e-ISSN 2280-6792

EL.LE
Vol. 13 — Num. 1 — Marzo 2024

Examinees' Four Skills Performance Balance in High Stake Foreign Language Examinations

Antonio Venturis, Julie Vaiopoulou
Aristotle University of Thessaloniki, Greece

Abstract The present study aims to investigate the balance of the examinees' performances in different skills tests as part of a foreign language examination. In many cases, the examinees' performances in the four skills present significant deviations. This deviation raises the question of whether it is an indication of an error in the construction or rating of the test or whether it is an expected result. In order to explore this question, a study was conducted using real data collected from 8715 candidates who participated in Italian language exams, administered by the Greek State Certificate Language Proficiency System (KPG), between 2011 and 2015 (ten sessions). The data were analysed with Latent Class Analysis (LCA) and indicated that there is no balance between the performances of the examinees in the different skills tests.

Keywords Foreign language testing. Four skills performance balance. Skills integration. Skills segregation. Language certification.

Edizioni Ca'Foscari

Peer review

Submitted 2023-02-01
Accepted 2024-01-24
Published 2024-03-22

Citation Venturis, A. (2024). "Examinees' Four Skills Performance Balance in High Stake Foreign Language Examinations". *EL.LE*, 13(1), 27-48.

DOI 10.30687/ELLE/2280-6792/2024/01/002

1 Introduction

The communicative ability includes the four skills in a relation of interdependence. This means that, in many cases, it is difficult to discern the limits of each skill and the range of involvement that each one has with the other in the final communicative output. In authentic communicative settings, separating receptive skills from productive skills and oral and written speech is almost always impossible. As Harmer states,

> it is very often true that one skill cannot be performed without another. It is impossible to speak in a conversation if you do not listen as well, and people seldom write without reading – even if they only read what they have just written. (2015, 52)

In traditional foreign language education, though, after more than 40 years from whole-language approaches proposal by numerous scholars (Rigg 1991; Harste, Burke 1977; Goodman, Goodman 1982; Watson 1989; Goodman, Goodman, Hook 1989), the practice of teaching the four skills separately is still present today, asking students to concentrate on only one skill at a time (Oxford 2001; Vernier et al. 2008; Vera et al. 2019). This approach, referred to by Oxford (2001) as Segregated-Skill Instruction (SSI), is based on the assumption that mastery of discrete language skills can lead to complete language learning. The more recent proposal for an integrated approach to language learning contradicted SSI, arguing that, in real life, different language skills are used in combination and the one affects the other during a communication event.

Despite the wide acceptance of the Integrated-Skills Approach (ISA) from the linguistic community (Oxford 2001; Gautam 2019; Pardede 2019), the SSI seems to be still present (Pardede 2019, 148) in foreign language education. There are still language courses focused on only one skill, claiming to satisfy specific needs or to be easier and more effective for the learners (Oxford 2001). However, even if skills in real life are almost always used in combination, this educational choice of skill segregation might be associated with an actual need of the students. In many cases, language assessment treats the four skills separately, dividing language tests into writing, reading, listening, and speaking sections. Since language examinations are crucial in many social environments, especially those of certification, this segregation directly affects language courses, making them exam-focused. This approach affects not only the goals of the language course, disorienting them from communicative skills acquisition to test achievement, but, in many cases, determines its content and methodology. Consequently, many language courses adopt the SSI with exam preparation in mind to ensure the successful performance

of the students or candidates. The separate approach can, however, sometimes lead to differences in acquiring these four skills. This could result from differences in the learning material, the methodology, the social environment (Kondo-Brown 2005; Firmansyah 2018), or the students' individual characteristics (Sparks, Ganschow 1991). The unbalanced development of the four skills inevitably affects students' performance in examinations, with severe consequences in the case of high-stake exams. Considering the above, the present study aims to investigate the correlation between the candidates' performance in the different skill tests of the Greek state certification system for the Italian language. More precisely, it intends to show whether the candidates' scores in the four units of the certification examination (each one corresponding to one skill) present deviation and whether they can be grouped based on a specific criterion (receptive-productive, oral-written speech skills). According to the research results, conclusions will be drawn based on the evidence concerning the accuracy and fairness of the assessment, and a point of reference for future relative decisions will be made in case of differences between the scores in the four skills units of a foreign language test.

2 Literature Review

The idea of ISA emerged at the end of the 1970s when communicative language teaching appeared. Up until then, structural linguistics and behaviourist learning theories established an extensive focus on speaking, grammar drills, and listening comprehension. These principles served as the basis of the approaches to language learning and teaching known as the 'oral method', the 'aural-oral method', the 'structural method', and later the 'audiolingual method' (Hinkel 2010). Even when the influx of foreign workers and students in the UK during the 1960s created new demands for language teaching and learning, the emerging need for integrated teaching of discrete skills did not lead to an essential change of practice. As Howatt and Widdowson stated, the main idea about language learning and teaching between the 1950s and 1970s was that "all four language skills (listening, speaking, reading, and writing) should be taught, but the spoken skills should be given priority" (2004, 299-300).

The new perspective of language learning and teaching proposed by the communicative approach radically changed many of the most basic and widespread practices in language education. The recognition of language as a social asset aiming to satisfy the specific social needs of students and the shift of language course focus from the structure of a language to communication, considering many of the factors affecting it, created a completely different educational context. As the communicative approach aimed at the acquisition

of communicative skills by the learners to be used in real-life situations, many researchers considered skills segregation a contradiction to the social reality (Corder 1971; Crystal 1971; Snow, Met, Genesee 1989). Hinkel asserts, based on the conclusions drawn by early 1970s scholars, that

> [i]n reality, it is rare for language skills to be used in isolation; e.g. both speaking and listening comprehension are needed in a conversation and, in some contexts, reading or listening and making notes is likely to be almost as common as having a conversation. (2010, 115)

One of the first linguists who suggested an integrative skills approach in language instruction was Widdowson (1978), claiming that language comprehension and production are combined in real-life communication; therefore, learners need to develop receptive and productive skills in both spoken and written discourse in order to acquire communicative proficiency. According to these considerations, he noted that:

> even though a particular exercise may focus on a particular skill or ability, its effectiveness will often require the learner to make reference to other aspects of his communicative competence. […] I have represented the learner's task as essentially one which involves acquiring a communicative competence in the language, that is to say, an ability to interpret discourse, whether the emphasis is on productive or receptive behaviour. If this definition of the learner's aim is accepted, it would seem to follow that any approach directed at achieving it should avoid treating the different skills and abilities that constitute competence in isolation from each other, as ends in themselves. What the learner needs to know how to do is to compose in the act of writing, comprehend in the act of reading, and to lean techniques of reading by writing and techniques of writing by reading. (144)

The need for skills integration in language instruction became more evident in the 1980s and 1990s, leading to interaction-centred and authentic activity planning. This need generated the task-based teaching approach, a practice that requires the engagement of a group of learners in an activity, simulating a real-life situation, like playing a game or organizing a trip through which the participants will have to seek information from a written source, discuss among themselves, take notes, and perhaps listen to information from media that contain spoken language, e.g., YouTube, using the target language. Nunan, in his effort to provide a complete and synthetic definition of task, states that it is

> a piece of classroom work that involves learners in comprehending, manipulating, producing, or interacting in the target language while their attention is focused on mobilizing their grammatical knowledge to express meaning, and in which the intention is to convey meaning rather than to manipulate form. The task should also have a sense of completeness, being able to stand alone as a communicative act in its own right with a beginning, a middle, and an end. (2004, 4)[1]

On this basis, language skills are taught and practiced according to students learning objectives, combining oral and written speech when needed, and engaging receptive and productive skills in the communication events of the tasks.

Another method of teaching that proposes the integrated skills approach is the content-based approach, inspired by Widdowson's (1978) and Halliday's (1978) ideas about language teaching and linguistic analysis. Content-based teaching is an approach in which the target language is taught within the study of a non-linguistic subject, such as geography, science, or technology (Lyster 2018). In this approach, the language is the medium of teaching and learning communication, meaning that the learners use the target language to seek information from the learning materials, communicate with the teacher and their classmates, and produce essays or implement other types of classwork or homework. In this context, learners acquire the foreign language rather than being taught it (Creese 2005). Skills integration comes naturally since studying a specific subject in an educational context requires a similar use of the language as in authentic everyday communication. Content-based teaching appears in the literature with various versions, such as content-based instruction (CBI) and content and language-integrated learning (CLIL).[2] Even though there are differences between them, especially in the role of the language (Met 1998), they share the integrated skills practice as a means of communication and a language learning approach.

Both models are widely accepted and widespread in foreign and second language education, and they motivate students to develop all four communicative skills. However, some scholars point out various disadvantages of integrating skills in language teaching. First, segregated skills instruction is dominant and highly appreciated in many countries or regions, so students and teachers resist skills

1 For a detailed discussion on task definition see Littlewood 2004. For more information about task-based approach see Nunan 2004; 2010; Skehan 2003.

2 In the relative literature are proposed various other models of content-based language teaching such as theme-based language teaching model, adjunct language teaching model, sheltered model (Oxford 2001). For additional information see Cenoz 2015; Met 1998; Stryker, Leaver 1997.

integration (Richards, Rodgers 2014). In addition, as Hinkel (2010) claims, a curriculum focusing on one language skill at a time favours intensive learning. Furthermore, Hinkel expresses her reservations about integrated instruction with more than two language skills addressed in tandem because it is more demanding for both teachers and learners and requires teachers to be well-trained. Another disadvantage of Integrated Skills Instruction (ISI) is that, in the case of large classes, it may create practical problems. Finally, an obstacle presented by ISI implementation could be unevenly developed proficiencies across the four skills from some learners, which is relatively frequent (Hinkel 2010; Speece et al. 1999). This fact affects language assessment practices where skills integration remains almost out of the question, as many researchers and practitioners consider it not applicable (Hinkel 2010).

3 The Four Skills in Language Tests

The wide acceptance of ISI in the language education community permits the hypothesis that it would clearly impact language testing. Nevertheless, despite some sporadic attempts to adopt an integrated approach to language testing, language testing methods and test designs continue to employ skills separation practices, believing that this is the most appropriate choice.

One of the first institutions that transitioned from a segregated to an integrated language test design was the University of Cambridge Local Examination Syndicate (UCLES). However, the experience was rather negative because they were found to be unreliable and error-prone. Spolsky (1995) identified the following factors as contributing to measurement error:

a. The discrete-point listening and reading comprehension sections did not show any consistency, with coefficients so low that they cannot be considered statistically valid.
b. Grading of the written productions subjectively without establishing consistency between raters.
c. Variation in the form of the test and the scoring methods of examiners.

Accordingly, UCLES eliminated the integration between reading and speaking in 1989 and between reading and writing in 1995. Since then, even if many researchers have continued to emphasize the pedagogical, social, and linguistic advantages of integrated skills assessment (Plakans 2012; Cumming 2013a; 2013b; Gebril 2018), most language assessment institutes and certification systems find separated skills tests more practical and suitable for testing purposes. The criticisms against this choice and the plethora of research prov-

ing the benefits of integrated-skills assessment do not seem to have a notable impact on large-scale test constructors' choices. Examples of certification systems that use integrated tasks, such as TOEFL iBT and DALF for English and AP Spanish from the College Board for the US, are limited.

However, although some certification systems, such as the Greek State Certificate in Language Proficiency, nationally and internationally known as the KPG (Dendrinos, Karavas 2013), examine skills separately with each of the four modules within each language level covering a single skill, in reality, candidates have to combine two or more skills in order to complete the task in a given module. For example, in the fourth module for speaking, testees have to listen and understand a task presented by the examiner, sometimes read a text and use its content as a stimulus or as an information pool, and for some tasks, take notes to be able to refer to the requested information, included in the source text. Of course, this practice could raise questions about assessment validity because it may be difficult to distinguish the limits between the skills involved in a task and their contribution to the test taker's performance (Bachman 1990). The tasks' construction and content, as much as the assessment criteria and tools, such as evaluation grids, could help the assessor focus on the target skill (Ventouris 2022). However, a clear isolation of a skill in an authentic task could be considered impossible.

In light of this discussion and considering the characteristics of a separated skills examination, a critical concern is whether a testee can present a notable difference in performance across skills tests. Certain conditions, such as the kind of language acquisition or learning, could affect the development of some skills. For example, if someone studies a language out of its context, they will have fewer opportunities to come into contact with oral speech than someone who learns it within the target language community (Nan 2018). This fact could lead to a major development in reading and writing. Migrants in the USA often achieve a high language level in speaking and listening, but their competence in writing and reading are not always equal because of the limited opportunities they have to develop them or due to restricting educational policies (Christensen 2000; Ladson-Billings 2016). A relative observation is that of De Las Fuentes Gutiérrez concerning migrants in Madrid (2020) and Chiswick, Lee, Miller (2004) about migrants in Australia. On the other hand, some language learners cannot accept test results that show an extended difference between their performance across different language skill tests (Ventouris 2019).

This question inspired the research presented below, which aimed to examine the relation between examinees' performance in tests assessing different skills (reading, writing, speaking, listening).

4 Research

4.1 Aim and Research Questions

The purpose of this study was to investigate the balance between the assessees' performance in different skills, which can be classified within the taxonomy:

1. type of the skill: receptive-productive;
2. channel of the speech: written-oral.

The main hypothesis concerns the empirical evidence of the independence of the underlying latent constructs corresponding to Reading, Writing, Listening, and Speaking. Thus, the research question posited was: "Is the performance of the test takers across the four skills tests balanced?".

In light of this, the following research questions can be posed:

a. Do skill type and speech channel affect possible balances?
b. Can any skill deviate from the rest in terms of performance balance?

4.2 Participants and Data Collection Procedures

The data were collected from the test results of 8715 adult candidates who participated in KPG Italian language certification examinations during ten sessions (2011-15). The sessions were selected due to the uniformity of the examination specifications they were based on (Dendrinos, Karavas 2013) to avoid the generation of regulatory variables that could affect the candidates' performances. The data were gathered from all thirty-eight examination centres of the KPG system located across the country.

In total, four pen and paper tests were given to the assessees over two days, each testing a different skill at graded levels (B1-B2). The modules of the examination were:

1. Reading comprehension and language awareness.
2. Writing and mediation.
3. Listening comprehension.
4. Oral production and mediation.

From the candidates examined, 176 (2%) failed the exam, 4247 (48.7%) achieved B1 level, and 4292 (49.2%) achieved B2 level.

4.3 Data Analysis

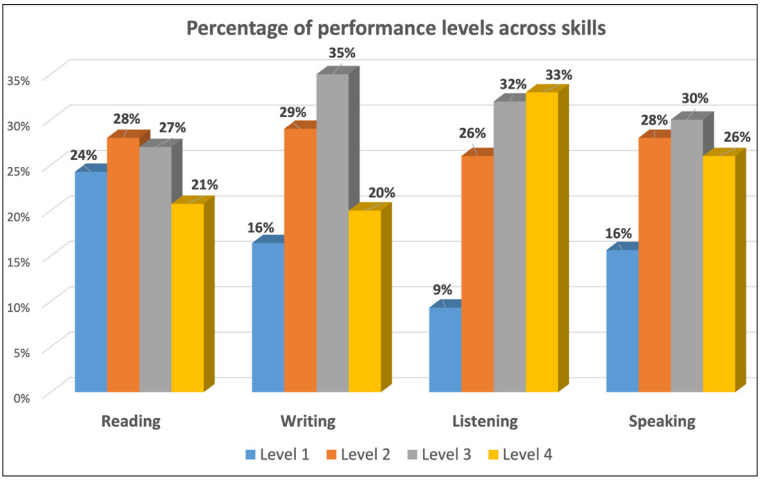
The main research question related to whether the four skills are balanced across tests was investigated through Latent Class Analysis (LCA) (Clogg 1995; Dayton 1998; Stamovlasis et al. 2018). LCA is a person-centred psychometric method and a measurement model that assumes that both latent and observable variables are categorical. As a cluster analysis, LCA divides the sample into segments or latent classes (LC) using an input set of categorical variables. The classification procedure is based on the similarity of responses (input categories) and, more specifically, on the probabilities of response patterns. Statistical goodness-of-fit indices are used to evaluate the classification model based on how well the results from clustering are accounted for. These indices are the number of parameters (Npar), likelihood ratio statistic (L2), Bayesian Information Criterion (BIC), Akaike's Information Criterion (AIC), degrees of freedom (df), and bootstrapped p-value, where the BIC is the most important to decide the number of the resulted clusters. LCA is a robust statistical methodology that has been applied to various disciplines and research fields. More specifically, in educational research, LCA has been implemented in a wide variety of recent research endeavours and is an established tool for the identification of participants' profiles, contributing to scientific dialogue in challenging theoretical issues, such as the nature of students' knowledge (e.g. Straatemeier, van der Maas, Jansen 2008; Vaiopoulou, Papageorgiou 2018; Stamovlasis, Vaiopoulou, Papageorgiou 2020). Moreover, it should be highlighted that LCA, as a psychometric approach, demonstrates several advantages compared to the traditional cluster analysis procedures (Magidson, Vermunt 2004).

In order to facilitate the application of LCA in this endeavour, some data transformations were performed. The four variables operationalizing the skills of reading, writing, listening, and speaking that were measured on an interval scale were converted to four-level ordinal-scale variables. This was achieved via the application of a two-step cluster procedure using the corresponding Z-scores. The resulting hierarchical levels were marked as Level 1 (lowest performance), Level 2, Level 3, and Level 4 (highest performance). Each participant was attributed a Level according to their performance in the skill tests. This means that an individual might have performed according to Level 1 in reading, but he/she might be at Level 2 in writing, etc., implying an unbalanced performance. If, however, all participants who are classified as Level 1 in reading are also classified as Level 1 in the other skill tests, and if the same holds for Level 2 in reading, etc., then the answer to the research question is that the performance of the test takers across all four skills tests is balanced. This thought leads to the implementation of LCA in a sophisticated way to answer the primary

research question since LCA can detect the sample segments consistent with a specific level of performance. This analysis aims to identify at least one latent class/cluster whose members would be at the same level (i) in all skill tests. The LCA was performed separately for each subsample of the corresponding result in KPG (B1 level, B2 level). The LatentGOLD_5.1 software was used.

4.4 Results

A descriptive statistics analysis was conducted to show the skill levels at which the candidates had performed better. The results indicate that regarding reading, 48% of participants performed at Levels 3 and 4, whereas the corresponding percentage was 55% in writing, 65% in listening, and 56% in speaking **[graph 1]**, implying that there are individuals who performed vastly differently across the skills.



Graph 1 Descriptive statistics of the performance Levels across the skills tested in KPG

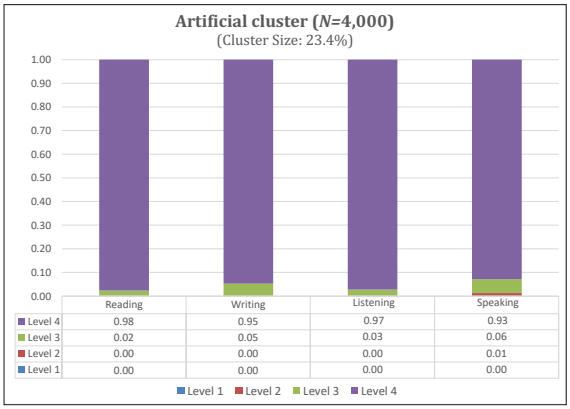
Table 1 shows the correlation coefficients among the initial interval scale variables. All skills were correlated with each other at a $p<0.001$ significance level. More specifically, reading is correlated with writing ($r=0.670$), listening ($r=0.637$), and speaking ($r=0.430$). Writing was correlated with listening ($r=0.519$) and speaking ($r=0.438$), whereas listening was also correlated with speaking ($r=0.357$). The moderate and low correlations comprise evidence implying that the four skills that are assessed in KPG are possibly not equally developed in individuals.

Table 1 Pearson's Correlations between skills

Variable	R	W		L		
1. R						
2. W	0.670	***				
3. L	0.637	***	0.519	***		
4. S	0.430	***	0.438	***	0.357	***

* p <.05, ** p <.01, *** p <.001

Subsequently, the results of the LCA are presented to demonstrate the above-mentioned unbalanced development of skills. To explicate how the LCA can test the hypotheses related to the development of the skills, a semi-simulation experiment was performed. In the empirical data set, an artificial segment of data was added, including 2,000 cases (23%), all attributed to Level 4 across the four skills of reading, writing, listening, and speaking. Five latent classes emerged from the following analysis of the modified data (original data plus the artificial segment), meaning that LCA identified the artificial cluster along with some very small portions derived from the original data. The properties of this cluster (LC) were as intended, i.e., the conditional probabilities tend to be equal to one for the members to perform at Level 4 across all skills **[graph 2]**. Note that 0.4% of the participants with a similar performance pattern were added to the artificial LC.



Graph 2
Artificial cluster

The homogeneous picture in graph 2 is expected to be found in the LCA of the empirical data, if a balanced development of skills indeed occurs, at least in some participants. The analysis was performed separately for participants who attained B1 and B2 levels.

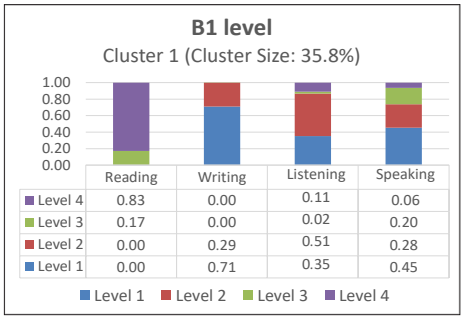
4.4.1 B1 Level

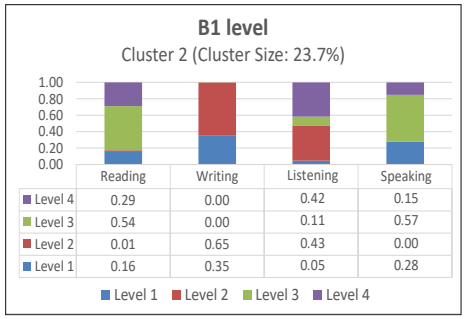
The LCA results for the sub-sample with participants who attained the B1 level (n=4247) are shown in Table 2. Among the calculated models, the four-cluster model solution had the minimum value of BIC and was chosen as the best parsimonious model (BIC = 38092.43, Npar = 51, df = 204, Class. Err.= 0.2582).

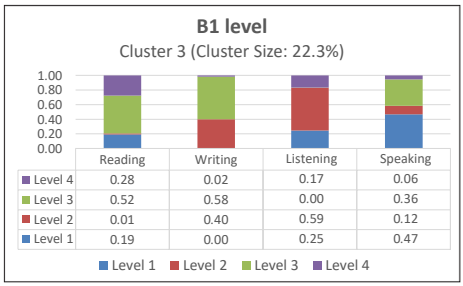
Table 2 LCA (B1 level, n=4247)

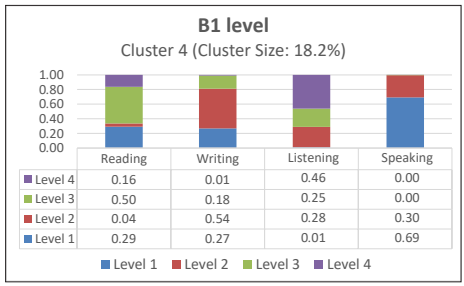
	LL	BIC(LL)	Npar	L^2	df	p-value	Class. Err.
1-Cluster	−19310.3	38720.83	12	1339.696	243	6.90E-151	0
2-Cluster	−18979.8	38168.53	25	678.796	230	7.50E-46	0.1314
3-Cluster	−18893.2	38103.95	38	505.6087	217	4.40E-25	0.2328
4-Cluster	−18833.2	38092.43*	51	385.4879	204	2.60E-13	0.2582
5-Cluster	−18806.5	38147.71	64	332.1658	191	1.10E-09	0.2626

The four clusters are depicted in the following graphs **[graphs 3a-d]**. Cluster 1 (35.8% of the participants) contained the individuals who had a very high probability (almost 100%) to perform the highest, i.e., at Levels 3 and 4 in reading, whilst it was most likely that they would perform the lowest, i.e., at Levels 1 and 2 in writing (almost 100%), listening (86%) and speaking (74%). Next, Cluster 2 (23.7% of the participants) demonstrated higher achievement in reading (83%) and speaking (72%). At the same time, the probability of performing the lowest in writing was 100%. It is worth mentioning that Cluster 2 members share an almost equal probability of performing at Levels 2 and 4 (43% and 42%, respectively). Cluster 3 (22.3%) is characterized by participants who are most likely to perform at Levels 3 and 4 in reading (80%), had a pretty low performance in listening (84%) and speaking (59%), but moderate performance in writing (98% likelihood for achievement at Levels 2 and 3). The last cluster was equally fragmented. The analysis revealed that its members had an almost 100% chance of demonstrating a lower performance in speaking and 75% in writing. On the contrary, in the case of reading and listening, they were more likely to have higher performance (66% and 71% respectively). Table 3 summarizes the skill attainment levels in reading, writing, listening, and speaking that prevail in each cluster.







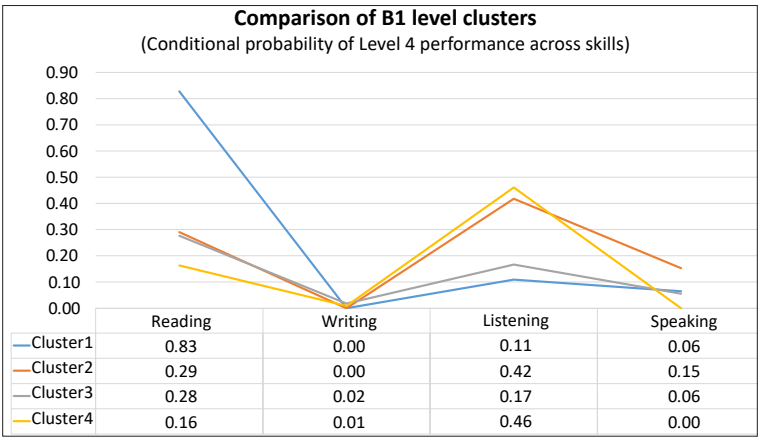


Graphs 3a-d Clusters derived from LCA (B1 level)

Table 3 The skill attainment levels in reading, writing, listening, and speaking that prevail in each cluster (B1 Level)

	Cluster 1 35.8%	Cluster 2 23.7%	Cluster 3 22.3%	Cluster 4 18.2%
Reading	Level 4	Level 3	Level 3	Level 3
Writing	Level 1	Level 2	Level 3	Level 2
Listening	Level 2	Level 2	Level 2	Level 4
Speaking	Level 1	Level 3	Level 1	Level 1

Graph 4 compares the conditional probabilities of demonstrating the highest achievement, i.e., Level 4, across the skills for each cluster. These conditional probabilities are completely inhomogeneous and vary across the four skills, highlighting that in language acquisition, as tested by the foreign language certification examinations, the different skill-level attainment is unbalanced and uneven for reading, writing, listening, and speaking.



	Reading	Writing	Listening	Speaking
Cluster1	0.83	0.00	0.11	0.06
Cluster2	0.29	0.00	0.42	0.15
Cluster3	0.28	0.02	0.17	0.06
Cluster4	0.16	0.01	0.46	0.00

Graph 4 Comparison of B1 level clusters to perform at Level 4 across skills

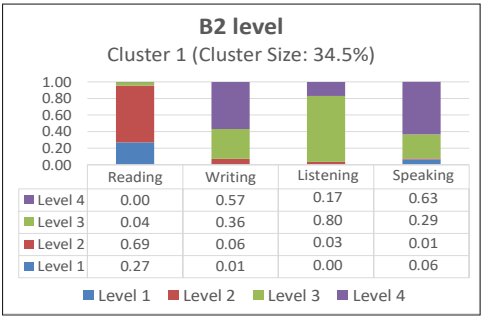
4.4.2 B2 Level

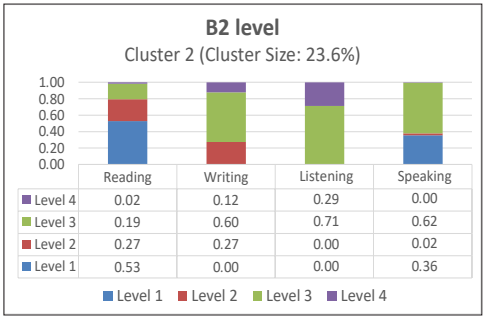
The LCA results for the sub-sample with participants who attained the B2 level (n=4292) are shown in Table 4. Among the calculated models, the 4-cluster model solution had the minimum value of BIC and was chosen as the best parsimonious model (BIC = 35967.41, Npar = 51, df = 204, Class. Err.= 0.2728).

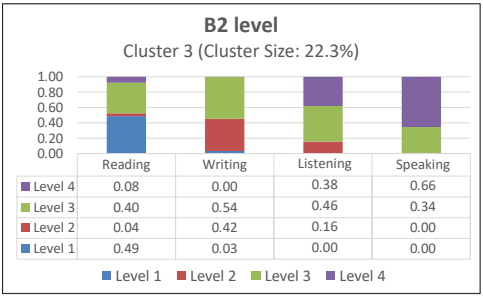
Table 4 LCA (B2, n=4292)

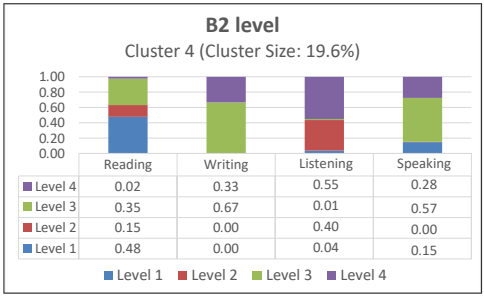
	LL	BIC(LL)	Npar	L^2	df	p-value	Class. Err.
1-Cluster	−18235.8	36571.93	12	1181.779	243	3.80E-123	0
2-Cluster	−17921.1	36051.33	25	552.441	230	1.50E-28	0.1205
3-Cluster	−17841.7	36001.19	38	393.5633	217	2.40E-12	0.2077
4-Cluster	−17770.4	35967.41*	51	251.0381	204	0.014	0.2728
5-Cluster	−17750.2	36035.7	64	210.5937	191	0.16	0.2893

According to the results, Cluster 1 (34.5% of the participants) had a high conditional probability of performing at the highest levels in listening (97%), as well as in writing and speaking (both 93%). On the other hand, their achievement was low in reading (96% likelihood to perform on Levels 1 and 2). Cluster 2 (23.6% of the participants) had an almost 100% likelihood of achieving the highest performance levels in listening whilst underperforming in reading (80% likelihood for Levels 1 and 2). In writing, the performance was somewhat moderate (87% for Levels 2 and 3). In speaking, it









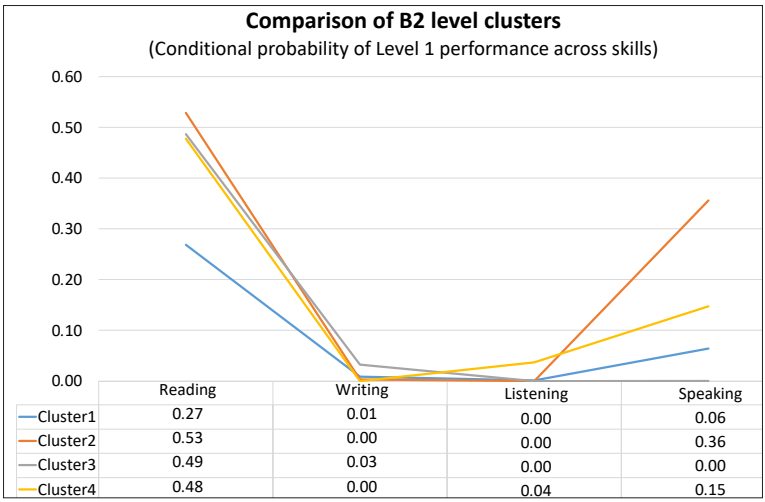
Graphs 5a-d Clusters derived from LCA (B2 level)

is worth mentioning that the conditional probability was 36% for performing at Level 1 and 62% at Level 3. The conditional probabilities for Cluster 3 (22.3% of the participants) were almost 100% of achieving Levels 3 and 4 in speaking and 84% for listening. For writing, there was a 96% likelihood of performing at Levels 2 and 3, and for reading, equally likely to perform at Level 1 (49%) or Level 3 (40%) appears equally likely. For Cluster 4 (19.6% of the participants), the conditional probability of achieving Levels 3 and 4 is almost 100% for writing and 85% for speaking. In contrast, the members of this cluster had a greater likelihood of performing lower in reading (65%). Interestingly, the conditional probabilities in listening were 40% for performing at Level 2 and 55% for Level 4. Table 5 summarizes the skill attainment levels in reading, writing, listening, and speaking that prevail in each cluster.

Table 5 The skill attainment levels in reading, writing, listening, and speaking that prevail in each cluster (B2 Level)

	Cluster 134,5%	Cluster 223,6%	Cluster 322,3%	Cluster 419,6%
Reading	Level 2	Level 1	Level 1	Level 1
Writing	Level 4	Level 3	Level 3	Level 3
Listening	Level 3	Level 3	Level 3	Level 4
Speaking	Level 4	Level 3	Level 4	Level 3

Graph 6 depicts the comparison of the conditional probabilities of performing at Level 4 across the skills for each cluster. Despite being expected to approach unity, these conditional probabilities were dramatically inhomogeneous and varied across the four skills, highlighting that in language acquisition, as tested by the foreign language certification examinations, the different skill-level attainment is unbalanced and uneven for reading, writing, listening, and speaking.



Graph 6 Comparison of B2 level clusters to perform at Level 4 across skills

5 Discussion

By analysing data collected from the KPG language certification exams in Greece, this study sought to determine whether the four skills tested in the exams (reading, writing, listening, and speaking) would appear to be balanced in the examinees' performance. The answer to this question could contribute to the discussion about high-stakes test administration and provide evidence about skills development in foreign language education. The analysis results were obtained using a person-centred approach, Latent Class Analysis, which identified distinct groups of participants with different conditional probabilities of attaining the four levels of performance in reading, writing, listening, and speaking. Both B1 and B2 level participants were analysed separately, and four profiles were identified, but neither had a homogeneous profile in terms of conditional probabilities. On the contrary, some profiles emphasized the performance discrepancies across the modules. All clusters were heterogeneous regarding the patterns of these probabilities, denoting inconsistent development rhythms across skills.

The present research contributes to the relevant literature by providing empirical evidence of the inconsistencies of testees' performance in the skills examined in foreign language certification examinations. The data used were the scores for each module in a real examination setting, where the participants typically try their hardest to perform the best they can, thus enhancing the reliability of the findings. The conclusion concerning the expected imbalance of testees' performance in foreign language examinations can contribute to the fairer and more appropriate administration of test results, especially in high-stakes exams. If the candidates' performance in the different skill tests is not expected to be balanced, the differences frequently noticed in examination results should not be perceived as evidence of evaluation error. According to this, it seems possible to observe even large differences in the various skills performance even if, according to the relative literature, the development of the four skills in a language learner is not segregated.

Another observation corresponding to the first secondary research question is that the analysis did not indicate a possible grouping of assessees' performances according to a specific criterion, such as the type of skill (receptive-productive) and the speech channel (oral-written). Given this, a candidate's performance in a specific skill test cannot serve as a measure of validity and reliability in other tests related to it. The answer to the last secondary research question cannot be directly based on specific findings of the present research since the analysis revealed that no skill-test performance can serve as a measure of general language proficiency.

6 Limitations and Directions for Future Research

Some limitations constrain the implications of the findings of the present research. One of these is the cross-sectional nature of the data, even though the sample is large, and the procedure is highly reliable. Longitudinal studies might offer a better understanding of language skills development or acquisition. Moreover, the lack of independent variables does not explain the characteristics of the observed profiles. The identification of crucial individual differences is expected to lead to associations that are meaningful and theoretically interpretable. Lastly, the fragmented patterns observed along with cluster identification might indicate that the relationships and the underlying processes of development of the various language skills are nonlinear in nature. The complex adaptive systems perspective is a meta-theory that has opened new avenues of investigation in educational research (e.g., Koopmans, Stamovlasis 2016; Vaiopoulou et al. 2021) that could prove very useful in foreign language testing research.

References

Bachman, L.F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press. https://doi.org/10.1017/CBO9781107415324.004.

Cenoz, J. (2015). "Content-Based Instruction and Content and Language Integrated Learning: The Same or Different?". *Language, Culture and Curriculum*, 28(1), 8-24. https://doi.org/10.1080/07908318.2014.1000922.

Chiswick, B.R.; Lee, Y.L.; Miller, P.W. (2004). "Immigrants' Language Skills: The Australian Experience in a Longitudinal Survey". *International Migration Review*, 38(2), 611-54. https://doi.org/10.1111/j.1747–7379.2004.tb00211.x.

Christensen, L. (2000). *Reading, Writing, and Rising Up*. Milwaukee: Rethinking Schools.

Clogg, C.C. (1995). "Latent Class Models". Arminger, G.; Clogg, C.C.; Sobel, M.E. (eds), *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. Boston: Springer US, 311359. https://doi.org/10.1007/978–1–4899–1292–3_6.

Corder, S.P. (1971). "Describing the Language Learner's Language". Perren et al 1971, 57-64.

Creese, A. (2005). "Is This Content-Based Language Teaching?". *Linguistics and Education*, 16(2), 188-204. https://doi.org/10.1016/j.linged.2006.01.007.

Crystal, D. (1971). "Stylistics, Fluency, and Language Teaching". Perren et al 1971, 34-53.

Cumming, A. (2013a). "Assessing Integrated Skills". *The Companion to Language Assessment*. Hoboken: Wiley. 216-29. https://doi.org/10.1002/9781118411360.wbcla131.

Cumming, A. (2013b). "Assessing Integrated Writing Tasks for Academic Purposes: Promises and Perils". *Language Assessment Quarterly*, 10(1), 1-8. https://doi.org/10.1080/15434303.2011.622016.

Dayton, M.C. (1998). *Latent Class Scaling Analysis*. London: Sage Publications.

Dendrinos, B.; Karavas, K. (2013). *The Greek Foreign Language Examinations for the State Certificate of Language Proficiency. The KPG Handbook*. Athens: RCeL publications.

Firmansyah, D. (2018). "Analysis of Language Skills in Primary School Children (Study Development of Child Psychology of Language)". *PrimaryEdu – Journal of Primary Education*, 2(1), 35. https://doi.org/10.22460/pej.v1i1.668.

Gautam, P. (2019). "Integrated and Segregated Teaching of Language Skills: An Exploration". *Journal of NELTA Gandaki*, 1, 100-7. https://doi.org/10.3126/jong.v1i0.24464.

Gebril, A. (2018). "Integrated-Skills Assessment". *The TESOL Encyclopedia of English Language Teaching*, 1-7. https://doi.org/10.1002/9781118784235.eelt0544.

Goodman, K.S.; Goodman, Y. (1982). "A Whole Language Comprehension Centered View of Reading Development". *Basic skills: Issues and choices*, 2, 125-1351.

Goodman, K.S.; Goodman, Y.; Hood, W. (1989). *The Whole Language Evaluation Book*. Portsmouth, NH: Heinemann.

Halliday, K. (1978). "Ideas about Language". *Arts (Sydney, NSW)*, 11, 20-38.

Harmer, J. (2015). *The Practice of English Language Teaching*. 5th ed. London; New York: Pearson.

Harste, J.C.; Burke, C.L. (1977). "A New Hypothesis for Reading Teacher Education Research: Both the Teaching and Learning of Reading are Theoretically Based". Pearson, P.D. (ed.), *Reading: Research, theory, and practice: = Twenty-sixth Yearbook of the National Reading Conference*. Chicago: MasonPpublishing.

Hinkel, E. (2010). "Integrating the Four Skills: Current and Historical Perspectives". Kaplan 2010, 110-216.

Howatt, A.P.R.; Widdowson, H.G. (2004). *A History of ELT*. 2nd ed. Oxford; New York: Oxford University Press.

Kaplan, R.D. (ed.) (2010). *Oxford Handbook of Applied Linguistics*. Oxford: Oxford University Press.

Kondo-Brown, K. (2005). "Differences in Language Skills: Heritage Language Learner Subgroups and Foreign Language Learners". *Modern Language Journal*, 89(4), 563-81. https://doi.org/10.1111/j.1540-4781.2005.00330.x.

Koopmans, M.; Stamovlasis, D. (eds) (2016). *Complex Dynamical Systems in Education: Concepts, Methods and Applications*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-27577-2.

Ladson-Billings, G. (2016). "Literate Lives Matter". *Literacy Research: Theory, Method, and Practice*, 65(1), 141-51. https://doi.org/10.1177/2381336916661526.

De Las Fuentes Gutiérrez, E. (2020). "The Development of Written Expression in Immigrant Children from 6 to 9 Years Old". *Open Linguistics*, 6(1), 109-31. https://doi.org/10.1515/opli-2020-0008.

Littlewood, W. (2004). "The Task-Based Approach: Some Questions and Suggestions". *ELT Journal*, 58(4), 319-26. https://doi.org/10.1093/elt/58.4.319.

Lyster, R. (2018). *Content-Based Language Teaching*. New York; London: Routledge.

Magidson, J.; Vermunt, J.K. (2004). "Latent Class Models". Kaplan 2010, 175-98.

Met, M. (1998). "Curriculum Decision-Making in Content-Based Language Teaching". *Beyond Bilingualism. Multilingualism and Multilingual Education*. Clevedon; Philadelphia: Multilingual Matters LTD, 35-63.

Nan, C. (2018). "Implications of Interrelationship among Four Language Skills for High School English Teaching". http://dx.doi.org/10.17507/jltr.0902.26.

Nunan, D. (2004). *Task-Based Language Teaching*, *Task-Based Language Teaching*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511667336.

Nunan, D. (2010). "A Task-Based Approach to Materials Development". *Advances in Language and Literary Studies*, 1(2), 135-60. https://doi.org/10.30762/jeels.v10i2.1751.

Oxford, R. (2001). "Integrated Skills in the ESL/EFL Classroom. ERIC Digest". *ESL magazine*, 4(1), 18-25.

Pardede, P. (2019). "Integrated Skill Approach in EFL Classrooms: A Literature Review". *EFL Theory & Practice: Voice of EED UKI*, 147-59. https://www.researchgate.net/publication/332607443_Integrated_Skills_Approach_in_EFL_Classrooms_A_Literature_Review.

Perren, G. et al. (1974). *Interdisciplinary Approaches to Language*. *CILT Reports and Papers 6*.

Plakans, L. (2012). "Assessment of Integrated Skills". Chapelle, C. (ed.), *The Encyclopedia of Applied Linguistics*. London: Blackwell Publishing Ltd, 1-8. https://doi.org/10.1002/9781405198431.wbeal0046.pub2.

Richards, J.; Rodgers, T. (2014). *Approaches and Methods in Language Teaching*. 3rd ed. Cambridge: Cambridge University Press.

Rigg, P.A.T. (1991). "Whole Language in TESOL". *TESOL Quarterly*, 25(3), 521-42.

Skehan, P. (2003). "Task-based instruction". *Language Teaching*, 36, 1-14.

Snow, M.A.; Met, M.; Genesee, F. (1989). "A Conceptual Framework for the Integration of Language and Content in Second/Foreign Language Instruction', *TESOL Quarterly*, 23(2), 201-17. https://doi.org/10.2307/3587333.

Sparks, R.L.; Ganschow, L. (1991). "Foreign Language Learning Differences: Affective or Native Language Aptitude Differences?". *The Modern Language Journal*, 75(1), 3-16.

Speece, D.L.; Roth, F.P.; Cooper, D.H.; De La Paz, S. (1999). "The Relevance of Oral Language Skills to Early Literacy: A Multivariate Analysis". *Applied Psycholinguistics*, 20(2), 167-90. https://doi.org/10.1017/S0142716499002015.

Spolsky, B. (1995). *Measured Words*. Oxford: Oxford University Press.

Stamovlasis, D.; Papageorgiou, G.; Tsitsipis, G.; Tsikalas, T.; Vaiopoulou, J. (2018). "Illustration of Step-Wise Latent Class Modeling with Covariates and Taxometric Analysis in Research Probing Children's Mental Models in Learning Sciences". *Frontiers in Psychology*, 9, 1-20. https://doi.org/10.3389/fpsyg.2018.00532.

Stamovlasis, D.; Vaiopoulou, J.; Papageorgiou, G. (2020). "A Comparative Evaluation of Dissimilarity-Based and Model-Based Clustering in Science Education Research: The Case of Children's Mental Models of the Earth". *International Journal of Data Analysis Techniques and Strategies,* 12, 247-61.

Straatemeier, M.; van der Maas, H.L.J.; Jansen, B.R.J. (2008). "Children's Knowledge of the Earth: A New Methodological and Statistical Approach". *Journal of Experimental Child Psychology*, 100(4), 276-96. https://doi.org/10.1016/j.jecp.2008.03.004.

Stryker, S.B.; Leaver, B.L. (1997). "Content-Based Instruction: From Theory to Practice". Stryker, S.B.; Leaver, B.L. (eds), *Content-Based Instruction in Foreign Language Education: Models and Methods*. 1st ed. Georgetown: Georgetown University Press, 3-51.

Vaiopoulou, J.; Tsikalas, T.; Stamovlasis, D.; Papageorgiou, G. (2021). "Nonlinear Dynamic Effects of Convergent and Divergent Thinking in Conceptual Change Process: Empirical Evidence from Primary Education". *Nonlinear Dynamics, Psychology, and Life Sciences*, 25(3), 335-55.

Vaiopoulou, J.; Papageorgiou, G. (2018). "Primary Students' Conceptions of the Earth: Re-Examining a Fundamental Research Hypothesis on Mental Models". *Preschool and Primary Education*, 6(1), 23. https://doi.org/10.12681/ppej.14210.

Ventouris, A. (2019). *Report on Canditates' Assessment*. Athens: Ministry of Education, Religion and Sports.

Ventouris, A. (2022). "The Effectiveness of Behaviorally Anchored Rating Scales in Writing Skill Evaluation". *IOSR Journal of Research & Method in Education*, 12(4).

Vera, V.D.G.; Valencia, B.J.; Cardona, N.B.; Cifuentes, M.A.; Herrera, S.J.; Martínez, L.M. (2019). "Should an Effective Language Learning Be through the Development of Just One Language Skill?". *The Qualitative Report*, 24(11), 2778-93. https://doi.org/10.46743/2160–3715/2019.4015.

Vernier, S.; Barbuzza, S.; Giusti, S.D.; Moral, G.D. (2008). "The Five Language Skills in the EFL Classroom". *Nueva Revista de Lenguas Extranjeras*, 10(1), 263-91.

Watson, D. (1989). "Defining and Describing Whole Language". *Whole Language*, Special issue, *Elementary School Journal*, 90(2), 129-41.

Widdowson, H.G. (1978). *Teaching Language as Communication*. Oxford; New York: Oxford University Press.