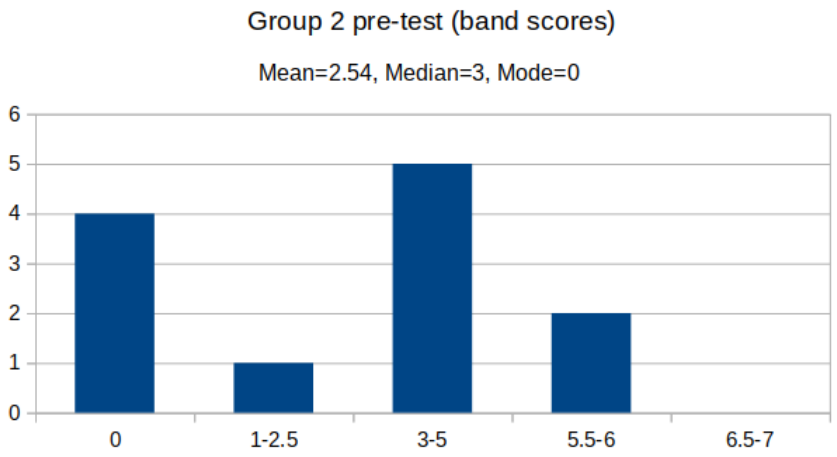

Figure 15. Central tendency parameters for Group 2 (pre-test).

With a mean of 2.54, which is slightly lower than the mean for Group 1 (2.76), it might seem tempting to have claimed a nearly identical general study pattern with a start on a very low average level of the tasks awareness. The mode figure of 0 (count = 4) also resembles Group 1 distribution (See Figure 15). Moreover, the median benchmark standing at 3.0 has solidified the assumption according to which the pre-test phase for both groups (Group 1 and Group 2) can be regarded as low-level mastery in terms of fulfilling Task 3 and Task 4 of the EGE Speaking part.

The post-test figures for Groups 2 are to be differential for making far-reaching conclusions on the group dynamics. However, even descriptive analysis might be revealing for this sake as it is easier to calibrate the scope of the sample and the most vivid deviations from the mean. The bar chart view enables the research author to claim that the post-test phase for Group 2 features a peaked cluster closer to a top-level performance band. Despite being identical on mode (4.5) and median (4.5) dynamics in comparison to Group 1 on the post-test level, it is worth noting that the post-test mean = 4.42 overlaps the pre-test mean by about 0.2, although the starting point (pre-test phase) had conversely favored the Group 1 representatives. Therefore, it can be assumed that the noticed differential rooted in study attempts quantity in terms of the task performance takes a clear upward direction within the analyzed sub-samples (Group 1 and Group 2).

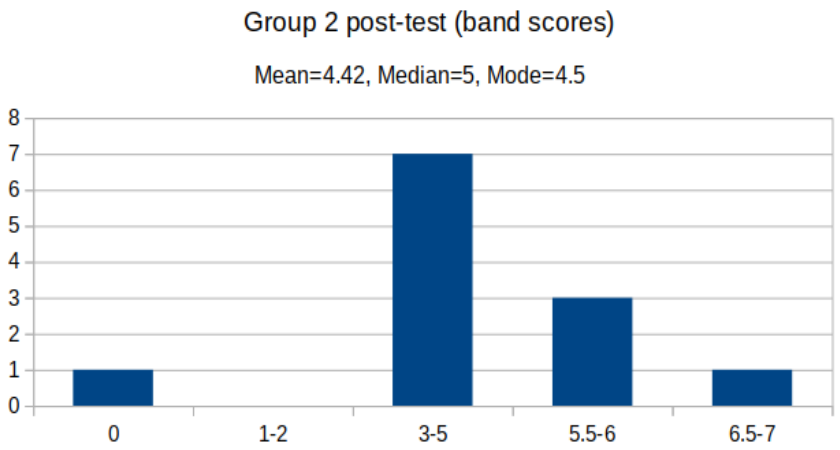

Figure 16. Central tendency parameters for Group 2 (post-test).

The group 3 outlook visually supports the recent claim about the marked difference that could be triggered by the quantity of study attempts: there is a huge span between the pre-test indicator of 1.79 and the post-test outcome of 4.5. The previously made claim on a considerably low level of study performance for Group 1 and Group 2 can be dissolved as the level of Group 3 awareness is substantially lower (See Figure 17).

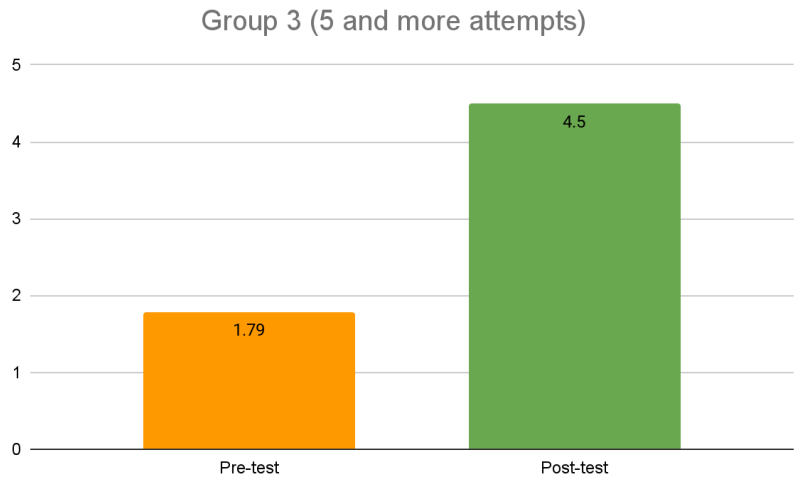

Figure 17. Comparison of Groups 3 means.

The visual disproportion shown on the bar chart above is best explained by the other two central tendency figures depicted and elaborated on further in the text. Nevertheless, it is still reasonable to claim that the students' cohort have demonstrated quite different study outcomes in the pre- and post-test phases in comparison to Group 1 and Group 2 test-takers.

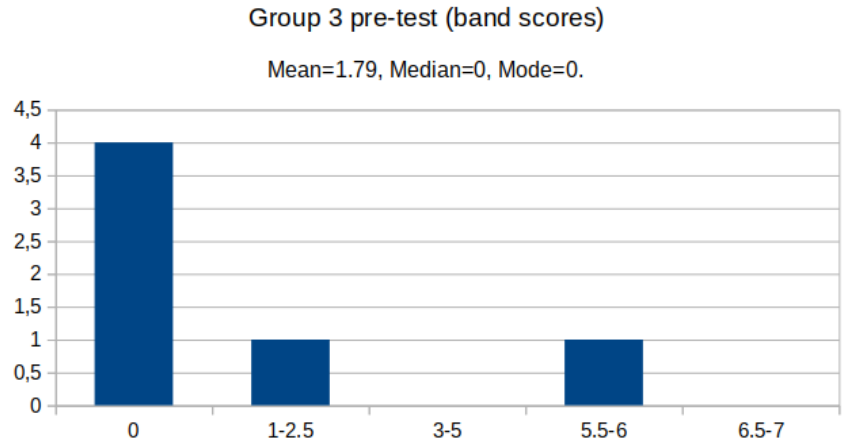

Figure 18. Central tendency parameters for Group 3 (pre-test).

The mode and the median for Group 3 coincide at 0 making it the least successful group type (See Figure 18). So huge diversity in participants' scores can be explained by both subjective and objective factors. The size of the group and its semi-randomized status impact the sample distribution. On the other hand, such a mixed cohort is manifested by the very different participants' profiles including low-achievers and high-achievers, who both possess a shared feature — an intent to practise in an intensive fashion. Thus, it is feasible to treat representatives of this group as learning seekers in comparison to the rest of the sample who have entered the ITS as a testing tool. Also, it should be mentioned that the increased mean for both pre-test and post-test study outcomes cannot be justified only by the low-base effect, which raises questions about the mode and median distribution on the post-test phase.

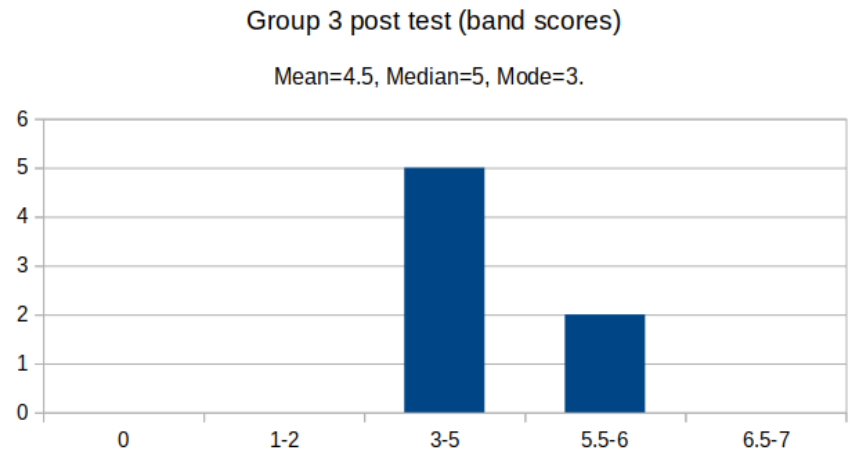

Figure 19. Central tendency parameters for Group 3 (post-test).

Visually, the dissemination of study performances for the post-test phase shows off its bimodal look (See Figure 19). Additionally, the median hitting 5.0 defines a high-achieving cluster within the group.

5.3 Significance testing

Although the above described data clearly depicts a divergent trend of study performances dependent on increasing a number of study attempts, there is one really important parameter which is especially worth considering for the sake of understanding differences between the groups in question. It is confidence intervals (CI) for the mean. Having agreed to the considerations stating that CIs are not suggested to be used for a repeated measure research (Kalinowski, 2010) and the repeated measure design generally have greater statistical power (Brauer, 2018), the following confidence intervals analysis might be used as an additional element of descriptive statistics for the whole range of group accomplishments as well as a 'significance descriptor' for Group 3 study performances. Having calculated 95% confidence intervals for all the groups of the research via AI-Therapy online calculator, the following table has been issued and piled up for further analysis. The data obtained throughout the manipulation included the previously used parameters (mean, mode, median, sample size) as well as the standard error of the means and 95% confidence intervals presented below (see Table 1).

Table 1. Confidence intervals data for Groups 1-3

	Pre-test	Post-test
Group 1	[1.637, 3.889]	[3.306, 5.168]
Group 2	[1.184, 3.899]	[3.309, 5.524]
Group 3	[-0.509, 4.080]	[3.399, 5.601]

Assuming the idea of CI being 'an estimate of plausible values for the population mean' (Kalinowski, 2010), it is reasonable to see some regularities in CI data in order to foresee the population mean and its significance value. To the naked eye, it is obvious that Group 3 figures at a pre-test phase are in discordance with the rest of the data. On a practical level, the negative low endpoint can be replaced by zero without no effect on the confidence level (Stark, 1997). Therefore, we infer that zero can be the lowest possible study performance score for the population in question. However, this obvious inference cannot fully justify what such scattered distribution of the Group 3 pre-test might mean for the whole research.
In a case of the chosen research design based on repeated measure methodology a CI difference between the means within two samples has been suggested (De Muth, 2019). The

very idea of the method implies comparing two CI intervals and drawing a conclusion on whether the two samples have statistically significant differences. By following the methodology and applying a calculator for CI between the means (Georgiev, 2017), a quite unexpected conclusion has been drawn upon (key figures on the case are given below). It occurred that in Group 3 pre-test and post-test sub-samples there was a statistically significant difference in the means as the null (zero) was not crossed within the 95% confidence interval. Although the following grouping hasn't plainly replicated the ones of De Muth's research focusing on sample-population comparison and two independent groups comparison, it seems reasonable to assume that both sub-groups, the pre-test and the post-test ones, might be considered comparable in terms of their practical values, i.e. study performances. Given that Group 3 contains non-randomized entities (Task 4 scores), this statistical implication makes it quite justifiable to have considered all the groups together within the sample.

Table 2. Confidence intervals data for Group 3

Difference (B-A)	2.714286
95% Confidence Interval	[0.6759 , 4.7527]
Value ± 95% SE	2.7143 ±2.038
Mean A	1.785714
Mean B	4.50

At this point, returning to the analysis of Table 1 opens up a path to investigate confidence intervals more broadly in order to discover these long-awaiting regularities. According to the data in the table, post-test subgroups are all clustered quite narrowly within a calculated range of 1.862 - 2.202 (with the minimum and maximum endpoints of 3.306 — 5.601 at the table). Therefore, it is plausible to assume that the subgroups are quite homogeneous having no vast scattering. However, understanding of the following regularity might be done based on the means comparison without applying confidence interval figures. What has to be done with the parameter is implementing its function aiming at finding correspondence with the population in question. Luckily, the obtained statistics about EGE Speaking part performance on a national level allows the procedure to be conducted. Thus, if we pile together the population means for 2015, 2017, and 2018 exam years in the 0-7 range of the exam, the following digits are to be taken into account: 4.97, 4.76, 4.62. If the figures are regarded as true means, then all the detected confidence intervals include these means. Moreover, the sample means of 4.34 might be treated as a corresponding figure showcasing the sample's relevance to the

population. The specifics of such figures is elaborated in the Discussion as their value might affect the overall interpretation of the research (Task 3 and Task 4 maximum scores totalling 14 points contribute to 70% of the Speaking part, but there is still 30% remaining).

As for the pre-test conditions, the determined CIs represent widely scattered entities with an expanding range dependent on the number of study attempts being undertaken. On the opposite side of the interval scale there are nearly identical endpoints (3.889, 3.899, 4.080), which also provoke thoughts on identifying one more correlation pattern in the sample. In the table below a set of central tendency measurements have been added by standard deviation (SD) benchmark in order to study possible correlations within the main descriptive statistics parameters (see Table 3).

Table 3. Central tendency figures for the sub-groups of the sample

	Group 1 pre	Group 1 post	Group 2 pre	Group 2 post	Group 3 pre	Group 3 post
mean	2.76	4.23	2.54	4.41	1.78	4.5
standard deviation	2.33	1.93	2.13	1.74	2.48	1.19
median	3.5	4.5	3	4.5	0	5
mode	0	4.5	0	4.5	0	3
sample size	19	19	12	12	7	7

Although SD figures for the pre-test subgroups are in fluctuation with the biggest dispersion in the Group 3 condition, there might be possible to detect a correlation pattern: the downward trend for the mean is accompanied by the similar median tendency. On a practical level, this correlation accounts for varied homogeneity for pre-test subgroups. In other words, students' performance varies more significantly for the pre-test Group 3 than for those of Group 1 and Group 2. Nevertheless, the very similar right endpoints for CI are ‘contributions’ of the peaks - in all three subgroups maximum study performances belong to the highest cluster on the scale and truly represent close-to-maximum EGE Speaking part score such as 5.5 (for Group 1), 6 (for Group 2), and 6.5 (for Group 3).

The aforementioned statistics features highlight a positive trend in study performance for all groups of research participants. Judging by the means description, the most significant study

performance effects have been detected within Group 2 and Group 3. Also, the applied methods have contributed to defining the sample's close-to-normal distribution accounting for the specifics of the experiment: a significant part of the sample representatives have entered the study environment with limited mastery in the research tasks (№ 3 and №4).

Despite having quite diverse data for the experiment, a sample splitting can secure significance testing on all the studied cohorts as if they are treated independently. This approach seems to be of value as it can clearly articulate stronger and weaker points of significance within the sample. In order to see the following statistical outcomes, it is necessary to study the groups separately. Given the size of them, it is predictable to have used non-parametric methods with one-tail design as it is implied that improvement is the expected effect of the EGEnglish.ru. Using the AI-therapy calculator in such data settings is limited to the use of Wilcoxon signed-rank test. For all following analyses the significance level is .05. The analysis (https://www.ai-therapy.com/psychology-statistics/results/20230502054836568) of Group 1 (n=19) states a statistically significant difference between pre-test and post-test conditions (see Figure 20). The calculated effect size reaches the level of .41 securing a moderate proportion of variance.

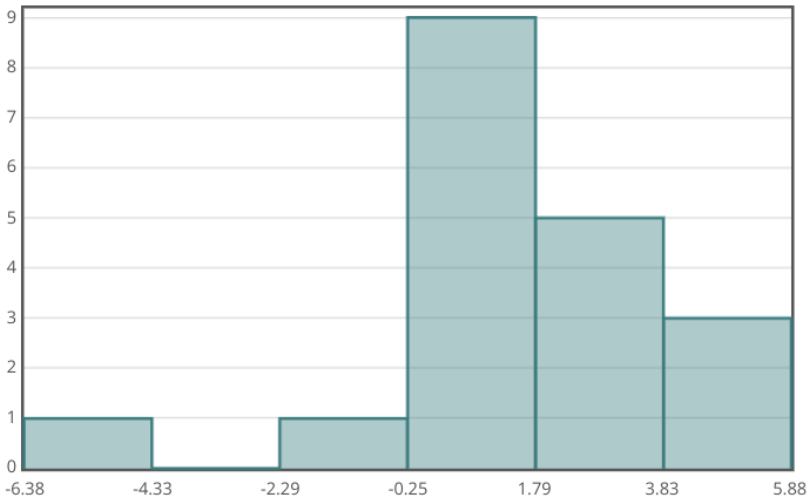

Figure 20. Group 1 paired difference value

The graph above (Figure 20) is a histogram of paired difference values within Group 1: (Post-test) - (Pre-test). There are 17 differences greater than or equal to zero, and 2 differences less than zero. The statistical manipulation uncovers the negative improvement for 2 users. This fact confirms the variance presence within learning settings, although the biggest degradation stands out of the cohort. In fact, this user demonstrated a high level of

achievement in the pre-test hitting a zero in the post-test. This particular case may have nothing to do with real degradation as a test-taker could have lost their interest in completing the task after performing well in the 1st attempt. The result of applying the same statistical measure for Group 2 (n=12) has also confirmed a statistically significant difference between pre-test and post-test conditions with the effect size of .35 (Figure 21). The graph below is a histogram of paired difference values: (Sample 2) - (Sample 1). There are 10 differences greater than or equal to zero, and 2 differences less than 0. For Group 2 similar degradations have taken place with top-performers losing the high-end results in their final attempt.

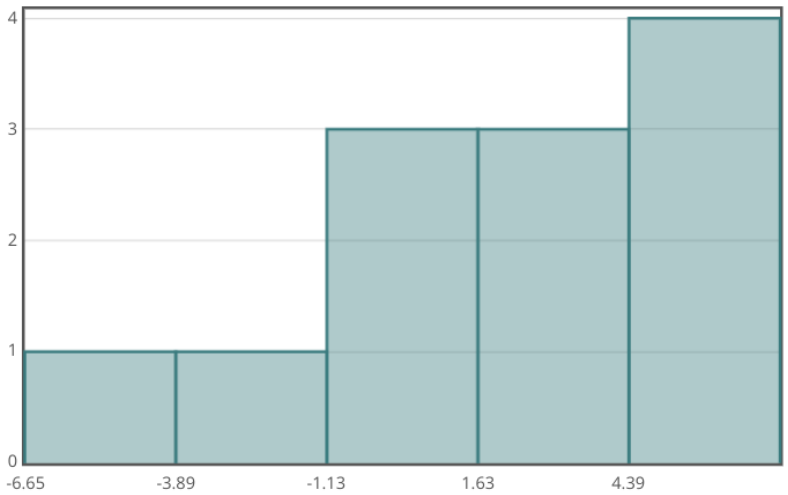

Figure 21. Group 2 paired difference value

The Group 3 (n=7) assessment reveals a slightly different picture in the group attainment (Figure 22). Although the statistical significance has also been confirmed (effect size -.59), no degradations have been detected. The graph below is a histogram of paired difference values: (Sample 2) - (Sample 1). There are 7 differences greater than or equal to zero, and 0 differences less than zero.

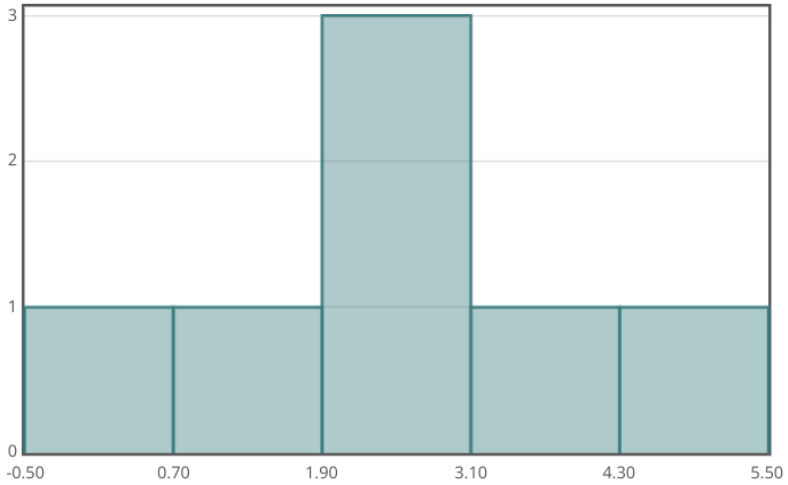

Figure 22. Group 3 paired difference value.

Although the research questions advocate for the sample splitting for reaching the research goals, it seems necessary to broaden outlook on the sample in order to secure statistical significance for the analyzed changes in the dataset. Given that the interval scale requirements and parametric test applicability (Beavens, 2022 & Bhandari 2022) have been fulfilled, there is an aspiration of making stronger inferences from the data than they could have been made in case of applying non-parametric tests. By applying the AI-Therapy calculator methodology for the research dataset, the following computing has been enacted: paired t-test processing was followed by the effect size measurement. The link provides a visualization of the applied methods: https://www.ai-therapy.com/psychology-statistics/results/20230326210119560 (also available at Appendix E) A short summary is given below.

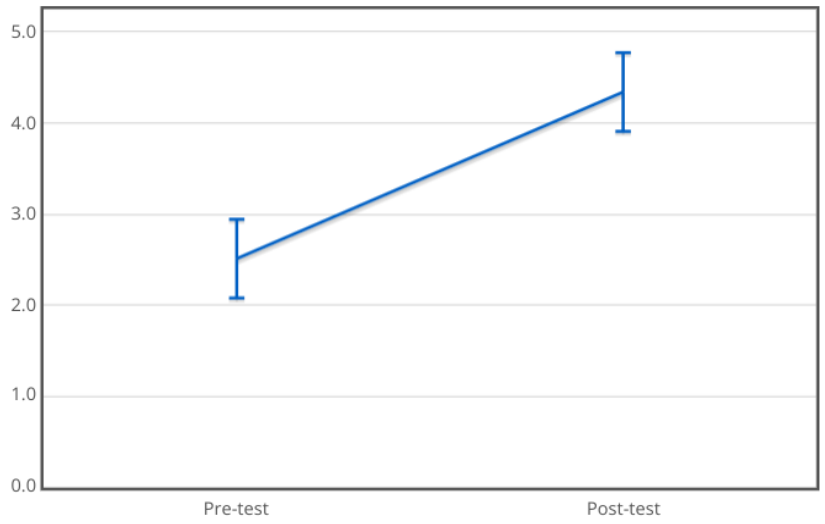

Figure 23. Pre-test and post-test performances for the whole sample

Based on a significance level of 0.05, there is a statistically significant difference between 'Pre-test' and 'Post-test' groups. The Pearson coefficient (r = .577) correlates with Cohen's d (a Pre-test variance = .807 and a pooled variance = .909), therefore postulating strong positive effect size of the study attempts for the whole sample representatives. The aforementioned inference solidifies the research design choice and highlights general efficiency of the ITS in question for reaching a short-term goal - practising for the EGE Speaking part. All in all, the calculated data also contributes to a positive reply to the research questions, which have been statistically confirmed.

Chapter 6. Results and Implications

This research was intended to investigate the frequency of using the chosen intelligent tutoring system among its real users in relation to the effects of this ITS on preparation for the EGE Speaking part - the most challenging section of the Russian high-stakes exam in the English language proficiency. In addition, the teachers' perception of applying non-conventional (online) tools for EGE preparation have been studied in support of the main goal of the research. Conducting the study in an ecological manner has provided opportunities to elicit thoughtful generalizations from observing real-life study conditions, which has led to deducting their implications on the usage of the ITS in question.

Possessing a status of one of the developers of the ITS in question and the only researcher in the longitude project has provided a lot of opportunities to see a full picture of the complexities of EGE Speaking part preparation. To make it more comprehensible and readable, the following storyline is suggested: starting unfolding the rationale for implementing a small-scale significance survey (teachers' questionnaires) might help to see reasoning for applying 'observational' tactics for data collection. Both procedures have significantly affected the research design as well as formulating the hypothesis. Also, some theoretical considerations are to be taken into account as long as the ITS in question cannot be aligned to just one type of computer-assisted tools.

Initiating a teachers' questionnaire was triggered by worsening scores of the exam participants as well as the author's desire to test teachers' perception of state-of-the-art learning tools. The positive remark to the latter might be correlated with teachers' concern over the issue and their goodwill to provide extra assistance to test-takers. Also, it was of value to have compared Russian teachers' perception of such sophisticated tools with their colleagues from some other countries. Basically, in the countries with similar socio-economic conditions teachers seem to be positive about implementing ICT tools, but have little exposure and instruction to such techniques (Aydin, 2013).

On the one hand, hitting a majority score in support of ITS usage in both iterations might be regarded as a positive sign showing teachers' determination to turn the tide. However, there is one fact that casts doubt on the possible takeaway as there was a 4-year span between the surveys, thus having no 'improvement' in terms of teachers' interest towards computer-based

techniques may be considered as a downward tendency as the EGE success rates were coming down throughout this span.

Another big issue is, undoubtedly, the EGEnglish status for all main stakeholders, primarily for test-takers and their teachers. After the EGE Speaking part introduction a few ‘training systems’ went into public offering specific ‘self-regulated courses’ for EGE Speaking part preparation. In fact, nearly all of these tools were ‘audio recorders with pictures’ without giving any useful feedback apart from providing recording for further listening. However, another feature of such training systems was implementing a timer telling a test-taker about the time limits of each task. In a way, this set of scaffolding techniques might be called a minimum simulator’s pack as the real-life EGE Speaking part is conducted by means of software possessing both features. In contrast, EGEnglish.ru has acquired varied learning capabilities, which altogether constitute a sophisticated online learning environment. In short, the ITS skillset presuppose having in-built domain specific knowledge (Brown & Sleeman, 1982), an assessment technology (automated rater) with an approximate success rate (human-machine correlation) of .8 (Evanini, 2015), and a set of suggested learning strategies for particular tasks (Andrade & Evans, 2013). EGEnglish.ru could remind some test-takers of ETS automated raters, but not with a holistic approach to testing as detailed reports (generated by automated speech recognition and natural language processing algorithms) were tailored to each user of the ITS in question.

Understanding the innovative specifics of the ITS, which were almost unfamiliar to the majority of key stakeholders, it was decided to provide continuous methodological and technical support in the form of open-to-all webinars to teachers as further agents raising awareness of EGEnglish.ru learning capabilities. Also, a few intervention groups were summoned in order to test teacher-student collaboration within the framework of the ITS in question. Not surprisingly, teachers’ direct involvement highly limited the use of EGEnglish.ru in-built capabilities due to teachers’ anxiety to automated speech recognition accuracy and quite low expectations towards the suggested learning environments, which promote studying in the self-regulated mode.

Promoting tutoring and control modules as learning tools has strongly altered the impression of users towards EGEnglish.ru, which deviates from conventional online tools not only because of the mentioned modules’ use, but also due to the presence of both audio and script

samples — a quite expected, yet frequently missing, part belonging to the domain module. All in all, the EGEnglish.ru framework can be regarded as a more diverse learning entity in comparison to competing systems. Therefore, it is reasonable to assume that some part of EGEnglish.ru users have been motivated to test a tool with more flexible learning opportunities. Moreover, the suggested diversity path can be seen as a more advanced tool, which could attract a population demanding more sophisticated and intense training, not exercised by the other online systems.

Grouping the students into the proposed cohorts was obviously an observational practice. Thanks to having access to all the meta-data within EGEnglish.ru it was possible to investigate study patterns of the ITS users. As it was said before, clustering around the real-life exam dates (in May and in June) was quite expectable due to the nature of the ITS, which is supposed to equip a learner with only exam strategies and helpful tips for arranging a task speech rather than exercises and techniques on improving spontaneous utterance production in general. However, the other three observations were certainly quite puzzling. First, the collected sample was relatively small due to the fact that there were very many registered users who attempted the tasks only once. It is reasonable to assume that for such users the ITS in question was only a ‘testing machine’ and they were not willing to use its study potential. Second, most of the users from the sample, 33 out of 38 to be precise, attempted the tasks within a span of 1 hour. Thus, it reveals with quite a high probability that they couldn’t get outside guidance. Also, it states that packing all these attempts in such a short time frame does not allow one to follow the EGEnglish.ru study strategies and guidelines, for instance an editing technique, aiming at detecting and preventing semantic and grammar mistakes. Third, 8 was the maximum number of attempts. It appears that exceeding this number might be treated by users as the ‘overload’ strategy, which is not worth it in terms of study outcome. After elaborating on the matter, it is plausible to conclude that the ITS in question has acquired a status of learning environment, but for quite a limited number of users, an approximate 10% share of the total population of EGEnglish.ru. Nearly the same ratio is seen in the sample of the research, in which only the high-intensity group appears to have exercised the learning-feedback-learning cycle. This compound term should not drive away from the basic feature of the ITS in question: users are expected to acquire specific exam skills. Therefore, this process can be treated as learning only in a limited way.
All the aforementioned considerations allow one to judge not only about the efficiency of the ITS in question, but also present study profiles of users, who are on the verge of attempting a

real-life high-stakes exam in a computer-based format. Group 1 sub-sample might be regarded as a ‘highly aware cohort’ with relatively high pre-test score and no intention to practice for a considerable amount of time. Group 2 sub-sample is made up of more motivated, or possibly more aware users (about their own weak points), who can get engaged in 3-4 study attempts in order to compensate for the lack of necessary skills. Group 3 sub-sample, although semi-randomized (Group 4 cohort of Task 4), consists of people that want to experience the task and use it extensively to increase their practice performance and/or preparedness for the real exam.

The accomplished scores of all groups of EGEnglish.ru users allow stakeholders to critically study the EGE Speaking part preparation strategies, which have to be tailor-made to the students’ learning needs and relevant environment as long as the real-life exam condition in a computer-based testing procedure with no human interference. In general, the process of preparation for the speaking part of various national and international high stakes-exam might incorporate individual study paths, in which test-takers have to be supported by teachers and tutors on the metacognitive level - the basic skill enabling one to calibrate an array of such state-of-the-art computer-based technologies as automated speech recognition and natural language processing.

Chapter 7. Discussions and limitations

Research works related to the field of intelligent tutoring systems have been published in immense quantities over the recent decades. Although language learning cannot boast of collecting extensive data on ITSs application to the study process, a considerable number of papers aiming at investigating narrow topics have received substantial interest among professional communities. The most attractive areas within the language learning domain are conversational dialogue (Graesser et. al, 2001), grammar (Virvou et al., 2000), vocabulary practice and reading (Heilman et al., 2006), writing (Mischaud et al., 2000). Despite raising the public's awareness of ITS applicability towards language studying, it appears that there is blank space regarding the speaking aspect, especially in terms of monologue-based tasks for high-stakes exams. Even the existing works of Educational Testing Service researchers dealing with TOEFL and similar exams cannot fully fill this gap.

The research threats which could be assigned to the chosen research design commonly include 2 kinds of effects: practice effects (performing differently in the second attempts due to familiarity with the experimental situation or the measures being used) and boredom effects (performing differently in the second condition because participants got bored or tired from having completed the first study attempts) (Field, 2013). These potential threats might be neglected in the current settings due to the research peculiarities. First, the ITS users must have had various exposure to the exam tasks prior to approaching the research conditions. So, the variation in the second and further attempts can be linked to the users' familiarity with received feedback (listening to a recording, studying a script or any follow-up analysis of the task within the ITS). Moreover, Discourse Completion Task (DCT) used as a measurement instrument target assessing quite an advanced monologue-oriented spontaneous performance which has to be practiced multiple times before succeeding in it. Second, the high-stakes exam condition, even in a simulation environment within the ITS, contributes to a relatively high motivation profile of the users thus almost eliminating boredom effects. As for potential tiredness of the procedure, the DCTs are aimed at being completed within a 3 min span (1 min for preparation + 2 min speech performance) so it is nearly impossible to expect a tedious-for-user scenario in a task driven by the self-directed deliberate action of the ITS user. At this point, it is extremely important to analyse the statement given in the previous passage as the idea of before-ITS exposure has to be fully clarified. Although this topic might be considered a deep ocean, it is still possible to detect pre-ITS study patterns. First of all,

assigning to the exam in question secures a test-taker general instruction on the task peculiarities, including Speaking part of the exam. The following stage cannot be regarded as training due to its goal — to familiriaze test-takers with specifics of the EGE. The next stage might involve some first-hand training, which is usually controlled by a teacher or a tutor. Moreover, this collaborated stage might take a form of more or less the same speaking attempts suggested for the computer-based mode within the studied ITS. The only difference is in the feedback agent — a human takes this responsibility. Both stages are clearly necessary for a test-taker as they frame a complete preparation path which, by nature, is inscribed into a well-planned study schedule. In other words, text takers are taken to the long-term route where the second stage (continuous teachers' feedback) is not limited by the timeframe virtually making it unlimited. In contrast, the suggested study path in the computer-based environment stands for reasonably short study periods. Therefore, it is fairly reasonable to treat the following stage as a third one in order, but with no direct connection, either cognitive or motivational, to the previously defined stages. Nonetheless, the suggested study paths might not be the only existing ones as the ITS in question could be approached by fully self-regulated learners who had not been exposed to the 'human' stages described above. These independent learners, certainly, lack some initial feedback, but the idea of self-regulated learning, as stated in the literature review, is rooted in acquiring learning strategies for performing a practical task (Andrade & Evans, 2013). In a way, self-regulated learning is seen as more practice-based and, arguably, more intense for learners. Hence, bearing in mind the aforementioned assumptions, the learning weight of the instructional stage appears to be quite low. As for the human-feedback stage, it definitely varies across the population making the computer-based stage as an 'equalizer' securing a standardized feedback for the users of the system.

The received post-test score of .62 for the research population, as it was said earlier, appears to be a match to the statistics of real exam performance within 2015-2018 years. However, this assumption has to be critically analysed due to the EGE Speaking part specifics and general features of the research, which tackled only the Task 3 and Task 4 of the exam in question. Task 1 and Task 2 are unanimously regarded as the easiest ones accounting for 6 points (out of total 20). Unfortunately, the all-Russian statistics on EGE exam performance does not provide a breakdown for each task casting doubts on the matter. Despite having no sufficient proof to the fact, it is still possible to deduce the approximate average score for both tasks. Teachers' first-hand statistics on the matter claim a much higher success rate for Task 1

and Task 2 thus pinpointing nearly maximum performance around 5 or 6 points. If one assumes that score, for instance 5, as a real average score for Task 1 and Task2, it is possible to see Task 3 and Task 4 joint shares within the overall performance rate. By presenting the overall success figures in the task scoring framework, one sees 13.2 points as an average overall score for all 4 tasks. A further calculation is to reveal a would-be share for Tasks 3 and 4, which total 8.2 points. A subsequent subtraction suggests an average of 4.1 points for each task. This figure might be used for stating the following: the suggested ITS as an additional tool human instruction might have a bigger impact on the exam performance score for the EGE Speaking part than the sole human instruction as it is reasonable to assume that only a minor part of EGE population might have used EGEnglish.ru (the total number of test-takers vary from year to year, but cluster around 100000 students while an annual EGEnglish.ru student population does not exceed 20 000 students annually).

The present study is not fully language-oriented as the marked interest is high-stakes exam performance, which is achieved not specifically through improving language skills, but via acquiring peculiar exam strategies and applying background linguistic knowledge. This inference resonates with the pragmatic competence concept presupposing realizing particular illocutions, knowledge of the sequential aspects of speech acts, and knowledge of the appropriate contextual use of the particular languages' linguistic resources (Barron, 2003). Therefore, the necessity of taking into account the pragmatic component, linguistic awareness of users, and technological ITS features constitute a multidisciplinary intersection, which has to be studied through the lens of various fields including human-computer interaction, applied linguistics, education psychology. The conducted research has become the first step along the path due to the holistic approach being exercised for the sake of understanding the major conditions and effects of the EGE Speaking part preparation. Apparently, a study with a framework comparing human via computer interaction appears to be the next step towards understanding the ITS usage as applied to the Speaking part of high-stakes exams. Also, it appears to be of interest to investigate a motivational aspect of test-takers incorporating ITS systems for exam purposes as alongside the motive for developing pragmatic (exam) skills there might be an array of motives triggering students to approach such self-regulated study environments.

In addition to the suggested research paths, there might be one more reasonable way of continuing studies, which can be dealt with estimating the effects of ITS systems on the

micro-skills acquisition as they match some of the exam criteria (in EGE and other language proficiency exams). The following investigation might compare the practical value of such ITS systems towards fulfilling pragmatic and linguistic goals.

The compiled list of limitations pinpoint not only certain difficulties questioning the research reliability and validity, but also highlight variables which have to be studied thoroughly in the forthcoming research papers dealing with the studied ITS or similar systems.

Limitations - Author

By being the author of the present research and a co-developer of EGEnglish.ru I am surely aware of the limitation regarding the validity of the experiment. Yet it is clear that the suggested methodology addresses the validity threat as the author is excluded from the assessment and instructional processes.

Offline support

It is quite predictable that students' might have some practice sessions for Speaking part outside the online environment. This doesn't seem counterproductive, although the exam in question is fully computer-based. Students might be getting continuous support from school teachers and private tutors (who are not using the tutoring system), parents, or former test-takers. The support might take a form of instructional support as well as training sessions with human feedback.

Competitive self-training systems

Some of such tutoring systems are parts of the paper-based preparation materials. Others are independent online resources. They all provide a set of sample responses and recording/playback feature for the EGE Speaking part.

Paper-based materials

Test-takers might take advantage of the English language course books for high schools which have some instructional and training features for the EGE exam. Also, there is a wide range of paper-based materials specified exclusively for the EGE Speaking part preparation.

Built-in instant feedback

Speech-to-text engine (automated speech recognition), by default, is activated by the platform users with a deliberate UX action – pressing a ‘Start recording’ button followed by clicking a ‘Stop recording’ button. Even in case of saving scripts within a user’s profile database it is impossible to assume that the received feedback has been thoroughly studied and used for further practice sessions.

On-site assistance

Test takers have been categorized based on their authorized profiles. Thus, we can assume with a considerably high probability that the logging process of the test takers was done deliberately for further practicing of their oral speaking skills themselves. Still, there exists a chance that execution of the test tasks was not done in a solo mode as even a double-check on the second voice presence does not eliminate a possibility of having an assistant encouraging a test taker in a silent fashion.

False start with the first attempt

Failing to make a meaningful attempt during the first session even though some criteria were applied.

Assessment variability

The assessment being applied in the research has replicated the procedure that has been in use in real-life exam settings. Although the framework is perceived as a quite objective evaluation tool, there is still an acceptable range of differentiation between examiners’ scores, which, undoubtedly, may vary raising questions about the procedure’s reliability.

References

Adnan, A. H. M., Ahmad, M. K., Yusof, A. A., Mohd Kamal, M. A., & Mustafa Kamal, N. N. (2020). *English Language Simulations Augmented with 360-degrees Spherical Videos (ELSA 360-Videos):'Virtual Reality'Real Life Learning!*. SSRN.

Andrade, M. S., & Evans, N. W. (2013). *Principles and Practices for Response in Second Language Writing: Developing Self-Regulated Learners*. London: Routledge.

Aston, G. (1995). Say 'Thank you': Some pragmatic constraints in conversational closings. *Applied linguistics*, *16*(1), 57-86.

Aydin, S. (2013). Teachers' perceptions about the use of computers in EFL teaching and learning: The case of Turkey. *Computer assisted language learning*, *26*(3), 214-233.

Barron, A. (2003). Acquisition in interlanguage pragmatics. *Acquisition in Interlanguage Pragmatics*, 1-416.

Bernacki, J. (2014). Creating collaborative learning groups in intelligent tutoring systems. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8671, pp. 184-193.

Bevans, R. (2022, December 05). Choosing the Right Statistical Test | Types & Examples. Scribbr. Retrieved March 22, 2023, from https://www.scribbr.com/statistics/statistical-tests/

Bhandari, P. (2022, November 17). Interval Data and How to Analyze It | Definitions & Examples. Scribbr. Retrieved March 20, 2023, from https://www.scribbr.com/statistics/interval-data/

Bloom, B. S., (1971). Mastery learning. In J. H. Block (Ed.), Mastery learning: Theory and practice. Rinehart & Winston, New York, pp. 47–63.

Brand-Gruwel, S. (2014). Learning ability development in flexible learning environments. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), Handbook of Research on Educational Communications and Technology (pp. 363- 372). New York: Springer.

Brauer, M. (2018). The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation.

Chien, S. Y., Hwang, G. J., & Jong, M. S. Y. (2020). Effects of peer assessment within the context of spherical video-based virtual reality on EFL students' English-Speaking performance and learning perceptions. *Computers & Education*, *146*, 103751.

Denisova-Schmidt, E., & Leontyeva, E. (2014). The Unified State Exam in Russia: Problems and Perspectives. *International Higher Education*, (76), 22-23. https://doi.org/10.6017/ihe.2014.76.5530

De Muth, J. E. (2019). Practical Statistics for Pharmaceutical Analysis with Minitab Applications. Springer.

D'Mello, S. (2007). Toward an Affect-Sensitive AutoTutor. IEEE Intelligent Systems, 22(4), pp. 53-61.

Evanini, K., Heilman, M., Wang, X., & Blanchard, D. (2015). Automated scoring for the TOEFL Junior® Comprehensive writing and speaking test. *ETS Research Report Series*, *2015*(1), 1-11.

Field, A. (2013). Discovering statistics using IBM SPSS statistics. sage.

Garito, M. A. (1991). Artificial intelligence in education: Evolution of the teaching—learning relationship. British Journal of Educational Technology, 22(1), pp. 41-47.

Georgiev, G.Z. (2017). "Confidence Interval Calculator", [online] Available at: https://www.gigacalculator.com/calculators/confidence-interval-calculator.php URL [Accessed Date: 24 Mar, 2023]

Golato, A. (2003). Studying compliment responses: A comparison of DCTs and recordings of naturally occurring talk. *Applied linguistics*, *24*(1), 90-121.

Graesser, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. IEEE Transactions on Education, 48(4), pp. 612-618.

Graesser, A. C., VanLehn, K., Rosé, C. P., Jordan, P. W., & Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI magazine*, *22*(4), 39-39.

Grobler, C. & Smits, T. F. H. (2017). Road map for the context-sensitive redesign of a technology-enhanced speaking practice environment. In Proceedings of Computer Assisted Language Learning (CALL) Conference, Berkeley, CA. Retrieved from http://call2017.language.berkeley.edu/wpcontent/uploads/2017/07/CALL2017_proceedings.pdf

Hattie, J.A.C. (2003). Teachers make a difference: What is the research evidence? Paper presented at the Building Teacher Quality: What does the research tell us ACER Research Conference, Melbourne, Australia. Retrieved from http://research.acer.edu.au/research_conference_2003/4/

Holden, C., & Sykes, J. (2013). Complex L2 pragmatic feedback via place-based mobile games. In N. Taguchi & J. Sykes (Eds.), Technology in interlanguage pragmatics research and teaching (pp. 155–184). Amsterdam: John Benjamins.

Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M. (2006). Classroom success of an Intelligent Tutoring System for lexical practice and reading comprehension. In *Ninth International Conference on Spoken Language Processing*.

Kalinowski, P. (2010). Understanding Confidence Intervals (CIs) and effect size estimation. *APS Observer*, *23*.

Kay, R., Knaack, L., & Petrarca, D. (2009). Exploring teachers' perceptions of web-based learning tools. *Interdisciplinary Journal of E-Learning and Learning Objects*, *5*(1), 27-50.

Kulik, J. A., Fletcher D., J. (2016). Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review. Review of Educational Research, 86(1), 42–78.

Lowyck, J. (2004). Students' perspectives on learning environments. International Journal of Educational Research, 41(6), pp. 401-406.

Ma, W. (2014). Intelligent Tutoring Systems and Learning Outcomes: A Meta-Analysis Journal of Educational Psychology, 106(4), pp. 901-918.

Michaud, L. N., McCoy, K. F., & Pennington, C. A. (2000). An intelligent tutoring system for deaf learners of written English. In *Proceedings of the fourth international ACM conference on Assistive technologies* (pp. 92-100).

Moreno, R. (2007). Interactive Multimodal Learning Environments. Educational Psychology Review, 19(3), pp. 309-326.

Neuendorf, K. A. (2002). The content analysis guidebook. Thousand Oaks, CA: Sage.

Nwana, H.S. (1990). Intelligent tutoring system: an overview, Artifical intelligence review. (pp. 251-277)

Ogiermann, E. (2018). Discourse completion tasks. In A. Jucker, K. Schneider, & W. Bublitz (Eds.), Methods in Pragmatics (pp. 229 - 255).

Olsen, J.K. (2014) Using an Intelligent Tutoring System to Support Collaborative as well as Individual Learning. In: Trausan-Matu S., Boyer K.E., Crosby M., Panourgia K. (eds) Intelligent Tutoring Systems. ITS 2014

Oxford, R. L. (2008). Hero with a thousand faces: Learner autonomy, learning strategies and learning tactics in independent language learning. Language learning strategies in independent settings, 33, 41.

Padayachee, I. (2002). Intelligent tutoring systems: Architecture and characteristics. Retrieved from https://www.researchgate.net/publication/228921731_Intelligent_tutoring_systems_Architecture_and_characteristics

Paris, S. G., Byrnes, J. P., & Paris, A. H. (2001). Constructing theories, identities, and actions of self-regulated learners. Self-regulated learning and academic achievement: Theoretical perspectives, 2, 253-287.

Pintrich, P. R., Wolters, C. A., & Baxter, G. P. (2000). Assessing metacognition and self-regulated learning. In G. Schraw, & J. C. Impara (Eds.), Issues in the Measurement of Metacognition (pp. 43–97). Lincoln, NE: University of Nebraska Press.

Pintrich, P.R. (2000). The role of goal orientation in self-regulated learning. In. M. Boekaerts, P.R. Pintrich & M. Zeidner (eds) Handbook of Self-regulation (pp. 451-502). San Diego: Academic Press.

Ramandalahy, T. (2010). An intelligent tutoring system supporting metacognition and sharing learners' experiences. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6095(2), pp. 402-404.

Roll, I. (2011). Improving Students' Help-Seeking Skills Using Metacognitive Feedback in an Intelligent Tutoring System. Learning and Instruction, 21(2), pp. 267-280.

Self, J. (1999). The defining characteristics of intelligent tutoring systems research: ITSs care, precisely. International Journal of Artificial Intelligence in Education (IJAIED), 10, pp.350-364.

Sydorenko, T., Daurio, P., & Thorne, S. (2018). Refining pragmatically appropriate oral communication via computer-simulated conversations. Computer Assisted Language Learning, 31, 157-180.

Stark, P. (1997) SticiGui. Statistics means never having to say you're certain. https://www.stat.berkeley.edu/~stark/SticiGui/Text/confidenceIntervals.htm

Sydorenko, T., Smits, T. F., Evanini, K., & Ramanarayanan, V. (2019). Simulated speaking environments for language learning: Insights from three cases. Computer Assisted Language Learning, 32(1-2), 17-48.

Taub, M. (2018) How Are Students' Emotions Associated with the Accuracy of Their Note Taking and Summarizing During Learning with ITSs?. In: Nkambou R., Azevedo R., Vassileva J. (eds) Intelligent Tutoring Systems. ITS 2018.

Tobin, D. R. (2000). All Learning is Self-Directed: How Organizations Can Support and Encourage Independent Learning. USA: ASTD.

Trevors, G. (2014). Note-Taking within MetaTutor: Interactions between an Intelligent Tutoring System and Prior Knowledge on Note-Taking and Learning. Educational Technology Research and Development, 62(5), pp. 507-528.

Tu, J. (2020). Learn to speak like a native: AI-powered chatbot simulating natural conversation for language tutoring. In Journal of Physics: Conference Series (Vol. 1693, No. 1, p. 012216). IOP Publishing.

Weinstein, C. E., Husman, J., & Dierking, D. R. (2005). Self-regulation interventions with a focus on learning strategies. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), Handbook of Self-Regulation Elsevier Academic Press. (pp. 727-747)

Vail A.K., Grafsgaard J.F., Boyer K.E., Wiebe E.N., Lester J.C. (2016) Predicting Learning from Student Affective Response to Tutor Questions. In: Micarelli A., Stamper J., Panourgia K. (eds) Intelligent Tutoring Systems. ITS 2016.

VanLehn, K. (2011) The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems, Educational Psychologist, 46:4, 197-221,

Virvou, M., Maras, D., & Tsiriga, V. (2000). Student modelling in an intelligent tutoring system for the passive voice of the English language. *Journal of Educational Technology & Society*, *3*(4), 139-150.

Wiklund-Engblom, A. (2015). Designing new learning experiences?: Exploring corporate e-learners' self-regulated learning. Åbo: Åbo Akademi University Press.

Appendix A

Participant	Survey #1	Survey #2 (Main - Stage1)
Teacher 1	1	0
Teacher 2	1	1
Teacher 3	1	0
Teacher 4	1	0
Teacher 5	1	1
Teacher 6	1	0
Teacher 7	1	1
Teacher 8	1	1
Teacher 9	1	0
Teacher 10	1	1
Teacher 11	1	0
Teacher 12	1	1
Teacher 13	1	1
Teacher 14	1	1
Teacher 15	1	1
Teacher 16	1	0
Teacher 17	1	1
Teacher 18	1	0
Teacher 19	1	1
Teacher 20	1	1
Teacher 21	2	1
Teacher 22	2	1
Teacher 23	2	1
Teacher 24	2	1
Teacher 25	2	1
Teacher 26	2	1
Teacher 27	2	0
Teacher 28	2	1
Teacher 29	2	1
Teacher 30	2	1
Teacher 31	2	0
Teacher 32	2	0
Teacher 33	2	0
Teacher 34	2	0
Teacher 35	2	1
Teacher 36	2	0
Teacher 37	2	0
Teacher 38	2	1
Teacher 39	2	1
Teacher 40	2	0
Teacher 41	2	1
Teacher 42	2	1
Teacher 43	2	1
Teacher 44	2	1
Teacher 45	2	0
Teacher 46	2	0
Teacher 47	2	1
Teacher 48	2	1
Teacher 49	2	1
Teacher 50	2	1
Teacher 51	2	1
Total		33

Answer code
0 - “Other Speaking training methods”
1 - "Tutoring system"
2 -"More practice at an exam format"

Appendix B - Final Survey data (Stage 2)

Participant	ITS need
Teacher #1	0
Teacher #2	0
Teacher #3	1
Teacher #4	1
Teacher #5	1
Teacher #6	0
Teacher #7	1
Teacher #8	0
Teacher #9	1
Teacher #10	0
Teacher #11	0
Teacher #12	1
Teacher #13	0
Teacher #14	1
Teacher #15	1
Teacher #16	1
Teacher #17	1
Teacher #18	0
Teacher #19	0
Teacher #20	1
Teacher #21	1
Teacher #22	1
Teacher #23	0
Teacher #24	0
Teacher #25	1
Teacher #26	1
Teacher #27	1
Teacher #28	0
Teacher #29	0
Teacher #30	1
Total	17

Answer code

1 - "Tutoring systems" (and attributes which can be linked to the notion)

0 -Other ways of Speaking exam preparation

Appendix C

Coding	Recording №	Task	Attempts
Tester 1	1 — 2	3	2
Tester 2	3 — 4	3	2
Tester 3	5 — 6	3	2
Tester 4	7 — 8	3	2
Tester 5	9 — 10	3	2
Tester 6	11 — 12	3	2
Tester 7	13 — 14	3	2
Tester 8	15 — 16	3	2
Tester 9	17 — 18	3	2
Tester 10	19 — 20	3	2
Tester 11	21 — 22	3	2
Tester 12	23 — 24	3	2
Tester 13	25 — 26	3	2
Tester 14	27 — 28	3	3
Tester 15	29 — 30	3	3
Tester 16	31 — 32	3	4
Tester 17	33 — 34	3	3
Tester 18	35 — 36	3	4
Tester 19	37 — 38	3	4
Tester 20	39 — 40	3	3
Tester 21	41 — 42	3	5
Tester 22	43 — 44	3	5
Tester 23	45 — 46	3	8
Tester 24	47 — 48	3	6
Tester 25	49 — 50	4	2
Tester 26	51 — 52	4	2
Tester 27	53 — 54	4	2
Tester 28	55 — 56	4	2
Tester 29	57 — 58	4	2
Tester 30	59 — 60	4	2
Tester 31	61 — 62	4	2
Tester 32	63 — 64	4	3
Tester 33	65 — 66	4	3
Tester 34	67 — 68	4	4
Tester 35	69 — 70	4	4
Tester 36	71 — 72	4	3
Tester 37	73 — 74	4	5
Tester 38	75 — 76	4	5
Tester 39	77 — 78	4	5
Tester 40	79 — 80	4	6
Tester 41	81 — 82	4	5

Appendix D

	Benchmark 1	Benchmark 2	Deviation	Category	Attempts	Task
Tester 6	5,5	6.5	1	low	2	T3
Tester 8	3,5	4	0.5	low	2	T3
Tester 11	4	6	2	low	2	T3
Tester 2	3	6	3	low	2	T3
Tester 4	0	4.5	4.5	low	2	T3
Tester 5	0	0	0	low	2	T3
Tester 7	6,5	7	0.5	low	2	T3
Tester 10	0	3	3	low	2	T3
Tester 1	4,5	4.5	0	low	2	T3
Tester 3	3,5	4.5	1	low	2	T3
Tester 9	0	3.5	3.5	low	2	T3
Tester 13	0	3	3	low	2	T3
Tester 27	5,5	6	0.5	low	2	T4
Tester 28	0	4.5	4.5	low	2	T4
Tester 25	4	5.5	0.5	low	2	T4
Tester 30	3,5	4.5	1	low	2	T4
Tester 31	3,5	2.5	-1	low	2	T4
Tester 29	0	5	5	low	2	T4
Tester 26	5,5	0	-5.5	low	2	T4
Tester 14	2,5	4.5	2	middle	3-4	T3
Tester 19	3	6.5	3.5	middle	3-4	T3
Tester 16	0	5	5	middle	3-4	T3
Tester 15	4	6	2	middle	3-4	T3
Tester 20	3	3.5	0.5	middle	3-4	T3
Tester 18	5,5	0	-5.5	middle	3-4	T3
Tester 17	0	5.5	5.5	middle	3-4	T3
Tester 40	3	3	0	middle	3-4	T4
Tester 32	0	6	6	middle	3-4	T4
Tester 35	3,5	4.5	1	middle	3-4	T4
Tester 33	6	4	-2	middle	3-4	T4
Tester 34	0	4.5	4.5	middle	3-4	T4
Tester 21	0	3	3	high	>=5	T3
Tester 23	5	6	1	high	>=5	T3
Tester 24	2	5	3	high	>=5	T3
Tester 22	0	5	5	high	>=5	T3
Tester 37	5,5	5.5	0	high	>=5	T4

Tester 39	0	3	3	high	>=5	T4
Tester 41	0	4	4	high	>=5	T4

Appendix E - Data set statistics

Sample name	Number of samples	Mean	Standard error of the mean	Standard deviation	Median
Pre-test	38	2.513	0.368	2.268	3.000
Post-test	38	4.342	0.279	1.721	4.500

Test results

Number of samples	N = 38
Normality of sampling distribution	Since the number of samples is relatively large (N > 30), the assumption of normality is likely to be satisfied.
Paired differences	• Mean difference = -1.829 • Standard deviation of differences = 2.626 • Standard error of differences = 0.426
Paired *t*-test	• t = -4.293 • df = 37 • Significance (2-tailed) $p < 0.001$ • Based on a significance level of 0.05, there is a statistically siginificant difference between 'Pre-test' and 'Post-test'.
Paired samples correlations	• r = 0.155 • Significance (2-tailed) p = 0.354
Effect size	• r = 0.577 • Cohen's d (using Pre-test variance) = 0.807 • Cohen's d (using pooled variance) = 0.909