The Comprehensive Learning System

Barry O'Sullivan

April 2020

BCP-1

The Comprehensive Learning System

Summary

This paper proposes that for learning programmes to function efficiently they should be seen as a system. Within the system the three core elements (curriculum, delivery, assessment) must be based on a single philosophy of learning supported by clearly defined models of language ability and progression and underpinned by a measurement model. Failure to ensure that all three are fully in harmony is likely to lead to the failure of the system.

The system itself does not exist in a vacuum, but instead, is situated within a specific educational and social context, which is defined by the key stakeholders who comprise that context.

In addition to meeting the academic and pedagogic requirements of the three core elements, for a system to work well and be accepted, a clearly described theory of action based on the needs and expectations of the key stakeholders is critical, as is the need to communicate with these groups and individuals in a timely and appropriate manner. In this way, the social consequences (both intended and unintended) of the implementation of the system can be closely monitored and effectively dealt with.

Note
While this paper has been written with English language education in mind, it is clear that the approach proposed here has equal relevance across all areas and levels of education. Indeed, successful learning programmes around the world tend to focus more on the system and the coherence of its central elements than on individual components.

The Author

Professor Barry O'Sullivan is the British Council's Head of Assessment Research & Development. He has worked on numerous test development and validation projects globally and advises ministries and institutions on assessment policy and practice. He has undertaken research across many areas on language testing and assessment and its history and has worked on the development and refinement of the socio-cognitive model of test development and validation since 2000. He has presented his work at many conferences around the world, while almost 100 of his publications have appeared in a range of international journals, books and technical reports.

He is the founding president of the UK Association of Language Testing and Assessment, holds a visiting professorship at the University of Reading, UK and is Advisory Professor at Shanghai Jiao Tong University, China. He was awarded Fellowship of the Academy of Social Sciences in the UK in 2016, and was elected to Fellowship of the Asian Association for Language Assessment in 2017. In 2019 he was awarded an OBE by the government of the UK for his contribution to language testing.

Table of Contents

The Comprehensive Learning System

Introduction

For many years, the role of assessment[1] within learning has been largely bypassed in the language learning literature. Learning was traditionally seen as being very much aligned with matters of teaching, with teachers typically used textbooks based on a locally recognised curriculum and developed either by boards of education or less usually by commercial publishers who use their own curriculum, adapt existing materials to fit a local context, or commission materials to reflect a local curriculum. In the meanwhile, assessment was typically left in the hands of 'experts' who spoke a different language to curriculum developers and teachers alike. Essentially, the different aspects of the education system were seen as being quite independent of each other. Even today it is not at all unusual to find that curriculum, materials development, teacher training, teaching practice and assessment are all treated as independent entities within the learning systems that are focused on language and indeed other subjects. Attend any major language conference and you will find 'streams' devoted to these *different* areas of interest!

This paper sets out an overview of an alternative to the traditional perspective, viewing language (and other subject) education as a system in which each element is informed by a single philosophy or approach. The emphasis, therefore, is on developing a comprehensive, interactive, and unified framework of learning. The elements of this framework are seen as forming the three vertices of a triangle and consist of the curriculum, all aspects of the delivery system and assessment. These three broad areas, when conceived from a single perspective, form what I see as a comprehensive learning system (CLS). The paper also argues that all of this does not take place in a vacuum. The context in which the CLS sits is conceived as being defined by the people who comprise this context (the stakeholders) who bring with them specific pressures, needs and expectations.

The focus has been on English language, though the underlying approach can be applied across all areas of education

Background

As I point out above, assessment and testing has often been viewed as somehow at odds with broader principles of learning. Birenbauma et al. (2006) argue against contemporary approaches to assessment and testing systems, maintaining that they lack meaning to learners in terms of essential life skills. They also argue that assessment of learning limits the curriculum and does little to contribute to learning development. This point is also implied by Kelly (2009) in his comprehensive (if quite Britain-focused) analysis and representation of the many faceted aspects of the curriculum. Kelly's view of the curriculum sees assessment as essentially external and disruptive, though of course he recognises the position of in-programme assessment, arguing that it should be matched to the curriculum (p.147). One solution, proposed by Adair-Hauck et al. (2006) is integrated performance assessment (IPA) – a system in which the authors set out to integrate assessment and learning from the design phase of a language programme. They argue that assessment should be seen as reinforcing learning and that an IPA driven approach would allow such assessment to be more explicitly linked to the real-world needs of learners, an argument similar to that of Birenbauma

[1] In this paper, I use the term assessment to describe both processes that are used to make judgements about learners (assessment of learning) and processes that are developmental in nature (assessment for learning or learner orientated assessment). I use the terms *developmental* and *judgemental* assessments as I feel they offer a more transparent perspective on the purpose of the processes and decisions made.

et al. (2006). While the studies referred to here recognise the need to integrate assessment *for* and *of* learning, they fail, in my opinion, to recognise the need to integrate assessment more explicitly into the entire system.

The first paper to argue coherently for an integration of assessment into the learning system was that of Frederiksen and Collins (1989). The notion that assessment should be considered at the development stage of any language programme was further articulated by Shohamy (1992) and perhaps most eloquently in terms of modern languages for specific purposes by Norris (2006).

In many cases, the practice of system development ignores the need for a specific connection between the core elements of curriculum, delivery and assessment. In fact, there are many examples of systems where there is a deliberate attempt to separate assessment from the rest of the system. The argument being that this ensures fairness. Shepard (2000) makes this very point when supporting a social-constructivist approach to assessment. The model of what she refers to as "an emergent, constructivist paradigm" (Shepard, 2000, p.8) is summarised here in Figure 1 (the original contains additional bulleted explanations of each element). This model pre-staged the later work of Norris (2006) and O'Sullivan (2006) in recognising the link between the three elements within the learning system at a theoretical level. While Shepard focused her attention on developmental assessment in the classroom, there has been considerable interest in the USA in establishing evidence that judgemental (often referred to as summative) tests reflect the curriculum, see for example Porter (2002) and Porter et al. (2008) in which argument are made for a formal psychometrically sound linking claim based on the judgements of expert raters. These raters, individuals described as subject experts, are asked to review tests item by item making judgements as to their 'fit' with the curriculum (or the standards upon which the curriculum is based).

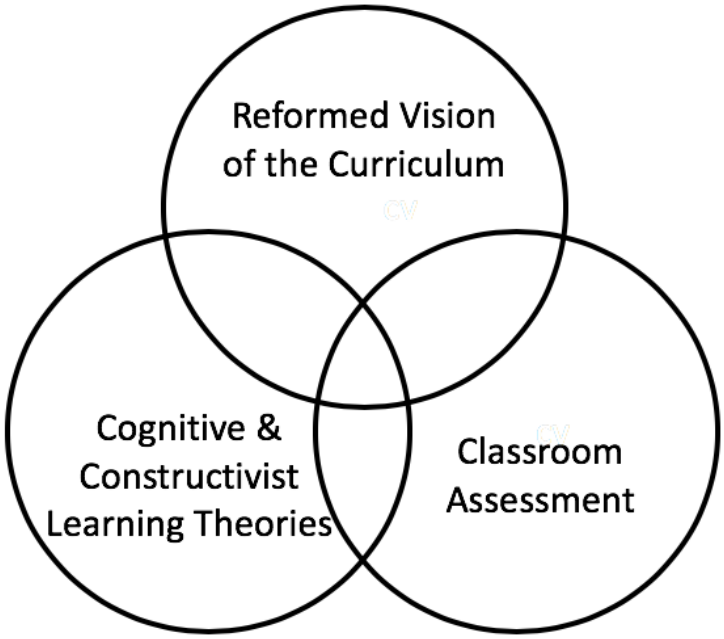

Figure 1: Overview of Shepard's (2000: 8) model of "an emergent, constructivist paradigm"

Bunch (2012: 1) begins his white paper for the USA-based Measurement Inc. with the following statement:

> *A key component of educational achievement test validation is alignment of the test to both curriculum and instruction. By alignment, we mean the degree to which the items of the test,*

> *both individually and collectively, match the structure and intent of the curriculum and instruction.*

He went on to propose a three-way model, with the curriculum, the classroom and testing forming the three elements. This model was designed to help test developers comply with the findings of a key US legal case:

> *The United States Court of Appeals (Fifth Circuit) ruled on May 4, 1981, that "The State may not deprive its high school seniors of the economic and educational benefits of a high school diploma until it has demonstrated that the SSATII (the Florida minimum competency test) is a fair test of that which is taught in its classrooms." (Debra P. v. Turlington 474 F. Supp. 244 (M.D. Fla., 1981)). (Bunch, 2012: 4)*

In an interesting final section to his white paper, Bunch (2012: 5-7) suggests that in addition to the sort of expert judgement-based approach suggested by Porter (2002) and Porter et al. (2008), we consider a more objective and observable criterion (in his case the later college or work readiness of school leavers). While this may be interesting, it may not be as easy as Bunch implies to find such evidence, given (in his example) the many variables that contribute to future success. It is certainly the case in the UK that school leaving examination results (A Levels) are, according to Murphy and Broadfoot (2017, p.71)

> *…by and large poor predictors of future educational success. There are plenty of understandable reasons for this: people mature at different rates, their interests and enthusiasms change and the subjects themselves make different intellectual demands at different levels.*

The emphasis in the work reviewed from the USA appears to be on the retro-fitting of tests to curricula and the classroom. I believe that a different approach has long been taken in the European context. The Common European Framework of Reference for Languages (CEFR) was published in 2001, following a period of discussion and development that stretched back almost three decades. The basis of the thinking around the CEFR was that it would influence how language learning systems across Europe might be designed. The area which embraced the CEFR most enthusiastically was that of testing and assessment, while the areas related to the design and delivery of the curriculum (see Table 3) lagged behind. The recently released CEFR Companion Volume (CEFR-CV) seeks to address this issue by focusing more on the learning and curriculum side – laudable in itself but again limiting thinking around language learning programmes and potentially exacerbating the current expert 'silos'.

The linking manual (Council of Europe, 2009) suggested a process through which a strong claim of a link between a test and the CEFR (Council of Europe, 2001) can be developed and supported. At a recent event in London in February, 2020, Kantarcıoğlu argued that the changes to the CEFR contained in the Companion Volume meant that the existing *Manual* needed to be updated. In the discussion that followed her talk, the point was made that any new *Manual* should not only focus on tests, but should also propose a linking process or processes for curricula, syllabi and course materials (textbooks in particular).

The Comprehensive Learning System

In a plenary address to the TESOL International conference in Tampa (2006), I suggested a model (presented graphically in Figure 2) in which all elements are considered with equal intensity *ab initio*, that is from the initiation of an education reform project. I believed then, and continue to believe, that a fully comprehensive learning system (CLS) can only hope to function successfully if its three

aspects (curriculum, delivery and assessment) are compatible[2]. By accepting that each of the elements of the CLS will have a significant impact on learning, developers of a programme should take all three into consideration when designing and developing a truly integrated learning system.

Norris (2006: 580) argues that "without a system for integrating assessment into program practice, FL [Foreign Language] educators will continue to do assessment where they must, though few will come to understand the value of using it." While his focus was on the role of assessment in English for Academic Purposes contexts, his stress on the need for such integration can, and should be broadened to the other elements of a CLS.

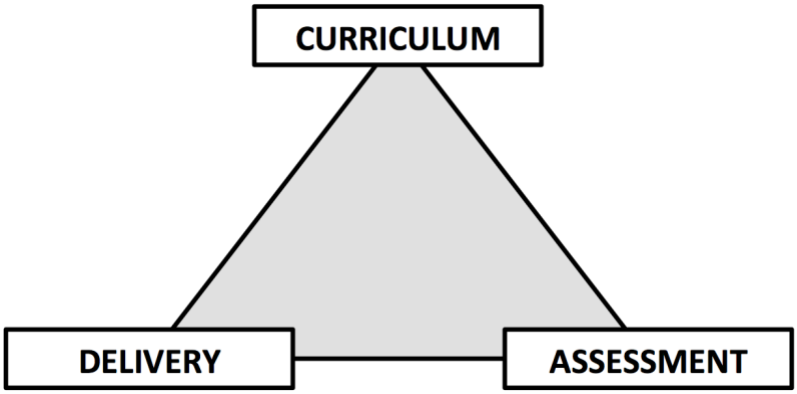

Figure 2: Modelling a Comprehensive Learning System (CLS)

As the name (and Figure 2) suggests, the comprehensive learning system marks a step away from our traditional models of learning programmes to a model in which all three of its major elements are closely connected. The essential underlying concept is that all elements and all of the detailed sub-elements should be informed by a single comprehensive philosophy of learning. This should then be recognised as a symbiotic system, in which any decisions or changes will have significant consequences across the model. It is my belief that this approach to learning system design is method agnostic. In other words, the need to fully integrate the different elements is critical irrespective of the theoretical approach to learning adopted by the developer. While I believe that this approach to learning system design is in essence method agnostic, it is important to ensure that the theoretical approach to learning adopted by the developer has a sound empirical basis and is context appropriate. This point drives the thinking behind the more elaborate model presented in Figure 5.

The different elements of the model are briefly discussed below. As the primary focus of this paper is on assessment within the learning system I will concentrate mostly on that area.

The Curriculum

Kelly (2009: 9), argues that the curriculum is more than "a statement about the knowledge content or merely the subjects which schooling is to 'teach' or 'transmit' or 'deliver'." Instead he argues that it should go beyond this to explain and justify precisely what is to be taught and the likely consequences inherent in the proposed system. In terms of a language curriculum, we would also expect the developer to consider not only aspects of the language and its use but also the social

[2] The concept of adopting a systems approach to education has been around for over 50 years, see Berger & Brunswic (1981) for a practical overview of the concept in practice.

consequences of teaching language or of teaching language in a particular way. Consideration should also be given at this design stage to value implications for delivery and assessment. This suggests that not all curricular aims are explicit, as in any curriculum there will be additional aims that are implied by factors associated with the context – e.g. the institution's language policy, the expectations and agendas of the learners etc. Another example of this has been referred to as the hidden curriculum. Kelly (2009: 10) sees the hidden curriculum as referring to learning that takes place or is partly provided for within the written curriculum, examples being social roles, gender roles and broader attitudes and values around these roles. It is important then that the system developer should take multiple perspectives on the curriculum, both formal and hidden, to identify the consequences, both educational and social, of all decisions made during the process. Messick's (1989: 20) matrix, Figure 3, suggests that the way in which we define a test construct (i.e. the underlying trait or ability we are testing) will involve questions of value, which, in turn, will impact on the way in which we interpret test scores.

	Test Interpretation	**Test Use**
Evidence Basis	Construct Validity	Construct Validity + Relevance/utility
Consequential Basis	Value Implications	Social Consequences

Figure 3: Messick's 'Facets of Validity' (1989: 20)

While the concept has never been fully operationalised in test practice, it is, nevertheless, relevant to the area of curriculum development (and of course to the decisions made in the area of delivery) in that the decisions that we make with regard to construct definition and operationalisation (i.e. describing exactly what aspects of language knowledge and use to target and how this is manifested across the curriculum) will clearly have significant implications for both the learning system and beyond to the society in which it is located. I will discuss this further in the following section.

The Delivery System

Traditionally, the delivery system has been seen as a process by which the formal curriculum is operationalised in specific learning contexts or domains. Within a CLS conceptualisation, this remains the case though additional consideration should be taken of a range of elements, these are outlined in Table 1.

The important point to take away from this table is that while other approaches have considered the classroom as the third pillar of the learning system, in the CLS approach we must take into consideration all aspects of the delivery of the curriculum

Table 1: The Delivery System

The physical environment	• the school building • the classroom • technology hardware • additional structures used in the formal or hidden/informal curriculum (e.g. gym, auditorium, outside nature areas) • playing areas or fields • surrounding community	
School staff	*Details*	*Relevant Areas*
	• leadership • teaching • administrative • management • support (technology, teaching etc.) • grounds staff • voluntary staff	• selection • training • continuing professional development • monitoring & evaluation
Learning materials	• formal • informal	• textbooks • reading material – books, articles, blogs etc. (hard or soft copies) • social media • technology software • games (online and real) • TV/films • Music/spoken word (radio/podcasts)

The Assessment System

As we have seen from previous sections, there have been calls for many years for a more holistic approach to assessment, whether this is *for* or *of* learning. However, some assessment practitioners were building what in effect amounted to CLS programmes twenty years ago, for example Weir for the teaching of English in Egypt in the late 1990s and O'Sullivan for English and Arabic in Jordan at around the same time. In each of these cases, an effort was made to create from a pre-existing curriculum and appropriately associated materials a system of assessment that fitted philosophically, culturally and methodologically. Nevertheless, as implied in the earlier sections of this paper, a fully functioning CLS requires considerable additional attention to the creation of the curriculum as the driving force, with all aspects of the delivery system taken into consideration. In addition, consideration should be given at the system design phase to the whole gamut of assessment opportunities that are associated with any programme. Figure 4 outlines these opportunities.

The most obvious reaction to this figure is to recoil in horror at the prospect of any programme developer taking advantage of all these opportunities – sadly it is not a rare occurrence and in many such programmes learners are heavily and needlessly over-assessed. However, these opportunities exist and it is the responsibility of the developer to design them into the system as *appropriate*. There should not be any room in the system for unplanned judgemental assessments that will impact on a learner's final grade or score.

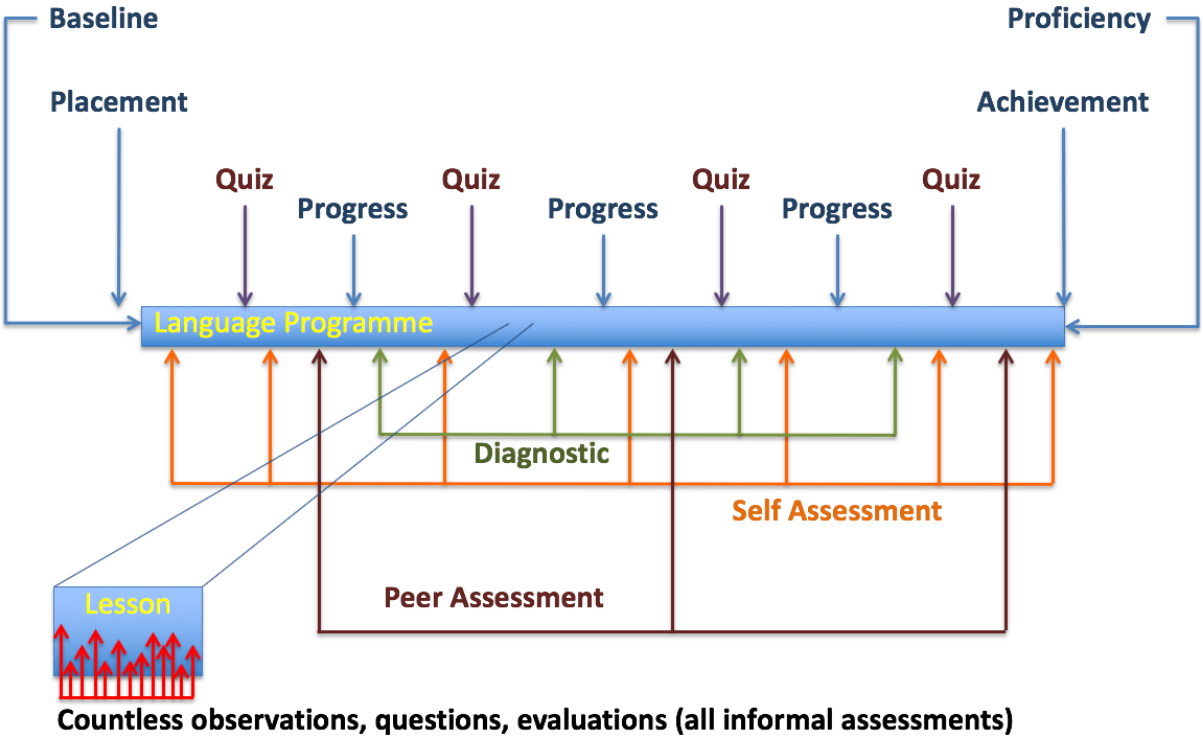

Figure 4: Opportunities for Assessment in Language Programmes

The opportunities for assessment are overviewed in Table 2

Table 2: Testing and Assessment Opportunities in Language Programmes

Opportunity	Description	Focus
Placement	Designed to give the teacher or school an idea of the level of language of incoming learners to place them appropriately into a class or level.	Developmental – the aim is to identify the level most appropriate to the learner
Benchmark	Here, results are used to form an opinion of the level of a group of learners (across a programme, school, district, region or country) prior to a programme. The results can be used as an indication of the level of incoming learners across the bracketed domains or across time. A benchmark test is commonly used in tandem with an exit version of the same instrument.	Developmental – no individual-specific judgements are made based on test performance. Instead the results ideally feed into programme development.
Progress	These are tests of what the learners have just been studying. They are used to ensure that all learners are on target to complete the course successfully. Remedial action may be recommended for learners who fail to demonstrate adequate learning. These tests are usually delivered at fixed points during the programme and can be either developmental or judgemental.	Judgemental, if the test outcomes contribute to programme grade. Even in this situation there is a developmental aspect, as outcomes are analysed on an individual basis to support learning. (though see the note at the foot of this table)

Opportunity	Description	Focus
Diagnostic	This can be a formal or structured assessment of how well a learner has grasped a concept that has just been taught. It is designed to identify a learner's strengths and areas for improvement and is always followed up with feedback (e.g. pointing the learner to a specific resource to help improve or consolidate learning). Without this element, the assessment cannot claim to be truly diagnostic (see Alderson, 2005).	Developmental only. Since these assessments are designed purely for developmental reasons they should never be scored even where the score does not contribute to a learner's overall programme grade.
Quiz	A quiz is a short, focused test that is designed to assess very specific knowledge – e.g. specific vocabulary. These are quite difficult to plan for as they are often delivered by a teacher in response to feedback from learners in the classroom situation. Used developmentally, such quizzes are a common feature of successful teaching. On the other hand, quizzes that are used for Judgemental assessment in a programme can be quite different. Such quizzes should be carefully considered and only administered in a pre-determined structured way to support the learning programme or system. Ad hoc judgemental quizzes should be avoided.	Can be either judgemental (if test scores contribute to overall programme grade) or developmental (if test performance informs individualised feedback to support learning). It is critical to ensure that learners are made aware of the nature of these quizzes (judgemental or developmental).
Achievement	This test comes at the end of a programme and is designed to look back at the language (or other content) studied as part of the programme. The expectation is that all students should perform well on this test. This assumption will have significant impact on the way we analyse and interpret test results – e.g. in a general proficiency test any items with a correct response rate of over 80% would be considered weak (as it tells us little about up to 80% of the candidates) while in and Achievement Test it would be considered fully acceptable.	Similar to Progress Tests, this is typically judgemental where the test outcomes contribute to programme grade. However, it is less likely to be developmental, as outcomes are rarely analysed on an individual basis to support learning.
Proficiency	This is an external measure of a learner's language and is independent of the programme. Such tests are sometimes the target of the programme or they are used as an external independent benchmark of ability. Examples of these range from certificated examinations such as Cambridge First to non-certificated (or institutionally certificated) tests such as the Institutional TOEFL.	Judgemental, as there is rarely if ever any developmental feedback offered in such tests. If overall performance contributes to programme development, then there is an argument that there can be some developmental outcome (though the same might be claimed of any test-type on this list).

Opportunity	Description	Focus
Self-Assessment	This is to be encouraged at all levels and at all stages of a programme. Indeed, the reflective nature of self-assessment is a critical aspect of learning.	Developmental – designed to encourage reflection on and the taking control of one's own learning.
Peer-Assessment	Another valuable reflective approach to understanding learning. By encouraging learners to peer assess, the teacher can instil an awareness of the complexities of the assessment and indeed learning process as it can be used to highlight critical parameters of language learning and use – e.g. the use of particular strategies or language structures. In terms of assessment, it can help learners raise their awareness and develop an understanding of the criteria that they are being judged on, thereby helping them to understand what makes a successful (or otherwise) performance. Note that for this work, learners must be explicitly taught how the different elements work.	Developmental – designed to encourage reflection on others' and own learning; also to encourage learners to take control of their learning.
Informal and Ongoing Assessment	All good teachers constantly assess the impact of their teaching in real time. This allows them to identify weaknesses in their own work and learners who are doing well and not so well. Without this on-going assessment, no successful teaching can take place.	Developmental – the assessment is dynamic (in that it takes place in real time response to real-time learning issues) and purposeful (in that it identifies specific issues and suggests specific solutions
Note It is critical that learners know *in advance* the purpose of an assessment/test. If they believe that it is in any way judgemental they will employ all available strategies to succeed, while if they believe that it is for developmental purposes they are far less likely to employ these strategies (thus giving a more accurate indication of their actual ability or knowledge). Policy makers and system developers should therefore be well aware of the limited developmental potential of judgemental tests, and of the dangers in confusing progress tests and diagnostic assessments.		

Within a learning system, any disconnect which isolates one corner of the triangle means that the system is immediately under threat. Some examples of what I mean here include using textbooks that are not a meaningful fit with the curriculum; introducing a significant change to the curriculum but not providing adequate training for teachers; selecting a test which does not test the language as it is conceived in the curriculum – e.g. a test of grammar when we are supposed to be focusing on speaking. Of course, ways in which the system can fail to function include the amount of testing carried out (as discussed above) or the over-importance placed on testing within the system. An example of the latter is the increased focus on accountability as measured by test performance. This type of evidence is extremely limited in that the measures are often poor in terms of the tests used (typically having a very narrowed construct definition – e.g. reading and/or grammar/vocabulary tested using multiple choice questions) and of the fact that this is typically the only evidence. This disregards the obvious fact that schools can be hugely successful even with relatively poor test results, for example in dealing with underprivileged children in areas of considerable deprivation.

The Learning System in Context

No CLS exists, or is developed, in a vacuum. In addition, the idea of creating one that is somehow learner or context agnostic (meaning it can be used anywhere with any learners) is equally preposterous. In order for any CLS to work, the philosophy that drives it must emerge from the context, and this context is defined by the stakeholders who populate it; that is the range of people who are likely to be affected by the system (see Figure 5). In this figure, which is informed by the socio-cognitive model of language assessment presented by O'Sullivan (2016) and modified by Chalhoub-Deville & O'Sullivan (2020), it should be clear that the theory of change (in our case a meaningful improvement in language learning) that drives the development and implementation of the system should itself be guided by the needs, expectations and hopes of the stakeholders.

In addition to understanding the important role of the stakeholders in shaping the CLA, it is also critical that developers understand the need to communicate appropriately with these stakeholders. While it is not unusual to find technical reports on how well learning systems are working, it is all too common to find that the vast majority of stakeholders, most especially those in the immediate firing line (e.g. learners, parents/guardians, teachers) are excluded from the interaction through the use of highly specialised technical language and dense, lengthy reports.

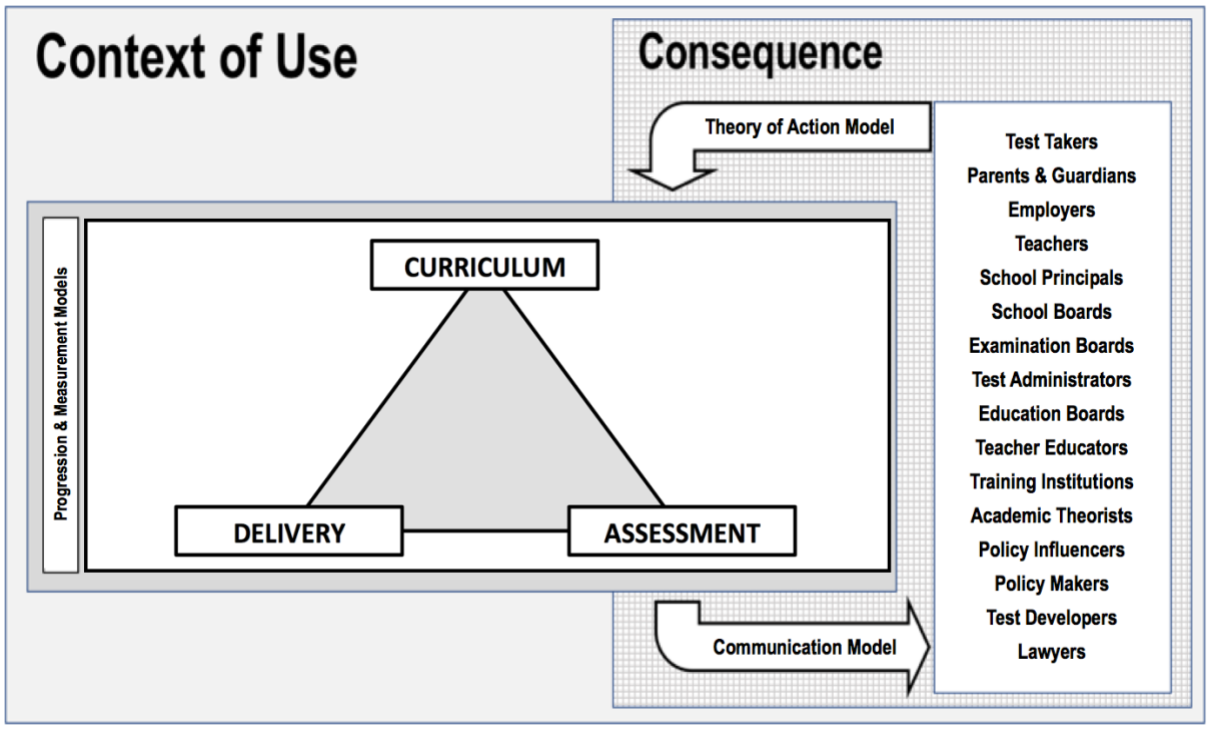

Figure 5: The CLS in Context (based on Chalhoub-Deville & O'Sullivan, 2020)

In the original model (Appendix A) the test system itself is represented by the three components within the central rectangle (Test Taker Model; Test Model; Scoring Model). This is informed by coherent models of language ability and progression and underpinned by a sound measurement model. For the model presented here, Figure 5 takes a similar approach with the learning system, placing it at the centre of a broader social context – the suggestion to do this comes from Dunlea (2016, 2018). This means that before considering the introduction of a significant reform project, the context in which it is to be introduced must be taken into account. This context is represented by the people who make it up. For the system to work, it must of course meet all the technical requirements across the three core elements. However, its success is equally dependent on the

appropriate communication of the intended change to all key stakeholders. Figure 5 gives an idea of the range of stakeholders who may need to be included. The history of education reform projects is littered with cases of failure due to its rejection by key stakeholders. This rejection often stems from the fact that key individuals and communities were not included in the initial thinking and rationale for the reform. Without the input of these individuals, no theory of action (or delivery plan for a theory of change) can be adequately prepared.

Making a Comprehensive Learning System Work

If we take just one aspect of language as an example, it should be relatively easy to demonstrate how a CLS should be envisaged. In this section, I will focus on reading at a specific language level (B1) to illustrate what I mean. Table 3 indicates what the CLS developer should be taking into consideration when working on this area.

The most important messages to take away from this table are:

- Without a clearly operationalised model of reading progression (i.e. the stages of reading learners pass through on their way towards the mastery level), the system will never work properly
- This model must drive all three elements of the system (curriculum, delivery and assessment)
- A full operationalisation of the model is dependent on as complete and accurate an understanding of the learner as is possible

Since the concept of suitability of reading texts is one of the critical elements of the operationalisation of the CLS, we would expect that the more personalised the system is to the individual learner the more likely it is that it will succeed. While this is the ideal, the traditional learning environment (the classroom) is probably not fully suitable as it is only at the higher proficiency levels that some degree of individualised learning is likely to occur. This is due to the number, application and motivation of learners.

Please note that while the example presented here focuses on reading, the same could be said of any skill area. If we are to move away from the traditional four skills approach (listening, reading, speaking and writing) to the approach proposed in the CEFR and the CEFR-CV (language production, reception as well as interaction and mediation in communication), we will need to develop, based presumably on the CEFR and the CEFR-CV, a series of comprehensive language models which can be operationalised within the CLS across all levels.

Vasager (2016) describes how a clearly defined philosophy of mathematics education in Singapore drives everything from the curriculum, to the way in which it is delivered in the classroom and assessed. The philosophy, summed up by Andreas Schleicher, the Director for the Directorate of Education and Skills at the OECD as “not about knowing everything. It’s about thinking like a mathematician.” The philosophy is operationalised through a problem-solving approach which has brought significant success, for example in the Trends in International Mathematics and Science Study (TIMMS).

Table 3: Operationalising CLS for Reading (broad outline) – Formal Curriculum Only

Expected Reading Performance (B1)		**Reading for comprehension at the paragraph level**
CLS Element	**Description**	**Focus**
Curriculum	A clearly stated model of reading progression is adopted Exactly how this is operationalised must be exemplified in the curriculum	The most practically useful model currently available is that of Khalifa & Weir (2009) In this model, the learner progresses from the word level, to sentences, paragraphs, multiple paragraph texts on concrete topics, multiple paragraph texts on more abstract topics, to intertextuality. At the B1 level, we expect that learners will be reading at the text level – though this will not be extensive in terms of length, see for example the British Council's Aptis test where the suggested length is approximately 140 to 160 words long (O'Sullivan and Dunlea, 2015). Other empirical estimates of text difficulty appropriate to B1 should also be clarified – reading age, grammatical and lexical complexity etc.
Delivery	Teacher Education	Selection and training of all teachers will focus on developing not only an understanding of subject knowledge (how reading comprehension is operationalised at B1) and teaching techniques, but also on the contribution to learning of the different elements of the CLS. This requires consideration of approach by training institutions and educators. An example of this is where teachers don't have some understanding of the principles of assessment literacy, they will struggle to achieve some of these steps (from fully understanding how the curriculum is implemented, to classroom delivery and assessment). For example, without some understanding of language testing theory, they will struggle to write appropriate test or assessment items, they may not understand the importance of having a text at a given level across cohorts (and how to assess the level of the text in the first place), as well as being unable to measure success accurately.
	Teaching – teachers are trained on the same underlying model.	Teachers can identify and/or prepare suitable texts and guide learning appropriately. Teachers can offer additional support to learners as they are comfortable with the expectations of the model.
	Teaching – indicators of successful learning are clearly stated	The learner and the teacher are equally aware of what is needed to demonstrate that comprehension has been achieved – as is the case with the British Council's Core Curriculum (and the same organisation's Aptis test service defines candidate ability as the comprehension of the structure and meaning of appropriate paragraph-level texts).

Expected Reading Performance (B1)		Reading for comprehension at the paragraph level
CLS Element	**Description**	**Focus**
	Materials – suitable texts are included in the learning materials.	Materials actively support the learning of approaches to reading comprehension at this level. This means that the texts are as learner-specific as possible – in terms of language level and complexity as well as topic.
	Materials – suitable tasks are set to encourage the learning and practice of appropriate reading skills.	This relates to what we ask the candidate/learner to do with a text. Simply answering a series of multiple choice questions or filling in blanks in the text does not guarantee comprehension (especially in the case of the latter), while re-telling the content or commenting on the attitude of the writer, in relatively concrete terms, is more likely to reflect the uses to which a learner might put a text in real life.
	Physical – the physical space is suitably quiet and laid out or the computers and internet bandwidth are adequate.	The design of the learning space takes the demands of quiet reading into account (e.g. in terms of space, time, noise, etc.). This will be the case regardless of the mode of delivery (face-to-face or via computer).
Assessment	The texts are suitable for the reading activity and for the learners.	The texts included in the test are screened to ensure that the are suitable in terms of language level and complexity as well as topic.
	Tasks are based on the model of reading progression.	The texts used in the text reflect those used in the learning context (class, online or blended).
	Success is identifiable and accurately measured	Appropriate standard setting (i.e. identifying what a test taker should do to demonstrate comprehension at this level) is undertaken. Test performance and score analysis is undertaken to ensure accuracy and consistency
	Score/Grade is meaningful and of value	The score or grade reflects the realistic expectations of the learner and/or teacher – evidenced from classroom behaviour and/or other appropriate test results.

Note: Suitable here refers to how well the text 'fits' with the expected learning and meets the needs of the target learners in terms of language, topic and complexity

Localisation and the CLS

The growing awareness of the need for localisation of assessment products where local decisions are made based on test performances suggests an additional complexity, see O'Sullivan (2011, 2019) and Weir (2019). Since test localisation refers to the act of ensuring the appropriateness of test approach and content to a specific group of test takers, we can assume that it would mean something very similar in a CLS approach to programme design and delivery. In practice, this entails making all the relevant decisions around curriculum, delivery and assessment with the local learner population and context in mind.

Of course, the ultimate in localisation is to move beyond the population to the individual. Personalising learning has always been an ideal rather than a reality – the traditional learning system sees learners grouped together in classes, so the most local the system can become is at the class level, and even this is down to the ability of the teacher to ensure that all materials and tasks are appropriate to the learners who comprise the class.

A technology-driven approach to personalisation is the most likely approach to work as it can be harnessed to select texts and/or tasks for learners that are suitable along several parameters, including topic (as identified by the individual learner), level (as identified by their success or lack of success with other suggested texts) and complexity or difficulty (as indicated by success or failure with earlier texts or tasks.

Conclusion

The concept of a comprehensive learning system is straightforward and easy to rationalise. It takes little technical knowhow to understand that the triangle of elements needs to be in harmony for a learning system to work as planned. However, it is clear from the presentation of the different elements in this paper that such a system can be quite complex to design and deliver. It is also clear that the complexity involved means that this approach will be most obviously beneficial to newly conceived education reform projects. It is possible to retro-fit tests into systems, as has been argued by Porter (2002) and Porter et al. (2008), though it has to be recognised that the incremental changes to the test that their proposal necessitates mean that it takes some time to get the test fully compliant with the curriculum. In the meantime, learners are losing out as their major judgemental tests are not fully focused on what they will have studied. It should always be recognised, of course, that there will be limits to what a retrofitted test might be able to achieve if, for example, the delivery and curriculum elements are very distant from the original test construct.

The approach proposed here is not new. We have always known of the importance of the elements described in the learning system. We have also known of the importance of establishing an empirically derived link between a judgemental end-of-programme test or developmental within-programme assessment and the curriculum. What is new is the idea that a learning programme:

- must be driven by a single underlying philosophy of language learning as operationalised by a set of clearly defined standards.
- should be recognised as a unified or integrated system, and not as a series of components.
- consists of three core elements which must be explicitly linked for the system to work.
- sits within a specified educational and social context – it is therefore highly unlikely that a system can be exported from one context to another due to different social, cultural and educational conditions.

Facilitating Successful Education Reform

At the time of writing this paper, many governments and institutions across the world are either considering the reform of their education system, developing policies around reform or actively engaged in the process of reform. While there will be cases in which full system reform is being considered, it is likely that there are many more in which single elements of the system are the focus of attention, for example changes to university entrance tests that are expected to lead to a greater emphasis on communication in language classes. The history of educational reform is littered with failure, and when we recognise the complex nature of reform (as indicated in Figure 5) it is not at all unexpected. While many reformers have a clear theory of change in mind (i.e. they are clear in what

they see as the ultimate goal of the reform), they fail to build a comprehensive theory of action through which this change can come about. By this I mean they fail to recognise the need to address all three core elements in the reform agenda.

Another significant issue lies in the failure of reformers to adequately communicate their agenda and their proposed solution. Different stakeholders have different needs and expectations and while it is not possible to satisfy all of these, the reformer must find ways to communicate in a meaningful way to these different stakeholders. While the underlying message should remain the same, the approach to communication will change depending on the stakeholder. For example, a long and highly technical report detailing the educational and economic benefits of pursuing a particular agenda in a particular may be well received by other ministries, academics and professionals, this mode of communication is unlikely to satisfy learners, parents and members of the general public.

In order to develop and operationalise a fully comprehensive learning system, the developing team should consider the following three sets of recommendations:

1. In order to facilitate the success of any educational reform initiative it is therefore necessary to:

a) Define clearly how the reform will change learning (Theory of Change)

b) Outline clearly, in terms of a CLS model, how this change will happen (Theory of Action)

c) Identify the needs and expectations of all key stakeholders in order to build a communication plan to engage meaningfully with these stakeholders

2. In order to build an appropriate comprehensive learning system, we should:

a) Start the process with a fully integrated comprehensive learning system as the goal.

b) Focus equally on all elements and not expect that the new system will function well with a limited or targeted focus – e.g. changing the test to promote change in the system.

c) Recognise that the current approach in which individual expert groups work in separate silos is unlikely to deliver – the development should be carried out by a team of people with expertise across the three core elements, and who are fully aware of the context in which the system is expected to work.

d) Consider, in light of localisation and personalisation, broadening the scope of the development team beyond language teaching and assessment to areas such as local educational and social expertise, educational technology expertise, artificial intelligence and machine learning experts, user experience experts (to help with interface design), and marketing and market insight experts (to understand the potential commercial value of a system where this is appropriate).

3. When attempting to retro-fit learning or tests to an existing curriculum:

a) Understand that the process may take some considerable time (and expertise) and accept that a negative outcome is possible.

b) Understand that simply matching materials (or tests) to a curriculum using an expert panel-based approach is never going to be enough. Classroom observations as well as focus groups with teachers and learners will also be needed in order to ensure that what is planned by materials developers to happen actually happens.

c) Apply, where appropriate, points 3 and 4 above.

A Note on Terminology

In this paper, I have used the terms *developmental* and *judgemental* assessment in place of more commonly used terms such as formative and summative assessment. I do this because I feel that the terms are more immediately transparent in terms of underlying meaning.

References

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. New York, NY: Continuum.

Adair-Hauck, B., Glisan, E. W., Koda, K., Swender, E. B. & Sandrock, P. (2006). The Integrated Performance Assessment (IPA): Connecting Assessment to Instruction and Learning. *Foreign Language Annals,* 39: 359–382.

Berger, G. & Brunswic, E. (1981). *A Systems Approach to Teaching and Learning Procedures*. Paris: The UNESCO Press. Accessed on March 11, 2020 from: https://unesdoc.unesco.org/ark:/48223/pf0000046025

Birenbauma, M., Breuer, K., Cascallar, E., Dochy, F., Dori, Y., Ridgway, J., Wiesemes, R. & Nickmans, G. (2006). A learning Integrated Assessment System. *Educational Research Review*, 1: 61–67.

Bunch, M. B. (2012). *Aligning Curriculum, Instruction, and Assessment*. Accessed on February 12th, 2020 from: https://measurementinc.com/sites/default/files/2017-08/Aligning%20Curriculum%2C%20Assessment%2C%20and%20Instruction%20%28M.%20Bunch%2C%202012%29.pdf

Chalhoub-Deville, M. & O'Sullivan, B. (2020). *Validity*: *Theoretical Development and Integrated Arguments*. British Council Monograph Series. London & Sheffield: British Council & Equinox Publishing.

Council of Europe (2001). *The Common European Framework of Reference for Languages*. Cambridge: Cambridge University Press.

Council of Europe (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR): A Manual*. Accessed on February 14, 2020 from: https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680667a2d

Dunlea, J. (2016). *Aligning Tests to the CSE: Validity Issues External Validity.* Paper presented at the International Conference on Language Testing and Assessment, Beijing, December 1-3.

Dunlea, J. (2018). *Linking Examinations to the CSE*. Paper Presented at the British Council New Directions Event, Kumming, China, April 23.

Frederiksen, J. R. & Collins, A. (1989). A Systems Approach to Educational Testing. *Educational Researcher*, 18: 27-32.

Kantarcıoğlu, E. (2020). The CEFR: A Road Map For Future Research and Development – Assessment Perspective. Joint UKALTA & EALTA conference on the CEFR. London, February 7 & 8.

Kelly, A. V. (2009). *The Curriculum: Thoery and Practice, 6th Edition*. London: Sage.

Khalifa, H. & Weir, C. J. (2009). *Examining Reading*. Cambridge University Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp.13-103). New York: Macmillan.

Murphy, R. & Broadfoot, P. (2017). *Effective Assessment and the Improvement of Education: A Tribute to Desmond Nuttall* (Routledge Revivals – first edition published in 1995). Oxford: Routledge.

Norris, J. (2006). The issue: the why (and how) of assessing student learning outcomes in college foreign language programmes. *Modern Language Journal*, 90, iv, 576-583.

O'Sullivan, B. (2006). *Daring to Lead or Failing to Follow: 40 Years of Language Testing*. Plenary Address, TESOL International Conference, Tampa Florida.

O'Sullivan, B. (2011). Introduction – Professionalization, Localisation & Fragmentation in Language Testing. In Barry O'Sullivan (ed) *Language Testing: Theory & Practice* (pp.1-12). Oxford: Palgrave.

O'Sullivan, B. (2016). Validity, Validation and Development: Building and Operationalizing a Comprehensive Model. *Japanese Language Testing Association Journal*, pp.25-33.

O'Sullivan, B. (2019). Forward - Localisation. In, Lily I-wen Su, Cyril Weir, Jessica R. W. Wu (Eds.) *English Language Proficiency Testing in Asia: A New Paradigm Bridging Global and Local* (pp.xiii-xxviii). New York: Routledge.

O'Sullivan, B. & J. Dunlea. 2015. *Aptis General Technical Manual: Version 1.0*. TR/2015/005. London: British Council. Accessed March 5th, 2020 from www.britishcouncil.org/sites/default/files/aptis_general_technical_manual_v-1.0.pdf

Porter, A. C. (2002). Measuring the Content of Instruction: Uses in Research and Practice. *Educational Researcher*, 31(7), pp. 3–14

Porter, A. C., Polikoff, M. S., Zeidner, T. & Smithson, J. (2008). The Quality of Content Analyses of State Student Achievement Tests and Content Standards. *Educational Measurement: Issues and Practice*, 27(4), pp. 2-14

Shepherd, L. A. (2000). The Role of Assessment in a Learning Culture. *Educational Researcher*, 29(7): 4-14.

Shohamy, E. (1992). Beyond proficiency testing: a diagnostic feedback testing model for assessing foreign language learning. *Modern Language Journal*, 76, iv, 513-521.

Vasager, J. (2016). Why Singapore's kids are so good at maths. London: Financial Times (July 16, 2016). Accessed on April 7, 2020 from: https://www.ft.com/content/2e4c61f2-4ec8-11e6-8172-e39ecd3b86fc

Weir, C. J. (2005). *Validity and Validation: an evidence-based approach*. Oxford: Palgrave.

Weir, C. J. (2019). Global, Local or "Glocal": Alternative Pathways in English Language Test Provision. In, Lily I-wen Su, Cyril Weir, Jessica R. W. Wu (Eds.) *English Language Proficiency Testing in Asia: A New Paradigm Bridging Global and Local* (pp.193-225). New York: Routledge.

Appendix A

Socio-Cognitive Model as Integrated Arguments (from Chalhoub Deville & O'Sullivan 2020)