Luke Harding, Tineke Brunfaut, Ari Huhta, Benjamin Kremmel

10 Conceptualising DIALANG 2.0: Is there a role for mediation in a theory-driven approach to computer-based diagnostic language assessment?

Abstract: This chapter provides a discussion of the extent to which mediation – a key element of dynamic assessment – could be integrated into a computer-based diagnostic language assessment system. The discussion focuses specifically on early conceptualisations of such a system, which we have named DIALANG 2.0. Following an introduction, we provide background information on the DIALANG 2.0 project, with reference to an earlier and widely used system called DIALANG. We describe the systems and provide a justification for the current plans. We then turn to a specific question: Is there a role for mediation in DIALANG 2.0? In answering this question, we think through ways in which mediation might be integrated into the DIALANG 2.0 system, considering roles and types of mediation, and proposing some potential modifications to the DIALANG 2.0 blueprint. Finally, feasibility and challenges are addressed, with a section weighing up the benefits of integrating mediation against the costs involved in a more complex and elaborate computer-based diagnostic assessment system. In a set of concluding remarks, we reflect on the process of writing this chapter.

10.1 Introduction

Amid growing interest in diagnostic language assessment – and the facilitating role that digital technology might play in diagnosis – DIALANG remains a unique example of a computer-based diagnostic language assessment (DiagA) system (Alderson 2005; Alderson and Huhta 2005). DIALANG was developed by a consortium of universities in the 1990s and, although the project ended in 2004, at the time of writing, the system remained available on an open-access platform (https://dialangweb.lancaster.ac.uk), offering diagnostic tests in 14 languages (as well as 18 instructional and feedback languages) targeting reading, listening, writ-

Luke Harding, **Tineke Brunfaut,** Lancaster University
Ari Huhta, University of Jyväskylä
Benjamin Kremmel, University of Innsbruck

https://doi.org/10.1515/9783111233918-010

ing, vocabulary, and grammar. DIALANG has been a popular tool for learners, teachers, and second language researchers (with the most recent figures estimating approximately 250,000 visitors to the DIALANG website each year). However, DIALANG faces two challenges. First, DIALANG is no longer funded and therefore has no sustained technical support (relying on the good will of its long-term technician for maintenance and troubleshooting, and with frequent outages and periods of inaccessability). Second, the field of DiagA has made numerous advances in the past two decades, and DIALANG no longer represents the most current thinking in theorising diagnostic language assessment (e.g. Alderson, Brunfaut, and Harding, 2015; Huhta et al., 2024). For these reasons, prior to the pandemic, a team of researchers from the universities of Lancaster, Jyväskylä and Innsbruck began work on conceptualising DIALANG 2.0: an online system designed to represent a radically different approach to diagnostic language assessment compared with the original DIALANG format.

Work on DIALANG 2.0 reached the conceptualisation stage but stalled during the pandemic due to teaching priorities, challenges in arranging meetings and gaining funding to support the project. At the same time, collaborations further developed between one key member of the DIALANG 2.0 team – Ari Huhta – and a group of researchers interested in dynamic assessment (principally Matthew Poehner and Dmitri Leontjev). There followed an invitation to present a paper at the 2021 AILA conference, contributing to a symposium (convened by Huhta and Leontjev, with contributions from Poehner) exploring the potential for commonalities between diagnostic and dynamic assessment. In that paper, Luke Harding explored the degree to which concepts and practices from dynamic assessment might fit within ongoing thinking around DIALANG 2.0. In the present chapter, we aim to extend that thinking, with a focus on a specific element of dynamic assessment that emerged as potentially useful: mediation.

In realising this aim, in this chapter, we will first provide an overview of DIALANG 2.0, charting its history and rationale for development, its theoretical basis and distinctiveness from the original DIALANG, and its most recent conceptualisation (including the challenges involved in moving those ideas forward). Second, we will consider potential synergies between diagnostic language assessment and dynamic assessment (DA), focusing specifically on mediation (as understood within DA), discussing its definition and role in the dynamic assessment paradigm, and considering the various approaches to mediation that have been implemented in practice. Third, we will apply this thinking to consider potential modifications to our initial plans for DIALANG 2.0, developing our thinking through a worked example specifically focusing on diagnostic assessment of listening. Finally, we will evaluate the feasibility of including mediation within DIALANG 2.0 and reflect on issues related to conceptual commensurability and practicality.

10.2 DIALANG 2.0

10 2.1 History and rationale

Early discussions about DIALANG 2.0 emerged from a general concern over the sustainability of DIALANG in 2017. Technical maintenance of DIALANG had, for many years, been offered without cost by Lancaster University's IT division, with the system being solely maintained by a web programmer, Adrian Fish, who had been part of the original DIALANG consortium team (see Alderson 2005). By 2017, it became clear that DIALANG would need some form of update to remain viable, and discussions began between Lancaster University's IT division (particularly Adrian Fish), academics from the Department of Linguistics and English Language (Charles Alderson, Tineke Brunfaut, and Luke Harding), and Ari Huhta (a key developer and "guardian" of DIALANG) around what shape any reform would take and where further funding might be sourced.

These talks about the future of DIALANG naturally included not only consideration of financial and technical viability, but also how DIALANG might respond to recent theoretical advances. Of relevance here is that, in 2013, and separate from discussions concerning DIALANG, Alderson, Brunfaut and Harding had begun to discuss the need for a more rigorous, coherent approach to diagnostic language assessment that was theory-based. Spurred by a symposium on diagnostic language assessment at the 2013 LTRC conference in Seoul, they published two papers (Alderson, Brunfaut, and Harding 2015; Harding, Alderson, and Brunfaut 2015) which attempted to conceptualise, first, how diagnostic assessment could be understood within a wider sphere of diagnostic activities across professional practice in diverse fields, and second, how the lessons learned from other professions might be applied to the diagnostic assessment of second/foreign language (henceforth L2) receptive skills (listening and reading).

In the first paper, Alderson, Brunfaut, and Harding (2015) carried out interviews with a range of diagnosticians working in the fields of car mechanics, computing, health and medicine, general education, and psychology. The interviews yielded insights into common features of diagnosis across these varied professions, from which the authors drew a set of five diagnostic principles:

1. It is not the test which diagnoses, it is the user of the test.
2. Instruments themselves should be designed to be user-friendly, targeted, discrete and efficient in order to assist the teacher in making a diagnosis. They should provide rich and detailed feedback. Most importantly, useful testing instruments need to be designed with a specific diagnostic purpose in mind.
3. The diagnostic assessment process should include diverse stakeholder views, including learners' self-assessments.

4. Diagnostic assessment should ideally be embedded within a system that allows for all four diagnostic stages: (1) listening/observation, (2) initial assessment, (3) use of tools, tests, and expert help, and (4) decision-making.
5. Diagnostic assessment should relate, if at all possible, to some future treatment.

   (Alderson, Brunfaut, and Harding 2015: 22; see also Brunfaut and Harding, 2025)

In the follow-up article, Harding, Alderson, and Brunfaut (2015) elaborated specifically on principle 4 – the four diagnostic stages – to imagine how a diagnostic assessment might be developed to reflect each step. A description of the four diagnostic stages is shown in Figure 10.1 (for a complementary perspective on the diagnostic cycle, see Figure 1.1, Chapter 1). Although the diagnostic process outlined in Figure 10.1 centres the teacher as the key diagnostician, Harding, Alderson, and Brunfaut (2015) also suggested that this role could be flexible, noting that “diagnosticians may also be an expert diagnostician, an applied linguistics/second language acquisition expert, learners’ peers, learners themselves, an adaptive computer-based diagnostic system or any combination of these working together” (Harding, Alderson, and Brunfaut, 2015: 333 – footnote).

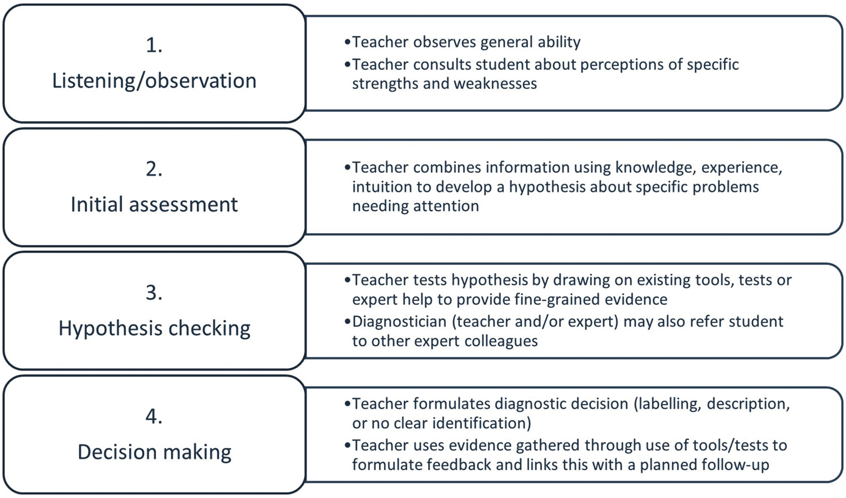
1.
Listening/observation
•Teacher observes general ability
•Teacher consults student about perceptions of specific strengths and weaknesses
2.
Initial assessment
•Teacher combines information using knowledge, experience, intuition to develop a hypothesis about specific problems needing attention
3.
Hypothesis checking
•Teacher tests hypothesis by drawing on existing tools, tests or expert help to provide fine-grained evidence
•Diagnostician (teacher and/or expert) may also refer student to other expert colleagues
4.
Decision making
•Teacher formulates diagnostic decision (labelling, description, or no clear identification)
•Teacher uses evidence gathered through use of tools/tests to formulate feedback and links this with a planned follow-up

Figure 10.1: The diagnostic process, drawn from Harding, Alderson, and Brunfaut (2015: 319).

Analysed against these proposals, it was apparent that the original DIALANG test aligned in some ways with the theory-driven approach put forward by Alderson, Brunfaut, and Harding (2015) and Harding, Alderson, and Brunfaut (2015); however, it was markedly different in other respects. To help explain this analysis, Figure 10.2 provides a simplified overview of the test-taker's "journey" through the original DIALANG system. A more detailed description is provided by Alderson and Huhta (2005).

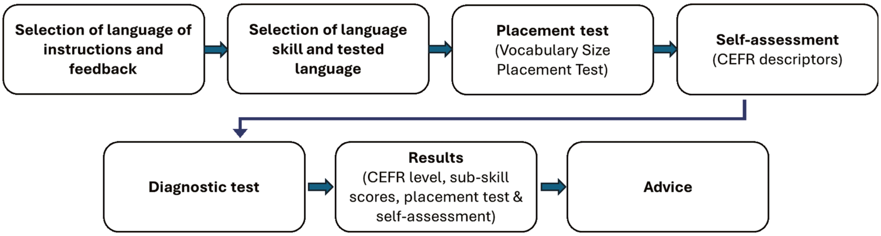
Selection of language of instructions and feedback
Selection of language skill and tested language
Placement test (Vocabulary Size Placement Test)
Self-assessment (CEFR descriptors)
Diagnostic test
Results (CEFR level, sub-skill scores, placement test & self-assessment)
Advice

Figure 10.2: Test-taker pathway in the original DIALANG test.

As shown in Figure 10.2, the DIALANG test-taker first selects the interface language, which is also the language in which feedback will be given to them at the end of the test session. They then choose the language and skill in which they would like to be assessed, selecting from Danish, Dutch, English, Finnish, French, German, Greek, Icelandic, Irish, Italian, Norwegian, Portuguese, Spanish and Swedish, to take a test in reading, listening, writing, vocabulary, or grammar. Test-takers then proceed to a placement test – the Vocabulary Size Placement Test (VSPT) – which provides an initial level screening of a test-taker's language ability. Performance on the VSPT is used to inform the computerised system about the level at which the diagnostic test should be pitched (note: the VSPT is optional). If the test-taker has chosen to be assessed on the skill of listening, reading or writing, they are then prompted to self-assess their ability against the relevant set of CEFR descriptors. Again, this level placement contributes to the selection of a suitable level for the subsequent diagnostic test (if they have chosen to be assessed on grammar or vocabulary, the system moves them from the VSPT directly to the diagnostic test). The test-taker then begins the diagnostic test.

In order to generate diagnostic information, each test contains items designed to target specific dimensions/subskills. Alderson (2005) and the DIALANG Assessment Specifications (DAS 1998) define sets of dimensions or subskills across the different skills within the exam, as shown in Table 10.1.

Upon completing the test, the test-taker is provided with a set of results: their CEFR level, sub-skill scores, and the results of their placement test and self-assessment (the latter includes a comparison between the CEFR levels based on the test-taker's test and self-assessment). The test-taker is then offered advice on

Table 10.1: DIALANG (original) subskills by test (drawn from Alderson 2005, and test specifications).

Test	Dimension/ subskill	Definition
Reading	Identifying main idea	Summarising/identifying the main idea(s), main information or main purpose of a piece of written discourse
	Reading for detail	Reading intensively for specific detail or specific information
	Inferencing	Making inferences on the basis of what is read and to be able to use context to infer the approximate meaning of an unfamiliar word
Listening	Identifying main idea	Summarising/identifying the main idea(s), main information or main purpose of a piece of spoken discourse (including the contribution of prosodic features)
	Listening for detail	Listening intensively for specific detail or specific information
	Inferencing	Making inferences on the basis of what was heard and to be able to use context to infer the approximate meaning of an unfamiliar word
Writing	Accuracy	Recognising or producing target-language like written texts in terms of grammar, vocabulary and spelling.
	Appropriacy	Features connected to style; determined by the context of writing and the roles of the writer and the reader(s). Includes features indicating politeness and distance.
	Textual organisation	Features connected with coherence and cohesion.
Vocabulary	Meaning	"Recognise/produce word meanings, including denotation, semantic fields, connotation, appropriateness" (Alderson, 2005: 193)
	Semantic relations	"Recognise/produce semantic relationships between words, including synonymy/antonymy/converses, hyponymy/hypernymy, polysemy" (Alderson, 2005: 193)
	Word combination	"Recognise/produce word combinations including collocations and idiomaticity" (Alderson, 2005: 193)
	Word formation	"Recognise/produce words by compounding and affixation" (Alderson, 2005: 193)

Table 10.1 (continued)

Test	Dimension/ subskill	Definition
Structures	Adjectives and adverbs	Inflection; comparison
	Nouns	Inflection – cases; definite/indefinite – articles; proper/common
	Numerals	Inflection; context
	Pronouns	Inflection; context
	Parts of speech	Word order, statements, questions, exclamations; agreement
	Punctuation	Punctuation
	Verbs	Inflection, tense, mood, person; active/passive voice
	Miscellaneous word grammar	Coordination; subordination; deixis

the basis of those results. For example, a test-taker who self-assessed their writing ability as B1 but who was placed at A2 on the basis of their performance on the DIALANG writing test would be shown a table illustrating key writing competences at A2 level (based on CEFR descriptors), compared with the adjacent A1 and B2 level competences. Suggestions to explain any gap between a test-taker's self-assessment and their performance on the diagnostic test can be explored in a further area of DIALANG labelled "About self-assessment".

The points of alignment between the original DIALANG and the proposals put forward in Alderson, Brunfaut and Harding (2015) and Harding, Alderson and Brunfaut (2015) include the following:

1. In DIALANG, a branching, computer-based testing system functions as the diagnostician. The system assesses the student based on their initial VSPT and self-assessment results, selects an appropriate test level, administers the main diagnostic test, compiles and presents results, and provides feedback based on those results.
2. Testing instruments were designed with a clear diagnostic function in mind. DIALANG is a purpose-built diagnostic test, and subskills theorised in advance (not *ex post facto* as with some methods of cognitive diagnostic assessment) are specifically targeted with a view to providing finer-grained feedback than a single test score.
3. Test-takers' self-assessments are gathered and fed into the level selection, so the test-taker plays an active role in the diagnostic process. Self-assessment is

also used to prompt the learner to reflect on their abilities and consider discrepancies after the test is completed.

4. The diagnostic process occurs over stages which broadly include an initial screening stage, a testing stage and a decision-making stage.
5. There is feedback provided; test-takers receive a score report, a description of typical abilities at their level, and guidance on how to improve further. DIALANG is oriented towards future learning. However, in this model the responsibility to act on recommendations rests with the learner.

The original DIALANG thus displays some characteristics of the theory-based work that has been conducted since its development. However, there are key elements of an ideal diagnostic procedure that are not currently captured in DIALANG.

First, the diagnostic procedures in the original DIALANG currently do not provide a filtering mechanism where an initial assessment leads to hypothesis checking of a more constrained set of priority areas. An efficient diagnostic procedure should ideally be designed to fine-tune and re-focus at each stage to ensure that hypothesis-checking tests are well-targeted and able to provide fine-grained evidence. The current DIALANG approach provides a point of selection at the initial screening stage (directing test-takers to one of three test levels based on their VSPT and self-assessment scores), but from that point on does not narrow the field of possible areas of challenge. A more sophisticated diagnostic system would make maximal use of time by identifying and prioritising problems identified in the listening/observation stage and the initial assessment.

Second, the feedback provided in DIALANG remains limited by (a) the broad categories embodied in the main skill (e.g., reading) and subskills (e.g., reading for detail) approach, and (b) the need for a more rigorous underpinning theory-based understanding of L2 development. Currently, DIALANG feedback includes descriptions of the CEFR levels in the tested main language skill or area and item-level feedback (correct vs incorrect) broken down by the subskill. The CEFR levels do not provide detailed diagnostic information and their basis on empirical research on L2 development has been critiqued (e.g., Hulstijn 2007). While DIALANG feedback includes scales describing language-specific CEFR levels in vocabulary and grammatical knowledge, these were created by the language teams in the project based on their expert opinions. For its part, item-level feedback is very detailed as it comes with every item, but its diagnostic value entirely depends on the test-taker's ability to figure out what was involved in responding to each item and why some of their answers were incorrect. The fact that item-level feedback is reported with reference to subskills probably gives that feedback some structure, but the relationship between the items and the subskills they presumably tap remains vague. As Huhta's (2010) study indicated, a certain proportion of DIALANG users find it difficult to interpret and use the feedback without a language teacher's help. The ambiguous wording of some DIALANG feedback was also critiqued by Chapelle (2006) in a test review.

It is clear that while DIALANG represented an impressive, purpose-built computer-based diagnostic assessment system at its inception, it would require a deep revision to align more closely with the proposals for diagnostic language assessment presented above that have emerged over the past decade. This revision would ideally also take into account theoretical and empirical developments in the wider fields of applied linguistics and second language research. For example, a considerable amount of diagnostically useful conceptual and empirical work has been carried out in the areas of listening and reading since DIALANG was designed (e.g., Alderson, Haapakangas et al. 2015; Grabe 2009; Grabe and Yamashita 2022). For these receptive skills, more fine-grained diagnostic instruments could potentially be built around processing models (e.g., Field 2013; Khalifa and Weir 2009), which would provide an empirical basis to draw inferences about how (and at what level of processing) comprehension broke down (see Harding, Alderson, and Brunfaut 2015, for a worked example). For written production, a body of research now exists on diagnostic writing assessment since the introduction of DIALANG (e.g., Huhta et al. 2024; Knoch 2009; Xie 2019), which would provide a strong basis for the development of assessment tools and feedback systems. For grammar and vocabulary, a key shift in recent years is the greater recognition of lexicogrammar as a unified construct (Alderson and Kremmel 2013; Kremmel, Brunfaut, and Alderson 2017; Paquot, Gries, and Yoder 2020), a change which would have a fundamental impact on approaches to diagnostic assessment of those areas. Finally, speaking – which was not included as a skill area in DIALANG due to logistical challenges – is now more feasible to assess in online environments, given developments in technology. There is a growing base of empirical work (though still nascent) to inform the diagnostic assessment of speaking (e.g., Isbell 2021; May et al. 2020), and potentially revolutionary capabilities being explored for diagnostic speaking assessment with conversational AI "agents" (e.g. Matsuyama et al. 2023).

Finally, a revision of DIALANG would need to respond to critiques made in the research literature concerning format and user-experience. For example, in a review of the DIALANG system, Zhang and Thompson (2004) noted that the range of item types used in DIALANG is relatively limited (multiple-choice, gap-filling, and short-answer questions), that speaking is not assessed, and writing is not tested directly. Kektisidou and Tsagari (2019), reporting on the use of DIALANG for a tracking study of English language learners in university settings, noted that DIALANG is long (when all components are taken together), and the website is "rather outdated and not particularly user-friendly" (Kektisidou and Tsagari 2019: 18). The affordances of current-day technology, as opposed to when DIALANG was developed, make it within reach to address many of the limitations related to user-experience.

10.2.2 Current conceptualisation of DIALANG 2.0

When discussions around a possible DIALANG 2.0 began in 2017, it quickly became clear that the project would need to take the form of a reconceptualization rather than a revision. The changes involved in meeting the vision of DiagA described above would require a radically different approach, albeit one that acknowledged and drew on the numerous strengths of the original DIALANG system. One of the key drivers of the reconceptualization of DIALANG 2.0 was to implement all stages of the listening-observation process identified in Alderson, Brunfaut, and Harding (2015) and elaborated in Harding, Alderson, and Brunfaut (2015), though to envisage these stages within an environment where the computer was the principal diagnostician. From this starting point, we embarked on two sets of intensive discussions to brainstorm and map out a macro-structure that captured how such a system might work.

Figure 10.3 provides the initial blueprint derived from these early DIALANG 2.0 discussions, illustrating the steps involved in the four key stages: listening/observation, initial assessment, hypothesis checking, decision making.

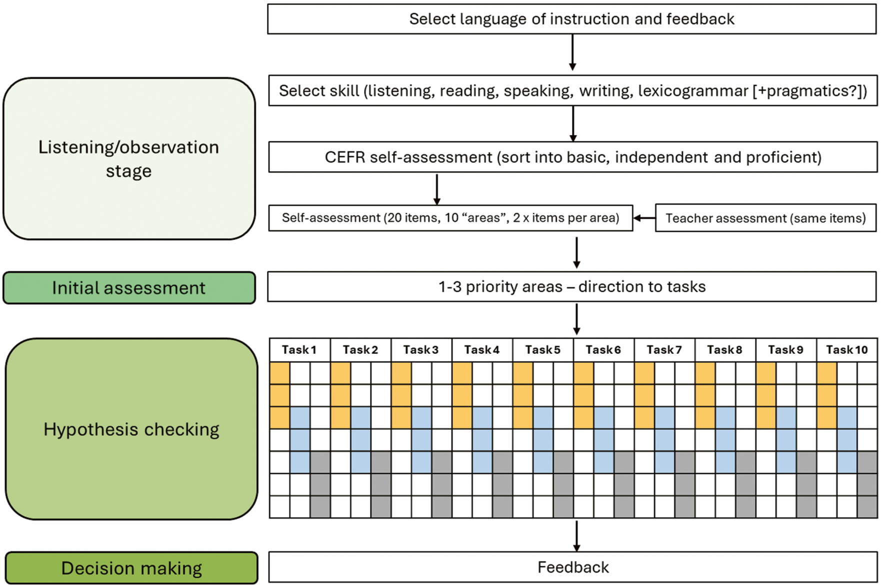
Select language of instruction and feedback
Listening/observation stage
Select skill (listening, reading, speaking, writing, lexicogrammar [+pragmatics?])
CEFR self-assessment (sort into basic, independent and proficient)
Self-assessment (20 items, 10 “areas”, 2 x items per area)
Teacher assessment (same items)
Initial assessment
1-3 priority areas – direction to tasks
Hypothesis checking
Task 1
Task 2
Task 3
Task 4
Task 5
Task 6
Task 7
Task 8
Task 9
Task 10
Decision making
Feedback

Figure 10.3: DIALANG 2.0: macro-level structural blueprint. © Harding, Brunfaut, Huhta, and Kremmel (2024).

Figure 10.3 illustrates that, in a similar way to the original DIALANG, test-takers in DIALANG 2.0 would first select an interface language and the skill area that they want to focus on. Notably, the skill areas we envisaged would now also include speaking, would combine grammar and vocabulary into a single lexicogrammar construct, and would potentially add pragmatics as a standalone skill area for diagnostic assessment (following increasing interest in assessing pragmatic competence: see Ross and Kasper 2013; Roever 2022). The test-taker would then proceed to a CEFR-based self-assessment – similar to the system of self-assessment in the original DIALANG – to be streamed into one of three levels: a basic user (A1–A2), an independent user (B1–B2) or a proficient user (C1–C2). At this stage, we would not include the VSPT given previous criticisms made by various vocabulary researchers concerning: (a) its narrow construct of word recognition, (b) its word family rather than lemma base, (c) issues around corpus frequency sampling rates, and (d) problems in its scoring algorithm and in the general interpretability of VSPT scores (Gyllstad, Vilkaitė, and Schmitt, 2015; Huhta 2007; Kremmel 2016; Read 2004). Vocabulary tests using this checklist format have also been found not to work particularly well with low-level learners (Meara 1996), who are, however, often among the main target audience of diagnostic tests. At this point in the process, self-assessment can be considered sufficient enough for the kind of broad branching that would be applied at this step (similar to its use in DIALANG); however, we intend to explore other efficient measures that might triangulate self-assessments at this stage in place of the VSPT.

An important step is then included in the listening/observation stage to help select the diagnostic pathway: test-takers are asked to complete a further self-assessment within their level (which was established with the initial self-assessment) on specific *areas* within their chosen skill. The *areas* at this stage would represent the fine-grained components of the skill, divided according to a theory of language ability. It is at this point that DIALANG 2.0 would depart from the original DIALANG and its focus on relatively broad dimensions/sub-skills. Rather, to give an example for listening (the skill area that remains the most developed area of our thinking thus far), at this point, the test-taker would be asked to self-assess their ability on the following ten discrete areas, based on Field's (2013) processing model as shown in Table 10.2 (for further discussion, see Harding, Alderson, and Brunfaut 2015: 330).

We envisage two self-assessment items per area, meaning a 20-item self-assessment for the chosen skill. We are currently exploring the best way to conduct fine-grained self-assessments of this kind, with current guidance suggesting that well-worded, explicit criteria in rubrics and some form of embedded self-assessment training would be important considerations (Li and Zhang 2021). Within the DIALANG 2.0 system, an important innovation would be the option for

Table 10.2: Proposed self-assessment areas for listening.

Level of processing	Self-assessment focus
Input decoding	1. Discriminating sounds/phonemes 2. Perceiving word and sentence stress
Lexical search	3. Recognising words 4. Understanding the meaning of words
Parsing	5. Understanding syntax 6. Understanding the literal meaning of utterances
Meaning representation	7. Understanding the pragmatic meaning of utterances 8. Understanding metaphor
Discourse construction	9. Distinguishing between relevant/irrelevant information 10. Understanding longer stretches of discourse

a teacher (or peer) to feed in their assessment of the learner on the same items based on listening/observation activities external to the DIALANG system. In this way, the learner's self-assessment could be supplemented and integrated with that of a teacher/peer, adding sophistication and, most likely, also precision.

An initial assessment would then be made based on the test-taker's response to the self-assessment tasks (in conjunction, if applicable, with the assessment of a teacher or peer). The initial assessment would take the form of the selection of one to three priority areas for which the learner appears most in need of further diagnostic information. Thus, the initial assessment is not a score or a particular piece of advice but rather a notification of priority areas for further investigation. This is akin to a general medical consultation that might yield two or three potential areas for further investigation, requiring more fine-grained tests.

The hypothesis-checking stage would then involve the administration of those fine-grained tests, selected according to the priority areas, to further investigate and potentially confirm the self-assessment. Again, we envisage ten sets of tasks, mapping onto the ten *areas* that have formed the basis for the self-assessment. For example, if the learner had identified word recognition as a potentially challenging area in their self-assessment, they would be directed to a word recognition task to confirm this initial assessment. Similarly, if they had identified pragmatic understanding as a particular challenge, they would be directed to an aural test of pragmatic comprehension. Performance on those discrete tests would provide further, more objective, information to feed into the diagnostic decision. These discrete, focused tasks would span three levels – reflecting the Basic, Independent and Proficient levels in the CEFR – and would overlap such that the portion of items at the lower end of the Independent task would be the same items that form the upper portion of the Basic task, and the items at the higher end of the Independent task would be the same items that form the lower portion of the Proficient task. In this way, the tasks across the three levels could be connected and vertically scaled (see Papageorgiou, Ginsburgh, and Gomez 2023).

Tasks at the hypothesis-checking stage would naturally vary in their type and format. For example, again focusing on listening, we would anticipate that tasks targeting input decoding and lexical search processes might resemble phoneme discrimination and word recognition tasks utilised in speech perception and processing research. In such cases, assessment methods are well-established, and the main challenge in developing tasks will be selecting target input that is level-appropriate and operationalising these through the DIALANG 2.0 system. By contrast, tasks designed to target higher-order skills might require purpose-built tests of pragmatic understanding or metaphor comprehension, drawing on theoretical work to establish appropriate constructs.

One key question that remains unresolved is how we might determine adequate construct coverage across the areas represented in more discrete tasks. For example, pragmatic competence has been approached in various ways within the literature on language assessment. Timpe-Laughlin, Wain, and Schmidgall (2015) draw on a range of prior literature – and specifically the work of Taguchi (2012) – to propose a componential view of pragmatic competence that includes sociocultural knowledge, pragmatic-functional knowledge, grammatical knowledge, discourse knowledge, and strategic knowledge. This means that in a standalone skill of “pragmatics”, there would be various ways in which *areas* could be identified. However, since pragmatics also forms an area within other skills (listening, reading, speaking and writing: e.g., “pragmatic understanding” in listening), the components of pragmatic knowledge that support this level of processing might only be represented in a crude way. These issues will need to be addressed at the design stage, taking into account the practicalities of the system.

The final stage in the diagnostic process is decision making. At this point the diagnostic decision would be communicated, and the test-taker would be provided with actionable feedback and advice. Feedback might include graphic representations of the learner’s performance in relation to their learning objective; for example, being able to interactively compare their ability profile with typical diagnostic profiles in specific areas or components of learners at particular CEFR levels, having feedback and areas for development exemplified rather than only metalinguistically described, and being directed to targeted (external) online materials and exercises to help improve certain ability areas. As DIALANG 2.0 develops, exploring innovations and feedback and advice will be a central area of focus.

10.3 Is there a role for mediation in DIALANG 2.0?

It will be clear from the preceding section that our thinking on DIALANG 2.0 remains at an early stage. We have not had the opportunity to take these ideas forward in concrete ways owing to the need for seed funding, the challenges in meeting brought about by the pandemic, and a need to focus on other priorities. This means, however, that our thinking on DIALANG 2.0 remains relatively open, and we expect the plans – and the blueprint itself – to evolve over time.

One area where rapid change in the field is almost certainly going to influence the ongoing development of DIALANG 2.0 is the use of artificial intelligence (AI). AI is already having a profound impact on the field of language assessment, and in a computer-delivered system such as DIALANG 2.0, we can see numerous opportunities to make use of AI technology to enhance diagnostic potential. In fact, current directions suggest that AI and digital technology are driving a renewed interest in diagnostic language assessment (evidenced, for example, by two symposia on that theme at the 2023 Language Testing Research Colloquium in New York City), with such technologies opening up possibilities for a more sustainable learning-oriented approach to classroom-based assessment.

Our thinking also continues to evolve in terms of theoretical matters. As stated at the beginning of this chapter, the collaborations of members of the authorship team with scholars working in the tradition of dynamic assessment have prompted new questions about the relationship between these two paradigms. Diagnostic assessment and dynamic assessment both share commonalities in their focus on supporting learning and in their orientation towards development. However, the two paradigms also differ in several important respects. First, one clear distinction is the extent to which both paradigms are underpinned by a theory of learning – dynamic assessment has a strong connection to Vygotskian Sociocultural Theory, whereas diagnostic assessment is less tied to a specific theory of learning and is often more concerned with language *constructs* (see Chapter 1; also Alderson, Brunfaut, and Harding, 2017). Second, the two paradigms differ in the extent to which they centralise learner development. As Antón and García (2022) explain, "in DA [dynamic assessment] the ultimate goal is not assisted performance, current-task completion, or evidence of actual ability but fostering potential development with the assistance of mediational tools" (Antón and García 2022: 174). This assertion would appear to put dynamic assessment at least partly at odds with diagnostic assessment, which is essentially focused on evidence of actual ability based on observations elicited from task completion. Both can be said, however, to share common ground with respect to "fostering potential development". Finally, both paradigms rest on different metatheoretical underpinnings; as discussed in Chapter 1, DA is more informed by dialectical materialism, while DiagA is more aligned with post-positivism.

We will return to this broader question of commensurability at the end of the chapter. However, in the spirit of trying to achieve common ground, our thinking has considered ways in which elements of a dynamic assessment approach might be integrated into diagnostic language assessment. We will not provide detail here on different ways of going about dynamic assessment as this has been covered in other parts of the book by scholars working in that tradition (see chapters by Levi; Poehner, Zhang, and Qin; Yu and Leontjev). In thinking through the potential synergies between the two approaches, however, we argue that the most promising element to be considered in DIALANG 2.0 is *mediation*. In the next section, we will provide further information on mediation in dynamic assessment and discuss its potential utility for diagnostic assessment.

10.3.1 Mediation in dynamic assessment: Roles and types

Mediation is a central element of dynamic assessment, defined by Antón and García (2022) as "the intervention of the assessor in order to select, amplify, and interpret objects and processes to the learner during the assessment". Similarly, Poehner and Wang (2021) see mediation as "particular kinds of support" that can be offered to the learner during their performance on an assessment task, which may take the form of "reminders, leading questions, hints, provision of a model, [or] feedback" (Poehner and Wang 2021: 472). According to Vygotskian Sociocultural Theory, it is through mediation that assessment and teaching are brought together in "dialectal unity" (Poehner 2011: 101), as the mediator (a teacher or an automated system of some kind) helps to draw a learner into their Zone of Proximal Development (ZPD), a fertile space in which the learner is maximally challenged just beyond their independent capabilities. Already it is evident that there is overlap in conceptualisations of who might be a mediator and who might be a diagnostician. Crucially, it is the learner's response to mediation that is of interest in dynamic assessment. As Poehner has stated, drawing on Sternberg and Grigorenko (2002), "DA targets what individuals are able to do in cooperation with others rather than what they can do alone" (2007: 324). Mediation thus holds promise for diagnostic assessment as it opens up the possibility of an *additional layer* of diagnostic information; the diagnostician could gain evidence of what the test-taker can do individually, *or* with intervention (diagnosis of ZPD). Depending on the context of the diagnostic procedure, both types of information may be useful for drawing conclusions about future action.

Mediation within dynamic assessment has taken many forms, though not all would necessarily be appropriate for application to diagnostic assessment. First, a key distinction is often made in referring to *when* mediating sessions take place in a process of dynamic assessment. Sternberg and Grigorenko (2002) refer to a *sandwich* format – where mediation occurs between two unmediated testing sessions – and a *cake* format – where mediation occurs within the main administration of a test itself (see also Poehner 2007). The cake format would seem to have more in common with the focus of most diagnostic assessment, which is to evaluate in a single sitting the strengths and areas for improvement of an individual learner. Another key distinction is between *interventionist* approaches to mediation and *interactionist* approaches to mediation. The former requires a more standardised, structured approach to the provision of prompts, hints, or clues, while in the latter, mediation is more dialogic and open (see also Poehner 2007). Both types of mediation may theoretically be applicable to diagnostic assessment generally, though the interventionist approach would have the most utility in a computer-delivered test environment.

Over the past decade, while there has been a considerable amount of research on dynamic assessment in L2 learning (see Poehner and Wang 2021), only a limited number of studies have explored dynamic language assessment in computer-based test environments. In considering the application of dynamic assessment to DIALANG 2.0, it is instructive to consider the ways in which mediation has been operationalised in such settings. In one highly-cited study, Poehner and Lantolf (2013) explored the integration of mediation into a computer-based test environment, focusing on L2 reading and listening assessment of Chinese, French, and Russian. All tests included four-option multiple-choice questions. An interventionist approach was designed for each task in which a wrong answer would precipitate a series of graded prompts. On the first incorrect response, a portion of the text was highlighted (in the case of reading) or replayed (in the case of listening) to direct the test-taker's attention. The test-taker was then invited to respond again. If the response was still incorrect, the range of text was narrowed down further (through highlighting or replay), with additional prompts giving further clues as required. When all prompts had been exhausted, the correct answer was revealed. Poehner and Lantolf (2013) were able to then analyse differences between the test-takers' "actual score" (their unmediated performance), their "mediated score" (their weighted score to indicate how responsive they were to feedback), and their "learning potential score" (a gain score indicating the difference between actual and mediated performance). This kind of interventionist procedure is, by nature, less time-efficient than an unmediated reading or listening test, but yields a considerable amount of information on which further diagnostic decisions might be made about learners' general strengths and areas for improvement. It is noteworthy that Poehner and Lantolf did not break down their analysis according to the particular sub-constructs targeted by specific items; however, had they done so, the diagnostic potential would have been further increased.

Since Lantolf and Poehner's study, some other computerised dynamic assessment (C-DA) studies have emerged. For example, Leontjev (2014) found facilitative effects on grammatical learning for adaptive corrective feedback (a type of mediation) delivered through the computer-based ICAnDoiT system. Poehner et al. (2015) reported on a study of a similar set of test materials to those featured in Poehner and Lantolf (2013), but with a slightly different approach to mediation through graduated prompts, which moved from implicit to explicit hints. Qin and van Compernolle (2021), building on this early work, provided specific advice for mediation approaches, which they argue should be "*contingent on learner need and graduated in explicitness*" (Qin and van Compernolle 2021: 59, italics theirs). Qin and van Compernolle explain that, "the form of support [test-takers] are provided during the test should be implicit at first, and should only become more explicit if test-takers are unable to benefit from a more implicit form of assistance" (Qin and van Compernolle 2021: 59). The authors operationalised this principle in a C-DA test, where graduated hints were provided within a test of aural implicature comprehension. In the test, listeners were provided with a context and then heard an aural prompt representing an indirect speech act delivered in the target language (Chinese). The task required listeners to select the correct interpretation of the utterance from multiple-choice options. At the most implicit level, the first prompt simply allowed listeners to hear the description of the context and the indirect speech act a second time, to provide another opportunity to answer. If the answer was still not correct, the second prompt introduced slightly more explicitness, allowing for replay of the indirect speech act together with "a hint highlighting the relevant content of the speech act (e.g., topic or vocabulary) (Qin and van Compernolle 2021: 62). Finally, at the third level, the prompt was another replay of the indirect speech act together with a "forced choice hint" utilising some common language with the answer key (Qin and van Compernolle 2021: 62). From a diagnostic perspective, graduated prompting of this kind can provide a diagnostic system with information about a learner's level of independence to understand/produce the language, again potentially enhancing the ultimate diagnostic decision.

These previous studies that sought to operationalise C-DA provide some useful guidance on how mediation might best be integrated into a computerized listening dynamic assessment like that proposed in DIALANG 2.0. First, mediation should resemble the cake format, and take an interventionist format. C-DA studies seem to agree that while an interactionist approach provides scope for greater sensitivity to a learners' ZPD, the interventionist approach makes most sense in C-DA contexts because it is "scalable" (Qin and van Compernolle, 2021: 58). Second, the use of standardised graduated prompts should move from implicit to explicit, allowing the test-taker to demonstrate their abilities using the least amount of support possible before moving on to the next level of explicit prompt. While this guidance is useful, it does require further thought concerning how mediation might be integrated in a concrete way in DIALANG 2.0. We turn to this issue in the next section.

10.3.2 Potential modifications to DIALANG 2.0

Given that DIALANG 2.0 would represent a multi-stage diagnostic procedure, the first question to address is *where* mediation might be embedded. Thinking this through requires us to depart from the existing guidance around C-DA because, to our knowledge, there is no existing application of DA to a system that resembles DIALANG 2.0. Theoretically, all stages of the diagnostic process could be reoriented if the ultimate aim is to add an additional layer of diagnostic information. Table 10.3 – adapted from a similar table in a recent chapter by Brunfaut and Harding (2025) – demonstrates how the key guiding questions for each stage of the diagnostic cycle could be broadened through a mediation orientation for the particular stages of DIALANG 2.0.

Table 10.3: Integration of mediation in approaches to diagnostic listening assessment (adapted from Brunfaut and Harding, 2025: 265).

Diagnostic stage	"Conventional" questions	Mediation questions
1. Listening/ observation	For a given skill, what does the learner self-report that they can do, and cannot do, independently?	For a given skill, what does the learner self-report that they can do independently, can do with assistance, and cannot do even with assistance?
2. Initial assessment	Does the learner have difficulty with X skill area(s) when working independently?	Does the learner have difficulty with X skill area(s) when working independently, but does the learner demonstrate emerging ability in X skill area(s) with mediation?
3. Hypothesis checking	Is the initial assessment confirmed or refuted according to independent task performance?	Is the initial assessment confirmed or refuted according to independent or mediated task performance?

Table 10.3 (continued)

Diagnostic stage	"Conventional" questions	Mediation questions
4. Decision making	What are the learner's current areas of strength and areas for improvement in independent task performance?	What are the learner's current areas of strength (in independent task performance), areas of nascent or emerging ability (in mediated task performance), and areas for improvement (following unsuccessful performance in mediated tasks)?

To address these questions, the mediated approach would require additional (and more complex) instruments at each stage of the process. First, at the listening/observation stage, self-assessments would need to include ratings both of what a learner believes they can do independently, and what they could do with assistance (see Poehner 2012). This would be feasible enough to operationalise through an extended set of self-assessment prompts/scales; however, it would create challenges in ensuring that learners understand what was meant by "with assistance". A variety of ways and degrees of providing "assistance" could be referred to here and the impact of different explanations of assistance at this stage would require further research. One option would be to tailor each self-assessment item to refer to assistance as a prompt in its most implicit format. For example, for listening, we might specify that assistance means being able to replay key portions of the text, or for reading, that assistance means narrowing down the field of text. This approach, however, underscores one of the main limitations of the interventionist approach to mediation which is its relative inflexibility. A learner may require a different kind of assistance from what can be offered through the proposed intervention but would not be able to express this adequately at the self-assessment stage. It also remains unclear to what extent learners would be able to gauge this in a reliable way through self-assessment, predicting the helpfulness of specific forms of assistance with precision.

Provided the self-assessment was made feasible, the initial assessment would ideally then focus on those skill areas that fall within the test-taker's ZPD. Thus, rather than selecting areas where test-takers expressed a particularly low level of ability (as in the current conception of DIALANG 2.0), the diagnosis would focus instead on the "sweet spot" just beyond the test-taker's independent capabilities. This would be a particular strength of the dynamic approach: it would provide a stronger rationale for the filtering of areas, giving the diagnostic procedure even greater potential to lead to more effective suggestions for treatment.

The hypothesis checking stage, however, is where mediation would be applied most conventionally (i.e., in accordance with approaches taken in the C-DA literature). As discussed earlier, we would anticipate an interventionist approach involving a range of graduated prompts moving from more implicit to more explicit. To illustrate for DIALANG 2.0, let us take the example of a learner who selects listening as their skill focus, is placed through self-assessment at the independent level, and who then indicates in their more fine-grained self-assessment that they struggle with elements of pragmatic comprehension, but can do it with some assistance (such as with opportunities for repeated plays). As in the existing plans for DIALANG 2.0, the learner would be directed towards a pragmatic comprehension diagnostic test to confirm whether or not this is the case. As we have not yet written items for DIALANG 2.0, we will draw here on an item from an existing test of pragmatic understanding developed by Roever (2006: 238), adapted from Bouton (1999), though with an additional answer option added to aid the mediation process. In the example below, the item assesses comprehension of pragmatic implicature:

Hypothetical item drawn from Roever (2006: 238), adapted from Bouton (1999), with additional answer option

Context provided to test-taker
Jack is talking to his housemate Sarah about another housemate, Frank.

Listening stimulus:
Jack: ‘Do you know where Frank is, Sarah?’
Sarah: ‘Well, I heard music from his room earlier.’

Multiple-choice item:
What does Sarah probably mean?

1. Frank forgot to turn the music off.
2. Frank’s loud music bothers Sarah.
3. Frank is probably in his room.
4. Sarah doesn’t know where Frank is.
5. Sarah will go and check on Frank.

In Roever’s implicature item, the answer is option 3. In this context, the statement “I heard music from his room earlier” is an indirect way of expressing the meaning that Frank is probably (still) in his room. In DIALANG 2.0 as originally conceptualised, an incorrect answer to this item would add to a tally of correct or incorrect item responses across the diagnostic test, leading to a decision about whether pragmatic understanding is indeed an area of challenge for the learner. Taking a mediation approach, however, we could extract further evidence about the learner from their performance on this item. In this case, we could apply Qin and van Compernolle's (2021) three levels of mediation, with prompts increasing in explicitness after each incorrect response. The following example describes the full set of prompts a learner would receive if they selected an incorrect answer option, illustrating a scenario where the learner is not able to get the correct answer after three prompts.

(1) Learner selects option 1 – Frank forgot to turn the music off.
(2) Prompt 1: *Listen again*
(3) Learner selects option 4 – Sarah doesn't know where Frank is.
(4) Prompt 2: *Listen again* + *hint 'why was there music in Frank's room earlier?'*
(5) Learner selects option 2 – Frank's loud music bothers Sarah.
(6) Prompt 3: *Do you think Frank is in his room?*
(7) Learner selects option 5 – Sarah will go and check on Frank.
(8) Learner is shown the correct response is option 3 – Frank is probably in his room.

Importantly, if a learner gets the correct answer after the first, second or third prompt – and a similar pattern is replicated across all the items in the diagnostic test – there is evidence to confirm the learner's initial self-assessment: that they can do listening for pragmatic understanding, but only with assistance.

The decision, based on this procedure, becomes more informative for the diagnostic system. If the candidate found the diagnostic test very easy, then feedback would be that this skill area is secure. However, contrary to the original conception of DIALANG 2.0, if the learner experienced significant challenge, and could not ably respond to the items even with assistance, then the learner would not be recommended (yet) to work further on this area as it would be deemed beyond their ZPD. Instead, the system would prioritise feedback on and further recommendations for areas where the learner consistently showed sensitivity to assistance as this would indicate that the learner is well-poised to develop further in that skill area at that level. The mediation approach is truly learner-centric in that it is concerned only with what the learner currently can just accomplish with assistance; it does not require a norming system of syllabus goals, targets or wider frameworks (see further discussion below). Feedback systems would need to be developed accordingly and tailored for the multiple outcomes.

10.4 Feasibility and challenges

In the previous section, we systematically thought through how plans for DIALANG 2.0 might change following the integration of a mediation approach. Mediation would bring two clear benefits. First, as discussed above, graduated prompts would provide additional layers of diagnostic information. Put simply, each item on each diagnostic test might yield three or four data points (showing independent performance and level of sensitivity to prompt) rather than just two (correct vs incorrect). This would greatly enhance the level of diagnostic precision. Second, and perhaps more importantly, the mediation approach provides a neat solution to setting priorities at the decision stage. In their article describing a theory of diagnostic language assessment, Alderson, Brunfaut and Harding (2015) noted that one of the key differences between diagnosis in surveyed professions and in language assessment was that:

> [s]everal of the professionals interviewed work in fields with a clear normative model on which to base diagnostic decisions (e.g. a healthy human body; a fully functioning computer system). It is much more difficult to locate a clear normative model for second/foreign language development, and this presents further challenges to developing a comprehensive theory of SFL diagnosis. (Alderson, Brunfaut, and Harding, 2015: 258).

In practice, normative models in language education may come from syllabi, frameworks, curriculum targets or norm referencing; however, DIALANG 2.0 needs to remain useful across multiple jurisdictions and contexts of learning. Further, the whole notion of comparison against a set of normative targets seems to run counter to the ideals embodied in DIALANG 2.0, which is to centralise the learner and their needs, and to be guided by a deep investigation into their language abilities. Mediation, and the potential to identify through diagnostic tools the skill areas where learners are in their ZPD, provides an elegant way to sidestep questions about normative models and, instead, to keep learners (and teachers) firmly focused on extending their own abilities according to personalised profiles.

Notwithstanding these benefits, though, we have also identified several challenges and areas where further thinking and empirical research would be required before a mediation approach could be implemented in DIALANG 2.0. We classify these as: (1) theory challenges, and (2) practical/resource challenges.

10.4.1 Theory challenges

Thus far, particularly in the context of C-DA, mediation has been applied more frequently in the skill areas of listening, reading and writing, and with broader proficiency-style assessment tasks. What is less clear for us in considering the role of mediation in DIALANG 2.0 is how prompts would best be integrated within the system for speaking, and also how they would work in the assessment of lower-level processes for the receptive skills. On the first point, for speaking we see great potential in interactional approaches to mediation that could be implemented through AI agents as opposed to interventionist approaches where prompts would be static and much less connected to speech produced in real time (see Jeon 2023). Such technologies, however, remain at the cutting edge of research and may not be easy to implement within DIALANG 2.0 in the short term.

On the second point, one of the key innovations of DIALANG 2.0 is that it takes a very principled approach to construct-definition based on theories of language comprehension, knowledge and processing. Thus, for listening, we envisage one skill area to target phoneme discrimination or segmentation tasks. How might graduated prompts be designed and employed in such scenarios? In the case of listening, there may be options such as exploring multiple plays, slowing down input (controlled by the learner), or including visualisations. However, choosing appropriate prompts for these kinds of lower-order processes would require further research before they could be fully implemented. According to current practices, mediation seems to assume some level of metalinguistic awareness; for many of the skill areas that we would want to test in DIALANG 2.0, learners might not be able to verbalise or reflect on their performance because the processes are automatic.

10.4.2 Practical/resource challenges

Some of the practical/resource challenges have been alluded to above, but they require dedicated thought because introducing mediation into DIALANG 2.0 would add significantly more complexity to an already complex system. We have already noted that self-assessment for skills areas would need to be twice as long (to include assessments of both independent and assisted performance), and the form of assistance would need to be adequately explained and understood by learners. This could create a situation where self-assessment becomes overwhelming for learners because it is too cognitively demanding, or where it threatens motivation and engagement with the diagnostic system. Beyond that, we would need to develop not only items but also theoretically sound graduated prompts for each item within each skill area at three levels across all skills. This would perhaps pose the most substantial practical challenges at the test design stage. However, it also translates into a resource challenge as this requires more time, more funding (for test developers), and a more complex interface for the software engineer who will design the system. A larger, more sophisticated system would require more computing power, meaning that the system would be less sustainable, harder to maintain, and leave a more substantial carbon footprint (see https://www.websitecarbon.com/).

10.5 Concluding remarks

Based on the foregoing discussion, our current position is that we would seek to implement some elements of mediation where possible in DIALANG 2.0, but this will be – to begin with – in a restricted and optional way. Specifically, we would imagine mediation might work best in the higher-order skill areas for listening and reading, as this would allow us to draw on the most substantial body of existing work on C-DA and to implement what DA experts have already successfully trialled in their work. We would also seek to leverage advances made in automated writing evaluation to integrate mediation into writing assessment where possible.

It has been instructive to consider the potential role of mediation in DIALANG 2.0, but our discussions have also prompted thoughts about how DA and DiagA researchers/practitioners might further extend their joint endeavours to solve some mutual problems. First, we need to fine-tune our understanding of how specific types of mediation would work across skill areas and for different task types and levels of processing. In other words, practitioners of all kinds would benefit from a comprehensive mediation "toolkit" that provides examples of prompts with evidence-based gradation from implicit to explicit. Second, just as we encourage other DiagA researchers to engage with mediation, we encourage more focus in DA research and DA practice on constructs: for DA researchers to consider more fine-grained diagnostic score reports based on close analysis of test items, and for DA practitioners to consider carefully the precise nature of the skills being mediated. Finally, we see a need to continue to discuss the commensurability of the two paradigms. Even after this detailed process of considering the potential for DA to inform a theory-based approach to diagnostic assessment, it remains unclear whether the two paradigms are fundamentally aligned or fundamentally at odds. We can certainly draw inspiration and ideas from one another, but this kind of blending of ideas – such as the approach described above – remains at a relatively superficial, design-level rather than a deep synthesis of the two paradigms. The current DD-Lang project may find that fundamental synthesis, but until then, we continue to see mutual benefits from the exchange of ideas.

References

Alderson, John Charles. 2005. *Diagnosing Foreign Language Proficiency*. London/New York: Continuum.

Alderson, John Charles, Tineke Brunfaut & Luke Harding. 2015. Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. *Applied Linguistics* 36(2). 236–260. https://doi.org/10.1093/applin/amt046

Alderson, John Charles, Tineke Brunfaut & Luke Harding. 2017. Bridging assessment and learning: A view from second and foreign language assessment. *Assessment in Education: Principles*, *Policy & Practice* 24(3). 379–387. https://doi.org/10.1080/0969594X.2017.1331201

Alderson, John Charles, Eeva-Leena Haapakangas, Ari Huhta, Lea Nieminen & Riikka Ullakonoja. 2015. *The Diagnosis of Reading in a Second or Foreign Language*. New York/Oxon: Routledge.

Alderson, John Charles & Ari Huhta. 2005. The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing* 22(3). 301–320. https://doi.org/10.1191/0265532205lt310oa

Alderson, John Charles & Benjamin Kremmel. 2013. Re-examining the content validation of a grammar test: The (im)possibility of distinguishing vocabulary and structural knowledge. *Language Testing* 30(4). 535–556. https://doi.org/10.1177/0265532213489568

Antón, Marta & Próspero N. García. 2022. Dynamic assessment. In Glenn Fulcher & Luke Harding (eds.), *The Routledge Handbook of Language Testing*, 171–186. Oxon/New York: Routledge.

Bouton, Lawrence F. 1999. The amenability of implicature to focused classroom instruction. Paper presented at TESOL, New York, USA, 1999.

Brunfaut, Tineke & Luke Harding. 2025. Diagnostic approaches in teaching and assessing listening. In Elvis Wagner, Aaron Olaf Batty & Evelina Galaczi (eds.), *The Routledge Handbook of Second Language Acquisition and Listening*, 255–267. New York/Oxon: Routledge.

Chapelle, Carol A. 2006. Test review: Dialang. *Language Testing* 23(4). 544–550. https://doi.org/10.1191/0265532206lt341xx

DAS (*DIALANG Assessment Specifications*) 1998. (Unpublished documents). DIALANG Project.

Field, John. 2013. Cognitive validity. In Ardeshir Geranpayeh & Lynda Taylor (eds), *Examining Listening: Research and Practice in Assessing Second Language*, 77–151. Cambridge: Cambridge University Press.

Grabe, William. 2009. *Reading in a Second Language: Moving from Theory to Practice*. Cambridge: Cambridge University Press.

Grabe, William & Junko Yamashita. 2022. *Reading in a Second Language: Moving from Theory to Practice* (2nd ed.). Cambridge: Cambridge University Press.

Gyllstad, Henrik, Laura Vilkaitė & Norbert Schmitt. 2015. Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL – International Journal of Applied Linguistics* 166(2). 278–306. https://doi.org/10.1075/itl.166.2.04gyl

Harding, Luke, John Charles Alderson & Tineke Brunfaut. 2015. Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing* 32(3). 317–336. https://doi.org/10.1177/0265532214564505

Huhta, Ari. 2007. The vocabulary size placement test in DIALANG: Why do users love and hate it? In Cecilie Carlsen & Eli Moe (eds.), *A human touch to language testing*, 44–57. Oslo: Novus Press.

Huhta, Ari. 2010. *Innovations in diagnostic assessment and feedback: An analysis of the usefulness of the DIALANG language assessment system*. Jyväskylä: University of Jyväskylä dissertation.

Huhta, Ari, Claudia Harsch, Dmitri Leontjev & Lea Nieminen. 2024. *The Diagnosis of Writing in a Second or Foreign Language*. New York/Oxon: Routledge.

Huhta, Ari., Sari Luoma, Matt Oscarson, Kari Sajavaara, Sauli Takala & Alex Teasdale. 2002. DIALANG: a diagnostic language assessment system for learners. In John Charles Alderson (ed.), *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Case Studies*, 130–145. Council of Europe.

Hulstijn, Jan H. 2007. The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal* 91(4). 663–667. https://www.jstor.org/stable/4626094

Isbell, Daniel R. 2021. Can the test support student learning? Validating the use of a second language pronunciation diagnostic. *Language Assessment Quarterly* 18(4). 331–356. https://doi.org/10.1080/15434303.2021.1874382

Jeon, Jaeho. 2023. Chatbot-assisted dynamic assessment (CA-DA) for L2 vocabulary learning and diagnosis. *Computer Assisted Language Learning* 36(7). 1338–1364. https://doi.org/10.1080/09588221.2021.1987272

Kektsidou, Nantia & Dina Tsagari. 2019. Using DIALANG to track English language learners' progress over time. *Papers in Language Testing and Assessment* 8(1). 1–30.

Khalifa, Hanan & Cyril Weir. 2009. *Examining Reading: Research and Practice in Assessing Second Language Reading*. Cambridge: Cambridge University Press.

Knoch, Ute. 2009. *Diagnostic Writing Assessment: The Development and Validation of a Rating Scale*. Frankfurt am Main/Berlin/Bern/Bruxelles/New York/Oxford/Wien: Peter Lang.

Kremmel, Benjamin. 2016. Word families and frequency bands in vocabulary tests: Challenging conventions. *TESOL Quarterly* 50(4). 976–987. https://doi.org/10.1002/tesq.329

Kremmel, Benjamin, Tineke Brunfaut & John Charles Alderson. 2017. Exploring the role of phraseological knowledge in foreign language reading. *Applied Linguistics* 38(6). 848–870. https://doi.org/10.1093/applin/amv070

Leontjev, Dmitri. 2014. The effect of automated adaptive corrective feedback: L2 English questions. *APPLES*: *Journal of applied language studies* 8(2). 43–66.

Li, Minzi & Xian Zhang. 2021. A meta-analysis of self-assessment and language performance in language testing and assessment. *Language Testing* 38(2). 189–218. https://doi.org/10.1177/0265532220932481

Matsuyama, Yoichi, Shungo Suzuki, Mao Saeki, Hiroaki Takatsu, Ryuki Matsuura & Yuya Arai. 2023. *Towards an explainable automated scoring of spoken interaction with a conversational AI agent*. Paper presented at the Language Testing Research Colloquium, New York City, USA, 5–9 June, 2023.

May, Lyn, Fumiyo Nakatsuhara, Daniel Lam & Evelina Galaczi. 2020. Developing tools for learning oriented assessment of interactional competence: Bridging theory and practice. *Language Testing* 37(2). 165–188. https://doi.org/10.1177/0265532219879044

Meara, Paul. 1996. The dimensions of lexical competence. In Gillian Brown, Kirsten Malmkjaer & John Williams (eds.), *Performance and Competence in Second Language Acquisition*, 35–53. Cambridge: Cambridge University Press.

Papageorgiou, Spiros, Mitchell J. Ginsburgh & Pablo Garcia Gomez. 2023. Assessment design issues in developing vertical scales for language tests. In Spiros Papageorgiou & Venessa F. Manna (eds.), *Meaningful Language Test Scores: Research to Enhance Score Interpretation*, 35–60. Amsterdam/Philadelphia: John Benjamins.

Paquot, Magali, Stefan T. Gries & Monique Yoder. 2020. Measuring lexicogrammar. In Paula Winke & Tineke Brunfaut (eds.), *The Routledge Handbook of Second Language Acquisition and Language Testing*, 223–232. New York/Oxon: Routledge.

Poehner, Matthew E. 2007. Beyond the test: L2 dynamic assessment and the transcendence of mediated learning. *The Modern Language Journal* 91(3). 323–340. https://doi.org/10.1111/j.1540-4781.2007.00583.x

Poehner, Matthew E. 2011. Dynamic assessment: Fairness through the prism of mediation. *Assessment in Education: Principles, Policy & Practice* 18(2). 99–112. https://doi.org/10.1080/0969594X.2011.567090

Poehner, Matthew E. 2012. The zone of proximal development and the genesis of self-assessment. *The Modern Language Journal* 96(4). 610–622. https://doi.org/10.1111/j.1540-4781.2012.01393.x

Poehner, Matthew E. & James P. Lantolf. 2013. Bringing the ZPD into the equation: Capturing L2 development during computerized dynamic assessment (C-DA). *Language Teaching Research* 17 (3). 323–342. https://doi.org/10.1177/1362168813482935

Poehner, Matthew E., Jie Zhang & Xiaofei Lu. 2015. Computerized dynamic assessment (C-DA): Diagnosing L2 development according to learner responsiveness to mediation. *Language Testing* 32(3), 337–357. https://doi.org/10.1177/0265532214560390

Poehner, Matthew E. & Zhaoyu Wang. 2021. Dynamic assessment and second language development. *Language Teaching 54*(4). 472–490. https://doi.org/10.1017/S0261444820000555

Qin, Tianyu & Rémi A. van Compernolle. 2021. Computerized dynamic assessment of implicature comprehension in L2 Chinese. *Language Learning & Technology* 25(2). 55–74. http://hdl.handle.net/10125/73433

Read, John. 2004. Plumbing the depths: How should the construct of vocabulary knowledge be defined? In Paul Bogaards & Batia Laufer (eds.), *Vocabulary in a Second Language: Selection, Acquisition, and Testing*, 209–227. Amsterdam/Philadelphia: John Benjamins.

Roever, Carsten. 2006. Validation of a web-based test of ESL pragmalinguistics. *Language Testing* 23 (2). 229–256. https://doi.org/10.1191/0265532206lt329oa

Roever, Carsten. 2022. *Teaching and Testing Second Language Pragmatics and Interaction*. New York/Oxon: Routledge.

Ross, Steven & Gabriele Kasper (eds.). 2013. *Assessing Second Language Pragmatics*. Basingstoke/New York: Palgrave.

Sternberg, Robert J. & Elena L. Grigorenko. 2002. *Dynamic Testing: The Nature and Measurement of Learning Potential*. Cambridge: Cambridge University Press.

Taguchi, Naoko. 2012. *Context, Individual Differences and Pragmatic Competence*. Bristol: Multilingual Matters.

Timpe-Laughlin, Veronika, Jennifer Wain & Jonathan Schmidgall. 2015. Defining and operationalizing the construct of pragmatic competence: Review and recommendations. *ETS Research Report Series* 2015(1). 1–43. https://doi.org/10.1002/ets2.12053

Xie, Qin. 2019. Error analysis and diagnostic assessment of linguistic accuracy: Construct specification and empirical verification. *Assessing Writing* 41. 47–62. https://doi.org/10.1016/j.asw.2019.05.002

Zhang, Su & Nancy Thompson. 2004. DIALANG: A diagnostic language assessment system. *The Canadian Modern Language Review/La revue canadienne des langues vivantes* 61(2). 290–293.