

Article

Washback Effects of Diagnostic Assessment in Greek as an SL: Primary School Teachers' Perceptions in Cyprus

Maria Mitsiaki ^{1,*} , Nansia Kyriakou ^{2,*}, Despo Kyprianou ³, Chrysovalanti Giannaka ⁴ and Pavlina Hadjitheodoulou ⁵

¹ Department of Greek Philology, Democritus University of Thrace, 69100 Komotini, Greece

² School of Education and Social Sciences, Frederick University, Nicosia 1036, Cyprus

³ School of Education, University of Nicosia, Engomi 2417, Cyprus; kyprianou.de@unic.ac.cy

⁴ Cyprus Ministry of Education, Culture, Sport and Youth, Nicosia 1434, Cyprus; chgiannaka@gmail.com

⁵ Pedagogical Institute Cyprus, Latsia 2238, Cyprus; hadjitheodoulou.p@cyearn.pi.ac.cy

* Correspondence: mmitsiaki@helit.duth.gr (M.M.); dledu.kn@frederick.ac.cy (N.K.)

Abstract: Washback of diagnostic tools targeted to young migrant learners has been an under-researched area in the language assessment field. This paper explores teachers' perceptions on the *Greek Diagnostic Language Assessment* (GDLA) tool recently introduced into the SL preparatory classes of the Cyprus primary education. The tool's implementation coincides with the launch of a new SL curriculum. The objective is fourfold: (1) to examine GDLA's washback effects on teaching/assessment, (2) to investigate washback's variability with respect to several contextual variables, (3) to collect feedback on the perceived credibility of the tool, and (4) to reflect on the use of the GDLA tool as a lever of instructional reform in support of curricular innovation. The study employs a mixed-methods approach and draws on (a) quantitative data (questionnaire, 234 informants) and (b) qualitative data (interviews, 6 participants). The results indicate a positive and quite strong washback on teaching and assessment. However, they bring to the surface several misconceptions on the purpose and the implementation of diagnostic assessment, pointing to gaps in the teachers' assessment literacy. They also bring into play school administration constraints. Finally, they imply that a diagnostic assessment aligned to a context-sensitive curriculum may bind the test to positive washback.

Keywords: diagnostic language assessment; teachers' perceptions; washback effects; assessment literacy; Greek as a second/additional language; learners with a migrant background; Cyprus primary schools; mixed-methods approach



Citation: Mitsiaki, Maria, Nansia Kyriakou, Despo Kyprianou, Chrysovalanti Giannaka, and Pavlina Hadjitheodoulou. 2021. Washback Effects of Diagnostic Assessment in Greek as an SL: Primary School Teachers' Perceptions in Cyprus. *Languages* 6: 195. <https://doi.org/10.3390/languages6040195>

Academic Editors: Dina Tsagari and Henrik Böhn

Received: 20 September 2021

Accepted: 15 November 2021

Published: 26 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Language assessment has been a major area of research in applied linguistics over the past 60 years (Davies 2014; Tsagari and Banerjee 2014). Among the approaches employed to monitor student language development, despite its limitations, testing has remained an overarching concern in education policies, mostly due to the legislative, authoritative, and transformative power of tests in the frame of national or school-based achievement and proficiency exams (Cheng and Curtis 2004; Shohamy 2017). At the same time, however, considerable discussion has arisen as to the necessity of dynamic—over static—assessment (Poehner et al. 2017), strengthening the argument that process-oriented approaches pose a unique challenge to language teaching and learning. In any case, recent initiatives taken worldwide signal a shift away from a narrow traditionally defined high-stakes focus to more holistic approaches, thus broadening our conception of assessment and even questioning the meaning of “high-stakes.”

In such a context, promising endeavors have recently been made that bring diagnostic language assessment into play (Alderson 2011; Lee 2015) and prompt us to reconsider the meaning of its stakes; compared with a national proficiency exam, diagnostic testing might

be seen as low- or no-stakes, but screening efforts in an educational context are considered “high stakes” for both the individual and the schooling system (Bailey 2017). Bailey’s claim can be easily supported if we take into consideration that an inaccurate diagnosis of a student’s needs may prove to be costly in terms of time, effort, and resources at a later point in time; falsely identified students may be at risk of failing courses/school because of their either limited or underestimated proficiency. Moreover, the meaning of stakes in young learners’ language assessment is another issue to be addressed, since low-stakes classroom-based assessment feeds high-stakes decisions (Rea-Dickins and Gardner 2000), which puts forward the argument that “assessments for young learners are potentially all high-stakes” (Butler 2016, p. 369).

These developments call for novel assessment tools or processes, especially when the changes brought about by globalization have revealed the presence of competent plurilingual users (Coste et al. 2009) with a diverse linguistic repertoire in many varieties and in different degrees. The challenge becomes even more demanding when it comes for students with a migrant background who enter the school of the host country and receive language instruction in the predominant majority language (Leung and Lewkowicz 2017). For them, being successful at school requires language proficiency and thus the conceptualization of “success” becomes a complicated issue tightly linked to assessment in a language other than their own but also through this language as a vehicle for content (Menken 2013; Mitsiaki and Lefkos 2018). However, assessing the language skills of migrant students in an unbiased and culturally valid way, immediately upon their arrival at school or during their school life, is a crucial and often neglected aspect in assessment literature and research. Of course, the recent geopolitical and economic changes and the migration patterns that emerged throughout the world have sparked a lot of attention on the assessment of additional/second languages (AL/SL), but in most cases the scholars provide arguments *against* language tests as criteria for obtaining citizenship or *for* a test fairness framework (i.e., Shohamy 2009; Kunnan 2013), giving an account of what should be carried out (or better, not carried out) and not necessarily what was actually carried out.

The previous observations highlight the importance of investigating the effect of assessment types on teaching practices and policies where migrant students are involved or, as it is commonly referred to in language assessment literature, the *washback effect*. It would be no exaggeration to say that washback has received the least attention in language minority learners’ assessment, maybe due to the tacitly held assumption that in their case the stakes are not that high. Such an assumption is implied in the low estimation of teaching effectiveness among teachers who offer courses to children with a migrant background (Kościółek 2020).

This paper attempts to pull together the aforementioned threads of language assessment, i.e., washback issues associated with school-based diagnostic assessment for young migrant/refugee learners of Greek in the Cyprus primary school setting. It takes the stand that diagnostic assessment is of vital importance for young minority learners who attend preparatory and/or mainstream classes, as it enables teachers to reflect on the learners’ strengths, interests, and areas of future development and leads to decision-making on the teaching content and methods.

The main purpose of the study is to examine the effect of the diagnostic *Greek Diagnostic Language Assessment* tool (GDLA, Mitsiaki et al. 2020a, 2020b) on public primary school teachers’ perceptions and practices. The timing of the research is considered to be critical, as it coincides with the implementation of the new *Curriculum for Greek as a Second Language* (CGSL, Mitsiaki 2020) and the teachers’ training programs on it, organized by the local education agency (Cyprus Pedagogical Institute).

The research objective is fourfold: (1) to investigate at what level the GDLA tool affects both *what* and *how* teachers teach and assess; (2) to determine any correlation between the washback effect and the teachers’ educational/academic background and motivation; (3) to gain feedback on the perceived credibility of the tool, i.e., the extent to which teachers view the tool as relevant to the test takers in the target situation (face validity); and (4) to reflect

on the use of the GDLA tool as a lever of instructional reform in support of curricular innovation.

To fulfil these objectives, we followed a mixed-methods design for data collection and analysis. First, a 26-item questionnaire was administered to 234 primary school SL teachers. Second, six semi-structured interviews were conducted to triangulate the data obtained.

The results reveal the teachers' positive attitudes and perceptions towards GDLA but at the same time they show that many other factors, beyond the test, might explain the washback effect, leading to several insights into the teachers' assessment literacy.

2. On Washback and Language Assessment Literacy

The concept of washback has evolved considerably since its first explicit mention as *backwash* by Hughes (1989). An important landmark is the foundational work of Alderson and Wall (1993), who define washback as the extent to which a test influences the *what* and *how* of the teaching–learning process assuming that it exists when “teachers and learners do things they would not necessarily otherwise do because of the test” (p. 117). Since then, the washback literature and research has blossomed. The rich literature available offers a wide range of terms used to denote washback with slight differences in shades of meaning, such as *impact*, *effect*, *consequences* etc. (see Tsagari and Cheng 2017 for a brief outline). Among them, *impact* is often used as the superordinate term while *washback* is used as the more specific classroom-based one. In this paper, the term *washback* is taken in its more inclusive sense (Rea-Dickins and Scott 2007) and seen as any interaction between tests or assessment tools and teaching (Green 2013).

An extended amount of work has discussed the aspects of washback (for a comprehensive overview see Cheng and Watanabe 2004; Tsagari 2007; Green 2013). The most common distinctions that empirical studies take into consideration are drawn from Watanabe (1997, 2008): (a) the *specificity* of the washback (general or related to specific aspect(s) of a test/a specific test type), (b) the *extent* (Bachman and Palmer 1996) or *intensity* (Cheng 2005) of washback (strong or weak), (c) its *direction* (beneficial or damaging), (d) its *length*, i.e., the influence of assessment in relation to its short- or long-term implementation, (e) its *intentionality* (unintended or intended), and (f) its *value* (positive or negative).

Both the terminological plurality and the various aspects of washback reveal its complex and multifaceted nature. Bachman and Palmer (1996, p. 35) highlight this complexity by stating that it should be “evaluated with reference to the contextual variables of society’s goals and values, the educational system in which the test is used, and the potential outcomes of its use”. Watanabe (1996), Cheng and Curtis (2004), and Tsagari (2009) further specify the various contextual factors, making mention of past education and academic background, personal beliefs, the status of the subject tested, management practices within the school, and the stakes of the test. Lastly, Green (2007) attempts to combine the aspects of washback with contextual factors, proposing a model that interrelates the test design considerations (*washback direction*) with (a) participant values, motivations, and resources (*washback variability*) and (b) the perceived importance and difficulty of the test (*washback intensity*).

Throughout its historical development, washback has been regarded as an effect found in high-stakes exams (Hamp-Lyons 1997). Such an assumption has important consequences on the range of assessment tools that have been considered as crucial to be investigated for washback. As a result, most research has concentrated on global EFL proficiency exams (Alderson 2011), whereas diagnosis seems to have been left aside for a long time (Jang 2013). However, the very recent work bears out the nascent recognition that the strong and well-attested effect of large-scale formal tests is “general, systemic, complex, and difficult to trace” (Lee 2015, p. 8) whereas the impact of diagnostic assessment is personalized and directly utilizable in teaching design and learning improvement within the classroom (Fan et al. 2021). This means that it is about time we assign the necessary importance to the effect(s) of diagnostic assessment and thus rethink the meaning of *teaching to the test* by examining it in specific diagnostic contexts.

Mainly evidenced in high-stakes contexts, *teaching to the test* is replete with negative connotations and it has been associated with negative washback or a tendency to narrow the curriculum (Shohamy 2001; Menken 2006). However, in classroom-based diagnostic assessment, which may play a less decisive role in the learners' life chances (at least in an explicit way) and have less detrimental effects on their professional career than a national proficiency or achievement exam (though an equally crucial role in their long-term linguistic/social development and school life), teaching to the test might be reconceptualized and potentially associated with a positive influence. In cases where the assessment is aligned to a new curriculum, it might offer a chance to further Andrews' claim that *teaching to the test* could be seen as a "strategy to promote curricular innovation . . . turning the apparently powerful effect of tests to advantage" (Andrews 2004, p. 39) so that the influential assessment tool is a diagnostic one.

Investigating washback becomes more interesting and demanding when the assessment of young school learners is the focus, as developmental, age-appropriate, and language level considerations must be addressed (Inbar et al. 2005; Bailey et al. 2013). The challenge is even tougher where learners with a migrant background are concerned, for whom the exposure to AL may start at any time during the primary school years. For SL teachers, cultural issues and issues related to familiarity with the assessment purpose are also raised, as in many school curricula the assessment of an additional language is still based on practices, content, and assessment criteria suitable for first language development or foreign language learning (Bailey 2017).

To our knowledge, no research has been conducted on the effects of diagnostic language assessment tools used for language minority learners in primary schools. There is a limited number of international studies on washback effects where language minority groups are involved which either focus on adult learners (Burrows 2004) or secondary students in high-stakes educational contexts (Stecher et al. 2004; Menken 2006). One of the very few studies in the context of Greek as an SL/FL was the one conducted by Antonopoulou (2004), which mainly reviews the positive washback effect of the exams for the certificate of attainment in Greek (a summative assessment tool published by the Center for the Greek Language used in tutoring lessons/courses) on teaching Greek as an SL/FL to teenagers and adults. The only study for the diagnostic assessment of migrant young learners in Cyprus is the one by Petridou and Karagiorgi (2017) which discusses the validation of a diagnostic test developed for the Greek school setting in the Cypriot one.

One last issue to be considered is the connection between washback and language assessment literacy (LAL). Within the past three decades, a growing body of literature and research has added new insights to our understanding of LAL as the professional knowledge in language testing or assessment that assessors/stakeholders are required to master (Fulcher 2012; Inbar-Lourie 2017). The LAL concept draws from its predecessor, general assessment literacy, a notion that for Popham (2009, p. 4) is regarded "a sine qua non for today's competent educator", irrespective of the subject taught and assessed. Thus, LAL emerges as a "multilayered entity" (Inbar-Lourie 2013, p. 304) that demands a fusion between different competencies on behalf of the teacher: (a) general and discipline-oriented pedagogical expertise, and (b) general and language-related assessment knowledge. This means that a language assessment literate teacher can merge between the different layers, i.e., content (language), pedagogy, and assessment, as she identifies the *why*, *what*, and *how* of the assessment process in general, and relates them to the language component (Inbar-Lourie 2013, pp. 305–6; Levi and Inbar-Lourie 2020, p. 2). In other words, she can develop, choose, implement, and evaluate a relevant/tailored assessment tool/process, administer/apply it, score/reflect on it, interpret its results in a pedagogically, socially, and culturally fair way, communicate them to her learners, integrate them to teaching to improve learning, and even identify potential misuses.

The previously sketched "(language) assessment knowhow" remains a nebulous concept, highly impacted by the testing-oriented focus that dominated general assessment and language assessment during the previous decades. As illustrated in Davies (2008),

such a knowhow comprises two components: (1) *knowledge and skills* in testing design, construction, analysis, and measurement, and (2) *principles* that revolve around issues of the assessment's proper use, fairness, and ethics. As broader theoretical and empirical assessment repertoires emerged, e.g., learner-centered and dynamic assessment approaches, a shift of focus from the testing content to the principles of assessment is observed. This shift was boosted by the emergence of CEFR and the suggested alternative forms of assessment, such as the European Language Portfolio and self- or peer-assessment (Inbar-Lourie 2017). The subsequent research that was conducted on teachers' assessment competencies verified their increased training needs in these "new" areas of assessment (Hasselgreen et al. 2004; Vogt and Tsagari 2014). It should be mentioned, though, that in Cyprus where the formal language of schooling is taught as an additional language to students with a migrant background, the emphasis has only recently been shifted from test design and measurement to alternative modes of assessment (Mitsiaki 2020). As it can be assumed, this development is in an infantile stage; at the same time, no research, to our knowledge, has examined Cypriot SL teachers' competencies in test design and measurement.

Given the multidimensionality of the LAL concept and its shifts in meaning and content, numerous scholars provide empirical evidence that a LAL culture has not been firmly established for teachers so far (i.e., Vogt and Tsagari 2014; Liu and Li 2020), without it being necessarily their fault, since assessment training in many cases is not delivered as a course in teacher education institutions (Lam 2015). More recently, LAL research has been targeted to EFL primary school teachers and it surveyed their lacking practices in alternative/formative assessment (Pehlivan Şişman and Büyükkarcı 2019) and their unwillingness to abstain from the more traditional assessment practices they were used to (Liu and Li 2020; Zhang and Soh 2016). Therefore, as put by Tsagari and Vogt (2017), LAL is an area in which professionalization of teachers should be enhanced. To proceed with this development, empirical studies turn their attention to the teachers and they acquire a "localized perspective" (Inbar-Lourie 2017, p. 263). This is exactly the point at which the interrelation between LAL and washback becomes evident, since the teachers' own knowledge base, beliefs, and practices on assessment might determine the perceived effect of any assessment tool or process. Moreover, the more assessment literate a teacher is, the more likely she will recognize the effects of an assessment on teaching and learning (Papakammenou 2020).

This is an orientation that not only empirical research but also teacher training programs and theoretical models on LAL takes. The training courses' content itself can emerge as a point of debate between the various stakeholders. In her research, Malone (2013) documents the opposing perceptions towards the focus of LAL training programs between testing experts, who lay the emphasis on theoretical aspects, and teachers, who express the need for classroom-focused assessment tasks. Scarino (2013) claims that by engaging the teachers into a reflective process that embraces their roles both as instructors and evaluators, the self-awareness of their own assessment literacy is strengthened, and as a result a shift in practices might be recorded in future. In a similar vein and without underestimating the importance of theory, Inbar-Lourie (2017) suggests that teacher-training programs should develop teacher activism and not a passive reception of information. Lastly, Fulcher (2020) emphasizes the necessity to move from practice to theory abstraction and puts forward a theory of pedagogy that operationalizes LAL through an Apprenticeship Model, which may help teachers to become literate in designing and building tests.

The previous literature discussion points to a broad LAL construct (the emphasis being laid either on testing or on alternative assessment forms), with different profiles, contents, and orientations and in different local contexts (Taylor and Banerjee 2013; Inbar-Lourie 2017). The SL teaching context where young learners with a migrant background are engaged in Cyprus makes an interesting research field to investigate the washback of a specific diagnostic assessment tool through the teachers' perceptions. We should keep in mind, though, that any washback effect observed might be intertwined with their degrees of LAL and might induce more or less positive consequences from the diagnostic process.

3. The Context of the Study

3.1. *The Integrating Policy in Cyprus Primary School Education*

Several socio-political changes and historical events that took place in Cyprus over the past few decades have increased the diversity of the country's population (Kyriakou 2014). Today, more than 15% of Cyprus' students in public primary schools are language minority students learning Greek and in Greek as an additional language (GAL). In fact, the numbers are slowly but steadily increasing; the percentage grew from 13.5% in 2015–2016 to 16.8% in 2019–2020 (Ministry of Education, Culture, Sports and Youth, MOECSY 2021).

GAL learners in Cyprus receive instruction both in mainstream and in preparatory courses within the regular school timetable. Newly arrived students who need support in the language of instruction attend extra lessons (separate classes or preparatory courses) per week, depending on their level of proficiency. Each learner receives support in GAL for up to two years. The process of assigning learners to CEFR levels and the number of the classes to be implemented is determined at the beginning of the school year (European Commission 2019a). It should be noted that teachers employed in these courses are not placed according to their specialty areas and thus they often have little or no expertise in teaching and assessing an SL/AL, which makes it questionable whether they can cater for the needs of their linguistically diverse learners (European Commission 2019a).

Besides the above-mentioned educational deficits, the highly centralized institutional policies lead to further complications as the years of employment is a determinant factor for the teachers' (re)appointment to schools¹. As a result, GAL teachers are often transferred just at the time when they have gained experience and competencies, which in turn leads to a loss of investment in qualified GAL teaching staff. The picture gets more complicated if we take into consideration the restricted teaching time allocated to GAL courses. This is mostly because the teaching hours are allocated based on the total number of GAL students per school, not the number of students per CEFR-related and level-appropriate class.

Over the past few years, MOECSY has launched several initiatives to support the integration of migrant students at school and their educational achievement. As a result of these efforts, a distinct *Curriculum for Greek as an SL* (CGSL) was developed and introduced a year ago (September 2020) with the main objective to provide a systematic CEFR-related and outcome-oriented framework of language learning and teaching. The curriculum promotes a policy that yields the concepts of plurilingual and intercultural education, taking into consideration both the learners' linguistic and cultural background and the state of diglossia or bilingualism in Cyprus. In such a framework, all linguistic varieties present at school are valued and, at the same time, learning the language of schooling, both as a subject in its own right and as the medium of curriculum content, is mapped out. Moreover, attention is paid to the three dimensions of continuity in language learning, i.e., biographical, thematic, and plurilingual continuity (European Commission 2019b). The CGSL is accompanied by detailed descriptors based on the first two proficiency levels of CEFR (A1 to A2 to guide instruction in the preparatory classes) and complemented with academic content "can do" statements to be used within the mainstream classroom (B1+). It is important to notice that the CEFR descriptors were specified to fit the needs of young migrant learners. Lastly, the CGSL provides guidance on the implementation of dynamic and process-based language assessment, language portfolio assessment, and culturally fair testing.

The recent curricular innovation raised a series of questions related to the teachers' readiness to shift from the traditional knowledge-based instruction usually applied to "monolingual" speakers to the skill-based and effect-driven one proposed by the SL curriculum, to embrace language in all its modalities (listening, speaking, reading, and writing), and, where relevant, to develop further sub-skills, such as phonological awareness. As it can be assumed, the curricular innovation opened up a whole new world for language assessment as well by introducing the process-based and dynamic forms of formative assessment proposed by the CEFR (portfolio, self- and peer-assessment).

To facilitate teachers' professional development, networks and online teacher communities were established. In addition, diagnostic, formative, and summative assessment tools were designed, published, and administered throughout the school year. Since they are relatively new in the Cyprus school setting, it is still unknown how they are perceived, used, and interpreted by the teachers.

3.2. The GDLA Tool

One of the major challenges in implementing the new SL curriculum is the establishment of a sound diagnostic assessment culture among teachers. This is not an easy task to accomplish, not only because teachers may lack SL assessment literacy but also for sociopolitical reasons: if teachers do not hold additional qualifications in SL instruction, nor do they teach in SL classes out of personal choice, the assessment process does not seem to matter; thus, there is little incentive to carry it out in a planned and valid way.

To face this challenge and provide support to the teachers, the Cyprus Pedagogical Institute proceeded with the development of the *Greek Diagnostic Language Assessment* tool (GDLA). GDLA is a collaborative work by developers well-immersed in the Cyprus educational system, designed to identify the strengths and weaknesses of emergent bilinguals from diverse migrant and refugee backgrounds in primary urban and rural schools of Cyprus as to their SL (Greek) communicative and linguistic competence. It consists of two components: one for first graders and one for those who enroll in the second to sixth grade. Each component is skill-oriented and graded according to age-specific literacy competencies, compliant with the CEFR descriptors offered in the curriculum and accompanied by culturally sensitive and age-appropriate illustration.

The GDLA's major innovative feature is the fact that it consists of a distinct component for first graders that takes into consideration their different maturational and literacy needs. This component includes ten tasks and can only be implemented in a private discussion with each child. It assesses the learners' oral competence as well as basic contextualized vocabulary based on the CEFR thematic areas.² The other component for second to sixth graders assess learners' competence and performance in all four language skills. Both components are graded in difficulty and in thematic relevance.³

From its first inception, GDLA was thought of as a dynamic process (and not only as a placement test), supported by elaborated guidelines for teachers who were not in any case experienced and literate in SL assessment processes. Considering that this is the first year that the GDLA tool has been implemented after the publication of the SL curriculum, a wide range of evaluation and feedback research has been planned. Our ultimate goal is to validate the GDLA tool, a process that is still ongoing due to the COVID-19 restrictions and the difficulty of having access to school data. However, investigating the washback effect of the tool as reflected on teachers' perceptions and attitudes was considered to be a crucial and feasible part of the validation process (face validity).

The assessment was administered in September 2020. The task sheets were handed out to learners in print to avoid possible deficiencies in digital skills. The duration of the second–sixth graders' assessment approached 2 h (split into more sessions) for all skills; however, the teachers were free to use time in a way that would best meet the assessment's purpose and minimize fatigue. To score and reflect on the young learners' performance, the teachers received support both from the analytical guidelines that accompanied the tool and the training programs offered by experts/collaborators of the Cyprus Pedagogical Institute.

Both the preceding literature review and the overview of the contextual background justify our collaborative research project by pointing to a gap in previous work.

4. Materials and Methods

4.1. Research Questions

The research questions (RQ) are formulated as follows:

RQ1: What attitudes do primary school GAL teachers hold towards the diagnostic language assessment in general?

RQ2: What are the ways they view the GDLA tool and interpret its features and demands?

RQ3: Are there any effects on teaching and assessing Greek as an SL/AL? Does GDLA encourage or discourage forms of teaching or assessment intended by the developers? Which aspects of teaching are affected (content, methodology, etc.)?

RQ4: Which are the contextual factors that may affect or differentiate the intensity and value of washback?

4.2. Study Design, Instrumentation, and Data Collection

The study followed a mixed-methods approach (quantitative and qualitative) to data collection and analysis. The quantitative data were collected through an online 26-item questionnaire administered to 234 public primary school SL teachers in Cyprus during the 2020–2021 school year. It is worth mentioning that a high percentage of the Cyprus SL teacher population (more than 80%) responded and returned the questionnaire (serving at 234 out of the 274 schools that offer preparatory courses for at least one hour per week). All participants served as language teachers and none of them were involved in the instrument design.

Since no other questionnaire is available, to our knowledge, for investigating the washback effect on teachers' perceptions of diagnostic assessment targeted to young school learners with a migrant background, we had to construct a new instrument, adopting, though, the generic dimensions (a) that are evident in theory and empirical evidence, and (b) of relevant EFL tools (e.g., [Collins and Miller 2018](#)).

The questionnaire was divided into two sections (see Appendix A). The first section, questions 1–6, includes background information: teachers' education, years of experience in SL classes, hours of SL teaching per week, types (first graders, second–sixth graders, both) and numbers of SL classes, and participation in SL teacher training networks (see Appendix A, Part A). The second and main part of the questionnaire includes 20 items which are scored on a 5-point Likert scale (with both numerical and verbal descriptors, e.g., 1-*Never*, 5-*Always*) and require teachers to take a stand towards diagnostic assessment, the GDLA tool, and the ways they perceive its implementation (effect on content, methodology, and assessment, alignment with the curriculum, personal beliefs, see Appendix A, Part B). A 5-point scale was chosen to increase response rate and quality and reduce the respondent's frustration ([Sachdev and Verma 2004](#)). To remove any possible flaws, the first draft was piloted with 20 GAL primary school teachers and revised mainly in wording.

In addition, semi-structured interviews were conducted via Microsoft Teams Calls based on guiding questions (see Appendix B). This qualitative part of the study set out to shed more light on any relationship between the GDLA tool and SL teaching or assessment, and thus to triangulate the findings obtained through the questionnaire. Six SL teachers were interviewed for 30–60 min each. Due to the restrictions of COVID-19 and the teachers' hectic program we used a convenience sample of interviewees who had shown willingness to participate in the study. All informants consented to participate. The interviews were conducted in Greek, the mother tongue of the respondents. Then, they were audio-recorded, transcribed using the Jefferson transcription system, and sent to the informants to receive agreement on their transcribed statements (communicative validity, [Dörnyei 2007](#)). The final versions were submitted for content analysis. The thematic areas of the analysis were drawn both in a deductive (by the interview guiding questions) and in an inductive way (emergent themes from the data). Both the survey and the interviews were conducted in February 2021 so that the effects of the GDLA tool would be integrated into the teaching and learning process and thus be better observed (see also [Green 2013](#)).

Due to the data protection law, the full database (responses and transcribed interviews containing sensitive personal data) cannot be shared. Authority for the research was provided by the Pedagogical Institute of Cyprus (P.I. 7.1.10.3.4./22-2-2021).

5. Results

5.1. Questionnaire

5.1.1. Construct Validity and Reliability

To examine the validity of the instrument and whether it represented all the important aspects of the intended construct, an exploratory factor analysis with varimax rotation was conducted (SPSS, v.27).

The results of the Bartlett's test indicated sampling adequacy ($p < 0.001$, KMO value = 0.88). Factor analysis yielded a 5-factor solution with an explained variance of 75.74%, as portrayed in Table 1. Items that didn't load strongly on a single factor were excluded (*Q.B7* and *Q.B8*, see Appendix A).

Table 1. Construct Validity.

Factors (F)	Items	1	2	3	4	5
F1: GDLA's Washback on Teaching and Assessment	<i>Q.B16</i> Classroom time management	0.884				
	<i>Q.B17</i> Teaching material	0.862				
	<i>Q.B18</i> Differentiated teaching	0.840				
	<i>Q.B19</i> Skills/modalities to emphasize on	0.736				
	<i>Q.B20</i> Language testing and assessment	0.710				
	<i>Q.B14</i> Needs analysis <i>Q.B15</i> Placement	0.626 0.610				
F2: GDLA's Usefulness and Credibility	<i>Q.B11</i> Useful guidelines		0.823			
	<i>Q.B12</i> Valid and reliable scoring		0.781			
	<i>Q.B13</i> Easy scoring		0.771			
F3: Feedback and Importance of Diagnostic Assessment	<i>Q.B4</i> Feedback on skills in the oral modality			0.801		
	<i>Q.B5</i> Feedback on skills in the written modality			0.786		
	<i>Q.B3</i> Importance of diagnostic assessment			0.636		
	<i>Q.B6</i> Implementation of diagnostic assessment			0.600		
F4: GDLA's Alignment with the SL Curriculum	<i>Q.B10</i> Alignment with the SL curriculum descriptors				0.838	
	<i>Q.B9</i> Alignment with the CGSL's programmatic text				0.821	
F5: Motivation in SL Teaching	<i>Q.B1</i> Creativity					0.904
	<i>Q.B2</i> Satisfaction					0.901
Variance Explained (%)		44.72	10.18	7.87	7.39	5.58
Cumulative Variance		44.72	54.90	62.77	70.16	75.74

F1 accounts for 44.72% of the common variance and comprises seven items, i.e., the GDLA's effect on classroom time management (*Q.B16*), teaching material (*Q.B17*), differentiated teaching (*Q.B18*), skills/modalities to emphasize on (*Q.B19*), language testing and assessment (*Q.B20*), needs analysis (*Q.B14*), and placement (*Q.B15*), and so it is labelled *GDLA's Washback on Teaching and Assessment*. F2, which explains 10.18% of the common variance, includes three loaded items that elicit ratings on the usefulness of the guidelines that accompanied the GDLA tool (*Q.B11*), the tool's valid and reliable scoring (*Q.B12*), and its easy scoring (*Q.B13*); thus, it is labelled *GDLA's Usefulness and Credibility*. F3 accounts for 7.87% of the common variance and it contains four more general items: two for the perceived feedback drawn from any diagnostic tool in the oral modality (*Q.B4*) and in the written modality (*Q.B5*), and two for the perceived significance of any diagnostic assessment in GAL contexts: importance of any diagnostic language assessment for the migrant learners (*Q.B3*) and implementation of the diagnostic assessment to all students (*Q.B6*); so it is named *Feedback and Importance of Diagnostic Assessment*. F4 explains 7.39% of the common variance and it groups two items on the tool's compliance with the new CGSL's programmatic text (*Q.B9*) and its descriptors (*Q.B10*), so it is named *GDLA's Alignment with the SL Curriculum*. F5 captures aspects that motivate teachers when teaching SL/AL courses (two items: creativity (*Q.B1*) and satisfaction (*Q.B2*)), and its label is *Motivation in SL Teaching*; this last factor explains 5.58% of the common variance. All five factors support the theoretical framework of the current study since they portray the main aspects of washback on teaching and learning as well as the contextual variables that may have an impact on teachers' perceptions.

Cronbach’s alpha was calculated as a measure of internal consistency. The instrument was found to have high internal consistency (alpha coefficient = 0.90). Alpha coefficients for the five factors were also high, ranging from 0.74 to 0.94.

5.1.2. Background Information

All participants hold a bachelor’s degree in Pedagogy/Education, 44.9% hold a master’s degree in General Education, and only 6% of them hold a master’s degree in Teaching Greek as an SL (Part A, Q.A1). Most participants (65.8%) have been teaching Greek as an SL for one or two years, 22.6% for up to four years, and 11.6% for five years or more, which means that most of them lack experience in the SL classroom (Q.A2). Moreover, it seems that in a significant percentage (65.4%) the informants teach preparatory courses up to six hours per week (Q.A3). Over 85% of the participants teach to learners of all grades (first–sixth) or learners from second to sixth grades, whereas 14.5% of them have only classes of first graders (Q.A4). The majority (57.2%) also stated that they teach Greek as an SL to one or two classes, despite the heterogeneity of learners per school that calls for more classes of bilinguals at different levels of proficiency (Q.A5). Lastly, less than half of the teachers (47.9%) stated that they have participated in the training networks organized by the Cyprus Pedagogical Institute during the last two years (Q.A6). Table 2 provides the background information for the participants of the survey.

Table 2. Respondents’ educational/academic background information.

		<i>n</i>	%
Q.A1 Education	Bachelor’s Degree in General Education	115	49.1
	Master’s Degree in General Education	105	44.9
	Master’s Degree in Greek as an SL	14	6.0
Q.A2 Years of Teaching in SL classes	Up to 2 years	154	65.8
	Up to 4 years	53	22.6
	5 or more years	27	11.6
Q.A3 Hours of SL teaching per week	0 to 6 h	153	65.4
	7 to 13 h	57	24.3
	14 or more hours	24	10.3
Q.A4 Types of SL classes	1st graders	34	14.5
	2nd–6th graders	97	41.5
	1st–6th graders	103	44.0
Q.A5 Number of SL classes	Up to 2 classes	134	57.2
	Up to 4 classes	65	27.8
	5 or more classes	35	15.0
Q.A6 Participation in SL Teacher Training Networks	Yes	112	47.9
	No	122	52.1

5.1.3. Analysis per Factor

In general, the results show the teachers’ quite positive effects of the GDLA tool on the teaching and assessment process, since from all the related factors mean scores higher than 3.50 in a five-point scale are obtained, as indicated in Table 3.

Table 3. Means (*M*) and Standard Deviations (*SD*) per factor.

Factors	<i>M</i>	<i>SD</i>
F1 GDLA’s Washback on Teaching and Assessment	3.56	0.75
F2 GDLA’s Usefulness and Credibility	3.78	0.85
F3 Feedback and Importance of Diagnostic Assessment	4.34	0.62
F4 GDLA’s Alignment with the SL curriculum	3.73	0.77
F5 Motivation in SL Teaching	3.59	0.98

GDLA’s Washback on Teaching and Assessment (F1). As portrayed in Table 3, F1 receives a quite high score (*M* = 3.56). If we examine each item that groups under F1 (see Appendix C for the means and standard deviations per item), we notice that the most intense effects

observed concern the skill-oriented teaching ($M = 3.74$) and the placement ($M = 3.75$) and needs analysis processes ($M = 3.76$). In other words, the GDLA tool has a more positive effect on teachers in planning their lessons so that they include all four skills, in placing their migrant learners to the level-appropriate class, and in diagnosing their strengths and weaknesses, but a less positive one in selecting or designing their teaching content ($M = 3.50$), in differentiating their teaching ($M = 3.44$), and in developing language tests and assessment tools ($M = 3.47$). An even less positive effect of the GDLA tool was observed on classroom time management ($M = 3.32$), which means that GAL teachers report to be less affected by the structure/emphasis of the tool (as suggested by the number and nature of the tasks included in each modality) on the time they dedicate to the development of each skill. It should be mentioned that all items that load on F1 exhibit a high standard deviation close to 1 (see Table A1, Appendix C).

GDLA's Usefulness and Credibility (F2). Moreover, GAL teachers' perceptions reveal a higher appreciation of the GDLA tool in terms of its usefulness and credibility ($M = 3.78$, see Table 3), i.e., its useful guidelines and valid, reliable, and easy scoring.

Feedback and Importance of Diagnostic Assessment (F3). At the same time, they acknowledge that diagnostic assessment (in general) is highly important ($M = 4.34$), as it offers valuable feedback in both modalities, and it must be implemented to all emergent bilingual learners with a migrant background. The low standard deviation ($SD = 0.62$) compared to the other factors is indicative of the agreement of the majority of teachers on the importance of the diagnostic language assessment.

Alignment with the SL Curriculum (F4). Despite the observed very high appreciation of the diagnostic assessment in AL teaching contexts where migrant students are engaged, GDLA's alignment with the SL curriculum appears to be less high ($M = 3.73$) but higher than its washback effect on teaching and assessment.

Motivation in SL Teaching (F5). Lastly, the results reveal interesting insights into the teachers' personal beliefs and more specifically into their motivation (creativity and work satisfaction) while teaching Greek as an SL. In particular, they seem to be quite motivated and content with their profession ($M = 3.59$) but with a high standard deviation close to 1.0, as illustrated in Table 3.

The previous findings raise some interesting issues on the perceived value of both diagnostic assessment in general and the GDLA tool in specific. First, the implementation of diagnostic assessment targeted to young migrant learners is undoubtedly welcomed by the Cypriot GAL teachers (F3). Second, the features of the specific diagnostic tool (F2) and its compliance with the new SL curriculum (F4) receives quite a high appreciation. Third, the results allow us to speak of a quite strong and positive washback effect of the GDLA tool on teaching and learning (F1), with skill-oriented teaching, placement, and needs analysis being the most affected aspects. However, both GDLA's usefulness and credibility and its alignment with the SL curriculum/its accompanying descriptors score higher than its impact on the participants' teaching and assessment practices. Lastly, GAL teachers' motivation (as reflected by their creativity and satisfaction responses) seems to problematize; though a quite high score is obtained, the high standard deviation indicates that the data points are spread out over a wider range of values, revealing both highly and poorly motivated teachers.

Between-group comparisons and correlations. Furthermore, a set of between-group comparisons is reported so that we find possible differences in the respondents' perceptions that are due to their demographic/biographical features. To check the normality distribution of the variables, the Kolmogorov–Smirnov statistical test was performed. The results obtained suggested that non-parametric tests should be applied.

Some very interesting findings on the differentiating factors of washback emerge from between-group comparisons. As indicated in Figure 1a, a more intense positive effect of the GDLA tool on teaching and assessment is observed in case the hours of teaching in SL classes increase ($\chi^2(2) = 6.911$, $p = 0.032$, mean rank scores: 109.47, 136.61, and 123.31); however, this effect drops a little when 14+-hour courses are assigned to the teachers. The

same holds for the perceived usefulness and credibility of the GDLA tool ($\chi^2(2) = 8.80$, $p = 0.012$, mean rank scores: 108.43, 138.53, and 125.40), as illustrated in Figure 1b.

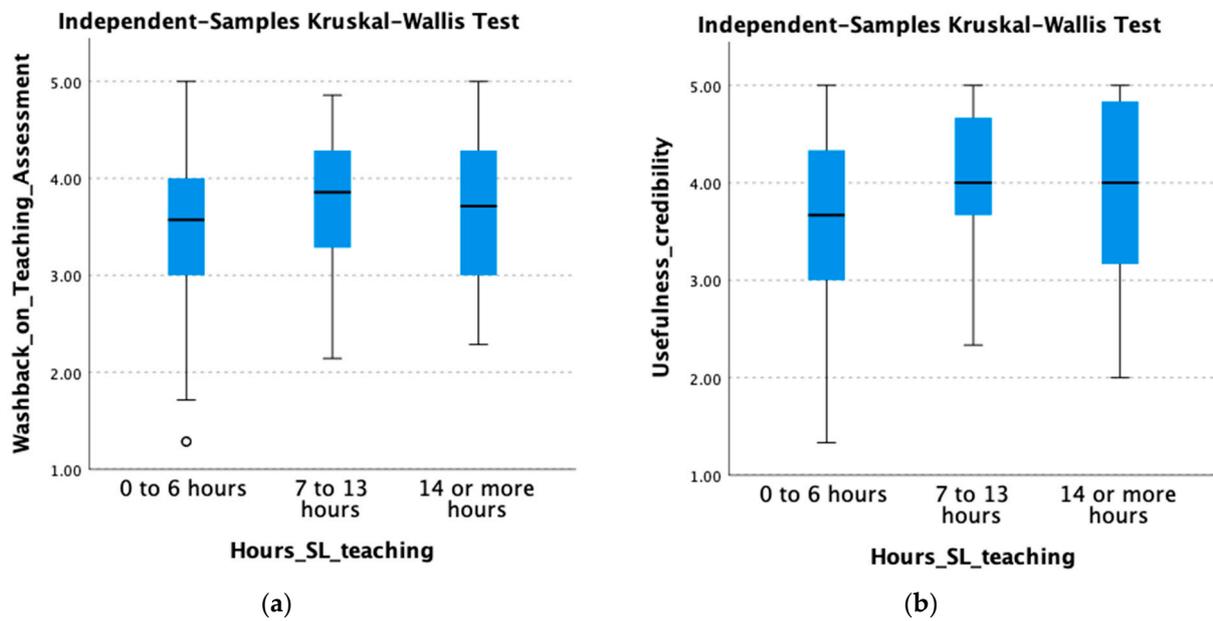


Figure 1. Between-group Differences: (a) Washback on Teaching and Assessment * Hours of SL Teaching; (b) Usefulness and Credibility * Hours of SL Teaching.

Moreover, a steady increase in positive effects is observed when it comes to the GDLA curriculum alignment with regard to (a) the hours of SL teaching ($\chi^2(2) = 10.57$, $p = 0.005$, mean rank scores: 107.67, 134.04, and 140.92), and (b) the numbers of the SL classes offered ($\chi^2(2) = 6.38$, $p = 0.041$, mean rank scores: 108.35, 129.40, and 130.43). The differences are illustrated in Figure 2a and 2b respectively.

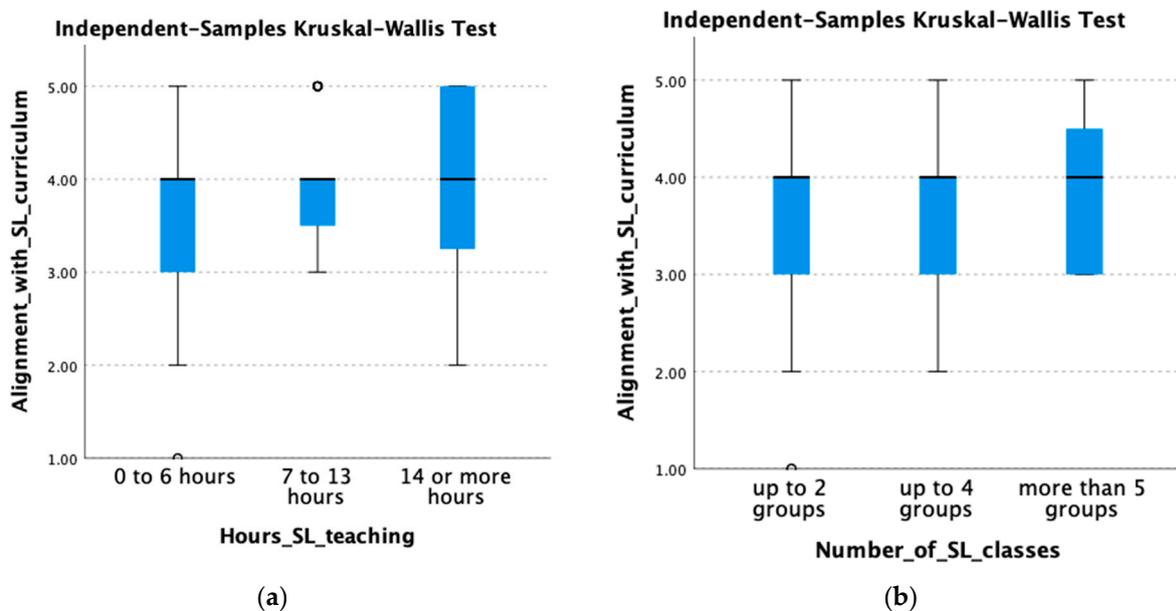


Figure 2. Between-group Differences: (a) Alignment with the SL curriculum * Hours of SL Teaching; (b) Alignment with the SL Curriculum * Numbers of SL Classes.

Motivation also appeared to vary when examined by some of the respondents' biographical features. Interestingly enough, it appeared to be higher for those who have

participated in SL teacher training networks ($M = 3.83$) during the past two years compared to those who haven't ($M = 3.37$). A Mann–Whitney test indicated that this difference was statistically significant, $U(N_{yes} = 112, N_{no} = 122) = 8658.00, z = 3.589, p < 0.001$ (see Figure 3a). The difference in work satisfaction and creativity was also significant in relation to the types of classes the respondents teach, as indicated by the Kruskal–Wallis test, $\chi^2(2) = 19.823, p < 0.001$, with a mean rank score 83.99 for those who only teach to first graders, 108.05 for those who teach to second–sixth graders, and 137.46 for those who teach in both types of classes (see Figure 3b). We should also mention that a positive moderate correlation between *washback on teaching and assessment* and *motivation* was observed ($r = 0.346, p < 0.001$).

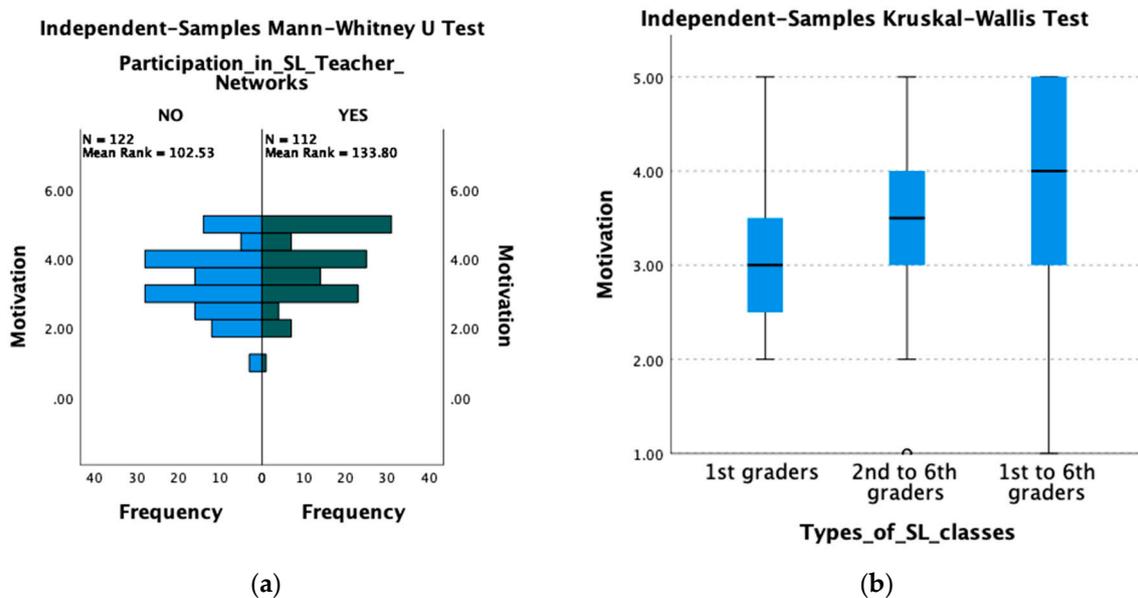


Figure 3. Between-group Differences: (a) Motivation * Participation in Teacher Training Networks; (b) Motivation * Types of Classes.

The previous findings raise further issues. First, the hours of teaching seem to matter as they differentiate the GDLA’s washback on teaching and assessment, its perceived usability and credibility, and its alignment to the SL curriculum and syllabus. Second, participation in training programs appears to increase motivation. Third, the number and types of preparatory classes turn out to be significant differentiating factors of the GDLA-curriculum alignment and the teachers’ motivation, respectively.

5.2. Interviews

The interviewees’ teaching experience ranged from one to eight years in SL classes. All of them hold a master’s degree in General Education and they attended the training seminars organized by the educational agency. Most of them have been GAL teachers for 2 or more years and they are employed in more than two preparatory classes for at least 7 h per week, as seen in Table 4.

Table 4. Interviewees' background information.

		<i>n</i>	Interviewees(I)
Education	Bachelor's Degree in General Education Master's Degree in General Education Master's Degree in Greek as an SL	6	I1, I2, I3, I4, I5, I6
Years of Teaching in SL classes	Up to 2 years	3	I3, I6
	Up to 4 years	2	I2, I4, I5
	5 or more years	1	I1
Hours of SL teaching per week	0 to 6 hours		
	7 to 13 hours	4	I1, I2, I3
	14 or more hours	2	I4, I5
Types of SL classes	1st graders		
	2nd–6th graders		
	1st–6th graders	6	I1, I2, I3, I4, I5, I6
Number of SL classes	Up to 2 classes	1	I2
	Up to 4 classes	3	I1, I3, I6
	5 or more classes	2	I4, I5
Participation in SL Teacher Training Networks	Yes	6	I1, I2, I3, I4, I5, I6
	No		

The interviews helped in triangulating the questionnaire responses and they also served as instances for reflection both on the teachers' background information/personal beliefs and on their influence by the GDLA tool. Some translated excerpts from the interviews are cited below, classified by thematic areas drawn from the analysis.

Motivation. The teachers' motivation (work creativity and satisfaction) when teaching Greek as an SL appears to vary a lot. Despite the differences observed in how intense the feelings of creativity and satisfaction were and how interlinked with the feelings of anxiousness, professionalism, and responsibility, a general positive effect is observed:

It's the love the kids give you. It's the immediate results you can have (.). It's what the learners share, they say "I learned that from you", the love, ↑<the results, what you see in children> (I2)

When you teach for the first time, it's kind of a stressful procedure (.). I'm telling you the truth about me and how it fits my personality. Of course, you get moral satisfaction from the learners and their progress. I believe if I do it for a second year in a row the satisfaction or the creativity would be greater. I need to first clear things in my head, understand the procedure (.). and then I can be creative too. (I3)

In some cases, the teachers report that they lack motivation and are unable to experience satisfaction or have space for their creativity to flourish because of the institutional difficulties they face, such as the limited time they have for teaching GAL learners:

↑The groups (.). in this case, we had a big problem. The available teaching hours were very few (. . .) for example, I had eight teaching hours per week, and I used five of them in the A1 group. But what was left for the A2 one? How could I work through the descriptors and cover the communication themes of the curriculum? When you experience such a stressful situation, I think that creativity is reduced. (I1)

Two teaching hours per week! (. . .) the teaching hours were the minimum (. . .) I had the students for two hours per week (.). but of course they were not enough. (I4)

Importance of diagnostic language assessment. Most teachers seemed quite convinced of the importance of diagnostic language assessment and the overall information they received related to teaching and assessing in SL classes:

Diagnostic assessment is of pivotal importance for young migrant learners: It shows you the way. You gain valuable feedback = especially in the oral modality. (I2)

GDLA's usefulness and credibility. As for the tool's usefulness and credibility, teachers generally acknowledge that the GDLA is equipped with useful guidelines, valid and

reliable scoring, and provides an easy scoring scheme. However, some teachers mentioned their need for further training:

The tool is well-structured and so are its guidelines. (I2)

↑I think I had a lot of help from the tool (.). I mean the rapport I developed with the students during the speaking component had helped me a lot to understand each one's personality. (I3)

I now turn back some months ago and I realize that the tool gave me a clear picture of the students' strengths and weaknesses. (I4)

That was a fair procedure with a <carefully planned scoring>. ↑ However, reflecting on the students' written productions gave me a hard::: time. I didn't know how to do it successfully (.). I need more detailed guidelines, more training on that. (I1)

We should mention, though, that in several cases some misconceptions were identified regarding the purpose of the GDLA tool or of diagnostic assessment in general. Teachers' comments revealed that through the use of the GDLA tool they were assessing students' learning based on benchmarks provided by the CEFR descriptors instead of approaching diagnostic assessment as a way to collect data on how much their students have developed each language skill at that particular point in time:

The kids could not respond to the tool's activities at the beginning of the year. It took them more time (. . .) we did not have such problems during the formative or summative assessment (.). At the final assessment things happened faster. (I2)

We had problems in the communication themes that the students had not yet been taught. For example, the weather forecast theme (.). That was difficult for them. (I2)

We implemented the test at the ↑beginning of the school year (.). and so we could see what some students remembered from the previous year. Summer holidays were in between (.). Maybe they needed some revision courses (. . .) maybe that was unfair for them. (I4)

I4: There was not enough time for the test (. . .) some children read in <a very slow pace>, some of them spelled out the words (.)

Researcher: Wasn't that indicative of their needs, though?

I4: It is a parameter, ↑ yes. I wrote that down back then (.). it is a parameter that I had to consider.

GDLA's Washback on Teaching and Assessment. As it emerges from the discussions with the teachers, in most cases a beneficial, strong, and positive washback of the GDLA tool is observed on teaching and assessment. The teachers report that the GDLA tool offers them useful feedback, sometimes in areas that they could not predict when they lean on other means (e.g., their experience or other tests) to do a principled *needs analysis*:

I now see my notes <after scoring and reflecting on the learners' performance> (.). and they give me information that might not come to my notice. (I2)

They also pointed out that the skill-based structure helped them in making crucial decisions on the formation of their SL groups at different levels of proficiency (*placement*) or in understanding the needs of each learner for future development, without relying on unreliable means, such as their intuition:

The tool gave us information, for sure. It helped us not to form groups based just on our intuitive judgements. (I2)

I could look at the results and say "↑Ok, good, here <in the reading section> some students could manage the task (.). and some other kids could not read at all: =so it helped me in forming my level-appropriate groups. (I4)

The four skills that we emphasize on were very clear (.). so we had help from the tool on that, to have a vivid picture of where the students are in listening and speaking or in reading and writing (. . .) the tool's contribution was crucial and it definitely helped us

a lot to proceed with a fair placement of the learners in the appropriate groups for the preparatory courses. (I2)

The interviews gave us a more detailed picture of the tool's impact on the *skill-oriented teaching*, the modalities to emphasize during their teaching, and on how to perform teaching meso planning during the school year:

I could give extra lessons to a student of mine, if I noticed from the tool that she needed help in listening skills (.) ↑I knew I needed to plan more listening activities. (I2)

The tool was a BIG help:: Till now, I used to plan my teaching in a theme-oriented way (.) The skills didn't really matter. I could have several lessons in a row teaching vocabulary and grammar (.) < e.g., clothes and accessories and how to build simple sentences on shopping>. =And as for listening (.) I was content with the teacher–student exchange within the classroom (.) I am involved in a change process now I search for age- and level-appropriate material to include it in listening tasks (.) real and authentic listening tasks:: (I3)

The GDLA tool seems to have a strong impact on the *teaching material* for some teachers (which text types to include in their teaching, which task types to avoid, and which others to use more extensively), but not so strong for others, whose decisions on teaching material did not seem to be influenced by the tool's content:

The tool gave me guidance on what activities and tasks to choose or plan (.) I moved forward one step at a time. (I4)

Yes, the tool guides me in preparing my lessons as well (.) Till now I used extended narratives and drill and practice activities(.) but ↑content matters. The tool has dialogues, posters, invitations, and other text types. (I5)

I can't say I looked at the test content (.) no, no (.) we weren't so far away from what the test contained (.) the test did not determine how I was going to handle the content (.) I didn't have the test as a model in order to decide on the material. (I2)

Few teachers mentioned that the GDLA tool could guide them in *designing their own tests or assessment*:

(. . .) I think, being based on this tool, I could design my own tests and plan other assessment processes for the evaluation of each thematic unit I teach. (I5)

Relatively few teachers reported an impact of the GDLA tool on differentiating instruction, and how this could be performed by including a purposefully chosen sequence of a range of activities of graded difficulty:

The activities are of graded difficulty (.) They function as a compass in grading the difficulty of the tasks we plan so that all students' needs are met. (I5)

The test is structured in a hierarchical way, progressing from the simple to the more difficult activities (.) I tried the same with my classes. (I3)

GDLA's alignment to the SL curriculum. Most teachers point out a clear assessment-curriculum alignment as well as the importance of this alignment in terms of them feeling confident and consistent in their teaching:

The test is based on the new curriculum (. . .) that means there is consistency and continuity. (I3)

I must admit that having such a material available, I felt safe. (.) There was a textbook, a guide, a curriculum to rely on; (.) I was confident to "transmit" the relevant knowledge and develop the appropriate skills. (I3)

6. Discussion

The results of this study provide us with valuable feedback on the way the participants view diagnostic language assessment. With regard to RQ1, both quantitative and qualitative

data reveal that GAL teachers afford great importance to this type of assessment mainly because it informs targeted intervention for learners with a migrant background in both modalities (oral and written); thus, it is seen as a prerequisite to customized needs-oriented courses. In proof of that, F3 *Feedback and Importance of Diagnostic Assessment* was the one with the highest mean score, close to 4.5. This finding is crucial to the scope of this study, since, as put by Green (2013, p. 40), “the importance afforded to a test has traditionally been regarded as the motivating force that drives washback” and determines its intensity. Information gleaned from the interviews also supports the quantitative data.

The second research question (RQ2) sought for more specificity as it addressed the GDLA’s perceived usefulness and credibility. This was also a highly appreciated factor (F2) but less appreciated than F3 ($M = 3.78$). Interview results are in line with that, since GAL teachers pointed out in many ways that the GDLA tool (a) gives them feedback on the skills that need to be developed, (b) is well-structured, (c) covers a wide range of communication themes, (d) provides a window into the learners’ personality and cultures, (e) is easy to score and valid, and (f) it is accompanied by detailed guidelines. In addition to that, they commented on the usefulness of the tool and its skill-based structure in taking crucial decisions on how to place the learners in SL classes of different proficiency levels or in understanding their needs.

However, the positive perceptions were laden with misconceptions on the purpose and implementation of the GDLA tool as reflected on the teachers’ comments. This might be a plausible explanation for the lower mean score on their ratings. Specifically, some of the interviewees raised questions on whether it would be fair to assess the language minority learners at the beginning of the school year, implying or directly stating that they perceived the whole process as a summative evaluation, a mid- or end-term test. As a result, they treated it as an administrative task of interest only with regard to its scoring results instead of focusing on what the learners can do and at the same time collect information to synthesize an action plan based on topics and skills they have not yet mastered. While alarming, this finding comes as no surprise since diagnostic assessment has not been mandatory in Cyprus primary education until recently and no other diagnostic tools were designed and administered. Up until now teachers have tended to rely on their intuitive judgments or on previous reports from colleagues. Some of them expressed their concerns on the communication themes included in the tool that were not previously taught while others on the time allocated for its implementation, pointing to a knowledge-based approach to diagnosis, with a clear impact from first language or *teaching to the test* practices, since their main concern was to raise the learners’ scores or to capture their progress. In many cases, GAL teachers admit that they lack competencies in performing the assessment or in interpreting the learners’ performance and they ask for even more detailed guidelines (especially in writing assessment) and for training courses.

The observed lack of knowledge agrees with numerous previous studies (see Section 2), but the teachers’ positive attitudes towards the diagnostic tool do not (Jimola and Ofodu 2019). These positive attitudes allow for further interpretation—beyond the test itself—based on this specific educational context. First, the assumed low-stakes of diagnosis free teachers from the constraints of other proficiency exams, e.g., summative end-term tests. Second, GAL teachers are rarely (if ever) assessed on their performance in preparatory classes. This is due to the low estimation in which teaching effectiveness is held by many school administrators where learners with a migrant background are concerned, which leads to lowering expectations. However, this fact perhaps lifts a great deal of “educational policy-musts” from the teachers’ shoulders and allows them to see the benefits of diagnosis, to focus on how to perform better at their jobs, and to crave training and overall guidance.

At the same time, the limited awareness of the GDLA’s main purpose raises issues related to the teachers’ assessment literacy and more specifically the expertise required to reflect on the results of a diagnostic tool. As inferred from both the quantitative and the qualitative data, GAL teachers felt the need to implement the GDLA tool, but they did not know how to interpret the gained information. There is also a strong possibility that,

since they were not fully assessment literate, the participants relied on previous assessment practices (Liu and Li 2020) experienced in the broader educational context of Cyprus and restricted on summative assessments, hence the misconceptions.

Addressing the cognitive and linguistic needs of young language minority students is an endeavor that poses a great challenge to GAL teachers in practice: it emphasizes the need for their own self-investment as a strategy to advance the development of their professionalization. From this study's findings, it is made clear that there is a dire need for the development of their assessment literacy. It should be mentioned, though, that to increase language assessment literacy, a multilayered entity as put by Inbar-Lourie (2013), GAL teachers should first improve their expertise in SL teaching in general and get acquainted with level-appropriate content, methodology, and strategies.

Having discussed the perceived importance of diagnostic assessment in general and GDLA's usefulness and credibility specifically, we move on to explore the washback effects on teaching and assessment which lie at the heart of this study (RQ3). Overall, the results of the survey suggest that there is a beneficial, positive, and quite strong washback of the GDLA tool on teachers' practices (F1, $M = 3.56$). As reflected in the interviews, the mean score of this effect could have been higher, provided that the teachers were more assessment literate and informed on SL teaching theory and practice. More specifically, the most affected dimensions are the placement/needs analysis processes and the skill-based teaching. However, it was noted that the washback effect was less positive when it came to differentiating instruction and designing their own teaching material and assessment tools. For reasons of clarity, each item that loaded on this factor is being discussed separately.

A positive washback effect on needs analysis and placement was supported both from the quantitative (mean scores close to 3.8) and the qualitative data. Teachers' comments expressed a clear message: that the GDLA tool shows them the way to make evidence-based choices throughout the teaching–learning process. This was quite expected since the tool was administered at the very beginning of the school year. The interlinked relation between a diagnostic tool and the placement process is also evident from the ongoing discussion in the SL/FL literature, reflected in Alderson (2005) who claims that there is “no clear distinction . . . between diagnostic tests and placement tests” (p. 12). However, when asked to give more details on their perceived impact of the tool on needs analysis, most teachers insisted on the benefits of appropriate placement and none of them mentioned a potential impact on designing and implementing dynamic diagnosis in various communication themes throughout the school year.

Teaching content (*what*) appeared to be one of the less (though positively) affected dimensions by the GDLA tool ($M = 3.50$), whereas skill-based teaching (*how*) was one of the most affected ones ($M = 3.74$). This unexpected finding does not seem to agree with previous research where the main effect was found on content (i.e., Alderson and Wall 1993; Alderson and Hamp-Lyons 1996; Watanabe 1996; Min and Park 2020). Further insights to this aspect are provided by the interviews. In their overwhelming majority, GAL teachers mentioned that the GDLA tool guides them in planning their courses so that all four skills are equally developed. Given the progress made in the field of applied linguistics and the insights gained during the past years, such a perception might point to a common practice in SL teaching. However, skill-based instruction has not been a well-established practice in GAL courses for young migrant learners in Cyprus until recently. This comes as no surprise if we consider that the preparatory courses are offered by general education teachers recently immersed in a multilingual teaching–learning setting that raises different expectations. This is also why, when they were asked about the GDLA's effect on teaching content, they confessed that they used to favor monomodal/continuous text types and narrative genres (under the influence of first language teaching practices) as well as drill and practice activities as compared to the communicative dialogic tasks suggested by the tool.

The less positive washback effect on differentiated teaching evident in the quantitative results ($M = 3.44$) was implicitly detected in the interviews. Teachers were able to directly

refer to the tool's structure and organization, claiming that it could potentially enable them to apply differentiation in their own teaching. They also reported on the linkage of the GDLA tool to the SL curriculum and how this is further enhancing the idea of differentiated teaching. This linkage agrees with [Brown and Hudson's \(1998\)](#) claim that when the assessment procedures are aligned to the course goals and objectives (curriculum), a positive washback effect occurs. Since both the GDLA tool and its accompanying rating scales (as well as the SL curriculum and its descriptors) cater for the varied learning needs of this group of students, and this is clearly acknowledged by the participants of the study, we can only assume that they are still not professionally ready to implement differentiated teaching, which calls for further training. So, we could safely claim that once again it is found that the teachers and the overall context mediates the extent of the washback effect.

The least positive effect of the GDLA tool was reported on the time dedicated within the SL classroom for the development of each skill ($M = 3.32$). This finding might be attributed to the recent—but not yet fully developed—familiarization of GAL teachers with skill-oriented teaching and its prerequisites and perspectives.

Both quantitative ($M = 3.47$) and qualitative findings report a less positive washback effect of the GDLA tool on designing other assessment tools and processes. It should be mentioned, though, that the preparatory classes for learners with a migrant background are the only courses delivered in the Cyprus primary education context where official diagnostic, formative, and summative assessment tools are provided by the Cyprus Pedagogical Institute. The lack of teaching experience in SL classrooms in combination with the gaps in language assessment literacy might again explain this less positive effect.

F4, which depicts GDLA's alignment to the curriculum, received a quite high mean score ($M = 3.73$). Most GAL teachers could comment on the relevance between the curriculum and the diagnostic tool both in content and in methodology. This finding reveals that the curriculum functioned as an important resource for guiding teachers while making them feel safe for their reflections on GAL students' performances during diagnosis. However, this is not the case in high-stakes exams research where curriculum and teaching seem to be narrowed to the material of the tests ([Menken 2017](#)). The findings of this study indicate quite the opposite: the obvious GDLA-curriculum alignment strengthened the teachers' consciousness and confidence in their new roles. They also advocate a different approach to *teaching to the test practices*, one where a low-stakes diagnostic tool leads the way to better and fairer teaching and assessment. In fact, as there is an overall positive washback of GDLA on teaching, in contrast to other studies in primary education (see Section 2), then in our case *teaching to the test* might be seen as something beneficial, stimulating instructional reform, as long as the test represents the curriculum.

So far, we have examined GDLA's effects on teaching and assessment in Cyprus GAL primary school classrooms as reflected on teachers' perceptions. A constant question hanging over this discussion is whether the obtained effects can be solely attributed to the tool itself or they emerge from a composite of several factors. Early enough, [Cheng and Curtis \(2004\)](#) pointed out that "the relationship between testing, teaching and learning does appear to be far more complicated and to involve much more than just the design of a "good" assessment" (p. 16). [Tsagari \(2011\)](#) adds that "there is not always a linear relationship between the design of a test and the teaching and learning that takes place in the classroom" (p. 431). Building on the previous concerns, we move on to the discussion of the contextual factors that may affect or differentiate the intensity and value of washback in our study (RQ4).

As indicated in Section 4, a more positive effect of the GDLA tool on teaching and assessment is observed in case the hours of teaching in SL classes increase, but interestingly this effect drops a little when 14+-hour courses are assigned to the teachers. One first explanation might be that when the teaching hours increase the working load is also heavier, as an increased number of GAL students have to be assessed. This explanation becomes even more favorable if we take into consideration that GAL teachers often find themselves in an awkward and stressful position when their mainstream classroom colleagues are reluctant

to cooperate and facilitate students' participation in preparatory classes or assessment processes. In addition, the official diagnostic assessment is carried out at the beginning of the school year when the staffing of the schools is not yet completed; as a consequence, GAL teachers are usually asked to substitute the teachers of the mainstream classrooms that have not yet been appointed and thus the implementation of the diagnostic assessment as well as the SL lessons lag behind. The same pattern of results holds for the perceived usefulness and credibility of the GDLA tool.

One of the most important (implicitly related to washback) findings in our study is the interrelation between teachers' motivation and their participation in training courses. Motivation (F5) appeared to be higher for those who have participated in SL teacher training networks compared with those who have not. In support of that, the interviewees of this study, who had been trained, expressed quite strong motivational feelings. This means that trained teachers exhibit more satisfaction and creativity about their professional role in multilingual classes. The difference in work satisfaction and creativity was also significant in relation to the types of classes the respondents teach. Teachers who teach in both types of classes (first graders and second–sixth graders) had the highest mean ranks maybe because they have a more complete picture of both the stages that GAL students pass through and the curriculum's gradation.

One last interesting finding is the steady increase in the GDLA-curriculum alignment mean scores with regard to (a) the hours of SL teaching and (b) the numbers of the SL courses offered. One plausible explanation traced in the interviews is that as involvement in GAL teaching increases, teachers seek for more information on SL teaching; hence, they consult the new curriculum, they are experientially engaged with its principles, and they see them reflected on the diagnostic tool. We should mention, though, that a circular issued by MOECSY in 2020 encouraged teachers, especially those who taught for 10+ hours, to participate in the SL teacher training networks. The data of this study confirm the previous observation: trained teachers were employed for many hours and in various SL classes. It goes without saying that this group of participants in our study was more capable of identifying the GDLA-curriculum alignment as they had received detailed training that highlighted their content, structure, and potential.

7. Conclusions

This study attempts to fill a gap in the literature by exploring the washback effects of diagnostic assessment on teaching and learning targeted to GAL young learners with a migrant background. A mixed-methods approach was employed that captured both qualitative and quantitative insights from GAL teachers in Cyprus regarding (a) the importance of diagnostic assessment in multilingual primary school settings, (b) the usefulness and credibility of the specific diagnostic tool (GDLA), (c) GDLA's potential washback on teaching and assessment, and (d) the contextual factors that may be responsible for washback intensity and variability.

The implementation of the diagnostic assessment was taken as a way of introducing teachers into the educational change brought about by the new Greek as an SL curriculum, launched in September 2020. Such an alignment was not meant as a control policy but as a preliminary implicit process to initiate general education teachers into practices that differentiate SL/AL learning from first language development and FL learning and to encourage behaviors compatible with the aims of the tool that reflect on the new SL curriculum.

The findings point to the teachers' positive perceptions towards the GDLA tool. The most positive stances were adopted with respect to the contribution of the GDLA tool in placement and needs analysis whereas the most intense positive washback effect was found on skill-based teaching. These results are understandable on the basis of the limited expertise and experience that novice GAL teachers in Cyprus have. In addition, the low stakes of this educational context, the lack of teachers' training, the management/teacher

appointment practices, and the observed gaps in assessment literacy might explain the positive but not extremely intense effects found.

These results provide multiple implications for teachers, test takers, and stakeholders not only in Cyprus but also around the world, indicating a more coherent linkage between language planning and integrating policies, instruction, learning, and assessment in diverse school settings. More specifically, this phenomenally low-stakes context provides evidence of language assessment training needs. This training should focus on how SL teachers interpret diagnostic assessment results (and assessments results in general), on SL teaching theory and practice, and how these are represented (or not) in each country's SL curriculum. Policy makers, apart from the predominant realization of teachers' training needs, can also focus their future decision-making on how to best mediate teachers' knowledge in diverse student populations instead of allowing constant teachers' transfers in multiple teaching contexts.

This study adds to empirical research that investigates the link between low-stakes assessment and positive washback. However, it should be acknowledged that despite the satisfactory number of participants in the quantitative part of the study (almost 80% of the Cyprus overall SL teacher population during the 2020–2021 school year), a relatively small number of participants were included in the qualitative part, mainly because of the restrictions caused by COVID-19 and teachers' hectic program. It should be mentioned that follow-up tests are being planned for the 2021–2022 school year, during which more SL teachers' voices will be heard and investigated. Despite this study's innovations and due to its limitations, it is recognized that a great deal of work remains to be performed: (1) research on diagnostic assessment tools, which take into consideration migrants students' full linguistic repertoire (Garcia and Li 2014) and teachers' ability to utilize translanguaging and to reconsider assessment through a multilingual lens; (2) longitudinal empirical studies that test washback's length; (3) further investigation into the relationship between diagnostic, formative and summative assessment, which will allow us to fully explore language minority students' trajectories; and (4) research conducted on the effects of assessment on learners.

To sum up, Shohamy (2001, p. 15) speaks of "the power of tests" and how tests can create "winners and losers, successes and failures, rejections and acceptances". The construction of the GDLA tool in addition to this study's findings provides a solid evidence-based beginning of a fairer educational chance being provided to this particular group of students.

Author Contributions: Conceptualization, M.M., N.K., and D.K.; methodology, M.M., N.K., and D.K.; formal analysis, M.M., N.K., and D.K.; investigation, M.M., N.K., D.K., and C.G.; resources, M.M., N.K., D.K., and C.G.; data curation, M.M., N.K., D.K., C.G., and P.H.; writing—original draft preparation, M.M., N.K., and D.K.; writing—review and editing, M.M., N.K., and D.K.; visualization, M.M., N.K., and D.K.; supervision, M.M.; project administration, M.M.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Pedagogical Institute of Cyprus (P.I. 7.1.10.3.4./22-2-2021).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding authors. The data are not publicly available due to fact that responses and transcribed interviews contain sensitive personal data.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Questionnaire

Part A

Q1. Education:

- Bachelor's degree from a Pedagogical Department
- Master's degree in Pedagogy or Didactics
- Master's degree in Teaching Greek as a Second Language

Q2. Years of teaching experience in preparatory courses of Greek as a second language:

- up to 2 years
- up to 4 years
- 5 years and more

Q3. Teaching hours of Greek as a second language per week during this school year:

- 0–6
- 7–13
- 14 and more

Q4. I teach preparatory classes of Greek as a second language to:

- 1st graders
- 2nd, 3rd, 4th, 5th, and 6th graders
- 1st to 6th graders

Q5. During this school year I have set up ... preparatory courses of Greek as a second language:

- up to 2
- up to 4
- 5 and more

Q6. I participated in the Teacher Training Networks for Greek as a second language.

- Yes
- No

Part B

Q1. Teaching Greek as a second language in preparatory classes gives space to my creativity.

- 1 (not at all) 2 (not really) 3 (somewhat) 4 (very) 5 (very much)
-

Q2. Teaching Greek as a second language in preparatory classes gives me satisfaction.

- 1 (not at all) 2 (not really) 3 (somewhat) 4 (very) 5 (very much)
-

Q3. I consider the provision of a diagnostic tool during the first contact with migrant students:

- | | | | | |
|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| 1 | 2 | 3 | 4 | 5 |
| <input type="radio"/> |
| Not at all | | | | Very |
| important | | | | important |

Q4. A diagnostic tool is useful to provide feedback on oral skills.

- 1 (not at all) 2 (not really) 3 (somewhat) 4 (very) 5 (very much)
-

Q5. A diagnostic tool is useful to provide feedback on written skills.

1 (not at all) 2 (not really) 3 (somewhat) 4 (very) 5 (very much)

Q6. I apply diagnostic assessment to all migrant students.

1 (always) 2 3 4 5 (never)

Q7. During the implementation of diagnostic assessment, I provide students with additional guidelines, each time they cannot understand the aim of a task.

1 (always) 2 3 4 5 (never)

Q8. The time needed to apply the GDLA tool was the one given in the guidelines.

1 (a lot less) 2 3 4 5 (a lot more)

Q9. The GDLA tool is aligned with the new curriculum for Greek as an SL.

1 (not at all) 2 (not really) 3 (somewhat) 4 (very) 5 (very much)

Q10. The GDLA tool is aligned with the curriculum's accompanying descriptors.

1 (not at all) 2 (not really) 3 (somewhat) 4 (very) 5 (very much)

Q11. The GDLA tool includes useful implementation guidelines.

1 (not at all) 2 (not really) 3 (somewhat) 4 (very) 5 (very much)

Q12. The GDLA tool includes a valid and reliable scoring system for the students.

1 (not at all) 2 (not really) 3 (somewhat) 4 (very) 5 (very much)

Q13. The GDLA tool includes an easy scoring system for the educator.

1 (not at all) 2 (not really) 3 (somewhat) 4 (very) 5 (very much)

Q14. The GDLA tool contributes to highlighting the strengths and needs of the students.

1 (not at all) 2 (not really) 3 (somewhat) 4 (very) 5 (very much)

Q15. The GDLA tool contributes to the successful placement of students in the respective preparatory classes based on their level of Greek proficiency.

1 (not at all) 2 (not really) 3 (somewhat) 4 (very) 5 (very much)

Q16. The GDLA tool helps me determine the duration of the teaching I dedicate to each skill.

1 (not at all) 2 (not really) 3 (somewhat) 4 (very) 5 (very much)

Q17. The GDLA tool guides me on my teaching material (content, worksheets, tasks, tools, etc.).

1 (not at all) 2 (not really) 3 (somewhat) 4 (very) 5 (very much)

Q18. The GDLA tool helps me to apply differentiated learning instruction.

1 (not at all) 2 (not really) 3 (somewhat) 4 (very) 5 (very much)

Q19. The GDLA tool shows me on which skills to focus during my teaching.

1 (not at all) 2 (not really) 3 (somewhat) 4 (very) 5 (very much)

Q20. The GDLA tool shows me how to best design language tests and assessments.

1 (not at all) 2 (not really) 3 (somewhat) 4 (very) 5 (very much)

Appendix B

Guiding questions for semi-structured interviews

1. Do you feel content or creative when teaching Greek as a second language in preparatory classes? Yes or no? Why do you think you feel that way?
2. Do you think that using a diagnostic tool is important when teaching migrant students? Yes or no? Why?
3. Which language skills are important to receive feedback on when using a diagnostic tool?
4. Do you see an alignment between the GDLA tool with the new SL curriculum and its accompanying descriptors? If yes, where do you see this alignment?
5. How do you find the implementation process of the GDLA tool?
6. How do you find the scoring system of the GDLA tool?
7. Do you believe that the GDLA tool may influence your teaching in preparatory classes? Yes or no? If yes, in what ways?
8. Do you believe that the GDLA tool may influence the assessment forms you use in the preparatory classes? If yes, in what ways?

Appendix C

Table A1. Means and Standard Deviations per item.

Factors (F)	Items	Mean	SD
F1: GDLA's Washback on Teaching and Assessment	Classroom time management	3.32	0.96
	Teaching material	3.50	0.92
	Differentiated teaching	3.44	0.93
	Skills/modalities to emphasize on	3.74	0.86
	Language testing and assessment	3.47	0.89
	Needs analysis	3.76	0.89
F2: GDLA's Usefulness and Credibility	Placement	3.75	0.90
	Useful guidelines	3.87	0.89
	Valid and reliable scoring	3.66	0.94
F3: Feedback and Importance of Diagnostic Assessment	Easy scoring	3.83	0.94
	Feedback on skills in the oral modality	4.29	0.87
	Feedback on skills in the written modality	4.17	0.85
	Importance of diagnostic assessment	4.38	0.85
F4: GDLA's Alignment with the SL Curriculum	Implementation of diagnostic assessment	4.55	0.74
	Alignment with the SL curriculum descriptors	3.71	0.83
F5: Motivation in SL Teaching	Alignment with the CGSL's programmatic text	3.76	0.77
	Creativity	3.51	1.05
	Satisfaction	3.68	1.03

Notes

- ¹ The following example illustrates such a policy. A qualified GAL teacher serves at school A, which has a high share of migrant students. Due to the current transfer model for teachers that applies in Cyprus primary schools, if a more experienced teacher (in total years of employment but not in teaching years in GAL preparatory classes) asks to be transferred to that specific school, school A, then the qualified GAL teacher might be obliged to transfer to any other school (that might or might not have a need for GAL teachers). It must also be noted that school leaders are not involved in teacher recruitment for their schools (European Commission 2019a).
- ² The GDLA's component for first graders is built upon an integrated approach to oral communication (listening and speaking). The students participate in a discussion with their teacher on everyday routines (school, home, siblings, hobbies, pets, etc.) and they are asked to talk about pictures tightly connected to their experiences and interests, e.g., listening to music, playing, etc. They also listen and point at the relevant picture (this might include greetings, comparisons, etc.). Some of the tasks aim at the understanding of contextualized or/and illustrated functional vocabulary ("show and tell" or "find the picture"). The GDLA's component for first graders is available at https://www.pi.ac.cy/pi/files/epimorfosi/entaxi/diagnostiko_dokimio_a_dimotikou.pdf (accessed on: 4 November 2021).
- ³ The GDLA's component for second to sixth graders assesses the learners' performance in all four language skills through authentic/genuine texts (both oral and written, mainly multimodal texts: i.e. posters, invitations, weather forecast, visiting a store, shopping at the school canteen, writing an email, etc.). Available at https://www.pi.ac.cy/pi/files/epimorfosi/entaxi/diagnostiko_dokimio_v_eos_st_taxeis.pdf (accessed on: 4 November 2021).

References

- Alderson, Charles J. 2005. *Diagnosing Foreign Language Proficiency: The Interface between Learning and Assessment*. London and New York: Continuum.
- Alderson, Charles J. 2011. Innovations and Challenges in Diagnostic Testing. In *Selected Papers from 2011 PAC/The Twentieth International Symposium on English Teaching*. Taipei: English Teachers' Association of the Republic of China.
- Alderson, Charles J., and Diane Wall. 1993. Does washback exist? *Applied Linguistics* 14: 115–29. [CrossRef]
- Alderson, Charles J., and Liz Hamp-Lyons. 1996. TOEFL Preparation Courses: A Study of Washback. *Language Testing* 13: 280–97. [CrossRef]
- Andrews, Stephen. 2004. Washback and Curriculum Innovation. In *Washback in Language Testing: Research Contexts and Methods*, 1st ed. Edited by Liying Cheng, Yoshinori Watanabe and Andy Curtis. Mahwah: Lawrence Erlbaum Associates, New York: Routledge.
- Antonopoulou, Niovi. 2004. The Washback Effect of Testing on Teaching Greek as a Second/Foreign Language. *Language Testing Update* 36: 100–4.
- Bachman, Lyle F., and Adrian S. Palmer. 1996. *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.
- Bailey, Alison L. 2017. Assessing the language of young learners. In *Encyclopedia of Language and Education*, 3rd ed. Edited by Elana Shohamy, Iair G. Or and Stephen May. Berlin: Springer, pp. 323–42.
- Bailey, Alison L., Margaret Heritage, and Frances A. Butler. 2013. Developmental considerations and curricular contexts in the assessment of young language learners. *The Companion to Language Assessment* 1: 421–39.
- Brown, James D., and Thom Hudson. 1998. The alternatives in language assessment: Advantages and disadvantages. *University of Hawai'i Working Papers in ESL* 16: 79–103.
- Burrows, Catherine. 2004. Washback in classroom-based assessment: A study of the washback effect in the Australian adult migrant English program. In *Washback in Language Testing*, 1st ed. Edited by Liying Cheng and Yoshimori Watanabe. Mahwah: Lawrence Erlbaum, pp. 113–28.
- Butler, Yuko Goto. 2016. Assessing young learners. In *Handbook of Second Language Assessment*, 1st ed. Edited by Dina Tsagari and Jayanti Banerjee. Berlin and Boston: De Gruyter Mouton, pp. 359–76.
- Cheng, Liying, and Andy Curtis. 2004. Washback or backwash: A review of the impact of testing on teaching and learning. In *Washback in Language Testing: Research Contexts and Methods*. Edited by Liying Cheng, Yoshinori Watanabe and Andy Curtis. Mahwah: Lawrence Erlbaum Associates, pp. 3–17.
- Cheng, Liying, and Yoshimori Watanabe. 2004. *Washback in Language Testing Research Contexts and Methods*. Mahwah: Lawrence Erlbaum Associates.
- Cheng, Liying. 2005. *Changing Language Teaching through Language Testing: A Washback Study*. Cambridge: Cambridge University Press.
- Collins, John B., and Nicholas H. Miller. 2018. The TOEFL (ITP): A survey of teacher perceptions. *Shiken* 22: 1–13.
- Coste, Daniel, Danièle Moore, and Geneviève Zarate. 2009. *Plurilingual and Pluricultural Competence. Studies towards a Common European Framework of Reference for Language Learning and Teaching*. Strasbourg: Council of Europe Publishing.
- Davies, Alan. 2008. Textbook trends in teaching language testing. *Language Testing* 25: 327–47. [CrossRef]
- Davies, Alan. 2014. Fifty Years of Language Assessment. In *The Companion to Language Assessment*. Edited by Antony John Kunnan. Hoboken: Wiley Online Library, vol. 1, pp. 1–21.
- Dörnyei, Zoltán. 2007. *Research Methods in Applied Linguistics*. Oxford: Oxford University Press.

- European Commission. 2019a. *Peer Counselling on Integration of Students with a Migrant Background into Schools*. Brussels: European Commission. Available online: https://www.pi.ac.cy/pi/files/epimorfosi/entaxi/Peer_counselling_integration_of_migrant_students_final_report.pdf (accessed on 19 September 2021).
- European Commission. 2019b. Proposal for a Council Recommendation on a comprehensive approach to the teaching and learning of languages. *European Journal of Language Policy* 11: 129–37.
- Fan, Tingting, Jieqing Song, and Zheshu Guan. 2021. Integrating Diagnostic Assessment into Curriculum: A Theoretical Framework and Teaching Practices. *Language Testing in Asia* 11: 1–23. [CrossRef]
- Fulcher, Glenn. 2012. Assessment Literacy for the Language Classroom. *Language Assessment Quarterly* 9: 113–32. [CrossRef]
- Fulcher, Glenn. 2020. Operationalizing Assessment Literacy. In *Assessment Literacy*. Edited by Dina Tsagari. Cambridge: Cambridge Scholars.
- Garcia, Ofelia, and Wei Li. 2014. *Translanguaging: Language, Bilingualism, and Education*. London: Palgrave Macmillan.
- Green, Anthony. 2007. Washback to Learning Outcomes: A Comparative Study of IELTS Preparation and University Pre-session Language Courses. *Assessment in Education: Principles, Policy & Practice* 14: 75–97.
- Green, Anthony. 2013. Washback in language assessment. *International Journal of English Studies* 13: 39–51. [CrossRef]
- Hamp-Lyons, Liz. 1997. Washback, Impact and Validity: Ethical Concerns. *Language Testing* 14: 295–303. [CrossRef]
- Hasselgreen, Angela, Cecile Carlsen, and Hidegunn Helness. 2004. European Survey of Language Testing and Assessment Needs. Report: Part One—General Findings. Available online: <http://www.ealta.eu.org/documents/resources/survey-report-pt1.pdf> (accessed on 29 October 2021).
- Hughes, Arthur. 1989. *Testing for Language Teachers*. Cambridge: Cambridge University Press.
- Inbar, Ofra, Elana Shohamy, and Claire Gordon. 2005. Considerations involved in the language assessment of young learners. *ILTA Online Newsletter* 2: 3.
- Inbar-Lourie, Ofra. 2013. Guest Editorial to the special issue on language assessment literacy. *Language Testing* 30: 301–7. [CrossRef]
- Inbar-Lourie, Ofra. 2017. Language assessment literacy. In *Language Testing and Assessment*, 3rd ed. Edited by Elana Shohamy, Iair G. Or and May Stephen. Cham: Springer, pp. 257–68.
- Jang, Eunice E. 2013. Diagnostic assessment in language classrooms. In *The Routledge Handbook of Language Testing*. Edited by Glenn Fulcher and Fred Davidson. Abingdon: Routledge, pp. 134–48.
- Jimola, Folasade Esther, and Graceful Onovughe Ofodu. 2019. ESL Teachers and Diagnostic Assessment: Perceptions and Practices. *ELOPE: English Language Overseas Perspectives and Enquires* 16: 33–48.
- Kościółek, Jakub. 2020. Children with migration backgrounds in Polish schools: Problems and challenges. *Annales. Series Historia et Sociologia* 30: 643–56.
- Kunnan, Antony John. 2013. Language assessment for immigration and citizenship. In *The Routledge Handbook of Language Testing*, 1st ed. Edited by Glenn Fulcher and Fred Davidson. New York: Routledge, pp. 176–91.
- Kyriakou, Nansia. 2014. Investigating Teaching and Learning Greek as an Additional Language in Public Primary Schools Is Cyprus. Paper presented at INTCESS14—International Conference on Education and Social Sciences, Istanbul, Turkey, February 3–5; Edited by Ferit Uslu. Istanbul: OCERINT, pp. 548–59.
- Lam, Ricky. 2015. Language Assessment Training in Hong Kong: Implications for Language Assessment Literacy. *Language Testing* 32: 169–97. [CrossRef]
- Lee, Yong-Won. 2015. Diagnosing Diagnostic Language Assessment. *Language Testing* 32: 299–316. [CrossRef]
- Leung, Constant, and Jo Lewkowicz. 2017. Assessing Second/Additional Language of Diverse Populations. In *Encyclopedia of Language and Education*, 3rd ed. Edited by Elana Shohamy, Iair G. Or and Stephen May. Berlin: Springer, pp. 343–58.
- Levi, Tziona, and Ofra Inbar-Lourie. 2020. Assessment literacy or language assessment literacy: Learning from the teachers. *Language Assessment Quarterly* 17: 168–82. [CrossRef]
- Liu, Jianda, and Ximei Li. 2020. Assessing Young English Learners: Language Assessment Literacy of Chinese Primary School English Teachers. *International Journal of TESOL Studies* 2: 36–49.
- Malone, Margaret E. 2013. The essentials of assessment literacy: Contrasts between testers and users. *Language Testing* 30: 329–44. [CrossRef]
- Menken, Kate. 2006. Teaching to the Test: How No Child Left Behind Impacts Language Policy, Curriculum, and Instruction for English Language Learners. *Bilingual Research Journal* 30: 521–46. [CrossRef]
- Menken, Kate. 2013. Emergent Bilingual Students in Secondary School: Along the Academic Language and Literacy Continuum. *Language Teaching* 46: 438–76. [CrossRef]
- Menken, Kate. 2017. High-Stakes Tests as De Facto Language Education Policies. In *Encyclopedia of Language and Education*, 3rd ed. Edited by Elana Shohamy, Iair G. Or and Stephen May. Berlin: Springer, pp. 385–96.
- Min, Jiayi, and Moonyoung Park. 2020. Investigating Test Practices and Washback Effects: Voices from Primary School English Teachers in Hong Kong. *The Korea English Language Testing Association* 15: 77–97. [CrossRef]
- Mitsiaki, Maria, and Ioannis Lefkos. 2018. ELFeS: A Greek Illustrated Science Dictionary for School. Paper presented at XVIII EURALEX International Congress, Ljubljana, Slovenia, July 17–21; Edited by Jaka Čibej, Vojko Gorjanc, Iztok Kosem and Simon Krek. Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 373–85.

- Mitsiaki, Maria, Chrysovalanti Giannaka, and Despo Kyprianou. 2020a. Greek Diagnostic Language Assessment tool: 1st Graders. Cyprus Pedagogical Institute. Available online: https://www.pi.ac.cy/pi/files/epimorfosi/entaxi/diagnostiko_dokimio_a_dimotikou.pdf (accessed on 14 November 2021).
- Mitsiaki, Maria, Chrysovalanti Giannaka, and Despo Kyprianou. 2020b. Greek Diagnostic Language Assessment Tool: 2nd to 6th Graders. Cyprus Pedagogical Institute. Available online: https://www.pi.ac.cy/pi/files/epimorfosi/entaxi/diagnostiko_dokimio_v_eos_st_taxeis.pdf (accessed on 14 November 2021).
- Mitsiaki, Maria. 2020. Curriculum of Greek as a Second Language. Cyprus Pedagogical Institute. Available online: https://www.pi.ac.cy/pi/files/epimorfosi/entaxi/aps_isbn.pdf (accessed on 14 November 2021).
- MOECSY. 2021. Intercultural Education. Statistics. Department of Primary Education. Available online: http://www.moec.gov.cy/dde/diapolitismiki/statistika_dimotiki.html (accessed on 20 September 2021).
- Papakammenou, Irini. 2020. The Importance of Washback Effect in Teachers' Assessment Literacy. The Stepping Stone for more Learner-centred Exam-classes. In *Language Assessment Literacy: From Theory to Practice*. Edited by Dina Tsagari. Cambridge: Cambridge Scholars Publishing, pp. 286–303.
- Pehlivan Şişman, Emine, and Kağan Büyükkarcı. 2019. A Review of Foreign Language Teachers' Assessment Literacy. *Sakarya University Journal of Education* 9: 628–50. [CrossRef]
- Petridou, Alexandra, and Yasemina Karagiorgi. 2017. Validation of Greek Language Proficiency Tests "Milas Ellinika I" in Cyprus Context. *Epistimes Tis Agogis* 3: 80–99.
- Poehner, Matthew E., Kristin J. Davin, and James P. Lantolf. 2017. Dynamic Assessment. In *Language Testing and Assessment*. Edited by Elana Shohamy, Iair G. Or and Stephen May. Cham: Springer International Publishing, pp. 243–56.
- Popham, James W. 2009. Assessment Literacy for Teachers: Faddish or Fundamental? *Theory into Practice* 48: 4–11. [CrossRef]
- Rea-Dickins, Pauline, and Catriona Scott. 2007. Washback from Language Tests on Teaching, Learning and Policy: Evidence from Diverse Settings. *Assessment in Education: Principles, Policy & Practice* 14: 1–7.
- Rea-Dickins, Pauline, and Sheena Gardner. 2000. Snares and Silver Bullets: Disentangling the Construct of Formative Assessment. *Language Testing* 17: 215–43. [CrossRef]
- Sachdev, Sheetal B., and Harsh V. Verma. 2004. Relative importance of service quality dimensions: A multisectoral study. *Journal of Services Research* 4: 93–116.
- Scarino, Angela. 2013. Language assessment literacy as self-awareness: Understanding the role of interpretation in assessment and in teacher learning. *Language Testing* 30: 309–27. [CrossRef]
- Shohamy, Elana G. 2009. Language Tests for Immigrants: Why Language? Why Tests? Why Citizenship? In *Discourse Approaches to Politics, Society and Culture*. Edited by Gabrielle Hogan-Brun, Clare Mar-Molinero and Patrick Stevenson. Amsterdam: John Benjamins Publishing Company, vol. 33, pp. 45–60.
- Shohamy, Elana G. 2017. Critical Language Testing. In *Language Testing and Assessment*, 3rd ed. Edited by Elana Shohamy, Iair G. Or and Stephen May. Cham: Springer, pp. 441–54.
- Shohamy, Elana. 2001. *The Power of Tests: A Critical Perspective on the Uses of Language Tests*. Language in Social Life Series; New York: Longman.
- Stecher, Brian, Tammi Chun, and Sheila Barron. 2004. The effects of assessment-driven reform on the teaching of writing in Washington State. In *Washback in Language Testing: Research Contexts and Methods*. Edited by Liying Cheng, Yoshinori Watanabe and Andy Curtis. Mahwah: Lawrence Erlbaum Associates, pp. 53–69.
- Taylor, Lynda, and Jayanti Banerjee. 2013. Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing* 30: 403–12.
- Tsagari, Dina, and Jayanti Banerjee. 2014. Language Assessment in the Educational Context. In *The Routledge Handbook of Educational Linguistics*, 1st ed. Edited by Martha Bigelow and Johanna Ennser-Kananen. New York: Routledge, pp. 339–52.
- Tsagari, Dina, and Karin Vogt. 2017. Assessment Literacy of Foreign Language Teachers around Europe: Research, Challenges and Future Prospects. *Papers in Language Testing and Assessment* 6: 41–63.
- Tsagari, Dina, and Liying Cheng. 2017. Washback, Impact, and Consequences Revisited. In *Language Testing and Assessment*. Edited by Elana Shohamy, Iair G. Or and Stephen May. Berlin: Springer, pp. 359–372.
- Tsagari, Dina. 2007. Review of Washback in Language Testing: What Has Been Done? What More Needs Doing? Available online: <https://files.eric.ed.gov/fulltext/ED497709.pdf> (accessed on 14 November 2021).
- Tsagari, Dina. 2009. *The Complexity of Test Washback: An Empirical Study*. Frankfurt am Main: Peter Lang GmbH.
- Tsagari, Dina. 2011. Investigating the 'assessment literacy' of EFL state school teachers in Greece. In *Classroom-Based Language Assessment*. Edited by Dina Tsangari and Ildikó Csépes. Frankfurt am Main: Peter Lang, pp. 169–90.
- Vogt, Karin, and Dina Tsagari. 2014. Assessment Literacy of Foreign Language Teachers: Findings of a European Study. *Language Assessment Quarterly* 11: 374–402. [CrossRef]
- Watanabe, Yoshinori. 1996. Investigating Washback in Japanese EFL Classrooms: Problems of Methodology. *Australian Review of Applied Linguistics. Supplement Series* 13: 208–39. [CrossRef]

-
- Watanabe, Yoshinori. 1997. The Washback Effects of the Japanese University Entrance Examinations of English: Classroom-Based Research. Ph.D. Thesis, University of Lancaster, Lancaster, UK.
- Watanabe, Yoshinori. 2008. Methodology in washback studies. In *Washback in Language Testing: Research Contexts and Methods*. Edited by Cheng Liying and Watanabe Yoshinori. Mahwah: Lawrence Erlbaum Associates, pp. 19–36.
- Zhang, Limei, and Kaycheng Soh. 2016. Assessment Literacy of Singapore Chinese Language Teachers in Primary and Secondary Schools. In *Teaching Chinese Language in Singapore*. Edited by Kaycheng Soh. Singapore: Springer Singapore, pp. 85–103.