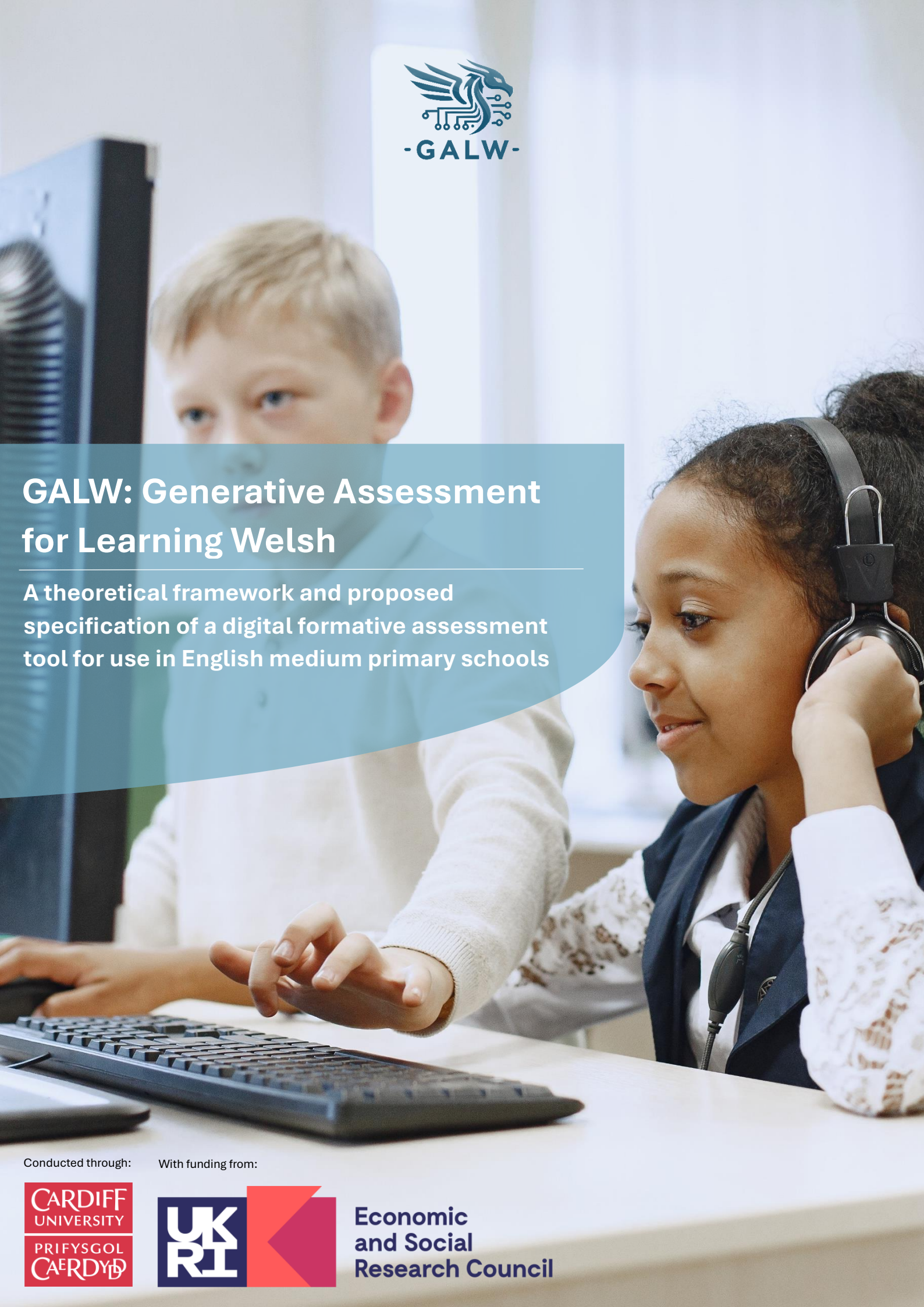

-GALW-

GALW: Generative Assessment for Learning Welsh

A theoretical framework and proposed specification of a digital formative assessment tool for use in English medium primary schools

Conducted through:

Cardiff University
Prifysgol Caerdydd

With funding from:

UKRI Economic and Social Research Council

GALW: Generative Assessment for Learning Welsh

A theoretical framework and proposed specification of a digital formative assessment tool for use in English medium primary schools

Author: Peter Russell

Research supervisor: Dr Sion Jones (Senior Lecturer, Cardiff University School of Social Sciences) and Dr Katy Jones (Senior Lecturer, Cardiff University School of English, Communication and Philosophy).

This document was generated as part of a doctoral thesis through Cardiff University's doctoral programme and the WGSSS.

Published: July 2025

Ethical approval for the scoping study included was granted by the School of Social Sciences Research Ethics Committee on the TBC (SREC No. 733)

The research was indirectly funded by an ESRC Studentship

Cover image: copyright free, credited to Gustavo Fring, available at: https://www.pexels.com/photo/boy-and-girl-sitting-in-front-of-a-computer-5621950/

Contents

Acronym List

ADHD/ADD – Attention Deficit and Hyperactivity Disorder / Attention Deficit Disorder
AEQ – Achievement Emotions Questionnaire
ALN – Additional Learning Needs
AoLE – Area of Learning and Experience
APD – Audio Processing Disorder
CDF – Cumulative Distribution Function
CEFR – Common European Framework of Reference
CfW – Curriculum for Wales
CIM – Constructive Integration Model
CSC – Central Sount Consortium
CTT – Classic Test Theory
DE – Distractor Efficiency
EAL – English as an Additional Language
EM – English Medium
HAL – Higher Achieving Learner
IELTS – International English Language Testing System
IRT – Item Response Theory
L1 – First Language
L2 – Second Language
LAL – Lower Achieving Learner
LLC – Literacy Language and Communication (one of the AoLEs of the CfW)
MAT – More Able and Talented
MCQ – Multi-Choice Question (AKA selected response items)
NAS – Negative Affective Schemata
NAW – National Assembly for Wales
PTE – Pearson Test of English
RLP – Reference Language Proficiency
SEM – Standard Error Measurement
TL – Target Language
TLU – Target Language Use
TOEFL – Test of English as a Foreign Language
TP – Target Population
ULI – Unique Learner Identifier
WAG – Welsh Assembly Government
WG – Welsh Government
WL – Welsh Language
WLC – Welsh Language Continuum
WLEB – Welsh Language and Education Bill
WM – Welsh Medium

1. Introduction

This document serves two key purposes, to outline the structure of the Generative Assessment for Learning Welsh (GALW), and to provide the theoretical basis for that structure. 'Galw' in Welsh means to 'call' and it seems a fitting acronym for an assessment intended to enhance the teaching and learning of Welsh in English medium schools.

The format of this document is intended to provide sufficient background information from the academic and policy literature that those without a background in Welsh language policy, linguistics, or assessment can understand both the structures of the GALW and the theoretical/pragmatic basis for the design decisions. This does mean that the specification here is more extensive that many comparable documents, and an executive summary is provided in section 1.2 to give an overview of the GALW, from which readers are able then to direct their reading more selectively, should they choose.

In this initial chapter the main purposes, context, and scope of the assessment will be explored. This is followed in chapter 2 by a detailed appraisal of the objectives of the assessment, with reference to the specific language constructs targeted, and the alignment of the GALW with broader measures of linguistic competency, such as the CEFR. Chapter 3, explores the characteristics of the target population, and considers how individual learner profiles interact with the assessment format and content. These initial three chapters form a contextual foundation upon which the assessment design is built, and against which it is appraised. Chapter 4 explains the specific features of the assessment design including its structure, delivery, and outputs. Chapter 5, concerns the source and curation of assessment content, including the relationship between the GALW and the WLC, and the development and refinement of distractors. Chapter 6 explores the validation process for the assessment, isolating cognitive, context, construct, consequence and scoring validity, and considering the way in which validation is considered in the assessment design, and its iterative refinement. Chapter 7 focuses on the scoring and feedback elements of the assessment beyond the remit of validity, including the presentation and interpretation of scores for learners, teachers and researchers.

1.1 Purpose of the Assessment

The GALW is designed to assist both learners and teachers in second language (L2) Welsh language learning in English medium (EM) primary settings, whilst providing insights into language acquisition for academic purposes. The goals of the provision therefore fall into three categories: learning, pedagogy, research.

Learning

- To ensure foundational language knowledge is established and maintained to provide a solid foundation for the acquisition of higher order communicative skills.
- To enhance learner motivation and self-efficacy through perception of progression and competence.
- To improve learning continuity in transitional periods (primary to secondary education, inter-key stage or progression step).

Pedagogy

- To allow teachers to easily carry out assessment for learning: identifying areas of the Welsh Language Continuum (WLC) which require additional or remedial attention.

- To improve differentiation and scaffolding based on a more accurate understanding of the specific class needs.
- To identify individual learners who may require additional support in their Welsh language learning.
- To gain an understanding of different elements of class and learner competencies and ensure appropriate progress is being made within each area.

Research

- To facilitate the collection of comparable data on learner progress in developing WLC knowledge. This data can then be tracked against other factors to determine the impact of variations in pedagogic approaches, learning provision, socio-economic factors, and individual learning strategies.
- To allow for the tracking of different cohorts and populations over time to understand early Welsh language development in the EM sector and how this aligns with the goals of Cymraeg 2050 and the Welsh Language and Education Bill (B2 on the CEFR).
- To provide a tool that empowers teachers to conduct their own action research. Such research can then inform individual practice, the practice of colleagues, and institution/cluster policy or approach to Welsh language teaching.

1.2 Executive Summary

This executive summary is intended to provide an overview of the key features of the GALW, its scope, functionality, application and outcomes. You will also find links to the sections in the main specification which provide background and more detail on each aspect of the assessment.

Scope of the GALW

Geographical

The GALW is intended as an open, free to use, AfL assessment tool. It will be available to all schools across Wales through the Hwb platform. The GALW includes WLC content blocks from ERW, CSC, GWE, and EAS, reflecting the dialectical diversity of the country, making it suitable for schools in all regions. This accommodation of all schools will allow for a holistic understanding of early language development across the whole EM sector. Chapter 6 details how these different dialectical forms are included without compromising assessment alignment with learning.

Institution and User profile

The GALW is intended for use in mainstream EM and dual-language schools. The format of the assessment makes it suitable for self-administration by learners in progression step 2+, but it could also be deployed in younger age groups with support. It's possible it could also be used in immersion units to track the early development of lexical knowledge in learners transferring into WM education. We would not anticipate the tool being used extensively beyond the start of secondary education, where the use of specialist teachers and a greater focus on communicative skills over content acquisition makes the GALW less appropriate. However, in cohorts with particularly low levels of content knowledge, or as a diagnostic tool to identify areas of lexical knowledge deficit, it is possible that the GALW could be used in a supplementary diagnostic fashion. A full discussion of the assessments outputs and their uses can be found in chapter 5.7.

The provision will incorporate accessibility features for learners with additional learning needs (font and text size adaptations, text to speech interface). Functionality has also been included to try to mitigate this effect of L1 proficiency (e.g. text and speech items, repeat options, culturally neutral content) to avoid disadvantaging EAL learners. A full discussion of how learn profiles are considered and accommodated can be found in chapter 4.

The format of the GALW

User interface

The GALW is a digital tool that can be accessed through a variety of devices (tablets, laptops, desk-top computers). The provision is self-administered with only minimal teacher support required once users are familiar with the provision. A single interaction will involve the learner completing four tasks, taking a total of 10-15 minutes. Use of the provision requires no planning or preparation, and feedback is automatically generated for review by staff when convenient. This ease of use and short duration is designed to encourage frequent use and minimize impact on learner routines. A full discussion of the user interface can be found in chapter 5.

Content selection

The content of the GALW is based on the WLC content lists produced by the educational consortia. This approach was selected to ensure maximum correlation between class and assessment content, mitigating negative back-wash effects. The patterns and content of the WLC are broken down into 'blocks' defined by grammatical or communicative function. The GALW uses a teacher-selective item list, drawing only on blocks that the learners have covered. This ensures that the data is only indicative of learning/retention, rather than coverage, and avoids potential de-motivation caused by uncalibrated norm-referenced models. A full discussion of content selection and design can be found in chapters 5 and chapter 6.

Constructs and Tasks

The MFLA consists of four tasks aimed at eliciting skills associated with two key constructs: receptive language knowledge, and lexical-syntactic parsing (chunking). Users are presented with:

- Task 1 - 10 multi-choice questions (MCQs) in L1 to L2 pattern translation
- Task 2 - 10 MCQs in L2 to L1 pattern translation
- Task 3 – up to 10 MCQs drawn from Task 1 and 2 asking users to identify key syntactic features within the item phrase.

A full discussion of the construct selection and validation can be found in chapter 3.1, and a more detailed description of the assessment tasks in chapter 5.

Scoring and Analysis

The MFLA produces data at three different levels – user data, teacher data, and research data.

- *User data* – User data is intended to support the learning process and develop improved motivation and self-efficacy. User data outputs are automated and qualitative/directive in nature: at the end of the 'quiz' learners are provided with details of whether they have improved their performance, and (based on that performance) which 'block' they should work-on before they try again.
- *Teacher data* – This is collated automatically and is presented as a class profile. Each learner will have 3 scores relating to the constructs and direction of translation (L1-2 score, L2-1 score, and a 'chunking' score) allowing teachers to identify specific

construct deficits (e.g. a formulaic rather than parsed understanding of content) in particular learners. Particular 'blocks' on which the class are performing poorly will also be highlighted in the data, helping to direct teaching and support more effectively. The scoring will also include two aggregated scores of learner competency. Firstly, a predictive vocabulary score that aims to infer the size of the learner's lexicon, and secondly, a CEFR level which score-links GALW outcomes with CEFR performance. For more discussion of these different elements see chapter 3.2 for details of CEFR linking, chapter 5.7 for details of the teacher facing outputs, and chapter 7.5 for information concerning scoring validity.

- *Researcher data* – This data set strips out identifying information allowing researchers to utilize information without compromising user confidentiality. It also includes latency data that may be indicative of learner competence/confidence that would not be of pedagogical use (response time, number of option changes, number of audio item plays). For information about the research applications of GALW data see chapter 5.7. For information on data security and user confidentiality see chapter 8.

Project outputs and impacts

Planned outputs

- A combined learning and assessment tool available to teachers across Wales that supports learning and teaching and can be easily integrated into existing approaches and provision.
- National level data revealing trends in Welsh language acquisition across the EM sector.
- An initial report exploring the learning outcomes data will be made in relation to different types of existing provision, approaches and socio-economic contexts, with an emphasis on the interactions with learning attitudes and motivation.

Anticipated impacts

- Improved learning outcomes for pupils in primary education through more responsive and informed pedagogy.
- Greater understanding of the provision and approaches that have a positive effect on the learning of pupils in different socio-economic contexts, leading to improved guidance for teaching staff.
- Improved learning continuity and consistency through transition periods (inter/intra institution), leading to enhanced learner motivation and teaching efficiency.
- A more accurate and objective understanding for teachers of the progress of their class/institution within the WLC, leading to improved allocation of time/resources in alignment with learning outcome aspirations.
- The future development of more effective learning resources through the facilitation of comparable assessment data, allowing for more effective control trials.
- A better understanding of the impact of different provision and training leading to the more efficient and effective allocation of resources and funding.
- A better understanding of Welsh language progression in EM schools, informing the development and monitoring of Welsh Government education policy.

2 Context and Background

The broader background of Welsh language education is beyond the remit of this specification. Here only context as it relates to assessment in the EM primary school sector will be explored. The only exception will be to highlight how current learning outcomes do not appear to facilitate learners reaching their potential (Estyn 2024) or align with the aspirations of Welsh Government policy (WG 2017).

Four key areas require exploration to contextualise the GALW: the Curriculum for Wales (CfW), the Welsh Language Continuum (WLC), current assessment practice, and the forthcoming Welsh Language and Education Bill (WLEB).

2.1 The Curriculum for Wales

Central to the Curriculum for Wales is the focus on learner progression, a process underpinned by clear assessment principles. The primary purpose of assessment is to support each individual learner's journey by identifying their strengths, areas for improvement, and informing subsequent teaching strategies (WG 2024a). This approach ensures that assessment is a formative part of the learning process, rather than a separate or summative activity.

Practitioners are encouraged to employ a variety of assessment methods that align with the curriculum's progression steps, which outline the expected learning milestones at different stages. This alignment aims to ensure that assessments are relevant and reflective of the curriculum's objectives.

To achieve these assessment objectives the CfW outlines several key assessment processes: ongoing observations to monitor learners' engagement and understanding during activities; timely and constructive feedback; encouraging learners to self-assess and reflect on their learning experiences to foster self-regulation and autonomy; and collaborative assessment, involving peers and the learners themselves in the assessment process.

The CfW represents a significant shift away from nationally standardized assessments in primary education. The new framework eliminates the requirement to formally assess attainment levels to learners, reducing the emphasis on summative assessments.

However, while the CfW emphasizes a formative and personalised approach to assessment, it also incorporates some mandated statutory assessments in reading, numeracy, science and Welsh (WG 2019). These assessments produced data for WG monitoring at a school level, and individual feedback for learners. However, whilst Welsh medium schools have tools for Welsh language assessment that produce comparable data, no such common assessment exists for Welsh as a L2 in EM schools. Instead, teacher assessment against CfW attainment goals form the basis of assessment practice to fulfil these statutory obligations (WG 2024d).

Details of how GALW aligns with the principles and structures of the CfW can be found in chapter 3.2.

2.2 Current Assessment Practice in EM Primary Schools

There is a significant gap in the research literature around assessment practices of Welsh in English medium schools. In response to this deficit, a scoping study was conducted as part of this specification development to assess current practice and establish if a need exists for Welsh language assessment tools.

The study consisted of surveys (n=81) and interviews (n=21) with teachers taking part in the Welsh Language Sabbatical Course in 2025. The 81 survey responses gathered represent the practices of 81 different schools (approximately 9% of the number of English medium primary schools in Wales). Participants were drawn from across the four WLS locations (Cardiff, Bangor, Swansea and Carmarthen) including teachers from a broad range of regions, consortia, and socio-economic contexts. In addition, the course tends to attract teachers who are more engaged with Welsh language provision within their schools and so offers a valuable insight into assessment practices across the whole of Wales.

A full copy of the scoping study can be found here. A summary has been included below highlighting the key findings pertinent to the development of the GALW.

The study highlights the lack of a widely used assessment tool for EM schools, which has led to a variety of localised assessment processes. In many cases, there was an absence of any assessment procedure at all. In others, monitoring was carried out of written work only ('book-looks/scrutiny') or by using entirely subjective teacher assessments. The study's key findings are summarised below, consisting of five main issues in current practice: partiality, impracticality, incomparability, inconsistency and inaccuracy.

Inconsistency and inaccuracy

Where assessment procedures do exist, they are often subjective, relying on teachers assessing the learners within their class. This can create inflated grades for learners resulting in unrealistic expectations of actual competency (Fleckenstein et al. 2018; Murphy & Wyness 2020). Such inflation could be in response to institutional pressures on teacher performance (Chowdhury 2018) but could equally be a product of intra-institutional norms (Koretz 2008; Marcenaro-Gutierrez & Vignoles 2015), where internal comparisons form the basis of assessment, make them unsuitable for inter-institutional or individual learner comparability. Assessor competency can also impact of accuracy: only around 9% of the teachers in EM schools are sufficiently competent in Welsh as to feel able to teach through the language (WG 2023). Such a skills deficit has been found to lead to inaccuracy in learner assessment (Lazaraton 2005).

Assessments are also often based on analysis of written work, which is unlikely to accurately reflect the learners' broader communicative competency (Azam 2021) and does not correspond to the emphasis on Welsh language oracy in the CfW (Thomas et al. 2023, p.57).

Such assessment procedures also reflect a 'snap-shot' of learner knowledge, failing to establish whether skills/knowledge are retained long-term (Baldwin 2018). Although some schools have started to use digital recordings to monitor learners' oracy development, these assessments often encourage the memorisation of rehearsed scripts, misrepresenting their actual communicative ability (Kim 2023), or allow learners to use extensive scaffolding which results in a misrepresentation of actual communicative ability in the language (Fulcher 2013).

Impartiality

Teachers work closely with their classes and endeavour to develop relationships with individuals that facilitate learning. This is especially true in primary settings, where classes tend to spend most of their time with just one teacher. However, this rapport can lead to problems when assessing language competency, with assessments involving subjective evaluations which may be distorted by other factors (Campbell 2015). Research shows that teacher judgments are often influenced by non-linguistic factors, including students' behaviour, personality, appearance, ethnicity, or perceived motivation (Cumming 2001; Rea-Dickins 2001).

Halo effects, where overall impressions of a student influence specific ratings, and leniency or severity biases are common sources of inconsistency (McNamara 1996). These forms of rater bias can result in unreliable or inequitable scoring, undermining both learner confidence and accountability in assessment systems.

Practicality

Teacher assessments of L2 competency require the observation of interaction with/between pupils for sufficient time to make an evaluation of their competence in the target constructs. Time constraints in the classroom often curtail opportunities for such detailed assessments, creating superficial results or infrequent assessment, and therefore limiting the utility for formative testing (Rea-Dickens 2004). In addition, designing such assessments and curating assessment content is a significant challenge for teaching staff who are unlikely to have linguistics backgrounds, or training in assessment design. This can lead to superficial/poorly calibrated assessments, or potentially disincentivise teachers from carrying out assessments at all (Scarino 2013). In contrast to this, it is possible that teachers become overly-engaged in assessment processes, to the extent that assessment data collection becomes a distraction from, rather than facilitator of, enhanced teaching (Carless 2005). Formative assessment therefore must achieve a balance between efficacy and practicality in order to achieve a positive impact on learning (Becker et al. 2017).

Comparability

The comparability of assessment data operates at two levels: intra-comparability, i.e. the comparability of longitudinal data formed by a series of scores for an individual; and inter-comparability, i.e. cross-sectional comparability of scored between individuals/cohorts/institutions (Robinson et al. 2005). Both types of comparability are important to assessment impact on improving teacher practice and individual learning outcomes (Zahner & Steedle 2015). However, in order to be comparable, the assessment design must ensure equivalence of constructs, establish validity, and maintain reliability (Bennett 2011). Simultaneously, the assessment must align with discrete learning objectives and broader curriculum goals to ensure learning is aligned to the expectations established in policy (WG 2024b). Such demands are challenging for teachers to meet in the design of classroom assessment. In addition, the emphasis on 'timeliness' in feedback (Black and Williams 1998) often results in a focus on the learning immediately prior to the assessment (Suskie 2008), and to a depth of assessment restricted by the time-limitations of the classroom context (Black 2003).

This lack of comparability in existing assessment may also have contributed to the report problems with learning continuity across transition points (Russell 2025, p.27), with many teachers reporting learners returning to the beginning of the WLC in year 7, negatively impacting on learner motivation and progression.

Staff training and Welsh language proficiency

Another theme that emerged from the scoping study was shortcomings in the training teachers receive in language teaching generally, and language assessment specifically (Russell 2025, p.21). This training deficit is compounded by low levels of Welsh language competency in the EM sector (Estyn 2023). Whilst the challenges of teacher competency in Welsh and language pedagogy are obviously important in the development of the language more generally, here the focus will be on how these relate to assessment.

To some extent language proficiency deficits are an inevitable consequence of a growing WM sector, with more proficient Welsh speakers taking on WM roles (Senedd Cymru 2024f), leading to difficulties recruiting Welsh speakers to EM schools. The Welsh Language Commissioner, NASUWT, WLGA and ADEW (Senedd Cymru 2024c; 2024d; 2024e) all highlighted the Welsh language skills deficit in EM schools as being an obvious barrier to the implementation of the WLEB (see chapter 2.5). Such skill deficits have implications for both assessment and learning. A lack of proficiency can have impacts on the teacher's ability to provide accurate instruction, modelling, and opportunities for spontaneous interaction for learners (Reves & Medgyes 1994; Kamhi-Stein 2000; Llurda 2005). Within assessment specifically, teachers with low L2 proficiency struggle to judge learner language use accurately, find it more challenging to devise valid assessments, and provide appropriate feedback (Vogt & Tsagari 2014).

The impact of this suboptimal language proficiency is compounded by a lack of training in language teaching pedagogy. Most teachers in the scoping study received less than 10 hours of Welsh tuition as part of the ITE, most of which was aimed at content knowledge and correct pronunciation. Many teachers reported having received no formal training in language teaching. This is a new trend: Estyn's (2024) annual report, found 'limited' understanding of language teaching pedagogy in EM teachers, and only in a few instances was this mitigated by suitable professional development. Whilst all consortia do offer support through CPD opportunities, these are elective, and schools in which Welsh has a low profile are unlikely to prioritise such courses for their staff.

Finally, less than 2% of teachers in the scoping study reported having received any formal training in Welsh language assessment. This lack of training compromises teachers' ability to monitor class needs and progress, and reflect on the efficacy of their own practice (Heritage 2007).

Conclusion of the Scoping Study into WL Assessment Practices

The principle of progression is central to the CfW and should equally apply to Welsh L2 acquisition as any other area of learning (WG 2025). However, at present due to the challenges detailed above, learner progression in Welsh in EM settings is difficult to both establish and monitor. This challenge applies at classroom level, but also at institutional, regional, and national levels, significantly hampering the ability of organisations to understand the needs of learners, teachers and schools. Such a knowledge deficit is likely to have negative impacts on the effective allocation of resources, the development of suitable provision, the identification of best practice, the planning of staff training and professional development, and the creation of policy. It is therefore one of the findings of the scoping study that there is a clear need for the development of suitable Welsh language assessment tools for use in English medium primary education (Russell 2025, p.39).

2.3 The Welsh Language Continuum

The Welsh Language Continuum (WLC) is a criterion reference framework of progression used to describe the varying levels of proficiency and use of the Welsh language in learners. The WLC adopts a 'multi-competence' approach (Lovell 2023, p.67) reflecting a holistic understanding of bilingualism as a spectrum into which all learners are situated. Hornberger (2003) warns against viewing such continuums as linear and finite, but instead as an indefinite process. Accordingly, the WLC in intended to be a framework for life-long Welsh language learning, not limited to compulsory education (WG 2024c). Valdés (2003) highlights that such continuums are not unidirectional, i.e. that it is possible for learners to regress towards monolingualism as well as

progress towards multilingualism. Indeed, with the WG's focus on education as a means to achieve the goals of Cymraeg 2050 (WG 2017), such regression post-education could pose a significant challenge.

The CfW is divided in Areas of Learning and Experience (AoLEs), with Welsh in EM schools falling into the 'Literacy, Language and Communication' AoLE. Welsh is therefore included in a broader framework of English language and multilingual development (WG 2024e). Within this structure the Welsh continuum of progress is divided into four levels: the 'what matters' statements, principles of progression, descriptions of learning, and language pattern content. The 'what matters' statements outline the key ideas essential for learning in each AoLE. They define broad concepts that shape a learner's journey. The 'Principles of Progression' describe how learners develop their understanding and skills over time in a broad sense, and are intended to ensure continuous and meaningful progress is considered in curriculum planning and assessment. 'Descriptions of Learning' provide specific descriptors showing how learners' knowledge and abilities evolve. They are divided into progression steps, though progression through these descriptors is unlikely to be linear (WG 2024e). Together, these initial three elements are intended to ensure a coherent, learner-focused progression. The final layer concerns specific language content adopted, the selection of which is delegated to local consortia and school clusters, rather than being decided collectively at national level. This localised approach to content development is intended to facilitate a greater level of freedom and responsiveness, allowing schools to differentiate to their learner community context and regional dialectical differences.

All schools appear to be using a pattern-based format of language content, with different stages (usually link to the CfW progression steps) of patterns delineated by grammatical complexity, topic, and/or functional category. These are sometimes composed by Consortia directly, or the local council, whilst in some areas local clusters decide on their own content (often based on resources developed before the introduction on the CfW. In addition to this variety of content sources, individual schools are then able to draw down items selectively to create their own curriculum offering. This individualised approach, whilst offering the schools greater flexibility and control, makes comparative assessment very challenging.

2.4 The Common European Framework of Reference for languages (CEFR)

The Common European Framework of Reference for Languages (CEFR) is a widely recognized standard for describing language proficiency. Developed by the Council of Europe (2001), it provides a comprehensive framework for evaluating and comparing language skills across different linguistic and educational contexts. The CEFR aims to facilitate transparency in language learning, teaching, and assessment and has been adopted by educational institutions, governments, and language testing bodies worldwide.

The CEFR is structured into seven proficiency levels, divided into four broad categories: Beginner (A0 or pre-A1), Basic User (A1, A2), Independent User (B1, B2), and Proficient User (C1, C2). Each level is defined by a set of descriptors outlining the communicative competencies expected of learners in reading, writing, listening, speaking, and mediation (Council of Europe 2025). These descriptors emphasize functional language use rather than grammatical mastery, reflecting a communicative approach to language learning (North 2014). Descriptors take the form of 'can-do' statements that provide practical descriptions of what learners at each level can accomplish in real-life communication (Little 2006).

The CEFR is designed as a descriptive reference tool rather than a rigid curriculum or testing system. It is used by external parties in curriculum development, language assessment, and teacher training. Many international language tests, such as the IELTS, TOEFL, and DELF/DALF, have aligned their scoring systems with the CEFR to ensure consistency in proficiency assessment (Papageorgiou 2010). In addition, it is commonly used as a guide for self-assessment, enabling learners to track their progress and communicate their level of proficiency clearly (Little 2005).

One of the key strengths of the CEFR is its flexibility and applicability across languages. Unlike national assessment systems, which are often tied to specific curricula, the CEFR provides a universal framework that facilitates comparability between different languages and educational settings (Council of Europe 2001). However, the CEFR has been criticized on a number of fronts: for lacking specificity, particularly in higher levels (Fulcher 2010; Foley 2019); for lacking the nuance for assessing early L2 development (Konrad et al. 2018; Kremmel et al. 2023); and for being poorly content-aligned to younger learners (Kahn-Horwitz & Goldstein 2024).

The accuracy of CEFR aligned assessments is also an area of concern, with Nagai (2020) highlighting the difficulty in accommodating uneven learner profiles in a reductive classification framework, and Green (2018) noting the score inconsistency of CEFR aligned assessments in rating individuals. In addition to these structural concerns, Hulstijn (2007) questions whether the CEFR descriptors adequately reflect the linguistic complexity or cultural nuances in language use.

However, despite these critiques, the CEFR remains the most widely used scale of language proficiency, and its ubiquity is a major factor in its utility for communicating proficiency (Heyworth 2013, p.297). Benigno and Jong (2019) suggest that we place an unreasonable level of expectation on the CEFR, assuming it is capable of the same clarity and consistency as measures in the natural sciences. Of course, increments of language development are not so clearly defined, and any tools for describing or measuring such units would require a degree of flexibility (Douglas 2010). Milton and Alexiou (2009) note that the inevitable price of such flexibility is imprecision, and advocate for the consideration of more targeted objective assessment tools to supplement direct assessments of CEFR descriptors.

2.5 Welsh Language and Education Bill (WLEB)

The forthcoming Welsh Language and Education Bill includes significant changes that will impact on the teaching, learning and assessment of Welsh in EM schools (WG 2024b). Whilst these impacts span many elements of policy, pedagogy, and provision, here the focus will be on the implications for assessment. The two main features of the WLEB affecting assessment are the adoption of the Common European Framework of Reference for languages (CEFR) as the basis for a new standardised framework for describing Welsh language ability, and the modification of expectations of learner proficiency in relation to this framework.

The CEFR is a widely used international standard for describing language proficiency. Though not itself an assessment, it is used as the basis for many language tests (IELTS, TOEFL, etc). Proficiency is divided into six levels (A1 to C2) that describe a learner's ability to understand, speak, read, and write in the target L2. The primary intention is therefore to create a standardised understanding of performative competency across languages, i.e. that a B2 speaker of German would have a similar level of communicative competence as a B2 speaker of Hungarian.

The WLEB proposes the adoption of the CEFR descriptors to create uniformity and continuity for learners, and consistency and comparability in assessments of proficiency (WG 2024b). In addition, it is hoped widespread adoption of the CEFR may go some way to help address the ambiguity in self-reported competency (e.g. census data, national survey for Wales data), moving away from a binary understanding of language ability, towards a commonly understood scale of competency (Senedd Commission 2024).

Of course, adopting the CEFR does not just entail the development of a Welsh framework, but also the alignment and integration of this with existing assessments and curriculum structures. The National Centre for Learning Welsh observes that there is significant continuity between the CEFR and the Literacy Language and Communication framework of the CfW, making alignment between the two 'fairly easy' (Senedd Commission 2024, p.68), whilst Mentrau Iaith Cymru highlight the need to map CEFR against existing assessment methods in both EM and WM schools (Senedd Cymru 2024a).

Such mapping poses a significant challenge as the WLEB seeks to both integrate CEFR descriptors in a way that aligns with current assessments and qualifications, whilst also raising the expected levels of attainment for pupils at these levels. Long term the Bill seeks to raise the goal for those leaving compulsory education to B2 on the CEFR (WG 2024b). B2 is described as a learner who can:

> "…interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party." (Council of Europe 2025)

According to Qualifications Wales, the L2 Welsh GCSE currently equates to A1/2 on the CEFR and elevating this standard to B2 will require substantial improvements in learning outcomes (Senedd Cymru 2024b). Despite these challenges, stakeholder organisations have generally welcomed the adoption of the CEFR descriptors.

However, the proposal has been critiqued by organisations such as Dyfodol i'r Iaith who point out that the CEFR was developed primarily for professional purposes and the assessment of adults (Senedd Commission 2024, p.63). Significant adaption of the framework descriptors would be required to make it fit for educational contexts and younger age groups. Heini Gruffudd, chair of Dyfodol i'r Iaith, also raises concerns about the broad nature of the CEFR (Senedd Cymru 2024a), advocating for EM schools to focus on oracy skills rather than attempt to developing holistic set of domain competencies. However, such divergence based on school type may undermine the WG goal of creating a single language continuum for Welsh language progression.

In addition to such structural concerns, Hasselgreen (2013) highlights the issue of scaling within the CEFR, with the progression from A1 to C2 potentially reflecting a lifetime of learning and therefor offering insufficient nuance/granularity for assessment during comparative short periods in compulsory education. Indeed, the CEFR descriptors themselves have also been critiqued for a lack of sufficient detail to capture the granular progress of learners. This is particularly pertinent at the early stages of language learning (Kremmel et al. 2023, p.78) where the development of lexical knowledge facilitating syntactic understanding is key (Slabakova 2016). These factors may create challenges in applying the CEFR as a measure of early L2 Welsh language development in primary school settings. In this instance, Milton & Alexiou (2009) advocate for supplementing direct assessments of the CEFR with more objective analytical measures of knowledge-based constructs to help delineate progress.

Though the impacts from the WLEB are unlikely to be felt in schools directly until after the Welsh in Education Strategic Plans in 2028, it is important to consider their impact on assessment and ensure that assessment provision is fit for purpose currently and after these changes come into force.

3. Assessment Objectives

The key goals of the GALW are explored in section 1.1. In this chapter the way these purposes are operationalised through the assessment design is explored in detail. The chapter starts with a specification of the specific constructs that will be selected to facilitate the assessment. How these constructs then align with existing frameworks for language assessment are then explored. Finally, the ways in which the GALW can contribute to research, both external academic studies and internal action research, are considered.

3.1 Construct Selection & Validity

A construct refers to the underlying theoretical concept or skill being measured by an assessment, encompassing the abilities, knowledge, or behaviours that the test aims to assess, including aspects such as vocabulary knowledge, reading comprehension, speaking proficiency, or grammatical accuracy (Bachman 2007). Ensuring that a test accurately reflects the construct is vital for its validity, meaning the assessment must align with the intended skills or competencies rather than measuring irrelevant or unrelated factors. However, equally important to the identification of the constructs intended to be included in the assessment, are those that are actively excluded to prevent distortion in measurement of the target constructs (Norris & Ortega 2012). Once these constructs are identified procedures can be formulated and tools created to operationalise the construct and analyse the construct-performance of the learner.

Constructs do not necessarily fall into traditional distinctions of linguistic domains (e.g. grammar, vocabulary, socio-linguistic knowledge), often these are 'layered' being composed of multiple domain elements and so require assessment that may include aspects of more than one domain (Ellis & Ferreira-Junior 2009; Ockey & Zhi 2015). As learners become more advanced, these constructs tend to become more entwined creating 'compound constructs', often necessitating a shift from the assessment of domain-based constructs to complex multi-faceted constructs (Kuiken 2023), or from explicit knowledge constructs to implicit skills constructs (Ellis 2005).

Lightbown and Spada (2013) note that, within such compound constructs, different elements may progress/regress/stagnate across the learner's development. Even within apparently mono-competence constructs such as vocabulary, studies have shown that lexical knowledge spans many linguistic and content elements of assessment tasks (Dodigovic 2005, Thornbury 2002) and is often in conflict with pragmatic factors in assessment design (Lewkowicz 2000). A balance therefore needs to be found between the context validity, practicality and reliability of assessment tasks (Becker et al. 2017; Kadir et al. 2019). For a more detailed exploration of this please see chapter 7.3.

The divergence between individual and compound constructs is also reflected in assessment design: whilst the underlying goal of language assessment is to understand what an individual can do with the language outside of the assessment context (Norris and Ortega 2012), assessments tend to adopt either a holistic, or analytic approach. Holistic assessments aim to assess the learners' communicative competence directly, usually through tasks imitating

authentic contexts which operationalise several layered constructs (e.g. vocabulary knowledge, syntactic/grammatic composition, socio-linguistic awareness, communicative repair) in which learners can demonstrate multiple language competencies (Gedera 2023). In contrast, analytic assessments seek to isolate a limited number of discrete constructs, and consequently assessment tends to be more abstract to avoid the inclusion of extraneous constructs (Hidri 2018). Although unable to directly observe holistic ability, analytic assessment may in some cases be used to infer broader competency (see the discussion on the work of Milton and Alexiou 2009 in section 3.2). The GALW seeks to mobilise this inferential format, being indicative of construct competency, rather than directly operationalising it. You can find a more thorough discussion of this in chapter 7.6.

Regardless of whether inferred or directly observed, a fundamental consideration in construct validity is whether assessment accurately represents the target construct. That is to say, to what extent is the representation of the construct legitimate (construct validity) and to what extent are the results an accurate reflection of the correspondence between learner ability and assessment output (assessment accuracy) (Norris and Ortega 2012). Construct validity can either be established independently, or through cross-validation with an already validated measure of the same construct.

Cross-validation is the alignment of a new assessment with an already validated assessment that targets the same construct. High levels of correlation (>r=0.7) between user performance in each assessment can be indicative of construct consistency i.e. that the assessments are measuring the same construct. Similar assessments do exist that would offer such correspondence to the GALW, such as the widely used 'yes/no' testing developed by Meara and Buxton (1987). However, cross validation against such vocabulary size measures are not possible in the Welsh EM primary context due to pedagogical practice: EM schools almost all use a stem sentence based system of content. This often leads to formulaic knowledge that doesn't distinguish individual word meanings required for yes/no type tests. Additionally, yes/no tests are also based on the assumption that knowledge of a lemmatised word is indicative of knowledge of the word's 'family' of derivative forms (Nation 2001), e.g. if a learner knew 'cysylltu' (connect) in Welsh that they would be able to construct associated forms, such as ymgysylltu, dadcysylltu, cysylltiad, cysylltiol, etc. Again, the prescriptive nature of the WLC content does not tend to develop vocabulary in this way at early stages, undermining the foundation for cross-validation against yes/no tests.

Without this option of cross-validation, construct validity must be established independently. This can be achieved in four stages: first, the construct is defined; second, how the construct is operationalised as a learner behaviour is identified; third, tasks are designed to elicit this behaviour; finally, the extent to which the results of the assessment express the construct is analysed (Vandergrift & Goh 2009).

Before moving on to applying this approach to the GALW's target construct of lexical knowledge, it is important to define the aspects of learner performance that are linked to the construct or separate from it. Assessment accuracy can be impeded by the impact of factors outside the construct ability-output correspondence. These are usually described as construct-irrelevant factors, whilst factors that are directly related to the target construct are described as construct-relevant. An important aspect of optimising construct validity is accommodating construct-relevant factors and mitigating or excluding construct-irrelevant factors. Construct-relevant factors include motivation, cognitive ability, and language aptitude. Construct-irrelevant factors may include emotional state, assessment delivery, and risk aversion.

Lexical knowledge as a target construct

Lexical knowledge is widely accepted as both the foundation for the development of linguistic skills (Nation 2001; Webb & Nation 2017) and an effective indicative measure of communicative competency (Milton & Alexiou 2020). Indeed, it has been shown that a lexical knowledge encompassing the most frequent two thousand words of a language is essential for learning less frequent words (Schmitt 2000; Nation 2006), whilst one thousand words are required for learners to begin to communicate independently (Milton & Alexiou 2009).

Lexical knowledge is however not simply a case of counting the number of L2 meaning-form correspondences known buy the learner. 'Knowing' a word is actually a complex multidimensional construct (Schmitt 2000), including a variety of different degrees and aspects (e.g. meaning, form, collocations, socio-linguistic usage, recombinant grammatical use). Greater specificity is therefore required when defining what is meant by the claim that a learner 'knows' a specific word, chunk, or phrase.

The GALW is designed to ascertain the learners' lexical knowledge of the WLC content. Whilst the descriptors of progression are consistent across Wales (WG 2024e), each Educational Consortium has a distinct version of specific WLC content, reflecting the dialectical differences particular to their constituent communities. Despite lexical differences, the structure of progression within the various content guides shows a high level of continuity. Each content continuum is composed of a series of phrase-based language pattern stems with specific recombinant elements that learners can exchange to extend communicative range (example available in Appendix Item 1). These patterns are usually structures in a communicative form, being composed of questions, associated answers, and auxiliary vocabulary. The continuums are arranged sequentially, with language patterns being presented in increasing order of complexity (lexically, conceptually and grammatically), and are notionally divided into progression steps. It should be noted that (aligning with the CfW principles of a learning continuum) these are not intended to be interpreted prescriptively, though it is possible this is occurring in some schools (see scoping report on WL assessment in EM primary schools here).

The most foundational aspect of vocabulary knowledge is the form-meaning link (Laufer 2004) Form-meaning knowledge has been found to correlate with comprehension of authentic spoke texts (Nation 2006), with a vocabulary size of 6000-7000 being correlated with a 98% comprehension rate. With regards to the assessment of such form-meaning constructs in the GALW, the individual learner's explicit knowledge of the language patterns will be further refined into two sub-constructs: recall and recognition (Read 2019):

- Recall knowledge – The learner's knowledge of the relationship of the L1 to the TL (for example, translation from L1 to TL)
- Recognition knowledge – The learner's knowledge of the relationship of the TL to their L1 (for example, translation from TL to L1)

Both productive and receptive knowledge are measures of linguistic knowledge 'breadth', i.e. the quantity of language patterns of which the learner has explicit knowledge.

However, such a quantitative understanding of vocabulary can obscure the nuances of lexical knowledge: subtleties of connotation and association, anaphoric and homophonic aspects, and appropriate socio-linguistic usage.

Simple measures of vocabulary size have been critiqued as 'superficial' in their failure to account for 'depth' of lexical knowledge (Nation 2001). Meara & Miralpeix (2017) highlight the

challenge here for assessment, as although vocabulary size can be expressed for an individual, vocabulary depth is specific to the individual lexical item. Any assessment of lexical depth will therefore require a significant level of inference and generalisation.

One aspect of lexical-knowledge depth is linguistic productivity/generativity, i.e. the capacity of learners to use lexical knowledge flexibly, to integrate it with other linguistic knowledge and to adapt it using know syntactic and grammatic frameworks (Chomsky 1965/2015). Meara (2009) conceptualises these frameworks as networks of words, collocations, and idiomatic language blocks, all interconnected. Knowledge of vocabulary depth is then a product of the 'organisation' of links between items rather than simple meaning-form association. Lexical development in L2 learning is therefore not simply about acquiring more words, but also strengthening and diversifying the links between them.

The phrase-based communicative structure of the WLC content (usually compromised of linked questions and responses) does offer advantages over a discrete word-based vocabulary assessment model (e.g. Yes/No tests) in forging these links, as it places lexical knowledge in contextualised communicative structures (Belgar 2013). However, there is a danger that emphasis is placed on a memorised, formulaic knowledge of the WLC content items, rather than developing authentic competency in using them communicatively.

This aspect is reflected by Lovel (2023), who raises concerns about how formulaic language knowledge may distort assessment inference of communicative competence: learners may appear more accurate (and therefore more holistically competent) when reproducing memorised formulaic phrases, but such lack of generativity results in far lower levels of communicative flexibility and spontaneity. Such concerns are reflected by Zeeland (2013) in a receptive context, with lexical knowledge not necessarily resulting in corresponding discourse comprehension.

In response to these concerns, the GALW aims to operationalise two constructs of lexical knowledge: a breadth measure (in two forms, recall/recognition), which explore knowledge of form-meaning correspondence; and a depth measure, that seeks to quantify the generativity associated with the lexical knowledge that the learner has acquired.

There are several approaches that could be adopted to operationalise generativity. For example, Read and Chapelle (2001) propose an 'interactionalist' approach, whereby assessment seeks to mimic contextualised communicative contexts relevant to the candidate. However, this would raise some of the construct conflation issues already discussed and would introduce further variation through the introduction of an interlocutor (Isaacs 2016). Norris and Ortega (2009) propose a complexity-accuracy-fluency framework which can be applied to learners' linguistic output to determine novel language use as an indicator of generativity. However, such an approach would be challenging to automate, and may be unsuitable for very early language development, where the capacity to produce a sufficient quantity of recorded material for analysis may be beyond the learners' competency. Lu (2010) addresses the challenge of automation through the deployment of computational analysis. However, this approach does not mitigate the need for extensive source material. In addition, its written format makes it unsuitable for younger learners (Alexiou and Milton 2020). Instead, a meaning-based parsing task, similar to that adopted by Boers and Lindstromberg (2012), will be included to operationalise the generativity construct (see chapter 5.2).

As content knowledge is the focus of the assessment, both phonological and lexical formats of each item will be provided. The inclusion of the phonological form improves accessibility for learners with additional learning needs (more details in chapter 4), who may be disadvantaged by a purely written format. Such an exclusive written format could also distort scores amongst primary aged children in Wales, for whom the majority of teaching focuses on the oral form (Thomas et al. 2023). Providing the lexical form alongside the auditory will improve accessibility for those with hearing impairment, whilst also mitigating the influence of disparities in higher-order skills such as syntactic parsing, phonological recognition/boundaries, working memory, interpretation of prosody, and accent divergence. It also introduces ambiguity about which cognitive processing skills are being deployed by the learner. The impact of these factors is considered in more details in chapter 7.2.

From a content perspective, these two assessment strands focus on propositional (literal) meaning without resort to what Laird (1983) characterises as 'complex meaning representation', e.g. use of sarcasm, intended inference, socio-culture contextual understanding. Items will also consist of discrete phraseological structures in isolation from contextual factors, therefore excluding consideration of learners' discourse representation/construction of meaning (Brown & Yule 1983; Cutler & Clifton 1999).

This phraseological approach to both WLC pedagogy and the assessment content also means that the assessment will not represent learners' lexical knowledge of variant forms, as such there can be no assumption that the knowledge of a lemmatised word form is indicative of a broader knowledge of the word in other forms, e.g. should a learner demonstrate awareness of 'cysylltu', it cannot be assumed that they would have the linguistic knowledge to infer forms such as 'datgysylltu', 'cysylltiad', 'cysylltiol', etc.

Constructs not included

It is important to acknowledge the constructs that are to be consciously excluded from consideration and the justification for this (Messick 1989; AERA 2014) in order to guard against construct underrepresentation or conflation.

- *Speed/automaticity* – Buck and Tatsuoka (1998) find that processing speed is not a discrete skill, but a product of competency and automaticity in other abilities. So, whilst speed can be measured, its composite nature makes it difficult to infer from which competencies it is either produced or inhibited. In addition, whilst speed/automaticity is widely accepted as essential long-term in real-time discursive efficacy (Canale & Swain 1980, Hui & Godfroid 2021), overt inclusion of latency data collection may incentivise guessing (see chapter 7.7) and increases test-anxiety (Hembree 1988; Putwain et al. 2010). Covert collection of latency data avoids these negative outcomes but can capture variance in user performance not associated with the target construct (Segalowitz 2010). Although faster reaction time has been found to be correlated with higher levels of accuracy in timed yes/no tests (Read 2019) it is not necessarily indicative of greater lexical knowledge (Mirapleix and Meara 2014).
- *Explicit grammatical knowledge* – In primary school WLC the focus of lessons tends to be on lexical content, with grammar learning assumed to be inductive or implicit, rather than declarative. The GALW seeks to align with this approach by adopting the same implicit phrase-based format. In addition, performance on explicit grammar tests has been found to correlate weakly with proficiency in spontaneous language production and listening comprehension (Norris & Ortega 2000), making it a relatively poor measure

of communicative competency. Finally, when considering user characteristics, explicit grammar testing is heavily reliant on metalinguistic knowledge (Ellis & Wulff 2019) that will be disproportionately underdeveloped in younger learners, rendering the construct unsuitable for primary school assessment.

- *Discourse competence* – Discourse competence is the ability to produce and comprehend extended, coherent, and cohesive texts (spoken or written) beyond the sentence level (Celce-Murcia et al. 1995). Littlewood (2004) notes that discourse competence represents a higher-order skill that typically develop later in the language acquisition process. Tasks demanding extended discourse production and comprehension require cognitive and linguistic maturity, making them less appropriate as the primary focus for younger or less proficient learners.
- *Pragmatic competence* – Pragmatic competence is the ability of the learner to utilise the language appropriately in a variety of social and professional context (Celce-Murcia et al. 1995), including aspects such as politeness, suitable use of sarcasm or irony, and suitable levels of familiarity and formality. However, Kasper and Rose (2002) highlight the emergent nature of pragmatic competence as a skill that develops from interactive experiences. Such a nuanced understanding of language use is therefore dependent on a foundational acquisition of lexical proficiency, the area of focus for the GALW.
- *Productive competence* – Productive competence is the learner's ability to produce the target language either verbally or in written form. It is a common and useful construct for language assessments as it more closely reflects authentic language use (Canale and Swain 1980), operationalises a host of sub-skills essential for communicative competency (grammar, lexicon, discourse strategies, and pragmatics) providing a more comprehensive picture of competency (Bachman & Palmer 1996), and can allow for diagnosis of specific communicative deficits (Kormos 2014). However, as previously explored, productive competency can hinder the isolation of more specific constructs (e.g. lexical knowledge), can result in increased subjectivity, is unsuitable for self-administered assessments and does not lend itself to automated marking, analysis and feedback. For these reasons productive competence is exclude from the GALW in preference of 'recall knowledge'.

3.2 Alignment with Existing Frameworks

It is important to consider any new assessment tool within the context of existing assessment and progression frameworks to ensure it achieves both its diagnostic and formative roles without causing misalignment or conflict. Accordingly, the GALW draws on Biggs (1996) 'constructive alignment' theory, which advocates for coherence between learning outcomes, teaching activities, and assessment methods (Genon and Torres 2020). Drawing on a constructivist approach, where students construct their own learning through meaningful engagement with course materials, constructive alignment requires assessments to correspond to both the learning goals and teaching approach. In this way constructive alignment itself corresponds to the principles of the CfW which emphasise student-centred learning and content aligned assessment (WG 2024d).

Despite such an approach, there are factors that can distort the alignment of teaching and assessment e.g. misaligned learning materials, teachers' pre-conceptions of language learning (Rouffet et al. 2023), and a failure to identify learning objectives clearly (Biggs and Tang 2007). These issues are explored in more detail in Chapter 7.5.

The two existing frameworks with which the GALW must align are the CfW and the CEFR. Below the alignment of the GALW with the assessment principles of the CfW and the structural character of the CEFR is considered in more detail.

Alignment with the Curriculum for Wales (CfW)

A foundational principle of the CfW is that the purpose of assessment is to facilitate the progression of the individual learner, providing information to direct pedagogical response (WG 2024d). Such a formative focus is reflected in GALW's design, which is primarily focused on providing diagnostic feedback to both users and teachers to enhance learning and teaching practice.

The CfW also highlights the longitudinal nature of effective assessment for learning, supporting an iterative process of progression, rather than a singular snapshot of performance (WG 2024d). On-going assessment should therefore be used to guide learning and teaching, with the impact of adaptions made based on previous assessments in turn evaluated through subsequent assessment data. In this way, learners develop their skills and knowledge, whilst teachers can refine, attune and adapt their pedagogy and provision to best meet the needs of learners (Raudenbush et al. 2020). To achieve this, assessments must be designed to be convenient, practical to deliver regularly, and provide comparable data that can facilitate appraisal of learning continuity. This should enable both individual and group/sub-group analysis of learning progression, allowing teachers to make informed decisions around provision, scaffolding, differentiation and support (WG 2024d). GALW provides both discrete and longitudinal data on learning progression in WLC content knowledge, allowing teachers to track progression and monitor for attrition. You can find more detailed information of how GALW aligns with these CfW aspirations in chapter 5.2.

Whilst the CfW stipulates that internal assessments should not be used for the purpose of accountability, the contribution of assessments to teacher and institutional self-evaluation can be valuable; creating an informed dialogue that can help direct support to enhance learner experience (WG 2024d). GALW's functionality in providing class and institutional level data can provide a valuable resource for facilitating such dialogues. Of course, using learner assessment as a measure for punitive staff evaluation can obviously be detrimental to both teachers and learners. This potential misuse of the GALW is considered in chapter 8.1.

Alongside the deployment of the CfW, regulations were introduced around the transition of learners from primary to secondary education, requiring schools to collaboratively create a transition plan (WG 2022a). A key element of these plans is how the 'continuity of learning' and 'individual progression' of transitioning pupils is achieved and monitored (WG 2022b, p.8). This was also emphasised in the WG report on Welsh language teaching (Fitzpatrick et al 2018, p.61), which highlights the importance of maintaining progress made in primary school when learners transition to primary settings. The GALW generates individual, class and cohort level data, allowing secondary schools to more easily meet these regulatory requirements and minimise negative impacts on learning.

In terms of content alignment with the CfW, the GALW draws content items directly from the WLC content lists developed by the Educational Consortia and used in schools to design their individual curriculum offering. Alignment of content is further assured by the GALW's functionality that allows teachers to select assessment content based on the coverage profile of the class or individual learner (for more details see chapter 6.1).

Alignment with the Common European Framework of Reference (CEFR)

The Common European Framework of Reference for Languages (CEFR) is a standardised criteria referenced framework developed by the Council of Europe to describe language proficiency across different languages. It is intended to provide an open, coherent, and holistic basis for creation of syllabi, assessments, and provision. See chapter 2.4 for more detailed information about the CEFR.

The CEFR is composed of six levels of proficiency: A1 and A2 (Basic User), B1 and B2 (Independent User), and C1 and C2 (Proficient User). Each level has a set of descriptors outlining what learners are expected to be able to do in each domain (listening, reading, speaking, writing, and mediating), offering a functional, competency focused approach.

Assessments associated with the CEFR take two forms, direct assessments and score-linked assessments. Direct assessments attempt to measure learner competency against the CEFR framework directly, creating a series of tasks aimed at operationalising the domain descriptors, and aggregating learner performance in order to assign an overall CEFR level. Examples of such direct assessments would include IELTS, TOEFL, and PTE. Milanovic (2009) highlights how the intentional lack of specificity in the CEFR descriptors, make it challenging for assessments to align closely with them. Assessments that seek such direct alignment risk becoming a 'blunt instrument' (Taylor and Galaczi 2011, p.178), useful in the intuitive accessibility of their classification, but limited in their diagnostic ability to inform specific learning/teaching objectives and approaches. As a result, many CEFR aligned assessments represent an attempt to combine the beneficial features of holistic and analytic assessment (Fulcher 2004).

In contrast, score-linked assessments seek to measure a discrete selection of domains or competencies and use that to predict what CEFR level the learner corresponds. As a discrete measure of lexical knowledge, the GALW functions as a score-linked CEFR assessment.

Score-linking the GALW measure of lexical knowledge to the CEFR

Whilst the CEFR is not prescriptive in its treatment of lexicon development, it does acknowledge the importance of developing vocabulary, including descriptions of lexical knowledge with the level descriptors. For example, at level A2, learners are expected to:

> 'Draw upon sufficient vocabulary to conduct routine everyday transactions involving familiar situations and topics.' (Council of Europe 2020)

Such descriptive, rather than prescriptive, criteria allow for flexibility across languages, learning contexts, and populations, but make it challenging to aligning objective measures of vocabulary development against the CEFR (Milton 2010, p.213). Consequently, it is inappropriate for assessments such as GALW to claim 'alignment' with the CEFR when the measures are clearly directed towards different goals. Instead, GALW seeks to score-link to the CEFR scale, connecting quantifiable progress in the WLC content to the broader competency descriptors of the CEFR. Although not a direct measure, such an approach is not antithetical to the CEFR: whilst it does not include prescriptive guidance on vocabulary, it would be expected that, as the learners' CEFR level increases, so too would the range and complexity of their lexical knowledge (Council of Europe 2001, p.150). This point further supported by Stæhr (2008), who highlights vocabulary knowledge as key to both comprehension and communicative competence.

Of course, assessments that aim to directly measure against the CEFR have a far stronger claim to be representative of CEFR scale proficiency (Figueras et al. 2005). However, although more abstracted, a score-linked assessment can offer significant pragmatic advantages. Direct

assessments produce a substantial amount of data on each learner which can be overwhelming and difficult for practitioners to interpret (Figueras et al. 2005), and assessors require a significant amount of specialist training in order to produce accuracy and consistency in assessment results (Nagai et al. 2020). Indeed, the complexity and detail of the CEFR descriptors can lead to ambiguity concerning which level correlates to a particular learner's competencies (Milton & Alexiou 2009).

Treffers-Daller et al. (2016) note that, no one measure will ever confidently distinguish between CEFR levels, but that this does not invalidate the use of discrete assessments. Tools such as the GALW don't seek to emulate the holistic nature of linked assessments, but offer the capacity to teachers to monitor and facilitate progress towards the content knowledge that would facilitate competency at each level. Bachman et al. (1995, p.99) highlights that score correspondences should not be mistaken for interchangeability: the GALW explicitly measures lexicon development, not CEFR level. Therefore, whilst it would be inappropriate to draw on GALW as a means of CEFR levelling, it is useful to draw on the CEFR as a way of quantifying GALW performance, i.e. one could not claim a score of 100 indicates a CEFR level of A2, but rather that one would expect an A2 learner to have a GALW score of, say 90 to 110.

This score-linking of lexical knowledge to CEFR level is well established in language assessment literature. Milton (2010) examined how vocabulary size correlates with CEFR proficiency levels, finding that vocabulary size significantly contributes to communicative performance, suggesting that as learners progress through CEFR levels, their vocabulary breadth expands in a predictable fashion. Benigno and De Jong (2019) explored the alignment of vocabulary knowledge with CEFR levels using a psychometric approach, noting that as learners' proficiency increases, their vocabulary size expands, aligning with higher CEFR levels. Meara and Milton (2003) developed an approximation of English vocabulary size mapped against the CEFR levels, ranging from around 1,500-2,500 words for an A2 learner, to 4,500-5,000 for a C2 user. Such measures however are likely to be language specific and not generalisable (Milton 2010). Alexiou (2021) explored such score-linking in a primary school context, developing the Pic-lex tool for measuring early L2 vocabulary development in younger learners. These scores were then mapped against CEFR levels creating an estimated score-level correlation table. However, it should be noted that Elliot (2013) warns against placing too much trust in the accuracy of CEFR to describe younger learners at higher levels due to the assumptions within the framework descriptors of educational and developmental stage.

In light of these examples, the GALW adopts a score-linked approach to the CEFR in order to enhance its functionality for teaching staff, allowing them to monitor individual and group progress towards lexical knowledge associated with particular levels of the CEFR. This will assist them in ensuring that pupils make progress towards the learning goals outlined in the WLEB and leave education as a competent user of the language, in alignment with the objectives of Cymraeg 2050.

The score-linking between GALW and the CEFR will be carried out using equipercentile linking (Muraki et al. 2000); a nonparametric method that matches scores based on their cumulative distribution functions. A sample of 120 learners who had taken both assessments will be used to improve accuracy in the calibration. First, the structural assumptions will be checked, with a Spearman's Rank correlation being computed as an indicator of correlation, whilst residual analysis is used to check for linearity and homoscedasticity. The scores on each assessment will be ranked and the cumulative distribution functions (CDFs) computed for both GALW and CEFR. Each score in GALW was then mapped to the score in CEFR that corresponded to the

same percentile rank, ensuring that equivalent performance levels were aligned across assessments. To account for variability and improve robustness, kernel density estimation will be used as a smoothing technique, to reduce noise in the CDFs. The sample will be split into a training set (80% of the data, n=96) for constructing the linking function and a validation set (20%, n=24) to make an initial assessment of generalizability. The validation set will be used to compute prediction errors and evaluate the stability of the linking function by comparing actual and predicted CEFR scores. To quantify uncertainty, a bootstrapping procedure will be used to estimate a 95% confidence interval for each linked score.

Grade thresholds in score-linking to CEFR

Benigno and Jong (2019) raise an important question: where within the confidence range do you situate the level threshold? As score linking between the GALW and the CEFR involves anchoring a continuous score to a categorical score, it will inevitably create equivalent score ranges e.g. a GALW score between 22 and 28 may be found to correspond to a CEFR grade of A2. However, for the purposes of score-linking, a decision must be made as to what point within that range we classify the learner into that CEFR grade. As the range represents a confidence interval (see chapter 7.6 for a full explanation of scoring validity), it is not necessarily accurate to include the whole range in the allocation of the equivalent CEFR grade. In addition, ranges will almost certainly overlap or have gaps, and so the establishing of grade thresholds must be considered.

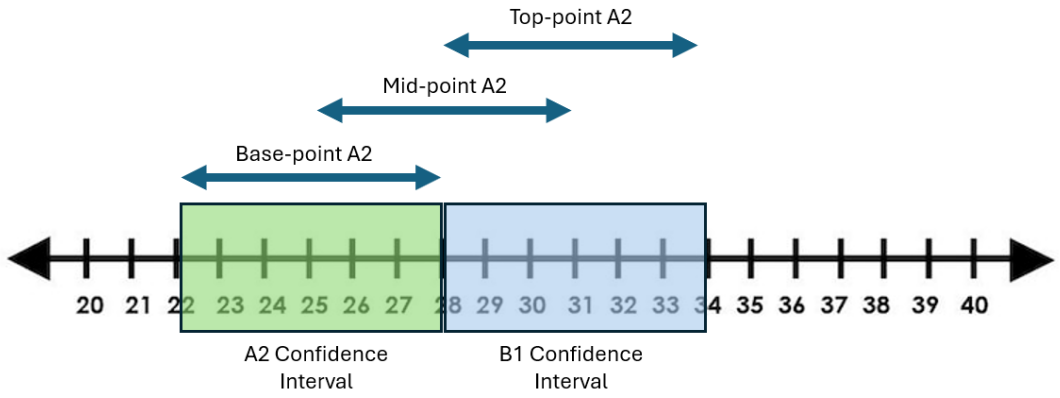

Fig.1: Visualisation of scoring linking options when determining grade boundaries.

Three broad options present themselves: base-point linking, mid-point linking, or top-point linking. Base-point linking, fixes the score-grade threshold to the bottom of the confidence interval, (e.g. if A2 = between 22 and 28, and B1 = between 28 and 34, a base-score link would set the threshold at 22 and extend to the base-point of the next grade, 28). Mid-point linking, would span from the median of the range to the median of the next range (e.g. in our example from the midpoint of A1, 25, to the midpoint of B1, 31). Finally, top-point linking would span from the top of the A1 range (28) to the top of the B1 range (34).

Base-point thresholds are the most generous; assuming that learners who reach the bottom of the confidence interval are equivalent to the corresponding CEFR level, therefore creating the possibility of positively misrepresenting their likely CEFR performance. Top-point linking offers the greatest degree of certainty that the learner has almost certainly reached/exceeded the equivalent score for the corresponding CEFR level (Cizek and Bunch 2007), but is likely to negatively misrepresent the ability of many learners.

Kolen and Brennan (2014) note that decisions on where to set a cut score within that interval should be guided by both statistical factors and by practical considerations, such as the purpose and context of the assessment. Lord (1980/2012) argues that, assuming the linking function is linear and the error distribution is symmetric, then the most robust score-grade link is best represented by the midpoint. However, if the linking function is asymmetric or if there are additional criterion-referenced anchors mandated, the threshold might need to be adjusted accordingly.

Thresholds also need to be aligned with the assessment purpose. For high stakes tests where assuring minimum competence may be critical, a top-point threshold may be suitable. However, for the GALW the emphasis is upon the guidance of learner and teacher towards progression. Whilst this would seem to support the simple application of a mid-point threshold, as providing the most robust score-grade link, this must be considered within the context of the confidence interval used. If a 0.95 value is applied (i.e. a score falling within the range will equate to the equivalent CEFR score in 95% of cases), it would seem unreasonable to exclude users who fall within the range, but below the mid-point. Given this high threshold and the low-stakes, formative nature of the GALW, a base-point approach seems justifiable.

The likelihood of gaps is also an important factor for consideration. In these instances, the simple approach would be to set the threshold to the median of the score-grade gap. However, this assumes a linear relationship between the score and grade correspondence, which is unlikely in language assessments generally (Bachman & Palmer 1996) and does not represent the nature of the CEFR (Figueras et al. 2005). Therefore, in order to reflect the likely uneven distribution of scores between two score-grade thresholds, the GALW will modify the confidence interval for the two grades to determine the most appropriate grade-score threshold. In the example below, the 0.95 interval leaves a 3-mark gap between the B1 range and the B2 range. Relaxing the confidence interval to 0.7 for both B1 and B2 accommodates the difference in distribution whilst indicating a suitable threshold between the two score-grade ranges.

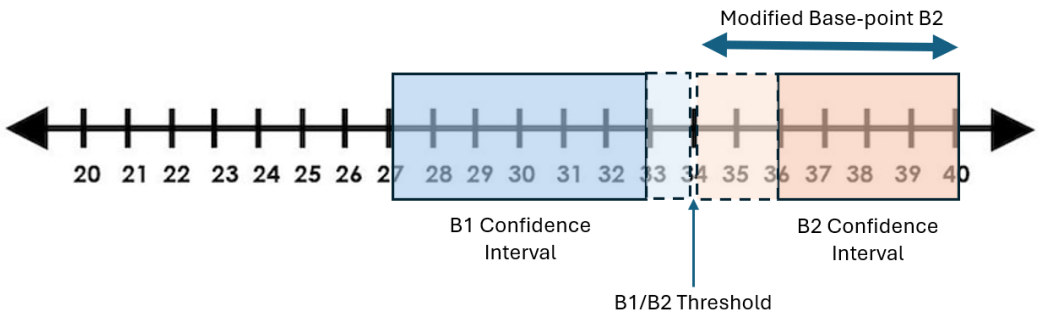

Fig.2: Visualisation of interval gaps addressed by modification of confidence.

In this way, scores from the GALW can be used in combination with CEFR based assessment to build a table of corresponding score-grades, with thresholds that offer the best prediction practicable in corresponding GALW scores to anticipated CEFR performance. Progress towards these score thresholds can be displayed to teachers to provide an indicator of progress towards policy-lead goals, and to demonstrate progression to learners in order to enhance self-efficacy and motivation.

Error propagation

The accuracy of the score-link assessment relies heavily on accurate identification of the CEFR level of the users who are used to calibrate the assessment (Milton 2010): inaccuracy in the CEFR calibration can lead to 'error propagation' (McLaughlin 1983) further reducing the accuracy of the GALW link-scores as they are a product of compounded measurement errors from both assessments.

This compound uncertainty can be expressed simply as:

$$\text{Total Uncertainty} = \sqrt{(\text{CEFR Interval})^2 + (\text{GALW Interval})^2}$$

In a simplified example where both scores are on a scale of 1-10 and the confidence interval for both score is 2 (p=0.95), we can see how the uncertainty in the CEFR interval exacerbates the uncertainty in the GALW interval (from 2 to 2.83):

$$\begin{aligned}\text{Total Uncertainty} &= \sqrt{(2)^2 + (2)^2} \\ &= \sqrt{4+4} \\ &= \sqrt{8} \approx 2.83\end{aligned}$$

Obviously, the greater the uncertainty in the CEFR assessment the greater the effect on the accuracy of the GALW's predictive CEFR level.

The effects of error propagation can be mitigated through continuous refinement and calibration of the assessment over time (see chapter 5.1 for details of this process), but ultimately ensuring high levels of accuracy in the direct CEFR assessment of Welsh will be the most important factor in minimising error propagation.

3.3 Facilitation of Research

The GALW aims to facilitate research into Welsh language teaching and learning in EM primary schools through the collection of comparable cross-sectional and longitudinal data on learner progress in developing WLC knowledge. This can then be combined in models including other variables to determine the impact of variations in factors such as pedagogic approaches, learning provision, socio-economic contexts, individual psychological traits, and individual learning strategies.

It is important to emphasize that the GALW is not a standardized assessment, due to the content selection process (see chapter 5.2). Instead, it seeks to create comparability, allowing learners in different contexts with different educational experiences to be measured against the same scale (Winter 2010). The lack of such score comparability in current practice is a significant impediment to conducting research into pedagogy, provision and their impact on learning outcomes (Russell 2025). The comparability offered by the GALW facilitates trial studies observing differences in outcomes that can be attributed to the intervention/ provision/pedagogy itself, rather than variations in the assessment tools (Hill et al. 2023).

Anonymized unique learner identifiers (ULIs) will also facilitate linking to other data collection tools (surveys, demographic data, other assessments) allowing researchers to use the GALW generated data in a range of multi-variate models. More details of this process can be found in chapter 8.3.

In addition to formal academic research, it is intended that the GALW will provide the means for teachers to carry out their own 'action research'. Also referred to as 'professional enquiry', action research is strongly advocated in the 'National Approach to Professional Learning' (WG 2018b) with numerous advantages identified. These include enhancing professional development (Gilchrist 2018), empowering teachers to challenge existing practice, and encourages greater participation from the school's community (Cardiff Met 2019). The GALW will allow teachers to measure the impact of adaptations to pedagogy, and the introduction of new provision, approaches, or interventions. This will enable teachers to develop their practice, tailor and differentiate provision, and target support more effectively, whilst giving them a clear way to demonstrate and communicate impact with other learning professionals.

4. Target Population

It is essential in assessment design, that in addition to what is to be assessed, and how it is to be assessed, we consider who is being assessed (Norris 7 Otega 2012). Defining the target population (TP) is important in ensuring that bias stemming from individual and sub-group features is avoided. It is important these factors are considered in an *a priori* manner (Elliot 2013), prior to empirical validation, in order to ascertain the efficacy of the theoretical mitigations adopted. To carry out this appraisal, O'Sullivan and Green's (2011) theoretical framework will be drawn upon, which identifies three categories of characteristics: physical, psychological and experiential (p. 38). The constituent factors within each category are detailed below in a table adapted from O'Sullivan (2000):

Physical	Psychological	Experiential
Age	Cognitive:	Education
Gender	○ Working memory	Assessment preparedness
Short-term ailments	○ Processing speed	Assessment experience
Long-term disabilities	○ Attention Span	Delivery language ability
	○ Language Aptitude	Topic knowledge
	○ Long-term memory	Broader knowledge
	Affective:	
	○ Personality	
	○ Affective schemata	
	○ Emotional disposition	

4.1 Physical Characteristics

The physical characteristics of an assessment user consist of persistent features (age, gender, disability) and transient features (injury, illness). Precautions must be taken to ensure that the assessment delivered in its standard format does not disproportionately disadvantage users who possess particular features, in isolation or in combination. This may include the provision of additional support or allowances to ensure equitable access to the assessment (Douglas 2011).

In terms of transient features, school attendance is a significant mitigating factor: the informal nature of the GALW makes it unlikely learners will make additional effort to attend school when unwell as they might for more high-stakes assessments. It is reasonable to assume that

attendance is indicative of an absence of any transient features that would significantly impede performance. Further mitigation is provided by the assessment rubric for administrators which indicates that additional support should be provided to any learners who have additional needs to access the assessment (e.g. the provision of a touch-screen interface for learners who may struggle to manipulate a mouse). A final modifying factor is the assessment intended use as a regular formative assessment tool. This means that retesting is likely to occur at regular intervals and so the impact of any transient feature will be ameliorated by the aggregation of cumulative data (Gnambs 2014).

Persistent features pose a greater challenge in terms of accessibility and score distortion. Schools are under a statutory obligation to accommodate the learning needs of students (WG 2018a; UK Gov. 2010) and as such, provision should be in place to ensure that all learners are supported to access to provision. Despite the WG adopting an inclusive and pupil centred approach which aims to deal with many learning needs through enhanced pedagogy in the classroom (WG 2018a), in practice some learners with additional needs do not receive the support they require (NAHT Cymru 2024), and so additional mitigation needs to be considered as part of the assessment design.

The following features seek to make the GALW more accessible for those with persistent features that may impact their performance:

- Text-to-speech functionality will be included on all items. This will benefit users with learning needs (e.g. dyslexia), as well as those with visual impairments. It will also act to decrease cognitive load and fatigue for those users with a lower reading proficiency.
- Text will be included on all items. This will benefit learners with hearing impairments or auditory processing disorders (APDs).
- Video instructions will be included alongside text rubrics to ensure all learners can understand the assessment format and process. Subtitles will be included in the videos to improve accessibility for those with hearing impairments or APDs.
- ALN friendly fonts and backgrounds will be used for all text to improve readability for those with dyslexia (Alexiou 2021; Yoliando 2020), with potential benefits for those with minor visual impairment, and ADHD/ADD (Phalke 2023).

4.2 Psychological Characteristics

Psychological factors can be categorised as either cognitive, or affective (O'Sullivan and Green 2011). Cognitive factors, such as memory, style, or concentration, are likely to directly affect the candidates' performance in the assessment tasks and are therefore associated factors of proficiency. Affective factors, such as personality, motivation, or emotional disposition, have a more indirect impact on performance (Zabihi 2018). A distinction therefore needs to be drawn between psychological characteristics that influence performance, and those that are part of the construct competency (Henning 1992).

Cognitive factors

We can distinguish five cognitive features that impact performance at an individual level: working memory, processing speed, attention control, language aptitude, and inferencing. These cognitive factors need to be considered within the context of the individual's cognitive development (Elliot 2013), as significant divergence in performance within each factor can be a product of age-related development. This is clearly pertinent to the GALW's target application in

primary school settings, generally including children aged 4 to 11 years old. It is therefore important to design tasks that do not exceed the cognitive abilities particular to this age range or consider guidance that mitigates the effects of cognitive development disparities.

- Working memory, i.e. the ability to hold and manipulate information in short-term memory. Although crucial for processing and producing language in real time (Baddeley 2003), the format of the assessment (providing text interfaces and repetition of audio) mitigates the effect of working memory of performance. In addition, the assessment's target constructs are centred on lexical knowledge which does not require the operationalisation of working memory.
- Processing speed, i.e. the speed at which an assessment user can perceive, analyse, and respond to language stimuli. Again, although an essential element of oral communicative competency, processing speed is not included in the GALW target constructs. The same text-based input and repeat function for audio input mitigates any impact of processing speed in order to isolate lexical knowledge.
- Attention control, i.e. the ability to focus on relevant input while ignoring distractions. Although particularly important in listening tasks, attention control will have a more limited impact on performance in the GALW due to the text/audio format already mentioned and the limited phrase length included in the WLC content. However, a more general factor in attention control must be considered; longer assessment formats can lead to learner fatigue and disengagement which can have a distorting effect on scores (Ackerman & Kanfer 2009). The design of the GALW needs to consider the potential for such distortion and the balance that must be struck between this and collecting a sufficient number of item responses for the purpose of scoring validity (see chapter 5.6).
- Language aptitude, i.e. vocabulary knowledge, grammatical sensitivity, and the ability to recognize language patterns, and long-term memory retrieval. Although both factors are highly influential on assessment scores, they are highly integrated in the GALW's target constructs, to the extent that mitigating the impact of language aptitude and long-term memory would obscure the measurement of the constructs under scrutiny (Weir 2005). As such, scores produced by the GALW must in part be considered a measure of individual language aptitude and long-term memory retrieval.
- Inferencing ability and strategies, i.e. the capacity to deduce meaning from context, which is key for reading and listening comprehension, especially when encountering unknown words. This could include inference drawn from L1 knowledge (e.g. cognates), partial knowledge of the L2, identification of narratives/themes, visual clues (e.g. images, body language), or auditory cues (e.g. inflection, tone) (Cohen 2011, p.663). Whilst inference is an important skill in development communicative competence, it can distort the measure of lexical knowledge (Pearson et al. 2007), especially in MCQ formats where guessing can already create inconsistency. The risk of inferential factors distorting scores is covered in more detail chapter 7.7.

Other higher order skills such as cognitive flexibility and executive function will have limited impact on the accuracy of the GALW, which targets lexical knowledge rather than these more communicative skills.

Affective factors

Affective factors may have a distorting effect on assessment results through their inhibition of motivation, effort, concentration or cognitive function (Nation 2007, Wise and Smith 2016). Learners may be bored, distressed, lethargic or combative during the test, inhibiting their

performance and misrepresenting their competence in the target construct. Elliot (2013) identifies three key affective factors in assessment performance: personality, schemata, and emotional state. As with cognitive features, the age/development of learners should be considered in mitigating the impact of these factors, whilst also ensuring that elements embedded in the target construct are not obscured.

- Personality factors (extroversion/introversion, risk taking, conscientiousness, agreeableness) tend to have a more exaggerate effect on speaking assessments, where such traits are likely to have a direct effect on both learning and assessment performance (Ghapanchi et al. 2011). Whilst the MCQ format of the GALW mitigates some of these distortions, others remain pertinent; chiefly, risk taking. A higher tolerance of risk is associated with a greater willingness to guess in MCQs (Rubio et al 2010) enhancing individual scores. This has also been observed to manifest differently across sexes, with male candidates being more predisposed to risk taking than female candidates (Coffman & Klinowski 2020). Further consideration and mitigation of this factor can be found in chapter 7.7.

  Whilst factors such as conscientiousness and agreeableness impact upon L2 learning (Chen et al. 2021), such learning outcomes are part of the construct under scrutiny in the GALW. Therefore, within the context of the GALW format, there are no obvious distorting effects on individual test performance beyond construct competency.

- Affective schemata are the emotional and attitudinal frameworks that effect how test-takers perceive, manage, and carry out assessment tasks. Such schemata are informed by experiences, beliefs, and emotions of language acquisition and assessment, and can have a significant impact on performance (Eliott 2013). Affective schemata largely fall into two categories, those affected by content, and those affected by format. Content factors would include the inclusion of emotionally resonant topics (e.g. gun control, hunting, bereavement) which may have detrimental effects on the language performance of participants (Bachmann & Palmer 1996), whilst format factors are more likely to include schemata relating to assessment anxiety more generally. Given the culturally neutral and age-appropriate nature of the WLC content, affective schemata relating to content will not be a contributing factor to assessment performance. However, there is the possibility that affective schemata relating to assessment anxiety could play a distorting role in the GALW. Mitigation of this factor are addressed in chapter 5.2 which explores the semi-covert assessment format of the GALW.

  Learners may also hold affective schemata regarding the Welsh language, with studies finding significant levels of disengagement from and/or disillusionment with the language in EM schools (Rhys & Smith 2022; Parry & Thomas 2024; WIZERD 2023; Gruffudd 2000). Such subject specific negative affective schemata (NAS) may result in reduced effort in assessment tasks, resulting in scores that do not accurately reflect user competence. Such NAS will also clearly have an impact on learning outcomes in the target construct (Yu 2022), making it challenging to disentangle situated impacts on the assessment performance, and more general impacts on construct competency. The ‘testing effect’ would potentially mitigate the effect of negative affective schemata (Gneezy et al 2019), however the covert formative approach of the GALW prohibits this.

Additional research into the effects of NAS will be required to assess the impact on assessment performance as distinct from construct competency. This analysis will entail the identification of learners with NAS through a survey integrating adapted elements of the Achievement Emotions Questionnaire (AEQ) (Pekrun et al. 2005). Identified learners will then be retested using the GALW format, but under more formal test conditions (invigilator present, informed that they are being assessed), and using additional motivational incentives agreed with teaching staff. A regression model can then be constructed, using the retest scores, the initial GALW scores and the AEQ score, to estimate a NAS coefficient. If significant, this can then be utilised in moderation of GALW scores, or considered within the context of broader qualitative data.

- 'Emotional state' refers to the transient emotions of the user at the point of assessment. Whilst potentially interrelated with NAS, this state is not necessarily a product of previous experiences or preconceived ideas around the assessment or language: for example, a learner may simply have had an argument on the day of the assessment or fallen out with a friend. Such temporary and discrete emotional events can have a significant impact on academic performance (Davis et al. 2003, p.9). The GALW includes a number of features to mitigate the distorting effect of the user's emotional state: firstly, the covert nature of the GALW seeks to decrease negative emotional responses associated with test anxiety; secondly, the capacity for multiple retesting decreases the impact of isolated emotional events on performance; finally, the GALW rubrics (chapter 5.6) offer guidance on the environmental and contextual factors to be considered in the administration of the assessment (e.g. ambient noise, proximity to play-times, proximity to emotional distress) which aim to mitigate or circumvent distortion caused by users' emotional state.

4.3 Experiential Characteristics

The experiences of assessment users can have a significant impact on their performance in multiple ways. As such, experiential characteristics of test takers should be considered in the development of the assessment in order to reduce the impact of these factors that may provide advantage/disadvantage to assessment users on the basis of factors distinct from construct competency. Precautions must therefore be taken to ensure that successful demonstration of competence does not inadvertently require a particular experiential profile (Elliot 2013).

O'Sullivan (2000) identifies five key aspects of user experience that should be considered in the context of assessment development: education, examination familiarity, target-language exposure, topic knowledge, and world knowledge. Each will be considered below within the context of the GALW.

Educational experience

Educational experience encompasses not just the topic specific instruction the user has encountered, but also users' awareness of broader educational practices and norms. The test format therefore needs to be concurrent with the users' educational experience to avoid familiarity bias (Peña & Quinn 1997). Due to the institutional setting on the GALW and the relatively simple structure of the WLC content, the potential for distortion through incompatibility of educational experience and test format seems unlikely.

Educational experiences may also be a key factor in the formation of NAS, as explored in section 4.2. Whilst the GALW cannot account for how education experiences have shaped users' schemata, mitigation is made in the assessment rubrics (see chapter 5.6).

Examination familiarity

Examination familiarity is a sub-category of educational experience, concerning the users' specific experience of testing, the norms and expectations associated with it. Higher levels of examination experience can allow users to cope better with the assessment pressure (assuming they do not have NAS associated with assessment), and develop the metacognitive skills to optimise their performance, e.g. time management, task-prioritisation, answer revision strategies (Dodeen 2008).

Examination experience is unlikely to have a strong effect on the GALW due to the similarity of users, who will generally be drawn from a relatively homogenous population of mainstream EM primary school students, who are therefore likely to have comparable levels of examination experience. In primary education, learners experience of formal examination should be very limited, as Wales moved away from high-stakes assessments for this age group in the early 2000s (NAW 2001). This should create greater consistence, as well as a decrease the likelihood of learners having NAS associated with assessment. In addition, the GALW's formative and semi-covert approach, untimed format, and mandatory completion structure go some way to mitigating the effects of examination experience and associated learner strategies.

Target-language exposure

Target-language exposure can advantage learners from areas, communities, or families with higher numbers of target-language users (Elliot 2013). Within a Welsh context, this is likely to disadvantage learners from more anglophone areas of Wales, and those from minority-ethnic communities. Accordingly, Cook (2016) defines two types of L2 speaker: 'L2 learners', which is to say those acquiring the language in an educational context, and 'L2 users', who utilise the language beyond the classroom setting. In much of Wales, learners in EM schools fall into the 'L2 learner' category, with only a minority progressing to L2 users (Lovell 2016). Whilst the additional exposure may advantage L2 users in assessments, such familiarity with the language is highly construct-relevant, and an attempt to mitigate for language exposure in the GALW would inevitably distort measures of construct competency. It should be noted that this issue is distinct from the user competence in the language of assessment delivery, which is addressed in chapter 5.6.

Topic knowledge

Topic knowledge refers to the subject matter of the assessment, rather than the language content, i.e. the assessment may include sections on the topic of sport, for example, as a useful format for exploring language skills or content. Disparities in individual topic knowledge can cause a misrepresentation of linguistic skills by conflating them with topic knowledge (Clapham 1996). The risk of distortion from variance in topic knowledge is mitigated by two factors in the GALW: firstly, the WLC content, although often delineated by topic (e.g. the weather as a format for exploring third person simple) is relatively generic and is directed to practical language patterns for the school context. As such, topic knowledge is unlikely to be a factor in construct competency. Secondly, due to the personalised content of the GALW, learners will only be exposed to assessment items that they have encountered in lesson settings.

Knowledge of the world

Finally, knowledge of the world is also cited by O'Sullivan (2000) as a potential source of individual score disparity. Whilst a potentially significant factor for more advanced learners using authentic texts (Elliot 2013), knowledge of the world is unlikely to have an impact on GALW scores, which draw on the relatively parochial content of the WLC.

5. Assessment Design

In this chapter how the contextual framework outlined in previous chapters forms the basis for the actual GALW assessment design will be considered. The GALW design is not a single event, but an iterative and recursive process, through which the assessment is refined and calibrated over time to improve accuracy and enhance validity. We start by outlining this process and the timelines surrounding it. Next, the format and functionality of GALW is outlined in detail, with sections on structure, user interface, item development, distractor development, excluded task formats, and test delivery processes. Finally, the focus of this chapter will shift to the assessment outputs and how they're intended to be deployed by teaching staff, learners, and researchers.

5.1 The GALW design process

Assessment development is not an isolated event, but an ongoing process, whereby the design, content and procedures are analyses and refined in response to performance and user need (Cumming 2012). In this way, the GALW continuously improves its capacity to both reflect learner proficiency in relation to the target constructs and adapt to changes in educational practice and learning context.

As shown in Fig.3, the design process of the GALW consists of five key phases: initial development, beta trialling, initial pilot rollout, new iteration development, and cyclical development.

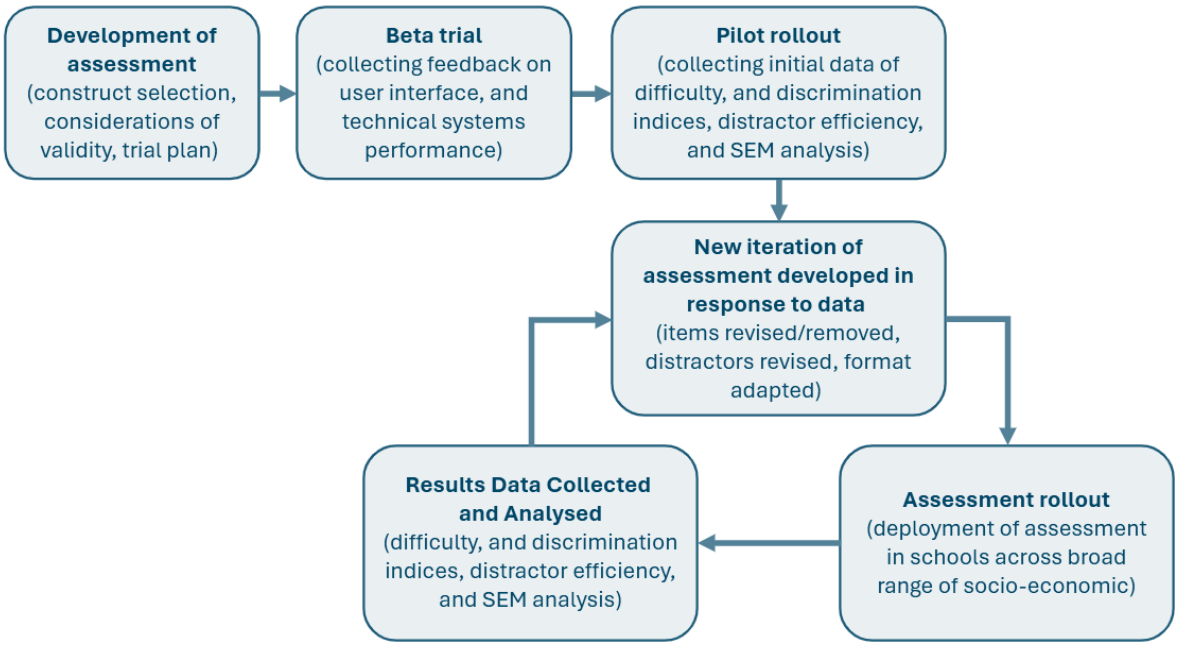

Fig.3: A flow chart of the phases of development for the GALW.

1. The development of the assessment

This phase is represented by this document, a process of compiling a wide range of information and sources to develop a cohesive and theoretically grounded assessment specification. Key elements of this phase include: a scoping study to assess existing provision and establish the need for an assessment tool such as the GALW; context evaluation of Welsh language education in English medium primary education; an exploration and appraisal of language assessment literature; and the development of a detailed specification.

2. Beta Trial

The beta trial is intended to highlight any functionality issues with the assessment tool prior to a wider rollout. It will consist of around 30 respondents, representative to the characteristics of the target population. Purposeful sampling is used to ensure that the beta trial includes perspectives of users who may face challenges to accessibility (visual/hearing impairment, EAL and ALN). Trial analysis will consist of three stages:

- Stage 1 – Live feedback and narration. A small group of users (n=10) will be asked to talk through their interaction with the GALW as they use it: what they're thinking and how they're feeling. This will highlight problems with the functionality, e.g. elements of the interaction with the tool they find confusing or unclear, and aspects that may impact them emotionally, such as the level of difficulty, test length, and type of feedback. Notes will be taken of issues raised and suitable mitigation or adaptation made to the assessment where possible.
- Stage 2 – Trial and survey feedback. A larger group of users (n=50) will be asked to complete the assessment and complete a short survey about their experience. In this they will be given the opportunity to highlight any problems they experienced whilst using the assessment tool. Data from this stage will be reviewed to identify any themes or discrete problems that can be addressed prior to rollout.
- Stage 3 - Data output analysis. This stage consists of an analysis of the data collected from stage 2. This includes ensuring that the data is being recorded correctly, stored appropriately, is generating appropriate feedback in both the staff dashboard and learners' automated feedback.

The information and feedback from these three stages guides final refinements to the user interface and functionality before the release of the GALW version 1.

3. Pilot roll-out

Version 1 of the GALW will be rolled out to schools, with efforts made to recruit from a wide variety of regional and demographic profiles. Anonymised data will be collected on user performance in each item, which will be used to analyse difficulty, discrimination, distractor efficiency, internal consistency, and SEM (see chapter 7.5 for more details).

Calculating sample size for this pilot is challenging. In standardised assessments with stable item blocks and assuming a battery of 1000 items, with each user interacting with 40 items, each item will be seen by approximately n / (1000/40) = N / 25 assessment responses. Whilst significance will increase in line with sample size (Blanchin et al. 2011), it is generally assumed that a minimum of 100 responses per item are required in CTT (Cappelleri et al. 2014). With this assumption, the minimum number of assessment responses would be 2500.

However, two factors complicate this assumption in the case of the GALW. Firstly, the content selection format of the assessment (see section 5.2) means that assessment responses are

unlikely to reflect an even distribution of the assessment items, with more common pattern blocks being overrepresented, and more esoteric blocks underrepresented. This means that the validity of the assessment is unlikely to be homogenous, with commonly selected items benefiting from substantially more analytic data. Assuming 2500 item responses, rather than 100 responses on each item, this is more likely to result in hundreds of responses on some blocks of content and very few on others. As the number of item responses grows through more extensive assessment rollout, this disparity should become less significant as the per-response impact of feedback data diminishes as it accumulates (Fig.4):

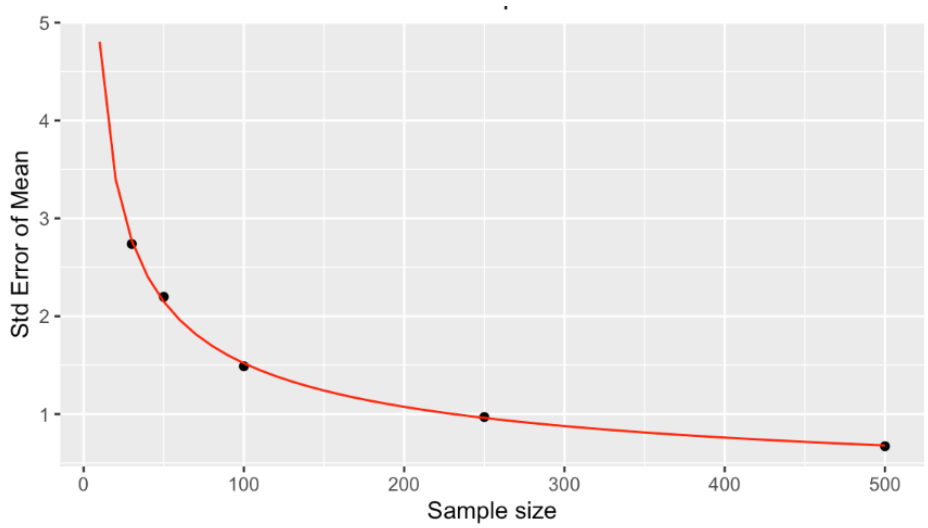

Fig.4: A visualisation of depreciating returns in reduced error in relation to sample size. (Magwene 2023)

However, in the short term this may mean that the assessment may be less accurate when learners are working on less commonly covered content blocks, this could be because the blocks link to a specific curriculum theme for their institution, or if they are working at a higher level than the majority of learners.

The second challenging in determining sample size, is that due to the formative focus of the GALW, a single participant is likely to make repeated engagements with the assessment over time. This means that the aim of collecting at least 2500 responses before developing version 2 of the GALW may be gathered from a much lower number of users (e.g. 500 users who each complete the assessment five times). In this case, caution must be taken to ensure that the sampling includes a sufficient number of individual users, and offers sufficient representation of groups that my experience assessment bias.

4. New iteration of assessment developed

The data from the initial roll-out is collected and subjected to a variety of statistical modelling, including difficulty, discrimination, distractor efficiency, and item bias analysis. You can find full details of the statistical approach to assessment development in chapter 7.5. Functionality will also be reviewed in light of any feedback received from users during the pilot period (e.g. error reports, information requests, technical support enquiries).

In response to this analysis a new version of the GALW is then developed. This may include the removal of difficulty based redundant items, the adaption of inefficient or over-efficient distractors, and the adaption of items offering low levels of discrimination. Where possible, continuity will be sought in order to maximalise comparability between version 1 and version 2 scores (Berman et al. 2020).

5. The cyclical phase

As already stated, the assessment design of the GALW is not a discrete event, but an ongoing process. As shown in Fig.3, the GALW design follows a cyclical model (Lam & McNaught 2008), which continuously repeats the process of assessment rollout, data collection, data analysis, assessment adaption, new version rollout. After phase 4, the GALW enters this cyclical phase, with new iterations being developed at pre-set response number thresholds (see Tab.2). This ensures that the development is responsive to the assessment usage, avoiding the revision of the GALW with insufficient user response data, or the delay of revising the assessment which may negatively impact upon its functionality.

Version #	No. or Responses	No. of Users
v.1	2500	>500
v.2	5000	>500
v.3	5000	>1000
V.4+	10000	>1000

Tab.2: Response number thresholds for development of new iterations of the GALW

5.2 Format and Structure of GALW

Taylor & Galaczi (2011) highlight the need to align the format and structure of an assessment with the nature of the constructs: as outlined in chapter 3, GALW adopts an analytical approach, seeking to assess different constructs relating to learners' lexical knowledge. Accordingly, the assessment tasks and structure seek to isolate and quantify these constructs, rather than seek a holistic measure of learner competency.

In addition to this overall analytic approach, the assessment format must be considered with respect to the overall assessment goals (see chapter 1). Before looking at these goals individually, it must be emphasised that the primary objective directing GALW's design is the enhancement of learning. This prioritisation is informed by both ethical factors, as assessment should be designed with the best interests of the learner in mind (Green et al. 2007), and a pragmatic one, as the widespread adoption necessary for generating useful data will only occur if staff find the assessment beneficial and practical.

Below we outline how GALW seeks to respond to each of the assessment goals:

1. *Identification of deficits in WLC knowledge that can be used by teaching staff to improve targeting of content coverage, remediation of previous content, differentiation of content, targeting of scaffolding and resources.*

   GALW acts as a diagnostic AfL tool, able to carry out a lexicon analysis of individual learners, classes and cohorts to ensure learning objectives are set appropriately and monitor ongoing progression (Alexiou & Stathopoulo 2021). A lack of such formative

assessment can lead to poor content selection, goal/expectation setting, differentiation and scaffolding (Guskey 2003; Milton & Alexiou 2023). To ensure accurate measurement of content knowledge (rather than conflating knowledge with coverage), GALW is content aligned by the class teacher. Consequently, it measures learning and retention providing a predicted content knowledge score based on a combination of teacher specified content coverage and assessed content retention (see chapter 7.5 for more details). Scores will be collated automatically and will be accessible in a teacher dashboard, giving a number of analytical insights to help inform pedagogical practice.

2. *Monitoring of WLC knowledge development over time, identifying periods where learners experience periods of enhanced acquisition/stagnation/attrition, allowing the targeting of additional support or adaptation of pedagogical approaches to mitigate any negative factors.*

   GALW does not only act as a cross-sectional assessment but seeks to further enhance learning and research insights through the collation and presentation of longitudinal performance. A record of longitudinal scores allows teachers to understand the developmental trajectories of learners, identifying critical periods for additional support or intervention (Kwok et al. 2018). The GALW will deploy a cumulative and tracked competency measurement to give insight into both the current user ability, and the learning trajectory (Fernandes et al. 2018). The cumulative competency measure will be composed of the average score of a learner over a rolling two-week period (time-frame based on analysis by Pedhazur and Schmelkin 1991). This increases the accuracy and stability of the predicted score, without excessively compromising the assessment's representation of competence through inclusion of redundant scores (i.e. those that no longer reflect the learner's competence due to additional learning progress). The tracked scores will be used to generate a time series plot graph, visually displaying the learners' scores over the time they have used GALW. An aggregated class score will also be provided allowing for a convenient method of monitoring group progress in response to particular units of teaching.

3. *To provide a research tool allowing investigation into the comparative efficacy of different resources, pedagogical approaches, learning experiences and interventions in developing WLC content knowledge.*

   In addition to its AfL functionality, GALW will produce data sets to facilitate research into Welsh language learning in the EM primary school context. Its combination of cross-sectional and longitudinal data can be used to indicate potential causal relationships between adaptations to pedagogy/provision and learning outcomes, as well as facilitating experimental trial-based studies. Such data can be used to develop recommendations around best practice, identify the impacts of various specific interventions or provision (Watts et al. 2019; Nese et al. 2013). GALW also offers a tool for empowering teachers in conducting their own action research to inform their practice, as advocated by the CfW (see section 5.7).

4. *To provide research insights into the impact of different institutional, socio-economic, and individual factors on learners' acquisition of WLC knowledge. Allowing an improved*

understanding of how such factors may contribute towards educational inequalities and how these may be addressed.

The datasets generated by GALW will include anonymised ULIs (for more detail chapter 8.3) allowing for the linking of GALW scores to demographic and survey-generated data. This data linking will allow researchers to investigate potential relationships between Welsh language performance and socio-economic or psychological variables of interest (e.g. national identity, material deprivation, motivation).

5. *To provide individual learners with a way of tracking and understanding their progress in WLC knowledge with the intent of developing improved self-efficacy and motivation*.

   GALW's functionality will include automated learner feedback. This is generated by an appraisal of the cross-sectional data to highlight specific areas of learner deficiency that the user needs to address, and longitudinal data to make a general response on progression. This feedback is designed to facilitate self-monitoring and progress perception in learners, with the goal of enhancing self-efficacy (Pajares & Schunk (2001; Schunk & Mullen 2012). Details of the user interface can be found in chapter 5.7).

6. *To provide standardised measure of WLC content knowledge to teaching professionals managing learner transitions (between classes, key stages, primary/secondary education, inter-institution) allowing improved learning continuity, differentiation, and content alignment, whilst mitigating the risk of reducing learner motivation through a regression to foundational language patterns*.

   The longitudinal data collected by GALW will be accessible to download as an individual learner profile. This will summarise the learners current predicted lexicon, the score history at the end of each term to demonstrate learning trajectory, a CEFR linked score, and a list of language patterns that require future development. Similar summaries can be generated at a class/cohort level and can be included in the school's transition plan. This data will help secondary schools to ensure support, continuity and progression of learning for pupils transitioning, which can enhance learner motivation and improve long term learning outcomes (Bolster 2009; Braund 2009).

To achieve these goals GALW aims to balance accuracy and practicality. As use of GALW is elective, it is essential to make the assessment both useful and user friendly if uptake is to be widespread. These requirements led to the following decisions being taken concerning the structure and format of the assessment: the use of multichoice questions, the use of English as a reference language, self-administration, semi-covert format, teacher selected content, generative feedback, integrated audio features, and item complexity.

Multi-choice questions (MCQs)

Laufer (2004) proposes a hierarchy of format difficulty in L2 assessment, incorporating the dichotomous pairs: passive/active, and recognition/recall. 'Active' here refers to the ability to produce or retrieve language forms and meanings independently, without external assistance or cues. In contrast, 'passive' refers to understanding or recognizing language without necessarily being able to produce it. Passive knowledge is about comprehension, whether through listening or reading, and is typically more receptive. Recall involves retrieving the word from memory

based on its meaning or context, without any cues, whilst 'recognition' involves identifying or matching a word with its meaning or form when given some kind of prompt or cue. Laufer pairs these characteristics and arranges them from easiest 'passive recognition', through 'active recognition', 'passive recall', and finally 'active recall', the most challenging. A holistic assessment, such as those directly aligned with the CEFR would be expected to include the full range of these pairs. However, it is possible for analytical assessments to be more restricted and still legitimately represent their more discrete construct competencies (Field 2013).

Active recall, often also referred to as active production, involves constructed responses, and it often considered the best indicator of lexical knowledge depth as it required the faculty to draw upon the word form-meaning without cues (Laufer 2004). However, productive assessment tasks are inherently less consistent and so more challenging to score objectively (Read 2019). Indeed, Norris and Ortega (2012) highlight the danger of constructed responses presenting features unconnected to the target constructs which can lead to an overestimation of the learners' communicative competence.

Multi-choice questions (MCQs, also referred to as 'selected response items') are often used to test receptive skills but can also be used in testing controlled-productive skills (Nagai 2020). Selected responses items have numerous practical advantages: they offer objectivity, consistency and quantifiability, can be automated in both selection and marking, and represent a familiar format for learners, thereby decreasing familiarity bias (Beerepoot 2023; Field 2013).

However, use of MCQs necessitates the adoption of an analytic, rather than holistic approach, that isolates competencies and doesn't take into account the effect of cognitive load on performance (Hughes 2003) i.e. whilst a learner may be able to perform certain language functions with a high degree of proficiency in isolation, the additional cognitive load of parallel processing these items diminishes the learner's holistic performance. Caution must be taken in inferring the learners' ability to integrate these discrete skills when estimating holistic competence (Taylor & Galaczi 2011). Although this does not impact the efficacy of the GALW as a comparative and diagnostic tool, the intention to score link the GALW to the CEFR levels means that the effect of cognitive load should be considered.

The effect of cognitive load is largely neutralised by the way in which score-linking is calibrated from established CEFR levels to GALW performance. In this way cognitive load is accounted for in the direct CEFR evaluation and reflected in the associated score-linking to the GALW. Of course, this does assume that the effect of cognitive load is evenly distributed. As a result, the GALW-CEFR score link is likely to be inaccurate for learners with unusually high or low capacity for managing cognitive load. As has already been made clear, the GALW is not intended as a direct measure of the CEFR, and the predictive CEFR level produced by the GALW is intended to facilitate progression, rather than act as any kind or summative measure. This will be made clear to teachers using the assessment to avoid confusion.

Use of English as a reference language

Another danger of the GALW format is the use of English as the primary reference language for assessment i.e. all the assessment tasks refer to the learners' English language knowledge as a means to ascertain their Welsh language knowledge. This perpetuates the use of English as the basis for production, rather than conceptual use of Welsh (Thornbury 2002, Dodigovic et al 2017). However, given that the GALW aims to cater for younger and less competent Welsh learners (see chapter 4), it is important to balance the benefit of promoting conceptual use of Welsh, with the risks of score distortion from increased cognitive load (Nation 2001, p.351),

increased anxiety, poor alignment with teaching approach (Hanif 2020), and increased difficulty (Field 2013) leading to a higher likelihood of guessing (see chapter 7.6)

Self-Administered Format

The GALW adopts a self-administered format. Once learners are familiarised with the functionality of the assessment provision, they are able to interact with the assessment independently of staff input or monitoring. Support mechanisms will be integrated into the pre-assessment rubric and assessment structure to help support this independent use, including a help tab, item text prompts, and an instructional video guide for users to reference. The self-administered format was selected due to several inherent advantages: practicality/usability, objectivity, and greater learner autonomy.

Self-administered assessments have a significant advantage in pedagogical practicality (Milton and Alexiou 2020), they allow assessment to occur alongside other activities and without the need for direct teacher input. This is particularly true of formative assessment tools that are most effective in generating useful data when used regularly by learners. As GALW is intended to serve such formative purposes and as it's use is elective rather than mandated (as with many standardised tests), ensuring that the assessment is practical and easy to use is a high priority.

An advantage of self-administration is the removal of 'examiner effects' (Tsoy et al. 2021). This can improve the consistency of scores by removing potential subjectivity or bias. Examiner effects can be particularly pronounced in classroom settings, where the teacher will have established relationships with learners that may influence scores in teacher-assessed tasks (Fleckenstein et al. 2018; Murphy & Wyness 2020).

The format also gives learners greater autonomy, allowing them to track their own progress and make informed decisions about their learning. This can not only improve learning outcomes but also enhance learner motivation (Nicol & Milligan 2006). This is further supported by the automated feedback generated by the assessment (see section 5.7).

Of course, self-administration also poses some challenges and disadvantages. Independent tasks such as the GALW can result in higher variance in levels of engagement, leading to score distortion (Saunders & Kulchitsky 2021). Such distortion can also result from a lack of technical competence to interact with the digital provision (Ercikan et al. 2018), or a misinterpretation of rubric instructions/item content (Dubins et al. 2016, p.601), all of which may have been mitigated in teacher-guided activities. To address these challenges, the teacher rubric will encourage teachers to introduce the provision to the whole class and demonstrate its functionality, to allow learners using the GALW to approach them and ask for support, and to encourage engagement through the use of existing classroom systems (see chapter 5.6).

Semi-covert assessment

Covert assessment refers to the practice of evaluating learners' knowledge, skills, or behaviours in contexts where they are unaware that assessment is taking place. This approach is often integrated into informal learning activities or everyday classroom interactions (Black & Wiliam 2009). One key advantage of covert assessment is that it can reduce performance anxiety and allow learners to demonstrate more authentic understanding, free from the pressure of formal testing environments (Torrance 2012). However, covert assessment also has limitations, including potential concerns about transparency and fairness, as learners may not have a clear opportunity to prepare or demonstrate their knowledge intentionally (Harlen 2007), and impacts upon learner motivation and effort (Harlen & James 1997).

The GALW seeks to operationalise some of the advantages of covert assessment, whilst mitigating the potential negative features, by adopting a semi-covert format. The GALW is presented as a 'quiz' rather than a 'test', and the informal qualitative nature of the feedback (see chapter 5.7) is designed to present users with the (accurate) impression that the GALW is part of the learning provision, rather than an assessment tool. Rubrics encourage this, suggesting that the assessment is not presented as a high-stakes or formal test, but rather as a way for learners to see how much progress they have made and work out what to focus on next. This impression is further supported by the repeated and self-administrated format of the GALW, which contrasts with the typical assessment environment (e.g. summative, strict, separate from other learning activities, monitored, timed).

However, some elements of overt assessment formats are maintained in the GALW. Users will be made aware through the assessment rubric that part of the GALWs purpose is to help teachers know what users need to learn. They will be informed that their teacher will be able to track their progress over time as a way of helping ensure they are progressing.

It is hoped that the GALW's largely informal semi-covert format, that avoids overtly misleading users, will contribute to engagement that is regular, authentic, and anxiety-free.

Selected content format

An extremely important feature of any assessment is content alignment, i.e. the extent to which the assessment content corresponds to the content and formats the learner has encountered in lessons and through learning provision. Failing to ensure content alignment can result in distortions in scores and damage to learner self-efficacy and motivation (Biggs 1996). From a research perspective, it also becomes challenging to discriminate whether assessment scores are reflective of coverage or learning and retention (Glaser et al. 2001).

Content alignment is particularly challenging within the context of the CfW, which encourages schools to develop their own curricula to meet the needs of its learners. This results in a broad range of different approaches to Welsh language teaching and learning provision, and inconsistency in which language patterns are drawn down from the WLC content (if such sources are referenced at all).

To ensure content alignment, the GALW uses a selected content format, where teachers customise the assessment depending on the coverage within their class. This targeting of content to items that have been covered ensures that learning and retention are assessed, rather than coverage (Milton and Alexiou 2020), and ensures that learner self-efficacy/motivation is protected from the potential effects of mis-calibrated assessment content (Poupore 2013).

Of course, such a bespoke approach to assessment content makes score comparability between classes/institutions more challenging (Berman et al. 2020). The GALW uses a predictive scoring system to address this problem, see chapter 7.5 for more details.

Generative Feedback

The GALW produces generative feedback at both learner and teacher levels. This feature is essential to the GALW's primary function as a formative assessment tool, aiming to produce diagnostic feedback that allows pupils to direct their own learning, and for teachers to adjust and augment their practice to best meet the needs of their pupils.

The advantages of automated feedback for learners include the timely-ness and consistency of the advice (Steinert et al. 2024) and the scalability of the provision (Messer et al. 2024), with large numbers of learners able to receive bespoke individual feedback simultaneously. Of course, generative feedback lacks the nuanced understanding of individual learners that is possessed by teaching staff. However, in the case of the GALW, the generative feedback does not seek to replace teacher feedback, but to augment it.

This augmentation is facilitated through the teacher dashboard of the GALW, where the teacher can find a summary of individual learner performance, and whole class performance. Both contemporary and longitudinal data is provided allowing the monitoring of performance and trajectory. This information allows teachers to supplement their lesson observation and learner interactions, to better understand the needs of both individual learners and their whole class.

More detailed information about the feedback content and format can be found in chapter 5.7.

Audio features: speech rate, accent, and repetition

When assessing comprehension skills in automated assessment, speech rate, accent and access to repetition are a contributing factor to item difficulty (Brindley & Slayter 2002) and must be considered in relation to assessment accuracy. Although the GALW includes sound files largely as a provision for accessibility, it is important that sound file users should not be disadvantaged compared to text-prompt users.

With regards to speech rate, the GALW aligns with research by Chiu and Chen (2023) that recommends audio files not to exceed 98 words per minute to ensure optimum user comprehensibility. This includes both text-to-speech features and pre-recorded item files.

Accent can also have a substantial impact on audio comprehension (Major et al. 2005). In addition to the dialectical accommodations for regional differences in WLC content, the GALW will also ensure that regionally appropriate pronunciation is used in the corresponding audio files to minimize a distortion of user performance as a result of accent divergence. Coupled with the reduced speech rate, this should mitigate any negative impact on user scores.

Finally, repetition functionality has been found to mitigate for the variance caused by speech rate, accent and the challenges posed by the 'online' nature of listening tasks (Field 2013). The GALW allows learners to listen to the audio file or text-to-speech options an unlimited number of times during the assessment. However, this does have implications for communicative authenticity: a single listening opportunity would more accurately reflect real-life interactions (Taylor 2013, p.26). However, real-life interactions are rarely context-free; significant information is contained in situational/relational context as well as established conversational themes (Sperber & Wilson 1986). Additionally, interactive dialogue often offers opportunities for repetition through explicit request, clarification, reformulation, or other communicative repair strategies (Chiang & Mi 2011; Fotovatnia & Dorri 2013). Given these considerations, and the GALW's target construct of lexical knowledge, rather than communicative competence, the inclusion of repetition functionality is warranted.

Item complexity

Field (2013) highlights a potential challenge of MCQ assessments: that of cognitive load on participants. Unlike with productive formats where learners only require one proposition on meaning, MCQs require users to process multiple options for selection. Even when these are presented in written format, it requires the assessment user to hold multiple options in their working memory in order to make a comparative analysis and select the most appropriate

response. Accordingly, Gierl et al. (2017) recommend the limitation of distractors in both length and complexity. Field (2013) also raises the ease of which disconfirmation can be discerned as a way of mitigating cognitive load. However, if distractors are not sufficiently plausible as to become 'non-functional' (Tarrant et al. 2009) this can compromise the assessment's ability to discriminate between levels of content knowledge.

To address these concerns, the GALW will carry out analysis of item difficulty/discrimination. In addition, as the GALW is targeted at younger and lower competency learners of Welsh, it is not inappropriate for the assessment items to focus on short one/two clause patterns, thereby reducing cognitive load. Such phrase-focused content is in alignment with the WLC content up to progression step 3 in the Consortia compile lists. More information on the design of distractors can be found in section 5.5.

Format overview

The flow chart below (Fig.5) shows the progression of a user's interaction with the GALW. Users will be able to access the assessment through their profile page, generated from the teacher dashboard. When accessing the assessment, the user is presented with their landing page, which includes a child-friendly assessment rubric in the form of an instructional video.

On clicking the 'start quiz' button, they will move into section 1: a series of 10 questions requiring them to translate from English to Welsh. Each item consists of an English phrase and 4 Welsh options, one of which is the accurate translation. Learners must select the correct option and ignore the distractors. Learners can not progress without selecting an option, this ensures that guessing as a variable is more evenly distributed, decreasing bias from psychological features such as risk taking and confidence. This section aims to operationalise the target construct of 'recall knowledge' (see section 3.1 for construct details).

Section 2 mirrors the section 1 content but reverses the translation direction, requiring learners to recognise the English option that corresponds to the Welsh item prompt. Both sections 1 and 2 draw randomly from the teacher-selected content 'bins'. Section 2 aims to operationalise the target construct of 'recognition knowledge'.

Section 3 is a part of the branched element of the assessment, only being displayed if learners answer a sufficient number of items correctly in sections 1 and 2. It draws upon these correct responses and seeks to establish the learners' parsing/chunking competence. Items are displayed with the prompt in Welsh, followed by the instruction to identify a particular chunk of text based on a meaning provided in English. This is used to estimate the learners' recombinant competency linked to the generativity construct.

The user then finishes the assessment on a feedback page, which provides qualitative performance feedback and recommendations for how users can advance their learning. More details of the format and content of the learner feedback can be found in chapter 5.7.

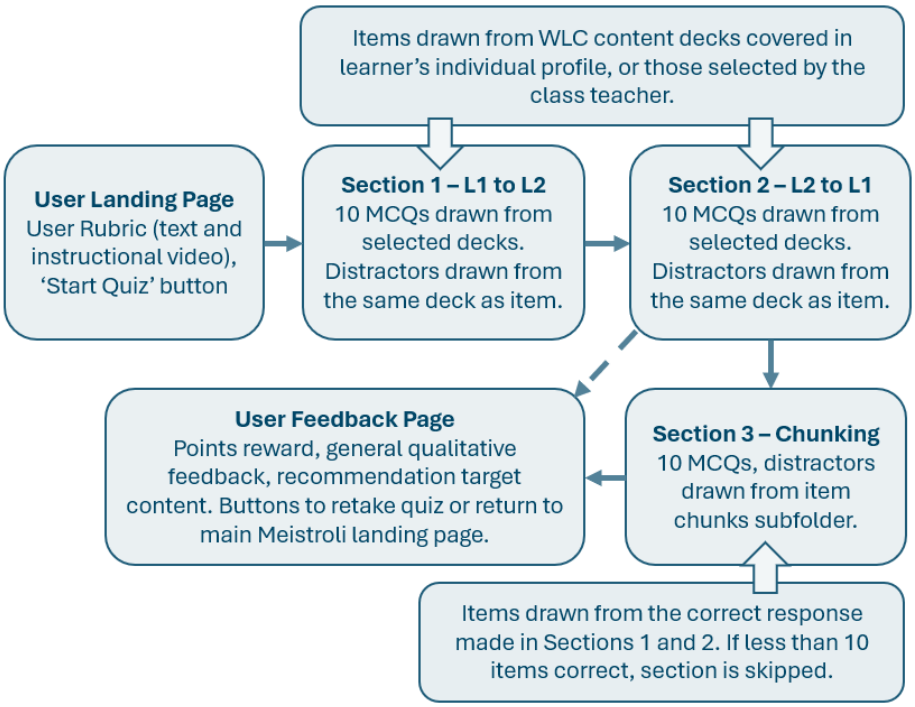

Fig.5: Flow chart showing user progression through the GALW

5.3 User interface

The user interface will be divided into two sections: the teacher dashboard, which gives teachers control of selecting content and access to the performance summaries generated by the GALW; and the learner dashboard that provides users with information about the assessment, access to the assessment and post assessment feedback. The learner dashboard is linked to the user's individual profile which is managed from the teacher dashboard.

The teacher dashboard

The landing page of the teacher dashboard (Fig.6) includes links to all the key functionality of the assessment and a 'quick-look' summary of class performance over the last 12 weeks. This allows teachers to appraise their class-progress regularly without navigating extensively.

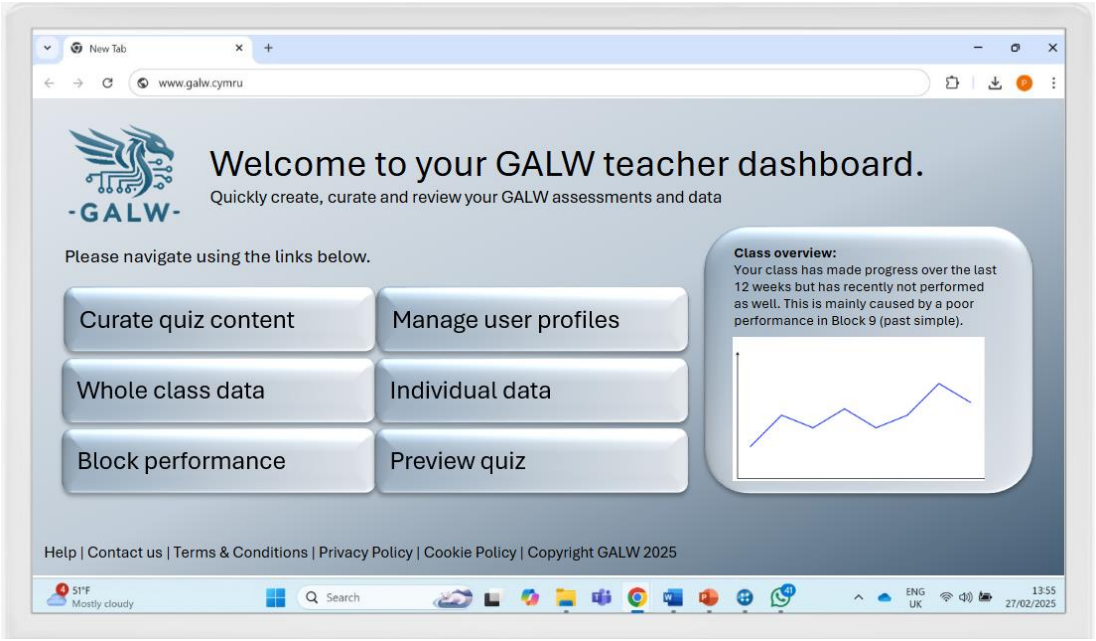

Fig.6: Mock-up of GALW teacher dashboard landing page

From this page, teachers will be able to:

- Curate which language blocks from their WLC content are included in the assessment.
- Manage individual learner profiles and download pdf summaries of individual learner performance (Fig.7).
- View statistical analysis of whole class performance to monitor for progression and content that requires remediation and identify learners who may need support.
- View individual learner data to identify learners that require additional support or challenge.
- View class performance in each of the selected content blocks and preview the learner experience of the assessment.
- Preview how the quiz will appear to their learners.

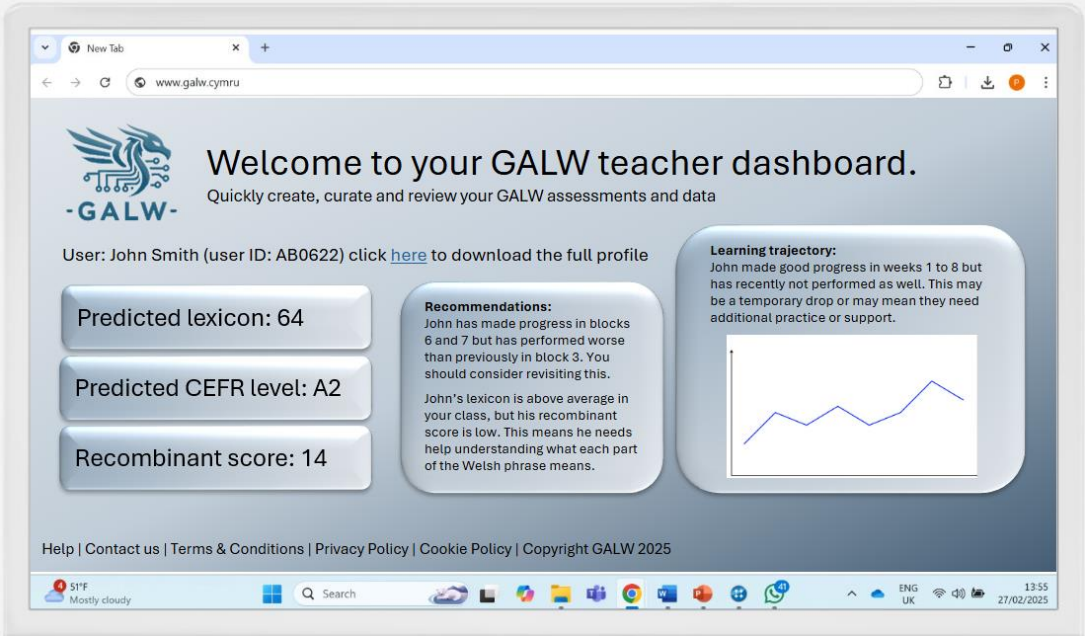

Fig.7: Mock-up of GALW individual user profile

There will also be a footer navigation bar that is displayed on all pages, allowing users to find support or information from anywhere on the site. The help section will include a video tutorial of how to use the GALW teacher dashboard, FAQ section and links to contact administrators and report errors.

Learner dashboard

The learner dashboard adopts a different aesthetic to the teacher dashboard. Lighter colours, bold simple text in accessible fonts, emoji icons to help improve accessibility and navigation, and a more informal tone. The landing page will include an embedded, child-friendly instructional video guide to using the functions on the page and taking the assessment. A simple three option menu allows learners to start a quiz, review the feedback from their last quiz, or review their overall progress (Fig.8).

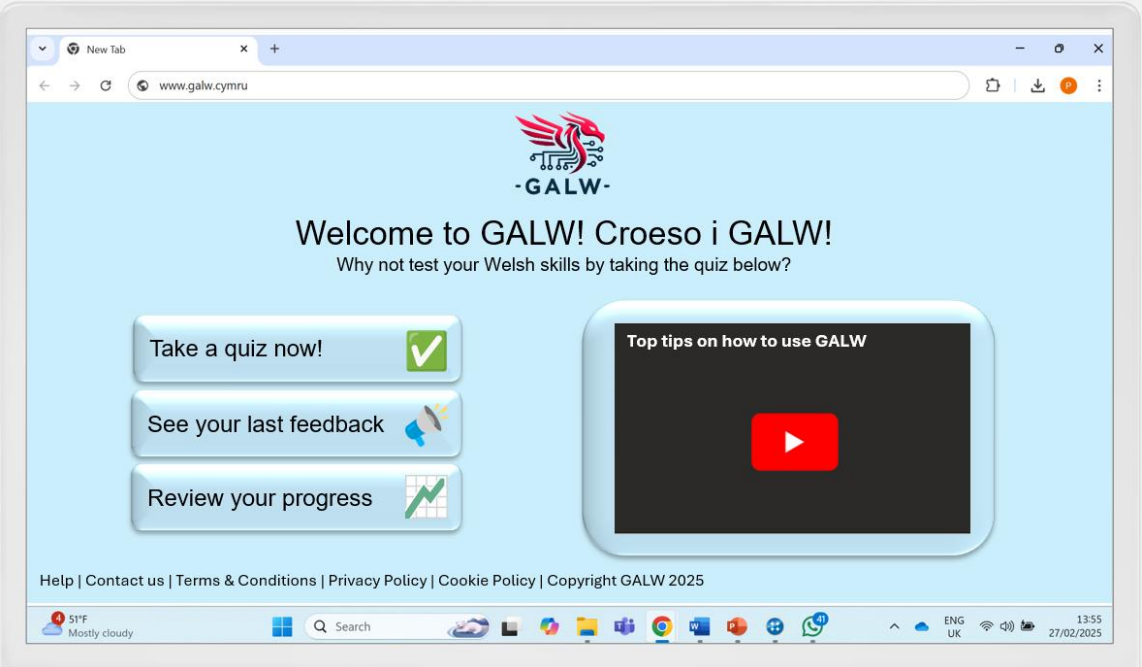

Fig.8: Mock-up of the GALW learner dashboard landing page

The quiz section of the learner experience adopts a consistent aesthetic (Fig.9). With the instructions in English using simplified vocabulary to avoid confusion that may distort performance (e.g. saying 'How do you say this in Welsh?' rather than the more complex 'Translate this from English to Welsh'). Sound files are available for all text items to support accessibility and account for phonological knowledge that exceeds grapheme knowledge. The background colour is selected to optimise accessibility for dyslexic and ADHD learners (see chapter 4).

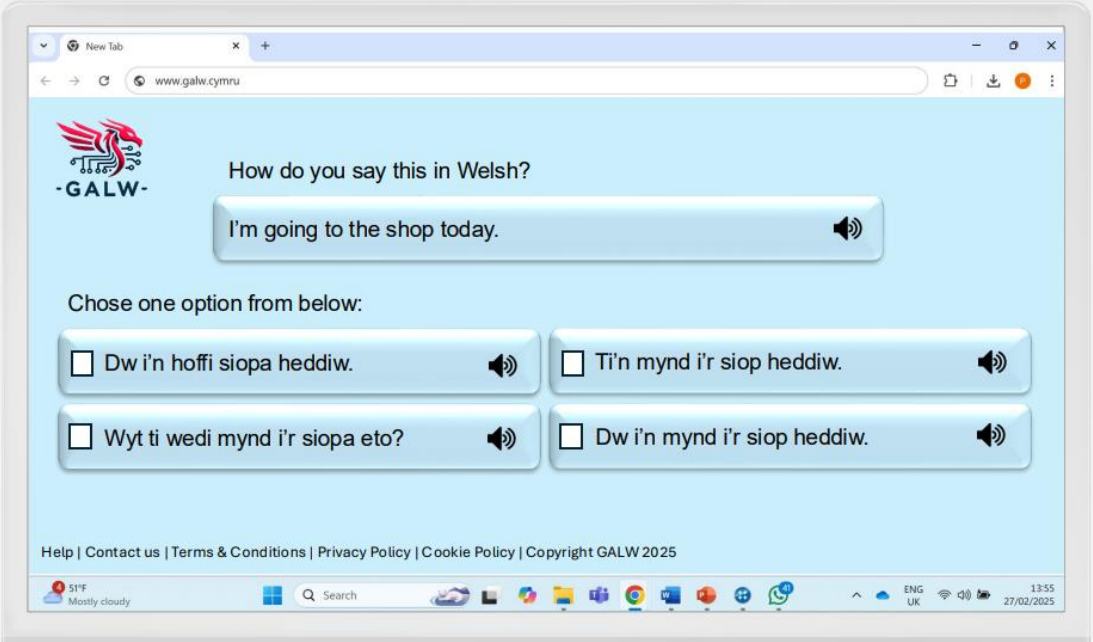

Fig.9: Mock-up of the GALW learner quiz question from Section 1 (L1 to L2 Translation)

User feedback will be displayed at the end of the assessment, along with options to return to the landing page, retake the assessment, and view an overview of their recent progress (Fig.10). The feedback is generated automatically by analysing the user performance in each content block and comparing this to their previous assessment data. The feedback will be in three sections: 1, highlighting a content block in which they have performed best; 2, comparing their performance to their previous results; 3, making a specific recommendation on what content to work on before taking the assessment again.

Fig.10: Mock-up of the GALW learner post-quiz feedback

5.4 Item selection

Items will initially be drawn from the entirety of each WLC content guide as produced by each educational consortium. This baseline item database will then be subjected to a series of reviews to ensure that the content included in the GALW is optimised for representation of the target constructs. This item review process consists of 4 phases:

1. An initial review will first identify items unique to each consortium and those shared by multiple consortia, to create a database of items unique to each organisation.
2. These databases will then be reviewed to identify excessive item repetition that could lead to pattern overrepresentation or redundancy, with items being removed in these cases.
3. Each database of items will then be reviewed for cognate/false-cognates. Where their removal does not affect the utility of the item, cognates will be removed and replaced with equivalent vocabulary from the same content block. Where cognates can not be removed without compromising content alignment, distractors will be modified to mitigate their influence.
4. Finally, items will be reviewed for any socio-cultural specificity that may disadvantage learners from a particular subgroup. Where possible items will be recreated with equivalent, but culturally neutral vocabulary from the same content block. Where this is not possible, items will be removed on condition that this does not negatively impact on content alignment. Where it is deemed necessary to retain items that do not reflect cultural neutrality, monitoring will be used to ensure that item bias does not emerge.

Full details of how items are managed within the GALW can be found in chapter 6.

This initial process of item selection will be followed by the on-going process of item development. Statistical analysis will be used to monitor and modify items to ensure they are performing effectively to demonstrate learner proficiency in the target constructs. You can find details of this process in chapter 7.5.

5.5 Distractor design

Distractors will initially be designed using patterns and vocabulary from within the content block. The emphasis in designing distractors is placed primarily on pattern recognition, and secondarily on vocabulary recognition, reflecting the pattern-based structure of the WLC content. Where possible, for each item the distractors will include 3 of the following forms selected at random:

- Assuming the correct response is a question, the statement form of the same sentence will be included as a distractor, or vice versa if the correct response is a statement. E.g. Correct response = Wyt ti'n mynd i'r siop? Distractor = **Ti'n** mynd i'r siop.
- A distractor will be created by maintaining the subject and object of the sentence, but changing the tense. E.g. Correct response = Dw i'n mynd i'r siop. Distractor = **Es i** i'r siop.
- A distractor will be created by maintaining the tense and object of the sentence but changing the subject pronoun. E.g. Correct response = Mae fe'n mynd i'r siop. Distractor = Mae **hi'n** mynd i'r siop.
- A distractor will be created by changing the object of the sentence. E.g. Correct response = Dw i'n mynd i'r siop. Distractor = Dw i'n mynd i'r **gwesty**.
- A distractor will be created by changing the sentential polarity of the sentence, i.e. whether it is positive of negative. E.g. Correct response = Dw i'n mynd i'r siop. Distractor = **Dw i ddim** yn mynd i'r siop.
- A distractor will be created by changing the verb or adjective within the sentence. E.g. Dw i'n mynd i'r siop yn gyflym. Distractor = Dw i'n mynd i'r siop yn **araf**.
- A distractor will be created by omitting a critical chunk from the sentence. E.g. Correct response = Dw i'n mynd i'r siop heddiw. Distractor: Dw i'n siop heddiw
- A distractor will be created by including the incorrect collocation. E.g. Correct response = Dw i'n mynd i'r siop heddiw. Distractor = Dw i'n mynd **ar** y siop heddiw.
- A distractor will be created by using incorrect syntax. E.g. Correct response = Dw i'n mynd i'r siop heddiw. Distractor = Dw i'n siop i'r mynd heddiw.

A bank of distractors will be developed for each item in the database. When an item is presented, three random distractors will be selected to accompany the correct option, thereby mitigating the risk of distortion from the practice effect (Duff et al. 2012).

As data is returned from each rollout of the GALW, distractors will be monitored for efficiency and their effect on item difficulty to ensure they do not negatively impact on the item's capacity for discrimination. Full details of the distractor development can be found in chapter 6.2.

5.6 Test Delivery – Instructions and Rubrics

Rubrics in this context refer to the guidance documents issued to staff (administrators) and learners (assessment users) using the assessment. They describe the correct procedures for the delivery and use of the assessment. These instructions include factors such as time restrictions, environmental considerations, support and scaffolding, mitigation of cheating, the desirability of guessing, and appropriate use of the results. Adherence to the assessment rubrics improves reliability and reduces bias from variation in these factors. From the learners' perspective, access to clear rubrics clarifies the assessment expectations, improving performance and mitigating test-anxiety (Weir 2005).

Full copies of the teacher and learner rubrics can be found in the appendix (item 2 & 3). In this section we will consider the theoretical basis for the structure and content of the rubrics and how these impact upon the efficacy of the GALW.

Elliot (2013) identifies four key features that are characteristic of an effective assessment rubric. They must be:

- Thorough, i.e. include all the information required for the assessment to be completed as expected.
- Concise, i.e. sufficiently short so as not to add significantly to the cognitive effort required to complete the assessment.
- Accessible, i.e. include differentiated language suitable for the age and ability of the user.
- Clear, i.e. explicit and unambiguous to avoid confusion.

An important element of the clarity is explicit guidance as to whether guessing is acceptable or even desirable (Read 2019). As explored in chapter 7.6, guessing can introduce a significant amount of distortion into the assessment scores, and clear guidance mitigates this by creating greater consistency.

It is with these four characteristics in mind that the rubrics for GALW were constructed. The beta trial of the provision will specifically seek feedback on the rubric using these factors as metrics of success. Feedback from the trial will be considered and any adaptations required will be made before the rollout of version 1.

With regards to rubric content Bachman and Palmer (1996, p.49) offer a comprehensive list that was used as a reference for the GALW rubrics:

- Characteristics of the setting (physical, participants, time of task)
- Instructions (language, channel, procedures)
- Structure (number, sequence, and importance of tasks)
- Time allotment
- Scoring method

Each of these content blocks is explored below with reference to the instructional rubric type (teacher/user).

Characteristics of the setting

This refers to not only the features of the location, but also those of the participants. Guidance on this factor includes how teachers should seek to accommodate both transitory and permanent physical characteristics (see chapter 4). Given the intention for GALW to be used at regular intervals, it is acceptable for teachers to delay the assessment of those with transitory features. However, for permanent features that may impact upon user performance, teachers should ensure suitable support is provided to ensure equitable access, in line with their responsibilities in the Welsh Government (2021a) ALN and Education Tribunal Act. The teacher rubric also recommends that EAL learners should not be asked to use the GALW until they have reached sufficient competence in the English language skills, as the assessment's reliance on English as a reference language is likely to distort results (Burgoyne et al. 2009).

In terms of the space used for delivering the GALW, the teacher rubric details how suitable provision should be made for the completion of the assessment. This includes mitigating the detrimental effects of excessive temperature and noise levels (Realyvásquez-Vargas et al.

2020), and overcrowding or locating the assessment spaces near distracting activities (Gilavand 2016).

Finally, in terms of timing or scheduling of the GALW, although there is some evidence that time of day can have an impact on academic performance (Smith 2013), too strict a prescription on when teachers can deploy the GALW is likely to lead to reduced usage, increasing other forms of potential bias (e.g. affective and physical transitive features). Teachers are encouraged however to avoid utilising time periods immediately abutting break periods or home-time, as this could encourage learners to rush and increase their propensity to guess (Wise 2017).

Instructions

Instructions are an essential part of the overall assessment rubric, that seeks to control how the tasks are performed to minimise bias and distortion and maximise learner performance (Bachman and Palmer 1996). Instructions fall into two categories, user instructions, and teacher instructions.

The primary purpose of learner instructions is to mitigate the error score (see chapter 7.5) caused by non-construct specific variables (e.g. guessing, effort, concentration, misconceptions). Although much variation may be mitigated by the familiarity of the MCQ format to learners (Lakin 2014), clear instruction still offers the opportunity to limit variation not associated with the target construct (Nation 2009). Glušac & Milić (2022) identify five key factors for consideration in the design of assessment instructions: length, language, type of sentence, informativeness, component parts, medium of communication.

- Length – there is a consensus that instructions should be as concise as possible whilst not impeding comprehension (Glaser & Silver 1994; Luoma 2009; Glušac & Milić 2022). All items and instructions in the GALW are designed on this principle, and are reviewed for compressibility in beta trialling, and for item-difficulty in each new iteration of the assessment (see chapter 5.1)
- Language – The best language for the delivery of assessment is not universally agreed, with some advocating for assessment through the L1 (Cox et al. 2019) whilst others raise concerns that this could result in negative washback on pedagogical use of the L2 (Rahman et al. 2021). Purpura (2004) and Heaton (1988) emphasise the importance of instruction comprehensibility in minimising error score. Given the prospective user level (most pre-A1 to A1) and the learning context (English medium) the GALW adopts English as the reference language to maximise user comprehension. In addition to language selection, the complexity and technicality of the language used is also considered, to ensure suitability for the target population (Bachman & Palmer 1996).
- Type of sentence – To aid user-comprehension, Bachman and Palmer (1996) advocate for instruction sentences to be kept simple, even when a longer series of instructions is required.
- Informativeness – Instructions need to be sufficiently detailed to ensure users carry out the tasks in a consistent manner (Cohen 1994; Weigle 2009), however such detail needs to be balanced against the previous mention emphasis on being concise and simplicity. Use of examples can be a valuable tool in communicating information clearly and concisely (Cohen 1994). Efficiency and appropriateness of language is therefore an important factor considered in the instruction design of the GALW.
- Medium of communication – The medium of communication can be an important element in instruction design. For example, the discrete use of verbal instructions

results in only attentive students with effectively short-term memory skills benefitting (Heaton 1988). Where possible instructions should be provided in written, visual and audio mediums to help ensure broader user comprehension (Wei 2024). The GALW rubrics include written instructions, and corresponding audio-visual presentation that aims to maximalise user comprehension.

Structure

The structure of the assessment and the user-awareness of this structure is an important element of rubric content. Such understanding can lead to enhanced/sustained effort, and task fidelity (Bachman & Palmer 1996). Learners should be made aware of the length of the assessment i.e. number of tasks (Pools & Monseur 2021), and the purpose of each task (Galaczi & Ffrench p.125).

The GALW rubric for learners includes a simple guide to the length of the assessment, along with an explanation of the tracking bar that will show the user their progress during the assessment itself. The rubric also includes a simplified explanation of the purpose of each task and how they help the user by informing the post-quiz feedback.

The explanation of the GALW structure is careful to not present the GALW as a formal or high-stakes summative assessment (see chapter 5.2) in order to mitigate the risk of test anxiety.

Length, Timing and Duration

As with assessment length, time duration can be a significant factor in sustaining user motivation and effort consistently throughout the constituent tasks (Tobin & Grondin 2012). Learners also benefit from understanding that they are not under a time restriction, this decreases test anxiety (Boaler 2014) and propensity to blind-guess in response to time-pressure (Dror et al. 1999). Accordingly, users of the GALW will be advised the approximate test length (informed by the beta-trial) but also assured that the test is not time-limited.

The teacher rubric will include more pragmatic guidance about the range of time learners may take in completing the tasks and how to manage this effectively. This includes scheduling use of the GALW so that it does not fall too close to break/home times, adopting a flexible approach, allowing users to engage with the assessment around other class-room tasks, and instruction to avoid splitting user sessions, i.e. allowing them to go to break during the assessment (except where this is done intentionally in aid of accessibility).

Whilst these factors motivate assessment design that limits assessment duration, it is important to balance these factors with the need to collect sufficient data during each interaction, and across multiple data collection events, to ensure that data accurately reflects the learners construct capability (Tomasello and Stahl 2004). The more regularly learners engage with the assessment, the better the quality of the generative output will be. Teachers will be encouraged to integrate the GALW into their regular provision and use the feedback formatively throughout the year, rather than treat the GALW as a summative test.

Interpreting scores and feedback

Rubrics can also include instructions on the use and interpretation of scores and feedback. Guidance on these elements can improve the learners' ability to interpret feedback and so direct their own learning, and teacher's ability to interpret the statistical information generated by the GALW so as to better direct their teaching, support and provision.

The user rubrics and presentation will give guidance on how to understand the feedback provided, and give a brief example of what users can expect. This will be supplemented by the teacher rubric, which will recommend that teachers monitor individual feedback and recommend remedial activity based on the deficits identified.

The teacher rubric will include a written and video guide to interacting with the teacher dashboard. This will include a breakdown of the auto-generated statistical feedback generated on each page and how this information can be used to direct their teaching practice. The guidance will avoid prescriptive recommendations (e.g. specific pedagogy or provision) that could be misaligned with the CfW's contextual student-centred principles but will instead focus on the inferential meaning of each data item, allowing staff to make informed decisions that account for their teaching context.

5.7 Assessment Outputs

Assessment outputs are the measurable results or data generated from an assessment, such as scores, grades, feedback, or performance indicators. Brown (2012) identifies two key types of assessment output: holistic, and analytic. Holistic outputs use a single scale to give an overall score, whilst analytic outputs consist of a number of discrete measures of isolated construct competencies. For example, Henriksen (1999) identifies different modalities through which lexical competence can be demonstrated (e.g. precision, depth, receptive/productive use), if analytic outputs are to be generated then the assessment must be able to isolate these different modalities.

In addition to this analytic aspect, the GALW assessment output also offers a longitudinal scope, allowing for the tracking of learner progress in different aspects of lexical knowledge as well as different areas of content (Laufer 2004). This approach aligns with the CfW's emphasis on assessment as a tool for facilitating ongoing development and learning, rather than as a summative output (WG 2024d). In the GALW this longitudinal aspect is integrated into the automated generative feedback produced to ensure that teachers are able to maximise their impact on learners without dedicating time to extensive data analysis.

Outputs are generated at three different levels: learner feedback, teacher data, and research data. At each level, the outputs generated aim to meet the needs of that user type, whilst presenting the data faithfully. For clarity, each of these levels of output will be explored separately below.

Learner facing feedback outputs

Learner facing formative assessment feedback should not be reductive (i.e. simply categorising performance), but should instead be supportive, timely, and specific (Shute 2008). Below each of the two learner focused goals outlined in chapter 1 are explored below with reference to these principles:

1. *To ensure foundational language knowledge is established and maintained to provide a solid foundation for the acquisition of higher order communicative skills.*

   As advocated by Harris & Brown (2018), the learner feedback outputs from the GALW are intended to help the user to direct and focus their learning more effectively. To facilitate this, learner feedback is displayed in two ways: immediate feedback on their most recent performance on the GALW, and a summary encapsulating previous performances within a set timeframe. In this way the GALW aims to provide both timely

feedback on immediate performance (Black & William 2009) and an understanding of progression over time. The immediate feedback is intended to be formative, offering qualitative and prescriptive advice, directing learners to specific content they should revise, highlighting specific skill deficits (e.g. direction dependent lexical knowledge, chunking/parsing), and drawing attention to areas in which they performed well. This formative approach seeks to assist in the maintenance of a recursive learning model, where vocabulary is revisited and extended to avoid attrition (Nation 2013), rather than the linear approach often seen currently (Russell 2025).

2. *To enhance learner motivation and self-efficacy through perception of progression and competence*.

   Schunk (1991) highlights the importance of competence perception in the development of self-efficacy, an essential aspect of successful learning (Dörnyei 2014), whilst Teng et al. (2024) observes that long-term progress perception is key to the maintenance of learner motivation. Accordingly, the GALW seeks to enhance and sustain learner self-efficacy and motivation by providing both timely and long-term performance data to learners. In addition to its formative value, the immediate formative feedback discussed above seeks to enhance the learners' sense of learning efficacy (Hattie 2008) and develop their 'L2 self' (Ushioda & Dörnyei 2009). The more longitudinal feedback seeks to convey to learners a sense of their long-term progression. An awareness of this kind of 'big picture' progress is important in sustaining learner motivation and self-regulation (Zimmerman 2013). The GALW provides learners with a visual representation of their performance over time, along with an auto generated explanation to support accessibility. Three categories of progression status (i.e. the learner showing progress / stagnation / regression) will be identified. Learners who show progress will receive praise as a means of enhancing motivation (Faulconer et al 2022). In addition to the specific performance feedback discussed, learners who show stagnation or regression will receive reassurance that periods of stagnation or regression are not unusual, that language learning is not a linear process, and that they can make progress over time. There is of course a risk that demonstrating a lack of progress can have a negative impact on learner motivation (Dörnyei & Henry 2022), however the importance of learners having genuine insights into their learning can justify the inclusion of this feature, as long as it is accompanied by constructive feedback advising the learner how to improve their performance (Cauley & McMillan 2010).

Teacher-facing feedback outputs

Teacher feedback is an essential feature of any assessment that aims to enhance pedagogy and provision, informing and empowering teachers to take more control of their practice (Fandiño 2010). The Teacher facing feedback outputs of the GALW seek to address the five teacher focused objectives outlined in chapter 1:

1. *To allow teachers to easily carry out assessment for learning: identifying areas of the Welsh Language Continuum (WLC) which require additional or remedial attention*.

   Many authors acknowledge the importance of a responsive approach to lesson content, based on learner needs, rather than a restrictive linear approach to progression

(Tomlinson 2001; Hattie 2008; Heacox 2012). Linear approaches can lead to content/skills gaps that inhibit future learning and progression by not anchoring new learning in previous learning (Ausubel et al. 1978; Schmidt & Prawat 2006). Key to such a responsive approach is the need to identify areas of strength and deficit in individual learners and groups. The GALW aims to address this by providing teachers with information about the lexical knowledge of learners both holistically (through a predictive vocabulary score) and analytically through the identification of vocabulary that has been covered but not learned/retained. This information allows teachers to plan lessons to accommodate the remediation of previous content or provide suitable individual remedial instruction where appropriate. This feedback will be generated automatically from a statistical analysis of user performance across the selected assessment content.

2. *To improve differentiation and scaffolding based on a more accurate understanding of the specific class needs*.

   Numerous authors have highlighted the benefits of effective differentiation and scaffolding on learning outcomes (Black and William 1998; Van de Pol et al. 2010; Tomlinson & Moon 2013). Adaptations to pacing, content and support require the identification of learner needs through diagnostic approaches, which can include both formal and informal assessments. The GALW aims to inform differentiation and scaffolding design by providing teachers with information concerning specific lexical knowledge deficits in their learners. This feedback will be automatically generated by statistical analysis of the learners' performance. Easily tracked scores for the different aspects of learner performance will be provided through the teacher dashboard, providing accessible information allowing them to target their pedagogy and provision more effectively. See chapter 5.7 for more information about the data outputs for teachers.

3. *To identify individual learners who may require additional support in their Welsh language learning*.

   Similarly to the data for targeting and differentiation, the GALW will allow teachers to identify learners who are not achieving anticipated levels of progress in the Welsh language lexical knowledge. This allows teachers to provide additional support to these learners, helping to ensure inclusive pedagogy, educational equity and cohort cohesion (OECD 2012). The teacher dashboard will include an option to view the predictive vocabulary scores of the whole class simultaneously, allowing teachers to quickly identify learners who are falling behind and may require additional support. The longitudinal functionality also allows teachers to inspect learner trajectories, identifying learners whose progress may have stalled, or those who have regressed in their lexical knowledge. This enables teachers to instigate preventative or remedial strategies to identify and address causes of learning stagnation.

4. *To improve learning continuity in transitional periods (primary to secondary education, inter-key stage or progression step)*.

Learning continuity is an important contributor to learning progression and the sharing of assessment information can help to facilitate consistency for learners in levelling, content, and support (Fletcher 2018). In line with the assessment goals of the CfW (see chapter 2.1) the GALW aims to support transition points in its teacher facing outputs through providing a consistent assessment practice across different institutions and transition points, and comparable/communicable descriptors of learner level through the generation of pupil and class profiles (see chapter 5.7). This allows post-transition staff to easily appraise an individual or groups lexical knowledge and use this to inform that curriculum planning, creating continuity for learners.

5. *To gain an understanding of different elements of class and learner competencies and ensure appropriate progress is being made within each area*.

   Whilst the GALW is relatively narrow in the scope of its constructs, it does include functionality that allows for exploration of specific competencies. The current version is able to identify translation directional disparities (e.g. differences in competence between receptive/productive recognition), and deficits in chunking knowledge (as indicated by the user's recombinant score) indicative of generativity. Data on performance in these different aspects of the GALW will be available through the teacher dashboard. There are adaptations that could be made to the GALW's functionality in future versions that would help identify domain specific competencies. This could include presenting text prompts only as a measure of reading ability or only providing audio prompts as a measure of aural comprehension. However, the current format is based on the GALW's intended focus on lexical knowledge in a non-domain specific context.

Research data outputs

Education research has the capacity to have profound beneficial impacts upon teaching and learning, and 'research led' education practice (Hargreaves 2011) and policy have been a prominent feature of the UK approach over recent decades. This largely consists of a 'what works' focus, with research often concerned with discerning the optimal approach that can inform practice and improve outcomes (Whitty 2006). Such research will usually adopt an experimental or quasi-experimental format borrowed from the medical/natural sciences. Whilst the degree of objectivity of the natural sciences may be an unrealistic aspiration, if the scientific method is to be applied to educational contexts, similar levels of veracity should be sought (Bailey 1999). Such veracity is dependent upon having comparable, reliable and consistent measures of learning success in order to make robust claims around the efficacy or otherwise of particular factors (pedagogy, provision, intervention, support) (Towne & Shavelson 2002; Swann 2003). The GALW seeks to provide an assessment tool that, in addition to supporting teaching and learning, generates such comparable data.

It should be noted that this practice orientated instrumentalist approach to education research has been challenged (Whitty 2006), and this is considered in more detail in chapter 8.

The way in which the data outputs from the GALW meet the three research goals in chapter 1 are considered below:

1. *To facilitate the collection of comparable data on learner progress in developing WLC knowledge. This data can then be tracked against other factors to determine the impact*

of variations in pedagogic approaches, learning provision, socio-economic factors, and individual learning strategies.

The GALW is designed to be used in schools across Wales. It is hoped that its functionality as an AfL tool will lead to widespread adoption and use. This extensive usage will generate larger and more useful data sets for research purposes. Comparability is a challenge as standardization is both impractical and contradictory to the principles of the CfW (see chapter 3.2). However, the GALW's predictive approach to measuring vocabulary size does allow for the comparison of learning outcomes across institutions, accommodating variation in coverage, content and retention. This aggregated data is easily collated for research purposes and provides an insight into the Welsh language development of the English medium sector, as well as potential regional variation. An anonymized summary of the data will be available to stakeholder organisations, though institution level data will not be released in line with the ethical usage policy (chapter 8.4).

In addition to this isolated data, the GALW assigns a ULI to each user. With the user's and institution's permission, this identifier can be used by researchers to link GALW scores to other data, including survey responses, demographic information, other educational performance metrics, or intervention status. Such data linking allows researchers to carry out descriptive and inferential statistical analysis, finding associations/correlations that can be used to speculate about causation. Teachers will be able to download an anonymized breakdown of the user performance data associated with their class and control who has access to this data.

2. *To allow for the tracking of different cohorts and populations over time to understand early Welsh language development in the EM sector and how this aligns with the goals of Cymraeg 2050 and the Welsh Language and Education Bill (B2 on the CEFR).*

   In addition to the cross-sectional data already discussed, the GALW is capable of collating longitudinal data that can be used to explore trends in user performance over time. The GALW is able to produce mean predicted vocabulary scores across all users and generate learning trend data. This can be modified to isolate a particular cohort or region to enhance the utility of the data. Such trend data will be valuable in tracking progress within EM primary education towards the goals of the WLEB and Cymraeg 2050 (see chapter 2).

3. *To provide a tool that empowers teachers to conduct their own action research. Such research can then inform individual practice, the practice of colleagues, and institution/cluster policy or approach to Welsh language teaching.*

   Although the CfW encourages teachers to carry out action research to enhance their practice, at present any inquiry into Welsh language development is likely to necessitate the development of a bespoke assessment tool. This not only adds to the workload of the teacher but also poses a technical challenge for which they are unlikely to have received specific training. The GALW offers teachers an easily accessible assessment tool that can be deployed in inquiries exploring the efficacy of their practice or provision in the development of Welsh content knowledge.

6. Content Specifications

The selection and curation of assessment content are pivotal in shaping effective assessment, as it ensures alignment between instructional objectives, lesson content, and evaluative measures (Martone & Sireci 2009). This is particularly pertinent to AfL tools which aim to act as a conduit between teaching and learning, providing feedback that informs both teachers and students (William 2013). When assessment content is effectively curated to reflect the school's curriculum goals and content, it creates a reciprocity between learning, teaching, and assessment. Conversely, poorly aligned assessments may lead to misrepresentations of student abilities, hinder the learning process and undermine learner confidence and motivation (Harlen et al. 2002).

6.1 Alignment with WLC content

The GALW content is drawn directly from the consortia developed WLC content guides (examples can be found in Appendix Item 1). The WLC content is laid out slightly differently by each consortium, but follows the same broad patterns: Content is arranged in order of lexical and grammatical complexity, e.g. early patterns consisting of imperatives that learners are likely to encounter and single clause simple present tense sentences. Content gradually increases in complexity as learners progress through the continuum, incorporating a greater variety of tenses, communicative functions, grammatical structures, and recasting. This progression is often subdivided into progression steps, though this is aspirational rather than prescriptive with the CfW stating:

> 'Progression steps are broadly related to age. However, learning is not a linear process and progress in language learning will not be the same for all learners. What is taught and the resources used will need to reflect both the age of the learners and where they have reached in their Welsh learning.' (WG 2024e)

Some consortia also have vocabulary lists that accompany the pattern continuum (CSC communications 2024). These lists are arranged by topic rather than complexity (e.g. animals, time, transport), and are designed to support thematic approaches to curriculum design. However, the elective nature of their use, inconsistency of availability (some consortia produce no vocabulary lists), and their disconnection from the continuum make them unsuitable for inclusion in the GALW.

Structurally the WLC content patterns are relatively stable, being largely composed of sentence stems, with a variety of interchangeable elements to increase communicative utility/flexibility. These are used by schools to design their own individual school curricula, in line with the guidance in the CfW (WG 2024e). This means that whilst individual items are likely to be stable within an individual consortium area, the variety of patterns actually taught may vary significantly in different institutions. To accommodate this variation the GALW adopts a tailored approach to content selection, with teachers able to specify what sections of the WLC content their learners have covered prior to their use of the GALW. Items are then selected from these selected 'bins' ensure alignment of teaching and assessment. More details of this selective content format can be found in chapter 5.2.

Dialectically the WLC content for each consortium does show some variation, especially between GWE and the more southerly consortia. To accommodate these differences each

consortium area will have a distinct set of items from which teachers will select their content 'bins'. Where there is consistency, items will be shared across multiple areas, whilst items distinct to their consortium will be separated. In this way the GALW is able to align with dialectically distinct regional variations without compromising the comparability of the predictive lexical knowledge scores (see chapter 7.5).

6.2 Purpose of distractors

In the context of a multiple-choice question (MCQ) assessments such as the GALW, distractors refer to the incorrect answer choices that are provided alongside the correct option. They are not merely filler options but integral components of the item that help to discriminate between varying levels of proficiency among test-takers (Haladyna 2004). Distractors are designed to resemble plausible answers, challenging the test taker's ability to distinguish the correct response, thereby indicating their level of competence (Gierl et al. 2017). Distractors therefore consist of phrases that are related to the correct answer but differ in meaning, usage, or contextual appropriateness. In the GALW, the emphasis on lexico-syntactic form-meaning knowledge necessitates distractors that equate closely enough in form to make the selection of the correct option indicative of genuine knowledge of meaning correspondence. A more detailed discussion of distractor content design specifically for the GALW can be found in chapter 5.5, but here we will focus on the theoretical basis of distractor use.

Well designed and calibrated distractors help differentiate between students who have a deeper understanding of the material and those who may be guessing or have a limited grasp of the content. The importance of distractor design to score validity is explored in more detail in chapter 7.6.

When designing distractors for an MCQ lexical knowledge test, several key considerations must be taken into account. First, distractors should be plausible and related to the correct answer in some way. According to McMillan (2017), distractors that are too obviously incorrect or unrelated to the target concept can make the test too easy, reducing its effectiveness. Distractors should therefore have some commonality in both elements of meaning and levels of complexity (Belgar 2012). To ensure this, distractors should be selected based on their relevance to the target lexical item and their potential to mislead learners who have incomplete or imprecise knowledge. For example, distractors might include words that are semantically similar but not synonymous, words that are morphologically similar, or words that share syntactic features with the correct answer.

Moreover, the distractors should reflect common misconceptions or errors that learners might make, allowing the test to identify gaps in knowledge or areas of confusion. The inclusion of distractors based on common learner errors can enhance the diagnostic value of the assessment. Additionally, the number of distractors must be considered carefully. While the traditional MCQ format uses four options (one correct answer and three distractors), research suggests that the optimal number of distractors may vary depending on the context (Weir, 2005). Too few distractors can make the test too easy, while too many can introduce unnecessary complexity and confusion. Finally, distractors should be balanced in terms of difficulty, so as not to skew the assessment results e.g. by enabling guessing.

Several challenges can arise in the design and implementation of distractors in lexical MCQ assessments. One common problem is the inclusion of distractors that are either too obviously incorrect or irrelevant to the target concept, which makes the question too easy and less

effective at measuring lexical knowledge (Gierl et al. 2017). For instance, using words that are completely unrelated to the target word's meaning may not provide any meaningful insight into the learner's knowledge, and such distractors fail to challenge students to differentiate between nuanced lexical meanings.

Another issue arises when distractors are too similar to the correct answer, making them misleading and potentially confusing for test-takers (Haladyna 2004). When distractors are too close in meaning to the correct response, they can introduce ambiguity that may mislead students who have partial knowledge but are not fully confident in their understanding. This could lead to an increased likelihood of guessing, rather than demonstrating genuine lexical proficiency.

Additionally, distractors may unintentionally favour certain groups of learners if the words used are too familiar or unfamiliar to specific subgroups of students (Solano-Flores & Trumbull 2003). For instance, if distractors are overly influenced by regional dialects, academic jargon, or culturally specific vocabulary, learners who are unfamiliar with these terms may struggle to answer correctly, despite having adequate general lexical knowledge. As a result, it is important to ensure that the vocabulary used in distractors is appropriate and accessible to the target population of learners. This is explored further in section 4.3.

6.3 Managing Enemy Items, Cognates and False Cognates

In addition to ensuring the efficacy of distractors, there are other potential distorting factors in content selection specific to language assessments. Enemy items, cognates and false friends all pose the risk of distorting lexicon representation by facilitating or inhibiting the accuracy of guessing.

'Enemy items' are pairs of test questions that should not appear together on the test presented to a user, as the presence of one item can provide hints or answers to another, potentially compromising the test's validity (Van der Linden 2005). For example, question 1 may ask the learner to translate 'dw i'n mynd' and provide four options including the correct translation, 'I am going'. Then, in question 11 the learner may be asked to translate 'I am going' to the Welsh. Clearly, their experience of question 1 could inform their response in question 11. Identifying and managing enemy items is important in ensuring each item's independence and the overall assessment's fairness.

The risk of enemy items in GALW is managed through a stratified item bank and a simultaneous non-repeatable item draw-down. In effect, the item bank for GALW is stratified into different item 'bins' reflecting a distinct section of the WLC content (usually defined by grammatical or communicative function). The selection of items is stratified across the bins selected in the coverage functionality of the teacher dashboard, e.g. If the teacher selects a coverage of bins 1, 2, 3, and 4, the programme will draw down 10 items randomly from each bin to populate both sections 1 and 2 of the test simultaneously, with logic in place to ensure the same item cannot be drawn more than once. This ensures that the same item cannot occur in both test sections. There is still the risk of comparable items appearing, as the WLC content includes a large number of 'interchangeable' items (CSC Communications 2024) and it is possible that items drawn from the same bin may include similar structures or vocabulary. This is mitigated through item curation, with redundant items removed from the WLC content, though this must be balanced against the need to align lesson and assessment content (Martone & Sireci 2009).

Cognates and false cognates also must be considered in the way they can distort individual learner performance in the GALW. Cognates are words in different languages that share similar spelling and meaning due to a common etymological origin, such as ‘cat’ in English and ‘cath’ in Welsh. In contrast, false cognates are words that appear similar across languages but differ significantly in meaning; for example, the English word "key" and the Welsh homophone "ci" (meaning ‘dog’) sound the same but have different meanings. In language assessments, both cognates and false cognates can influence performance by providing clues or misdirection. It is important to carefully consider the presence of cognates and false cognates to ensure the validity and reliability of language assessments.

To address the issue of cognates and false cognates in the GALW, the item banks drawn from the WLC content have been reviewed, with items containing superfluous cognates or false cognates revised to include comparable (in terms of difficulty and relevance) words that appear within the same content block. This is not always possible where comparable vocabulary is unavailable, and in these instances, distractors will include comparable cognates/false cognates to mitigate the distortion caused.

6.4 Register and Style

In language testing, register refers to the level of formality and appropriateness of language used in a given context, such as academic, conversational, or technical communication, whereas style refers to the linguistic choices and manner of expression, influenced by factors such as tone, audience, and purpose. Whilst important factors in more advanced assessment requiring more open and productive output, register and style are comparatively peripheral in the GALW: the focus on lexical recognition, the prescriptive content of the WLC, along with the controlled nature of MCQs, mitigate complexities introduced by varying language styles and registers.

6.5 Welsh Specific Linguistic Features

There are some features of the Welsh language that need to be considered in the creation and curation of content for the GALW. These include syncretism, T-V distinction, aspect, and mutation. Whilst these are to some extent accommodated by the alignment of the GALW with the WLC content, it is important to be mindful of these factors in item adaptation e.g. in response to cognate presence. Below, each of these features and their potential impact on the GALW are considered separately.

Syncretism and T-V Distinction in English

Modern English does not distinguish between the singular and plural forms of ‘you’. This phenomenon, known as syncretism (Huddleston and Pullum 2002), occurs when distinct morphological forms (such as singular and plural pronouns) merge into a single form. In Welsh, these forms have remained distinct, with ‘ti/chdi’ denoting singular ‘you’, and chi denoting the plural ‘you’. This often leads English to use pronominal reinforcement (Yule 2022) such as ‘yall’, ‘you lot’, etc, which are unnecessary in Welsh. This causes problems for teaching and assessment, as the imprecision in English can lead to confusion e.g. If a child is translating ‘I will see you later’, they often transfer the lack of difference from the English and use the wrong pronoun.

Like other languages, Welsh uses this plural form to convey respect/deference when communicating with someone of perceived higher status, or someone with whom they are unfamiliar. This structure is known as ‘T-V distinction’ (Brown and Gilman 1960). This causes

similar issues in teaching and assessment, with learners usually defaulting to the one pronoun form or the other to mirror their L1 pattern (Mella and Gutiérrez 2023). This has obvious implications for the development of pragmatic competence.

In order to minimise the effect of both syncretism and T-V distinction on learner scores, the GALW will seek consistency of use with the WLC content of the user's consortium region e.g. if the question form is included in the continuum with the formal/plural 'chi', the same form will be presented in the GALW. In cases where both forms appear in the WLC content, the option will include ti/chi to ensure users are able to distinguish either option.

Aspect in Welsh

In Welsh 'aspect', i.e. the way in which the verb expresses the flow of time in relation to the action it describes (Binnick 2011), is usually contextual in the present tense, rather than explicit, as in English (Comrie 1976). That is to say, Welsh tends not to linguistically differentiate the present continuous/progressive, i.e. expressions of actions in progress (e.g. 'I am eating' – 'Dw i'n bwyta'), from the simple present, i.e. the base form of the verb expressing general states (e.g. 'I eat' – 'Dw i'n bwyta'). The aspect is usually inferred contextually or made explicit through the use of time adverbs (e.g. now). This is only an issue with the present tense, with past/future tenses having separate aspect constructions (e.g. 'I ate' – 'Bwytais I' / 'I was eating' – 'Roeddwn i bwyta'). As much of early language acquisition is in the present tense this can create a dilemma for both SLT and L2 assessment, as learners will often seek distinct patterns to differentiate aspects when producing Welsh or fail to conflate terms when aspects are presented in English.

To mitigate any confusion caused by variance in aspect, the GALW questions are to be phrased in accordance with the consortium WLC, which selects on the basis of common usage. Where both aspects are present in the WLC content, both options will be listed in the item.

Mutations in Welsh

In Welsh the initial phoneme/grapheme of some words change depending on the grammatical structure (e.g. after certain prepositions/possessive pronouns/numbers) and lexical contexts (e.g. after particular adjectives or verbs). These changes are called mutations and pose a significant challenge to learners from an Anglophone background (Ball & Müller 2002). Mutations do not tend to be explored explicitly (i.e. through the learning of prescriptive grammar rules) at primary school, with implicit knowledge being developed through teacher modelling and resources. As such, this approach will be reflected in the GALW, with mutations consistent with the WLC content included, but with no expectation of explicit knowledge included within the assessment.

7. Validity and Reliability

This chapter explores a theoretical framework of validity and how it is applied to the specific context of the GALW. We start by exploring the definition of validity and the structural framework of validity that is adopted by the GALW. The chapter then proceeds with an appraisal of each dimension of validity (cognitive, context, construct, and consequential) in isolation, and an overview of how the GALW aligns with the requirements of each aspect. Reliability will then be considered more generally within the context of the target population and learning context. Finally, the chapter will conclude with a consideration of the role of piloting and how this process is both directed by and supportive of assessment validity.

7.1 Defining Validity

Validity, in its contemporary understanding, refers to the justification of specific interpretations made from test scores. It is therefore not the test itself that is deemed valid or invalid, but the conclusions and inferences derived from the results (Kane 2006). Reliability refers to the consistency and stability of assessment results across different contexts, cohorts, or times (Nagai et al. 2020). So, whilst reliability can be seen to refer to the quality of the data collected, validity concerns the inferential legitimacy of that data (Zumbo & Rupp 2004). Field (2013) observes a tension between validity (which seeks authenticity, requiring complexity and integration), and reliability (which is achieved through simplification and isolation). A balance of sorts must therefore be found between the demands of both aspects, as well as the further core factor of practicality (Weir 2005).

In the following sections, Weir's (2005) socio-cognitive framework (SCF) for assessment validity will be used to structure the chapter content. The SCF presents a comprehensive model for establishing the validity of language assessments, considering cognitive, social, and contextual factors, whilst acknowledging interaction between assessment users' cognitive processes and the social contexts in which assessments occur.

Weir (2005) argues that validity should be considered across three dimensions: cognitive validity, context validity, and scoring validity. Cognitive validity pertains to the extent to which a test engages the mental processes required in authentic language use (Weir 2005; Shaw & Weir 2007), aligning with earlier work by Bachman and Palmer (1996). Context validity concerns the extent to which an assessment reflects authentic communicative contexts, building on Messick's (1990) argument that test validity should account for the social and situational appropriateness of test tasks. Finally, scoring validity addresses the reliability and fairness of scoring procedures and interpretations, ensuring that assessments align with the target construct and are interpreted within its limitations (Weir, 2005).

Weir (2005) presents these dimensions of validity, not as distinct and discrete, but as an interrelated system; a 'triangle of validity'. For example, Weir argues that cognitive validity is related to context validity: if a test does not faithfully replicate real-world language use, then the cognitive demands placed on test-takers may deviate from genuine language processing. Likewise, the nature of cognitive processing influences how scoring criteria are designed. This means scoring must be sensitive to the depth of processing required, reinforcing Weir's argument that tests should seek to measure meaningful cognitive engagement rather than superficial task completion.

By framing validity as an integrated and interactive system rather than a series of discrete factors, Weir's SCF requires assessment design that reflects the complexities of authentic language use, engages authentic cognitive processes, and maintains fairness and reliability in scoring. Assessment design must therefore seek an optimal balance between these different dimensions within the context of the assessment.

However, in this specification one adaptation will be made to Weir's SCF; the inclusion of a distinct appraisal of consequential validity. Consequential validity refers to the extent to which the intended and unintended consequences of a test support the appropriateness of its use and interpretation (Messick 1990). While this is not excluded from Weir's (2005) model, being

integrated into his concept of context validity, it will be considered separately in the section 7.4 below for the sake of clarity.

Finally, it must be noted that validation is not a generic process, but one that is highly situated, and specific to a given context and the purposes of the assessment, which must be made explicit for assessment validity to be legitimate (Read 2019). The validity of the GALW is therefore developed within the socio-cultural context outlined in Chapter 4 and is only legitimised within these constraints.

7.2 Cognitive Validity

Cognitive validity is the extent to which a test engages the same cognitive processes and abilities that are deployed in real-world contexts by a proficient user (Field 2013). It concerns the extent to which the assessment elicits candidates to utilize the same mental operations they would employ in authentic communicative contexts. Seeking such alignment between the test design and models of language processing improves the predictive value of the assessment of language ability in real-world situations (Akbari 2012).

The two cognitive processes associated with the GALW constructs are receptive comprehension and productive processing.

In receptive comprehension, where the item is available in both auditory and lexical forms, Kintsch's (1988) Constructive Integration model (CIM) provides a structure for conceptualising the cognitive process, consisting of: the individual's perception of the linguistic information; linking the information with existing lexical, grammatical and syntactic knowledge (form-meaning correspondence, syntactic parsing and semantic integration); establishing provisional propositions of meaning, utilising contextual inference to refine propositions; and the iterative adjustment of these propositions in the context of further linguistic information.

In the case of productive linguistic processing, Levelt's model (1989) offers a similar cognitive framework: starting with the conceptualisation of the message-level representation to create a communicative goal; the formulation of linguistic form through lexical retrieval, grammatical and phonological encoding; followed by articulation through the production of the form; and self-monitoring of output for errors and communicative efficacy.

Although Levelt's model has been challenged by more interactionalist and parrel theories (Dell 1986; MacDonald 2013; Pickering & Strijkers 2024), and others have proposed integrated theories of language production and reception processes (Pickering and Garrod 2013, Hurley 2008), the modular models proposed by Levelt and Kintsch offer a more pragmatic structure for the comparative purposes of validating cognitive assessment features. The modularization of processes, whilst not necessarily being as accurate a representation as more 'layered' or 'interwoven' models, offers the opportunity to isolate different elements for appraisal, aligning better with the need to map test performance onto well-defined cognitive constructs (Drackert 2016).

Kintsch's CIM represents a blend of both top-down and bottom-up processing, allowing for the evaluation of the assessment's ability to elicit both integration of information and extraction of meaning. Whilst such separation of top-down/bottom-up processing has been critiqued as artificial and reductive (Rauss & Pourtois 2013) it again offers a practical isolation of skills for the purpose of cognitive validation.

Field (2013) identifies three main questions that should be applied to the assessment in light of the cognitive processes identified: similarity, i.e. to what extent do the processes of the assessment correspond to those of authentic use; comprehensiveness, i.e. to what extent does the assessment elicit a comprehensive array (or limited sub-set) of cognitive processes of authentic use; and calibration, i.e. to what extent does the assessment differentiate the cognitive demands at different levels in relation to the anticipated performance level of participants.

When considering these three questions within the context of the GALW assessment format it is possible to draw correlations and divergences between the processes that the assessment elicits and the authentic language-use context. The selected response format (MCQ) of the assessment has the advantage of giving control of item design in selecting a target process (Field 2013). However, the flipside of such precision is a loss of holism and creativity: items are restricted in scope and limited in response, only allowing the demonstration of competency within a very restricted field (Dinçer et al. 2022). The MCQ format does offer some parity with elements of the CIM, with the development of provisional propositions of meaning, and utilising contextual inference to refine propositions, somewhat mirroring the confirmation/disconfirmation aspect of discriminating between options. The critical missing stage is the initial construction of the propositional meaning. Although an essential element of the CIM, from a pragmatic perspective, variance in the exact wording of proposed meaning required to operationalise this construct would prohibit conventional automated marking.

Selected response formats also diverge from authentic linguistic interaction by the presence of distractor options. Unlike in communicative contexts where propositions/output is settled on individually, in SRFs the proposition/production is held as a theory against which learners seek evidence to disconfirm the selected response. This disconfirmation process is only weakly reflected in the cognitive processes of comprehension (discourse representation) and production (self-monitoring). The differentiation of item meaning as a process of confirmation or disconfirmation in relation to the item responses available is a competence-variable independent of the target construct.

Finally, calibration of cognitive demands across different levels is included within the GALW's construct context, with lexical syntactic and grammatical complexity increasing in alignment with learner exposure level (as determined by the GALW's selective content format detailed in section 5.2). Whilst this will not be representative of the holistic progression of cognitive ability associated with language competency, it does provide appropriate scope for progression in the cognitive management of lexical knowledge.

Unfortunately, the GALW's format does not, therefore, offer optimal cognitive correspondence to authentic language use, only partially operationalising cognitive processes of the target constructs. However, it provides a compromise between cognitive authenticity and practicality, mobilising some cognitive aspects whilst allowing for pragmatic features (automated marking, self-administration, generative feedback). A fuller discussion of this balance between ideals and pragmatic decisions can be found in chapter 5.2.

7.3 Context Validity

Context validity in linguistic assessment refers to the extent to which a test reflects the real-world language use situations it aims to measure. It concerns to what extent that the assessment aligns with the target language use (TLU) domain. High levels of context validity

mean the assessment replicates authentic language demands, enabling a greater degree of inference from the assessment about a test-taker's ability to function in real-world scenarios.

One of the early proponents of context validity, Spolsky (1985), highlighted the fundamental incongruence between authenticity and the assessment environment: the test-takers' goal is performative rather than communicative and so, inherently inauthentic. Leung and Lewkowicz (2006) go so far as to argue that for this reason authenticity is largely unmeasurable in testing scenarios, a point challenged by Field (2013), who makes a case for establishing 'parameters' for authenticity. Such parameters challenge a binary notion of a test either possessing or not possessing authenticity; rather, different defined aspects of an assessment can have authenticity to different degrees. It is therefore possible for assessments to have high levels of authenticity in one aspect, but low levels in another.

Such a compartmentalised understanding is reflected in Bachman's concept of 'interactional authenticity' (1991 p.691), where authenticity concerns the engagement of skills and strategies learners engage in the assessment task, as distinct from the language content used. Assessment tasks can therefore have little resemblance to an authentic context, but still mobilise the same cognitive functions and strategies. However, although content is still distinct, in this case the distinction between contextual and cognitive validity could be seen to have broken down to the extent where its usefulness must be questioned. Addressing this, Bachman and Palmer (1996) distinguish between situational and interactional authenticity: a task could be abstract in nature whilst still eliciting the same cognitive processes as the TLU.

It is such an interactional context validity that the GALW seeks to mobilise. Whilst the MCQ format is a significant abstraction of authentic communicative contexts, its operationalisation of cognitive processes (see discussed in 7.2) and lexicon size/access allows an inferential appraisal of potential communicative capability.

These three elements of context validity (situational, interactional and content) can be applied to the proposed assessment structure in order to understand to what extent it represents authentic language use. Such a modular approach to context validity is critiqued by Widdowson (2003), who argues such isolation of discrete aspects misrepresents the dynamic and interconnected nature of language use. However, Taylor (1994) argues that learners are capable of extracting communicative value from 'inauthentic' situations. Considering the potential for such learner inference and given the purpose of the GALW in establishing content acquisition, such a modular approach can be justified.

Of course, the idea of authenticity is a situated and malleable concept itself. Leung and Lewkowicz (2006) highlight the nature of language exposure can be a factor in perceived assessment authenticity. This is particularly pertinent for the Welsh language, where many learners will be situated in communities where they are exposed to little or no authentic exposure to the language. For such learners, who see Welsh as an academic rather than community language, the assessment context may be perceived as authentic. However, given the Welsh Government's explicit goal of extending the authentic use of Welsh beyond education (WG 2017), and in keeping with the principles of the CfW (WG 2024c), it is important for concepts communicative authenticity to be provided for them.

One finally element of context validity to be considered is the cultural neutrality of assessment content. If a test includes culturally specific content (e.g. references to the NFL in an English test for international learners), some test-takers may be unfairly disadvantaged—not due to

lack of ability in the target construct, but due to lack of cultural knowledge. It is therefore important for content to be culturally neutral in order to avoid cultural bias (Djiwandono 2006). There is however an obvious tension between cultural neutrality and context validity: authentic real-world contexts are inherently culturally situated, not neutral. Weir (2005) acknowledges this challenge, not advocating for complete cultural neutrality in test design, but instead for culturally accessible test content that maintains authenticity while minimizing unfair disadvantage for test-takers from different backgrounds. This factor is largely managed within the GALW through alignment to the WLC content which is designed to be culturally appropriate for the target population. See chapter 6.1 for a more detailed discussion of how alignment is managed.

7.4 Consequential Validity

Consequential validity is the consideration of the intended and unintended consequences of assessments, including the impact the assessment may have on individuals, groups and the educational system more broadly (Messick 1996). The way assessments are designed potentially impacts on class pedagogy, resource development, prioritisation, and education policy (Belgar 2013).

More specifically within language assessment, consequential validity refers to the impact that an assessment has on individuals, educational systems, and society (Field 2013). It considers intended and unintended effects of both the context on the assessment (e.g. socio-cultural wash-forward), and the assessment on the context (e.g. educational washback). Such consequences include effects on pedagogy, learning behaviours, and access to opportunities. Considering consequential validity can help ensure an assessment has positive effects, such as encouraging meaningful learning, and minimizing negative outcomes, such as discrimination, bias, or learner anxiety (Roever & McNamara 2006). Tsagari & Cheng (2017) distinguish two forms of consequence: 'washback', relating to the influence of the assessment on learning and teaching; and 'impact', which also encompasses broader consequences for the community or society more broadly. In addition to washback and impact, we will also consider the potential for 'wash-forwards' (Gordon 2020). Although not typically considered an aspect of consequential validity, it can be usefully juxtaposed with considerations of washback.

Wash-forward

Wash-forward considers the possibility and implications of existing societal perceptions and educational practices feeding into the design and implementation of the assessment (van Lier 1989). It is important that we consider how such societal and educational ideologies and practices have potentially impacted upon the development of the GALW. Acknowledging these influences allows for informed scrutiny of design decisions to ensure that external factors have not had a detrimental effect on the assessment's ability to meet its goals.

One source of wash-forward effects can be how society values certain languages. In the Welsh context, English is often seen to hold higher societal status and broader utility (May 2013), and it is possible that assessments may reflect these perceptions, with Welsh seen as a cultural rather than functional language. This can lead to assessments that prioritize performative or symbolic competence over communicative utility (Bourne 2001). Shohamy (2020) warns that such tensions can arise when government policy promotes a minority language, but the cultural majority language is prioritised. Whilst such discontinuity does exist in many anglophone areas of Wales, the content of the GALW resists any such performative focus: the patterns included in the WLC content are unambiguously selected for the purpose of authentic use in educational

settings. This is evident in the patterns included (e.g. May I have a pen? What time is lunch?) and the guidance provided to teachers, e.g.:

> ‘Consider where all language patterns can be reinforced across the AoLEs in meaningful, authentic and purposeful contexts’
>
> (Lewis 2025, p.1)

However, despite this aspiration in the content, Russell (2024) demonstrated that many teachers and learners do see Welsh as an isolated academic activity, rather than a practical language of communication, and that authentic use of Welsh is often very limited. In this context, it is important to consider that an abstract and dislocated assessment format may be a perpetuation of this linguistic ideology. However, lexical knowledge is a prerequisite of authentic language use (Nation 2001), and as such, through its formative purpose the GALW can be seen to be laying the foundations allowing learners to access more authentic communicative assessments in future.

Wash-forward also arises when broader curricular choices influence more targeted assessment design. For example, in the scoping study (Russell 2025) many teachers reported that the school’s emphasis on assessing written work in the English element of the LLC AoLE was often carried through to Welsh assessment. Given the text and audio format of the GALW, it is possible that the choice to include text in the assessment items was influenced by a preponderance of text-based assessment. However, given the non-domain specific nature of the target construct and the accessibility factors considered in chapter 4, the inclusion of text can be justified without recourse to wash-forward influence.

Teacher competence may also act as a wash-forward effect, leading to assessments that do not support Welsh use. Teachers who have lower L2 confidence are more likely to stick to rigid, rehearsed, or written tasks to mitigate for their skills deficit (Carless 2004). Such deficits in teacher competence could also be a compounding factor in reliability, should subjective qualitative assessment form part of the format. These factors have acted as a wash-forward effect on the GALW: to mitigate any such variance from teacher proficiency the GALW employs an automated and self-administered format (see section 5.2 for more details) which removes teacher proficiency as a factor.

The wash forward effect may also stem from how assessment is used to fulfil political mandates for bilingualism rather than purely for the promotion of language proficiency. Leung & Lewkowicz (2006) highlight how language tests often serve a symbolic function in multilingual societies. Thus, assessments might be designed to demonstrate compliance with language policy rather than measure actual learner competence. To some extent, this is true of the GALW, as part of the design is a product of the need to generate data in response to policy aspirations (WLEB and Cymraeg 2050). However, the primary goal of the GALW is to enhance teaching and learning, rather than generate performance data. Whilst the quantitative output facilitates both goals, the lack of a more qualitative approach is not a product of prioritising policy directed data, but as a way of reducing subjectivity and facilitating comparability (see chapter 5.2)

Wash-back

Washback refers to the influence that assessments can have on teaching and learning practices, with the potential to shape content, resources and pedagogy. Wash-back is caused by the natural desire of institutions, teachers and learners to maximise test scores. This desire manifests in teacher pedagogy, pupils’ learning strategies, curricular design and learning

provision (Messick 1996). Whilst the term itself is neutral, wash-back encompasses both positive and negative effects, depending on whether assessment supports or constrains meaningful educational outcomes (Alderson & Wall 1993; Taylor 2005; Green 2007). Whether an effect is positive or negative can be considered within the confines of the assessment goals (i.e. within the context of the target construct), or in a broader sense on the learner's journey towards communicative competency (Bailey 1996; Cheng 2005). Shohamy (2020) highlights that washback reflects the power of tests as social instruments that can dictate both learner behaviours and pedagogical priorities. It is therefore important to ensure that assessment washback aligns with intended learning outcomes (both immediate and long term) rather than distorting them.

The drive for such alignment means that holistic assessments that create more authentic communicative experiences are often considered preferable, as they incentivise similarly communicatively focused pedagogy (Brown & Hudson 1998). More specifically within lexicon assessment, Belgar (2013) notes that the assessment of vocabulary benefits from items being assessed within meaningful communicative constructs, rather than in a 'decontextualised' form, a finding supported by van Zeeland (2013) in an aural comprehension context. This contextual aspect is also acknowledged in the CfW (WG 2024a), which places an emphasis on authentic learning experiences and assessment as an essential element of learning progression.

Of course, authenticity poses its own problems with learners combining a broad range of skills and strategies to facilitate the impression of communicative competence. Field (2010) highlights the distortion of learning strategies that focus on using 'peripheral features' in decoding, potential obscuring the extent of actual knowledge. Although authentic, and desirable in communicative contexts, such inference strategies compromise a holistic assessment's ability to analytically measure an isolated aspect of language competency, as explored in section 5.2. This analytic functionality, along with factors such as administration and comparability, is significant in the conscious adoption of an abstract assessment format for the GALW.

Despite such pragmatic justifications, it is important to acknowledge the backwash risk of such decisions. Assessments that focus on abstract rather than authentic demonstrations of ability (such as the GALW's lexical knowledge focus) can lead to an emphasis on memorisation and mechanistic instruction rather than the development of communicative skills (Alderson & Wall 1993). Equally, the adoption of an MCQ format could lead to learner strategies focused on distinguishing (i.e. ruling out incorrect answers or identifying correct answer through clues) rather than comprehension of meaning (Brown & Abeywickrama 2019). This challenge is discussed in more detail in section 6.2. This negative washback is mitigated in the GALW in two ways: firstly, the embedded nature of the WLC content is helpful, as it prevents the dislocation of vocabulary from functional and syntactic context (i.e. learners are directed towards meaningful phrase-like structures, rather than isolated decontextualised words). Secondly, the low stakes and formative nature of the GALW means that staff and pupils are far less likely to adapt their behaviours in an attempt to artificially enhance their scores. This is particularly true of Welsh primary school settings where rote-learning and assessment strategies are rarely, if ever features of the teaching/learning approach.

A final aspect of washback is the effect that the results generated by the assessment have on teaching and learning. In the context of the GALW, it would be easy to assume that the only results washback would be positive: the assessment is designed to provide formative feedback

to improve the pedagogic differentiation, content selection, resource development, and support allocation, so it would be reasonable to assume these intended results would be positive. Cheng (2005) argues that the diagnostic nature of such assessments can justify the use of apparently abstract formats.

However, such formats are also accompanied by the risk of over-interpretation (Read 2000). As a result of the quantitative nature of the data produced by the GALW, there is a danger that numerical measures can create false confidence or over-estimation of ability (Porter 1996). For example, teachers may assume that learners scoring high on a block have achieved mastery of the language content contained, when in fact the learner only has a superficial knowledge of the pattern. They may be able to recognise it, but not produce/comprehend it in authentic communication, chunk/dechunk it to affect generativity, retrieve it with high levels of automaticity, or understand its socio-linguistic function. This could lead to complacency in teachers and learners, where predominantly shallow linguistic knowledge is developed, leading to a focus on lexical breadth, to the detriment of lexical depth. To mitigate this risk, it is essential that teachers are provided with guidance concerning how the GALW data is to be interpreted and its limitations. This data is included in the teacher rubric (appendix item 2) and the instructional video embedded in the teacher dashboard (see section 5.3).

Impact

Tsagari & Cheng (2017) build on earlier definitions (notably from Bachman & Palmer 1996) clarifying that washback refers specifically to the effects of assessment on teaching and learning within the classroom context, whilst impact refers to the broader societal, political, and educational consequences of testing beyond the classroom. The impact of the GALW can be explored in three areas, policy making, institutional dynamics, and long-term societal outcomes.

The most obvious impact of the widespread adoption of the GALW would be its potential policy implications. With broad detailed and comparable data of learning outcomes across different regions, and the facilitation of research into educational best practice, the GALW has the potential to inform policy making, resource allocation, and training provision. The GALW's longitudinal capacity allows the tracking of learner, cohort and regional trajectories facilitating monitoring of progress towards policy objectives. As mentioned above, it is important that data is contextualised, with guidance provided in all reports concerning the limitations of the data generated to ensure it is not over-interpreted.

It is also important to consider the impact the GALW could have on institutional dynamics beyond teaching practice, such as the scores being used for school funding decisions or teacher evaluations. Whether such impacts are positive or negative largely depends on the institutional approach to the data use, whether it is used to create a constructive dialogue with teaching professionals or used punitively as part of top-down evaluation. This is discussed more thoroughly, along with procedures to mitigate the potential for negative impact, in section 8.2.

Finally, it is possible to speculate about the long-term societal impacts of the GALW. The GALW aims to support the acquisition of lexical knowledge in the early stages of language development, before more nuanced and holistic approaches to assessment become appropriate. This transitory function mitigates many of the societal risks associated with high-stakes tests. For example, Shohamy's (2020) concerns around assessments as 'social control', and exclusionary gatekeeping power, and Russell et al.'s (2009) concerns around the

perpetuation of social inequality, are ameliorated by the temporary and formative nature of the GALW.

If the GALW's primary objective of enhancing learning through greater continuity and progress perception is achieved, there are potential positive societal impacts:

- Being able to speak Welsh is advantageous in education and the jobs market. Enabling EM educated learners to develop Welsh language competency provides greater equity in employment opportunities (Grosjean 2010; Lewis 2021), avoiding the potential for linguistic capital to exacerbate social inequality (Bourdieu 1991).
- A greater number of learners leaving education able to work and live through the medium of Welsh helps strengthen and expand the language, preserving Wales's linguistic cultural heritage (WG 2017, p.79).
- Higher level of Welsh language competence also provides individuals with access to Welsh cultural assets such as music, poetry and literature, leading to a more energised cultural sector (WG 2017, p 64).
- Bilingualism has been shown in multiple studies to have cognitive benefits for learners (Bialystok 2001; Bialystok et al. 2012).
- Bilingualism also has the potential to enhance a sense of belonging in populations through shared cultural knowledge (Farhan 2019; Cummins 2000) and potentially encouraging social inclusion and intercultural communication (Fishman et al 2008).

7.5 Scoring Validity

Score validity refers to the degree to which the scores from an assessment are meaningful, appropriate, and useful for the intended purpose (Khalifa & Weir 2009). Adequate scoring validity is essential, as a failure to generate or interpret scores appropriately not only compromises their utility but undermines other forms of validity that presuppose score validity (Weir 2005). Field (2013) identifies five key factors in scoring validity: difficulty, item bias, internal consistency, error measurement and grading. For the purpose of clarity, each element will be considered independently with reference to the GALW structure.

Assessment difficulty

Assessment difficulty is an important factor in overall scoring validity. Poor calibration of difficulty can decrease the sensitivity of the assessment leading to scores that are not representative of the candidates construct-ability (Naumann et al 2019). For example, an assessment where all items are of a difficulty level far beyond the capabilities of the candidates will score all of them consistently low; concealing more nuanced differences in ability, that would be revealed by more differentially sensitive items.

Such mis-calibrations are often explored statistically through Rasch's (1960/1993) Item Response Theory (IRT), or Lord and Novick's (2008) Classic Test Theory (CTT). A foundational concept of CTT is that a user's observed assessment score is made up of their 'true score' (i.e. their actual ability in the construct) and their 'error score' (i.e. the factors that distort the assessments representation of the 'true score'). The observed score is therefore a representation of variation in both the true score and the error score. Improved scoring validity is achieved through more accurate measurement of the true score and mitigation of the error score (Bachman 1990).

With specific reference to difficulty, statistical modelling can be applied to individual assessment items to create 3 measures of difficulty: an item difficult index, a discrimination index, and a distractor efficiency index.

The difficulty index is simply the proportion of correct responses over total attempts for each item, giving a simple index score (e.g. a score of 0.5 indicates half the candidates who attempted the questions selected the correct response). A well calibrated item is usually expected to have a difficult index score of 0.5-0.6, but faculty is generally ascribed to items between 0.35 and 0.85 (Field 2013).

Tests of item discrimination evaluate the ability of each test item to discriminate between 'low-achieving' and 'high-achieving' users (D'Sa & Visbal-Dionaldo 2017). Discrimination scores for each item are generate by dividing respondents into higher achieving learners (HALs) and lower achieving learners (LALs), usually defined by those falling into the top/bottom 27th percentiles. A ratio is then created based on how each group performed on the discrete item (HAL-LAL/n). A higher figure represents a greater difference between the performance of each group and so a greater sensitivity in the item. A discrimination value of between 0.30 and 0.85 is generally considered sufficient (Field 2013).

Finally, distractor efficiency refers to the effectiveness of incorrect options (distractors) in MCQs in attracting selection from low-performing assessment users. Well calibrated distractors should discriminate between HALs and LALs, being sufficiently plausible to mislead those who do not know the correct option, whilst being distinguishable for those that do (Hingorjo & Jaleel 2012). Distractors that attract few selections from LALs are considered 'inefficient' as they do not contribute to the discriminatory value of the item. It is possible for distractors to be too efficient: factors such as shared cognates/false-cognate (words with similar phonology but different meanings in the L2 and L2, e.g. 'moron'), or grammatical ambiguity (e.g. Welsh doesn't distinguishing between present simple and present continuous). This is discussed in more detail in section 5.5 and 6.2. Distractor efficiency is calculated for each item in two ways: firstly, a DE index showing the distractors ability to attract selection from LALs can be calculated in a similar way to item discrimination, but with the values inverted: DE=LAL-HAL/n. A good distractor efficiency is indicated by a positive score >0.2. Secondly, a simple frequency analysis can be used to identify distractors that attract very low levels of selection, this also indicates poor distractor efficiency.

There are weaknesses in using these approaches to assess item, and more broadly, test difficulty. Firstly, all three tools draw upon sample data in defining item difficult, discrimination and distractor efficiency, leaving the determining metrics open to distortion from sampling bias (Field 2013). Whilst this can be mitigated by large scale trialling of the assessment across the whole target population, in assessments with many items it may not be possible to collect sufficient data on each item. A revisionist approach to test difficult is therefore important, with the test being reappraised periodically as the amount of contributary data grows, improving difficulty calibration (see section 5.1 for a discussion of calibration sample size and iterative refinement). Secondly, these approaches are purely measures of statistical difficult, i.e. the chance of a user selecting the correct answer, not measures of cognitive load or required effort. Whilst some correlation has been shown between statistical difficulty and cognitive/conceptual difficult, they should still be considered as distinct aspects of item difficulty for the purpose of scoring validation (Rush et al. 2016; Noroozi & Karami 2022). You can find discussion about the mitigation of cognitive load as a distorting variable in section 5.2.

Item bias

We have already touched upon content bias from a socio-linguistic perspective in the section discussing context validity. Here, we will consider bias within the context of scoring validity, i.e. to what extent items measure construct-irrelevant features that advantage or disadvantage learners from a specific sub-group (e.g. gender, ethnicity) thereby increasing the error-score. To determine item bias, we can carry out a differential item function procedure (Holland & Thayer 1988), where the difficulty index of an item is consulted in conjunction with respondent background information specific to sub-groups of interest to determine if a statistically significant ($p<0.05$) association can be found between the two (Chen et al. 2024). Where such an association is identified, the item requires inspection for potential bias.

A significant association with a subgroup is not necessarily indictive of item bias, there are legitimate reasons why a particular subgroup may perform better/worse in a particular construct. For example, if EAL learners perform worse on multi-clause L1 to L2 translation tasks, this may indicate item bias caused by L1 knowledge may be a construct-irrelevant feature that needs to be mitigated (e.g. visual references, additional scaffolding, invigilator support). However, if the disparity was, say, between different genders, there may well be socio-linguistic reasons for such a discrepancy (Viriya & Sapsirin 2014) and evidence of these is a valuable feature of the data. In such cases, ensuring that the test is heterogeneous in its inspection of the target construct can help mitigate format distortion and improve composite reliability (Girolamo et al. 2022).

Item bias will be monitored in both the initial development of the GALW and its subsequent iterations. You can find details of how this process is integrated into the iterative assessment development plan in section 5.1.

Internal consistency

Internal consistency is an indicator of how well a set of items measures a single unidimensional construct (McCrae et al. 2011). It is usually measured using Cronbach's Alpha (Cronbach 1951), McDonald's Omega (McDonald 2013), or G-Theory (Cronbach et al. 1972). However, due to the binary nature of the item responses in GALW, KR-20 (Kuder & Richardson 1937) is a more appropriate measure. KR-20 evaluates how well the test items collectively measure a single construct, whilst being optimized for binary data (i.e. within the GALW, correct/incorrect). KR-20 considers the variability of scores on each item and the overall test score variance to estimate how well the items collectively contribute to the reliability of the test.

To perform a KR-20 analysis, we first calculate the item difficulty, then the variance of all scores. The variability of each item's responses is then calculated and aggregated to determine the contribution of all Items. By combining the variability of individual items, the total score variance, and the number of items, an estimate the test's internal consistency can be made. A higher KR-20 value indicates greater consistency, implying the test items consistently measure the same construct.

However, it is important to consider that tests with large numbers of items (such at the GALW) can produce inflated consistency estimates as a product of item quantity (Streiner 2003). To mitigate this, smaller sub-sets of randomly selected items should be subject to KR-20 testing to ensure that the overall score is not misleading. Discrete testing is also required to accommodate different assessment constructs, as KR-20 cannot manage multidimensionality within a single assessment calculation (Cortina 1993). Finally, item redundancy (e.g. due to similarity or repetition) can artificially create inflated scores of internal consistency (Schmitt

1996). To address this an inter-item correlation matrix can be used to identify items with high levels of redundancy which can then be modified or removed (DeVellis & Thorpe 2021). These elements are considered with reference to the GALW is sections 5.1 and 6.1.

Error Measurement

Error measurement allows us to analyse and quantify how accurately an assessment represents a user's true score by inferring the error score inherent in the test model. To achieve this two error measurement calculations are applied to the GALW: standard error measurement (SEM), and test-retest reliability (TRR) (Gulliksen 1950).

Having tested the internal consistency of the assessment, it is possible to estimate the SEM for the assessment. This provides an estimate of the error score, providing a range for any one individual user's score which should capture the user's true score (Bachman & Palmer 1996). This is calculated by multiplying the standard deviation of the observed scores by the square root of 1 minus the reliability coefficient (in the case of GALW, KR-20):

$$\sigma_E = \sigma_X \sqrt{1 - \rho_{XX'}}$$

Where, σE = Standard Error of Measurement, σx = Standard deviation of observed test scores, and ρxx' = Reliability coefficient of the test (KR-20).

A smaller SEM indicates a higher level of accuracy in test scores reflection of the user's true score. For example, a user score of 70 with a SEM of 10, means the users true score lies between 60 and 80.

Test-retest reliability (TRR) measures the consistency of test scores over time by administering the same test to the same group of test-takers on two separate occasions. It evaluates the stability of test scores and assesses whether the test produces similar results under similar conditions. To conduct a TRR measure, a representative group of test-users is selected and completes the assessment twice. The time interval between assessment should be 1-2 weeks, long enough to avoid memory effects, but without giving sufficient time for significant development of vocabulary knowledge. Pearson's R is then used to calculate a correlation coefficient, measuring the degree of similarity between the test scores of each individual, with a higher degree of correlation (>0.7) indicating a more stable and reliable measure (Brown 2005).

SEM and TRR are included in both the initial trialling of the GALW and the iterative development plan outlined in section 5.1.

Grading

Within scoring validity, grading is the process of assigning scores or classifications to a student's performance. Taylor & Galaczi (2011) highlight two key approaches to grading: holistic (sometime referred to as 'global'), where the candidates' performance is considered as whole; and analytic (or 'profile'), where different aspects of the performance are isolated and examined separately. As already mentioned in section 3.1, the GALW utilises an analytic approach to isolate different forms of content knowledge, and so accordingly, grading should reflect this same approach.

The GALW's primary goal of supporting teaching and learning means that analytic grading is essential in providing prescriptive/diagnostic feedback to users and administrators. However, the assessment also aims to make broader claims concerning how these discrete scores are

indicative of broader communicative competence (see chapter 3.2). The GALW therefore needs to adopt a multi-dimensional and multi-level grading framework that represents the data in different aspects and layers aligning with these different purposes.

The analytic elements of the GALW's grading are relatively easy to construct. The narrow focus on lexical knowledge and MCQ format makes the generation of diagnostic and prescriptive feedback relatively simple to automate. Analytic performance feedback is graded through norm-referencing, where user scores are contextualised with previous performance and discrete performance to generate qualitative and formative grading. This includes highlighting content blocks where the learner performed best, or made the most improvement, block in which their performance was poor and where they should focus their attention, and a review of their longitudinal performance through trajectory analysis. Descriptive grades would be used to help communicate learner performance, e.g. 'master' for a block with >90% correct, 'developing' for users who improve their previous score by >10%. A fuller discussion on this element can be found in section 5.7.

The more holistic grading incorporated into the GALW poses more significant challenges. In this instance, norm-referenced grading, where comparative performance is used to generated score (e.g. percentile performance within a cohort), is not possible, as the target construct concerns absolute content knowledge and inter-institutional comparability is key to several of the assessment goals. Usually, such comparability is achieved through standardisation, however one of the key challenges encountered when developing the GALW was the inconsistency in content and performance between different schools. This is compounded by the comparative breath of the WLC content guides from which question banks would be drawn, and the far narrower field of content usually included in individual school curricula. Any standardised content would either result in distortion from variations in class coverage or necessitate a quantity of items proportionate to the WLC content, leading to excessively long assessment duration. This would undermine the GALW's functionality and practicality as an AfL tool and restrict it's take-up by educators.

In other circumstances, such challenges could be addressed through the establishment of predictive validity or concurrent validity. Predictive validity refers to the extent to which a language test forecasts a test-taker's future performance or behaviour. The GALW scores could therefore be linked to learning outcomes at GCSE or some other summative assessment point, in order to create a predictive score matrix e.g. a GALW score of 20 is associated with a C at GCSE. However, as a newly developed assessment predictive validity will not be possible until cohorts engaging with the GALW reach a summative assessment point. Currently the only assessment that would meet the requirements for such predictive validation is the Welsh L2 GCSE, which learners do not take until they are 16.

Concurrent validity assesses how well a language test correlates with an established measure of the same construct at the same point in time. Unfortunately, there are no comparable assessments in use at the moment within the EM sector (Russell 2025). Whilst there are assessments in adult education (e.g. Dysgu Cymraeg examinations) that could be used to attempt concurrent validation, they do not correspond to the WLC content, child developmental needs, or pedagogic approach of the primary school context. More generic measures of vocabulary development face similar challenges (see section 3.1 for a discussion of cross-validation with yes/no tests) making them unsuitable for establishing concurrent grade validation.

Holistic grading in the GALW is therefore structured around predictive lexical knowledge. Adopting the same principles as the yes/no tests of vocabulary size (Meara and Buxton 1987) whereby user performance in a sample of items is used to estimate their vocabulary size. For example, in a yes/no test, learners may be presented with 100 items drawn randomly from a list of the 1000 most common words, a score of 60 would therefore be indicative of a knowledge of 600 of the words on the list. This format has the advantage of predicting a holistic score without requiring the user to complete every item in the battery, resulting in shorter test duration, improving practicality, and reducing variability from disengagement (Wise & Kingsbury 2022).

However, as pupils in EM primary schools learn almost entirely through a stem-sentence format (see examples in item 1 of the appendix), their discrete word knowledge is likely to be disproportionate to their phrase knowledge, and so of limited faculty in predicting WLC content knowledge. Alignment (as discussed in section 6.1) with the WLC content is therefore essential in ensuring the accuracy of predictive scores. This is achieved through the selected content format of the GALW (see section 5.2) thereby distinguishing between content learning and coverage in assessment scores.

GALW predictive scores are therefore a compound of two factors: the content coverage, and the content learning/retention. Content coverage is defined by the teacher and ensures that learner scores are not distorted by the inclusion of untaught material. In assessments of English this selective approach would itself be a distorting factor, as most learners would have acquired significant lexical knowledge outside of the educational setting (e.g. through film, online media, music). However, in the Welsh EM school context, learner exposure to the language outside of education contexts tends to be extremely limited (WG 2012, p.12; WG 2021b, p.7) making such an exclusionary approach suitable. Coverage is selected by teachers prior to learner engagement with the GALW, identifying blocks of language that their class has covered. This list expands over time as new content is introduced. From these blocks, assessment items are generated in accordance with the GALW structure (see section 5.2).

The second element of the predictive score is comprised of the learner performance in the selected items. As explored in the yes/no test, the proportion of known items allows for the inference of total lexicon. So, if a teacher selects blocks that cumulatively include 100 items, the learner when presented with 20 questions, and scores 14 (i.e. 70%), it can be inferred that their total lexical knowledge is 70. This is obviously a rather crude simplification, not accounting for guessing, or transient feature variables (tiredness, distraction, illness).

These distortions are addressed through the format of the GALW. Transient features are accommodated through the score aggregation across multiple user engagements (see section 5.2), thereby dispersing the influence of any temporary distortion. Teacher rubrics also include guidance on when to delay assessment in order to avoid aberrant scores.

Guessing is a more complex issue, here it is considered purely from the perspective of score validity, but a more wide-ranging discussion can be found in section 7.6. The MCQ format of the GALW means the simplistic approach describe above will inevitably lead to a misleading and inflated score for some learners due to guessing. Given the four options available, a learner with zero lexical knowledge is still likely to score around 25% based purely on chance. Given a selected coverage of 100 items, this would lead to a predicted score of 25, even though their actual lexicon is 0.

In order to statistically correct for such error caused by guessing in the predictive scoring, we can apply a corrective formula:

$$S_c = R - \frac{W}{k-1}$$

Where Sc is the corrected score, R is the number of correct responses, W is the number of incorrect responses, and k is the number of answer choices per question. This adjusts for guessing by subtracting a fraction of the incorrect responses from the number of correct responses. Because the correction is proportionate to the number of correct answers, this adjusts the score proportionately for learners of all levels.

Whilst this approach improves the accuracy of predicted scores, it is important that teachers are made aware that corrections for guessing are made as part of the automated analysis as this will help prevent misinterpretation of the scores (e.g. assuming correct hadn't been made for guessing and attempting to account for it in the post-correction scores).

The advantage of such a compound predictive score is its responsiveness to multiple factors in the learning journey: on going expansion of content coverage and flexible curricula are accounted for in the selective approach, quality of learning is accounted for in the assessment format, progression/regression/attrition/stagnation are evident in the longitudinal data collection. Perhaps more importantly, it provides a point of comparison between institutions. Two schools can have completely different curricula, drawn from different regional dialectic sources, and still produce comparable measures of learner lexical knowledge through the GALW.

7.6 Reliability

In assessment theory, reliability refers to the consistency and stability of an assessment in measuring the target construct over time, across different test forms, or between raters. A reliable assessment produces similar results under consistent conditions. Reliability is crucial because it underpins the fairness and accuracy of assessments; inconsistent results mean an assessment cannot be trusted to provide meaningful information about a learner's true ability (AERA 2014).

The MCQ format and prescriptive nature of the WLC content used in the GALW mitigates some common challenges of reliability (e.g. rater subjectivity, descriptor ambiguity). Consequently, there are five factors that are anticipated to impact on the reliability of the assessment: guessing, the practice effect, reference language proficiency, dishonest conduct, and the implementation of the assessment rubric.

Guessing

A particular problem to selected response items is candidates guessing answers (McLean et al. 2015; Gyllstad et al. 2015). Guessing can have a distorting effect on scores, reflecting individual characteristics such as risk taking, confidence, and score motivation, rather than the target construct (Milton 2010). Read (2018) distinguishes between 'blind' guessing and inferential guessing which is based on related L2 knowledge. Inferential guessing is a key skill in developing comprehension (Ramos & Dario 2015) and so construct relevant to some extent, whereas blind guessing is a source of inconsistency and distortion in assessment results. The challenge becomes how to mitigate the impact of 'blind' guessing without inhibiting or failing to credit inferential guessing.

Nation (2013) argues for encouraging guessing to ensure such inferential strategies are captured in the results and encouraged through backwash. However, others (Read 1998; Wise & DeMars 2006; Wise & Kong 2005) highlights the danger of such an approach: finding that low self-efficacy and poor individual motivation can influence candidates' willingness to make informed guesses. This becomes particularly relevant when looking at low frequency assessment items (i.e. those with which fewer candidates will be familiar), with McLean et al (2015) finding that large proportion of score-variance was attributable to guessing. Bolt et al. (2002) note the effect of time restrictions on the candidates' propensity to guess, finding that time pressure can encourage random guessing. Cao and Stokes (2008) find that item difficulty is also a factor, with more difficult items encouraging a greater degree of random guessing, as learners expend less effort if the items are considered too challenging.

Guessing can be reduced by explicit instructions to participants indicating whether it is desirable or not, including penalties for incorrect answers, or by the inclusion of a 'don't know' or 'skip' option (Read 2019). However, Stoeckel et al. (2016) found that the inclusion of such 'don't know' options increased variance from factors unrelated to vocabulary knowledge: score differences between students of the same level varied due to differential usage of the 'don't know' option.

Whilst there is a consensus that guessing should be considered in assessment design to avoid score over-estimation (Gyllstad et al. 2015), the GALW's emphasis of diagnostic and comparative data use somewhat ameliorates this. However, the variation in scores guessing can cause through variations in socio-cultural or psychological factors does need to be mitigated to ensure comparative data use is robust. Rather than prohibitive approaches that reduce variation from guessing, but also disincentivise inference skills that are desirable in L2 learning, the GALW adopts a permissive approach that encourages all assessment users to make guesses consistently limits the variation from non-linguistic factors.

The GALW seeks to do this through a legitimisation of guessing in user guidance, administrator rubrics and assessment structure.

Firstly, the user and teacher rubrics will make it clear that guessing is permissible, and learners should use their language knowledge to try and work out the correct option in each question. This mitigates any sub-group effect that could bias results and potentially deceases learner anxiety from instruction ambiguity (Golvardi et al. 2021).

Secondly, the format of the assessment aims to ensure that guessing is evenly distributed through the use of a linear mandatory response format. This means that learners can not progress through the assessment until they select a response. Whilst this does not mitigate blind guessing, it at least ensures that guessing is not underrepresented in any sub-group and allows for a corrective approach outlined in section 7.5.

Practice effect

The practice effect, whereby familiarity/unfamiliarity with the assessment format influences the accuracy with which the assessment represents learner competence, must also be considered and mitigated to enhance assessment accuracy (Ockey and Zhi 2015). This is particularly pertinent to younger users for whom the effect of unfamiliarity can be exaggerated, therefore simplicity and similarity with classroom-based activities should be sought (Alexiou & Milton 2020).

One of the key advantages of the MCQ format adopted by the GALW is its ubiquity in educational settings and the high degree of familiarity learners tend to have with the format (Field 2013) thereby reducing the likelihood of any practice effect distorting user scores. This is further addressed through the repeat format of the GALW, where learners are encouraged to engage with the assessment tool at regular intervals. This repeat exposure will develop familiarity even if the format is initial novel to the user, mitigating any distortion to long term scores. Finally, clear instructions and demonstrations will be included in the learner rubric and landing page, helping ensure that learners understand the assessment format prior to engaging with it.

Reference language proficiency

Language assessments for children necessitates careful consideration of their proficiency in the reference language of the assessment. In the case of GALW this is English, and the justification for this decision can be found in section 5.2. Deficits in reference language proficiency (RLP) can significantly influence the accuracy and fairness of the evaluation, as it becomes difficult to distinguish between the effects of construct-proficiency and the impact of poor comprehension of the reference language. Children's RLP can vary widely across different age group, particularly among immigrant, ALN or EAL learners. Standardized assessments often fail to account for this linguistic diversity leading to sub-group bias in the results.

Administering language assessments in the target language is advocated by several scholars to mitigate RLP and enhance the context validity of the evaluation (Hasrol et al. 2022). This is preferable when dealing with a multi-lingual class without a shared L1. However, in the case of the GALW almost all learners will have a high proficiency in English, whilst their Welsh language proficiency is almost certainly inadequate for the purpose of assessment instruction. Nation (2001) argues that a bilingual test format (i.e. with definitions and instructions in the users' L1) is a more accurate measure of L2 knowledge, as it removes cognitive load of decoding definitions and is better suited to lower-level learners.

In the GALW, deficits in RLP are addressed in two ways: firstly, the administrator rubric (appendix item 2) states that learners with insufficient English language competency are to be temporarily withheld from engaging with the GALW; secondly, additional accessibility features such as text-to-speech (see section 4) are included in the assessment functionality to support learners who may have a domain-specific deficit.

Dishonest conduct

Dishonest conduct, commonly referred to as cheating, poses a significant threat to the reliability of assessments, undermining the accuracy and consistency of test scores, resulting in misrepresentations of construct-competency (Jacob and Levitt 2003). This misrepresentation can lead to erroneous conclusions about student performance and the effectiveness of pedagogical approaches, resources, and interventions. It can lead to the misallocation of teaching time and resources and can misdirect identification of learner needs (Tight 2024) Addressing cheating is therefore crucial to maintain the credibility of assessments and ensure that they serve their intended evaluative purposes.

However, Dawson et al. (2024) cautions against a moralistic approach cheating, advocating for the consideration of cheating and its mitigation as an aspect of assessment validity, with a broad focus on assuring learning, rather than punitive responses. In fact, many of the tools used to address cheating inadvertently undermine other aspects of the assessment validity and functionality, compromising its functionality even for honest users (p. 1010). Any mechanisms

designed to prohibit teaching must therefore be considered within the broader assessment structure, and the implications considered for all users.

It is important to recognise that the propensity to dishonest conduct is not evenly distributed. Potential sub-group factors indicating an increased likelihood of cheating include gender (Özcan et al 2019), academic performance (Brown et al 2020), and personality (Lee et al. 2020). The effect of cheating is therefore a factor that could bias results, compromising the use of data in secondary research.

The GALW addresses the risk of dishonest conduct through formatting decisions and rubric design. There are four forms of dishonest conduct that could be relevant to the GALW: copying, impersonation, collaboration, and unauthorised support material. These will be considered individually below with reference to both the likelihood and significance of each type.

Copying is simply when a user observes a response made by another user and mimics that response. In standardised assessments, copying is generally rare but where it occurs it is extremely influential (i.e. it has the capacity to completely misrepresent a learner's construct competence). To address this, the format of the GALW includes a random draw down of items from the selected content bins (see section 5.2), this means that no two assessments are likely to be the same. This variability of items makes copying an unlikely feature to impact on GALW scores.

Impersonation is when one user performs the assessment under the name of another user. This form of cheating is usually impossible in primary education, but the digital nature of the GALW means that it is possible for a user to log-in under an assumed identity. Despite this, impersonation is likely to be an extremely rare form of cheating as it involves no clear personal gain for the individual. Of course, should impersonation occur it would have a total misrepresentative impact on score legitimacy. In the GALW, impersonation is mitigated through the use of individual user accounts, that require users to log-in individually. Given the level of risk this is considered sufficient mitigation for a low-stakes formative assessment.

Collaboration is when more than one user works together of the assessment tasks. This could be sporadic cooperation on a limited number of items, or comprehensive collaboration for the whole duration of the assessment. This form of dishonesty is far more likely that impersonation and can be more common in low-stakes informal assessments such as the GALW. Depending on the extent of the collaboration, the effect of collaboration on individual scores may be anything from superficial to extreme. The risk of collaboration is addressed in the GALW through teacher/administrator and user rubrics. In the learner rubric, it is explained why helping each other is not allowed and how it could negatively impact on their learning. The teacher/ administrator rubric advises that users are spaced so as to disincentivise collaboration and monitored when possible.

Unauthorised support material could include textbooks, cheat-sheets, digital devices, or audio materials. Any resource that artificially supports or replaces lexical retrieval will have a significant distorting effect on user performance. Unauthorised material may be used intentionally but can also be inadvertently drawn on. Many classrooms will include displays and materials intended to enhance learning which include Welsh material. It is possible for learners to use these universal provisions in completing assessment tasks. Use of unauthorised materials is addressed in three ways: firstly, the GALW teacher/administrator rubric offers guidance prohibiting the use of additional resources during the assessment and advising where

possible that users are situated away from any universal Welsh language provision. Secondly, the randomised draw down of items makes it unlikely that the universal in-class provision will consistently impact on performance. Finally, the repeated and low-stakes nature of the GALW makes it less likely that learners will attempt the subversive use of unauthorised materials (Cizek 1999).

Although the GALW offers a low-stakes context for learners, it is possible that it creates a high-stakes context for teachers. If misused as a punitive measure of teacher performance (see section 8.2) it is conceivable that teachers may be incentivised to facilitate dishonest conduct in their learners in order to enhance scores. This could take the form or teacher/peer-collaboration, misreporting of results, or the provision of unauthorised materials. Whilst such unethical practice would be extremely rare, should it occur it would obviously have a significant impact on the whole class. Should the GALW be misused as a measure of performance for institutions, this effect could even be extended to school leadership impacting on whole cohorts or institutions. It is beyond the remit of this specification to police misconduct of this type, but mitigation through clear guidance concerning the appropriate use of assessment data can help ensure that such perverse incentives do not emerge.

Rubric design and implementation

Rubric design is also a factor that can impact upon the reliability of language assessments. There is some ambiguity around the term 'rubric': traditionally, it referred to a set of instructions or guidelines for test participants and administrators, but more recently the term has been used to denote a scoring guide used by learners to evaluate the quality of constructed responses (Allen & Tanner 2006). In the context of the GALW it is the more traditional interpretation that is used.

Field (2013) highlights two key ways in which rubric design can distort candidate performance: rubrics can be too long or complex, diverting attention and cognitive resources away from task completion; rubrics can be unclear, misleading candidates or administrators in ways that distorts results. From a candidate perspective, rubrics need to clearly and concisely include the information candidates require to complete the task, without the use of language unsuitable for the level/age of the candidate. From an administrator perspective, rubrics need to clearly outline the purposes (why the test is used), procedures (how the test is delivered), and application (how the results are used) of the assessment (Cohen & Wollack 2006). A full discussion of rubric design can be found in section 5.6.

7.7 Piloting, Development and the Validation

Assessment validity is highly situated, both contextually and chronologically i.e. an assessment's validity is dependent on its alignment with its setting. Whilst the contextual factors have been considered in detail, it is also important to acknowledge the time dependency of assessment validity: assessment are situated in the socio-political-cultural environments and these environments are not static, but shift and evolve over time. Assessments must be similarly dynamic, adapting to their users and the context of their use.

Piloting is an essential element of this adaptive validity, highlighting technical/pragmatic/functional deficits in the assessment's structure. Regular piloting and reviewing of assessment performance allows for an iterative approach to assessment development, ensuring that validity is maintained and enhanced over time.

To achieve this, the GALW adopts a cyclical approach to assessment development, integrating a series of rollouts, analytic appraisals, and adaptations. A full discussion of the piloting process and how it integrates aspects of validation can be found in section 5.1.

8. Ethical Considerations & Data Security

Ethical factors are an important in the development of language assessments to ensure fairness, validity, and inclusivity across diverse learner populations, whilst ensure learner data is protected and used appropriately. Language assessments often have significant consequences for test-takers, such as academic progression, employment opportunities, or immigration outcomes. Although the GALW purposefully seeks a low-stakes AfL function, it is still necessary to consider the possible implications of its design and use on learners and staff. As such, ethical factors must be considered in the assessment design to avoid unintended negative consequences (McNamara 2000).

A key ethical concern lies in test fairness, which encompasses the equitable treatment of all test-takers and the avoidance of discriminatory content or structures (Kunnan 2004). Without careful attention to ethical principles, language assessments risk privileging certain linguistic, cultural, or socioeconomic groups while marginalizing others. This factor is considered in detail chapter 7).

The validity of interpretations and uses of test scores is an ethical issue. Messick (1990) emphasized that test validity is not merely a technical property but also a moral one: if scores are used to make decisions that impact individuals' lives, their interpretive accuracy and appropriateness must be scrutinized to ensure no inadvertent harm is caused. Test washback, or the influence of testing on teaching and learning, is another ethical dimension. Assessments should promote beneficial pedagogical practices rather than narrowing curricula to teach to the test (Shohamy 2020). These factors and their influence of the GALW's design are discussed in detail in chapter 7.4.

There are four key areas of assessment ethics that have not been covered in previous chapters: the theoretical justification for education assessment; surveillance and monitoring through assessment data outputs; informed consent; and data security/ownership. These will be considered separately below:

8.1 The ethical justification for language assessment

It may appear inarguable that the assessment of language is not only desirable but essential. Language testing plays a central role in educational systems, both compulsory and further/higher education. As discussed, it serves to evaluate learners' proficiency, guide pedagogy and curriculum planning, and enable discursive approaches to accountability. Language assessment is often conceived as a neutral and objective process of measurement (Bachman & Palmer 1996), an unambiguously beneficial practice that promotes transparency, comparability, and fairness in language education systems.

However, this position has been challenged by authors who argue that language assessment is not a neutral or purely technical process, but one that is socially, politically, and ideologically loaded. Shohamy's (2020) *Critical Language Testing* provides one of the most influential critiques of the assumed desirability and objectivity of language assessment, arguing that language tests often function as 'mechanisms of control' (p.1) and gatekeeping, serving political

and institutional agendas rather than educational goals. They demonstrate how language assessments have been used to impose linguistic ideologies, marginalise certain populations, and restrict access to education or employment (p.117). Assessment may not merely be a pedagogical tool, but an instrument of social regulation and exclusion.

The GALW is by no means free of such political influence, both the motivation and formation of the assessment is impacted by government policy and existing educational context. Whilst the GALW has been designed with mitigation of bias towards any specific user group and the goal of improving educational equity, its existence is a product of an ideological conception of Welsh language education as a positive aspiration for all children in Wales. It is beyond the remit of this specification to challenge such assumptions, and it can be argued that such political influence is what empowers research rather than constrains it (Whitty 2006). It is inarguable that Welsh language skills are advantageous to learners in the current socio-political context, and therefore any assessment that seeks to improve accessibility to such skills can be argued to reflect a pragmatic moral good (Dewey & Tufts 1908).

Similarly, Roever and McNamara (2006) emphasise the social dimension of language testing, arguing that assessment must be understood in terms of its consequences for individuals and communities. They argue that traditional models often fail to account for the complex, contextual nature of language use and the ways in which language proficiency is socially constructed. Language assessment should therefore be critically examined in terms of its fairness, its impact on learners, and the ideologies it reinforces. They propose a "critical language testing" paradigm that foregrounds ethics, social justice, and contextual appropriateness, moving beyond a narrow focus on validity and reliability.

Further theoretical contributions support this rethinking of language assessment. Norton and Toohey (2001), drawing on sociocultural theories of language learning, argue that language proficiency cannot be meaningfully assessed without attention to identity, power relations, and social practices, a view reflected by Douglas' (2000) concept of specific purpose language testing, emphasising the need for authenticity and situational relevance, which complicates assumptions about standardized general proficiency measures.

While assessment may serve legitimate educational functions, it also risks reinforcing social inequities and misrepresenting learners' abilities if not designed and interpreted with careful attention to context, power, and impact. The challenge, then, is to ensure that assessment is employed ethically, inclusively, and reflexively whilst maintaining an awareness of its broader social and political consequences. This approach has not been treated discretely, but instead has been drawn upon throughout the development of the GALW, with attention paid to both the format and contextual application of the assessment.

8.2 Surveillance and monitoring

There is a potential for education assessments to be used as a form of surveillance/monitoring by institutions or organisations of teaching staff. Whilst this can be done with the intention of identifying staff who need support or training (Isoré 2009), such practice can lead to assessments being used punitively against staff, particularly for performance management or accountability. This can have negative consequences for teachers, such as increased stress, demoralization, and professional burnout among educators (Day & Gu 2010). Such consequences inevitably result in negative outcomes for learners, including elevated levels of

stress (Oberle & Schonert-Reichl 2016), lower learning outcomes (Arens & Morin 2016), and poorer mental health (Harding et al. 2019).

Such systems also risk having a negative impact on school environment, fostering a culture of blame, where assessment becomes a tool for surveillance rather than improvement (Ball 2003). This environment can damage professional relationships within schools, as competition and fear replace collaboration and trust (Kelchtermans 2005). Additionally, punitive assessment regimes can deter educators from working in more challenging contexts where results are likely to be lower, thereby potentially exacerbating educational inequalities (Sahlberg 2021).

For the GALW these potential negative impacts are particularly pertinent, as the assessment is intended to be used electively by the teacher within the routine classroom setting. As a result of this context, any teacher anxiety around monitoring or surveillance of assessment data is likely to result in reduced usage, manipulation increase apparent learner performance (e.g. by providing inappropriate scaffolding), or an abandonment of the provision entirely.

Of course, there are numerous examples of assessment having a positive impact on teacher management, institutional culture, and learning outcomes (Shepard 2000; Earl & Katz 2002; Fullan 2009). However, it is important to consider the potential negative consequences when designing an assessment in order to develop mitigation where possible. In the case of the GALW the mitigation consists of two elements: data aggregation levels, and assessment guidance.

The aggregation of data (beyond the generated output on the teacher dashboard) aims to provide useful information for institutions and stakeholder/government organisations (Consortia, Welsh Government). School level data will be anonymised in data releases which will report on holistic performance at a regional level. Whilst individual institutions may choose to share their data with external organisations, this will not be available in the collated data, limiting the potential for GALW to form a basis for inter-institutional competition or ranking.

At an intra-institution level, schools will be encouraged to avoid using the GALW as part of teacher performance management. A summary of the research showing the dangers of learner assessment as part of professional monitoring will be presented and the focus of the GALW as a formative tool for teacher emphasised. The importance of regular assessment points to effective use of the GALW in directing learning progress will be explained to emphasise the detrimental effect punitive use could have on the provision's efficacy.

Whilst it is not possible to control how school elect to utilise the GALW once it is made available, it is hoped that the guidance provided will encourage schools to maintain focus on its intended primary function as a formative assessment tool.

8.3 Informed Consent

Shohamy (2020) argues that language tests must be transparent in their purpose and use, as lack of transparency can lead to misuse, unintended consequences, and negative washback. She critiques how language tests are often presented as neutral tools but are actually used to enforce policy agendas, often without learners or educators fully understanding their objectives. Transparency is therefore essential to ensure that assessments serve educational rather than political or gatekeeping functions, and ethical test development involves consideration of informed consent and clarity about how results will be used (Taylor 2013).

In the GALW informed consent operates at three levels: institutional, teacher, and user. These are considered separately below as different factors are pertinent to each context.

Institutional consent will be obtained as part of the registration process. School leaders will receive an information pack detailing the functionality and purpose of the assessment, alongside guidance around use of the data generated. This includes notification that the data generated by their institution will be included in aggregated data supplied to external or government organisations. It will also include guidance around the internal use of the data generated (e.g. for action research).

Teacher consent will be obtained when the teacher account is registered through their institution. During this progress teachers will be informed about the functionality of the GALW, what data is collected, and how it is intended to be used in the classroom setting. Teachers will be informed that institutions will have access to aggregated class data, and how they have been advised to use this data. They will also be made aware that the data from their class will be collated in institution level data that may be made available to external and government organisations. They will receive information of how learner and teacher identity will be protected in these data sets.

Learner consent is more complex, in primary education assessments are usually administered to learners without informed consent (Shohamy 2020, p.143), reflecting the 'assumed compliance' of the education system, where learner consent is implicit in their participation in the system (Freire 1996). Such a stance is not without justification, Biesta (2004) argues that education has social and collective goals that cannot be entirely subject to the will of the individual. Additionally, Newton (2007) highlights the impracticalities inherent in an elective approach to educational tasks, whilst others (Lundy 2007; Brighouse 2006) argue that learners (especially in younger learners) may lack the capacity to make informed decisions in their best interests for developmental reasons.

Despite these arguments against the necessity for learner consent, many authors continue to advocate the importance of informing assessment users about the nature and purpose of assessments. This acknowledges learner agency whilst not submitting to it, can enhance motivation (Shohamy 2020), and aligns with guidelines around the rights of the child (United Nations 1989). Accordingly, the GALW will include child friendly information as part of the user rubric giving a simple explanation of the purposes of the GALW. This will be provided in video form to help ensure accessibility for all users.

8.4 Confidentiality and Data Protection

When designing digital language assessments, confidentiality and data protection are critical to ensuring ethical integrity and safeguarding learner rights. The collection, storage, and sharing of personal data, including test responses, biometric data, and user behaviour—raise significant concerns about privacy and security (Taylor 2013). Assessment users must have confidence that their personal and performance data will be handled responsibly and used solely for intended assessment purposes. Failure to ensure secure data practices can lead to data misuse, identity risks, and erosion of test-taker trust (Eignor 2014). Digital assessments particularly are subject to legal and ethical frameworks, such as the General Data Protection Regulation (GDPR) in Europe, which require transparent data practices, explicit purpose specification, and user consent (Williamson 2017). Protecting learner data is not only a legal obligation but also a dimension of test equity, as inappropriate disclosure of test scores or personal data can lead to stigmatization, discrimination, or unjust consequences, especially for vulnerable populations (Kunnan 2004). It is therefore essential to ensure that data systems are secure, and access is restricted appropriately.

The GALW is designed with these legal and ethical obligations in mind. The design features aimed at ensuring these obligations are met fall into three categories: protecting user identity, securing data, and data usage.

To ensure that the identity of users is protected, any identifying information will be stripped out of data sets automatically when downloaded (Williamson 2017). Only the teacher dashboard will include usernames, all other outputs will have responses tagged with anonymized unique learner identifiers (ULIs). These ULIs can be used to link data generated by the GALW to other research instruments (surveys, assessments, demographic data) for multi-variable analysis, but only if opt-in consent is sought from the individual institution for specific studies.

The GALW uses a role-based access system for local data to maintain data security whilst allowing accessibility for users that facilitates functionality. Users will have access to their own data through their Hwb accounts. Teachers will have access to the data of all the users within their class. Schools will have access to all the data of classes within the institution. Administrators will have access to all the data from across registered institutions.

To ensure that user data is secure it will be encrypted in both transmission and storage, and held on the WG Hwb server and protected by two factor authentication. Access to the whole data set will be restricted to the assessment administrators.

Finally, holistic data will only be stored and used in accordance with the GALW user agreements (school and teacher level). Any change in the use or storage of data would require the gaining of renewed consent from schools/teachers on the basis of the new conditions. Such renewed consent would not act retrospectively on data collected under the old agreements.

9. Concluding Remarks

This specification proposes the creation of a digital self-administered generative assessment for the monitoring of Welsh language development in the EM primary sector. The scoping study (Russell 2025) demonstrates a clear and present need for such a tool, with current assessment practices suffering from bias, inconsistency, inaccuracy and impracticality. It is reasonable to assume that this lack of high-quality assessment is having a negative impact on teaching and learning, as well as limiting the capacity for research into this area. Such effects risk compromising the ability of the EM sector to meet the aspirations of Cymraeg 2050, and potential exacerbate educational inequity.

The GALW does not offer a panacea for Welsh development in EM schools. The challenges faced are complex and multifaceted, and no one provision is capable of having a revolutionary impact in and of itself. However, the GALW's capacity to provide essential data to inform learners, teachers, researchers and policy makers makes it capable of instigating development at multiple levels. Such functionality can compound the impact of the assessment beyond its direct influence on teaching and learning: effecting long-term change in practice, empowering teachers to research and adapt their own pedagogy, facilitating on-going academic research, and informing policy and funding decisions to more effectively support learning. In this way the GALW has the potential to have a profound impact on the success of Welsh language learning in EM settings.

10. Bibliography

Ackerman, P.L. and Kanfer, R. 2009. Test length and cognitive fatigue: an empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied*, *15*(2), p.163.

Akbari, R. 2012. Validity in language testing. *The Cambridge guide to second language assessment*, pp.30-36.

Alderson, J.C. and Wall, D. 1993. Does washback exist?. *Applied linguistics*, *14*(2), pp. 115-129.

Alexiou, T. and Milton, J. 2020. Pic-Lex: A new tool of measuring receptive vocabulary for very young learners. *Zoghbor, W. & Alexiou, T. Advancing ELT Education*, pp. 103-113.

Allen, D. and Tanner, K. 2006. Rubrics: Tools for making learning goals and evaluation criteria explicit for both teachers and learners. *CBE—Life Sciences Education*, *5*(3), pp.197-203.

American Education Research Association. 2014. *Standards for educational and psychological testing*. Washington: AERA

Arens, A.K. and Morin, A.J. 2016. Relations between teachers' emotional exhaustion and students' educational outcomes. *Journal of educational psychology*, *108*(6), pp. 800-813.

Ausubel, D.P., Novak, J.D. and Hanesian, H. 1978. Educational psychology: A cognitive view. London: Holt, Rinehart and Winston.

Ayhan, Ü. and Türkyılmaz, M.U. 2015. Key of language assessment: Rubrics and rubric design. *International Journal of Language and Linguistics*, *2*(2), pp.82-92.

Azam, T.N., Hossain, M.S. and Shah, M.A.H. 2021. Barriers of Achieving Communicative Competence in English Language through the Existing Policy: Time to Research for a New Parameter. *IUB Journal of Social Sciences*, *3*(2), pp. 1-9.

Bachman, L. F., Davidson, F., Ryan, K. and Choi, I-C. 1995. An investigation into the comparability of two tests of English as a foreign language. In: *Language Testing volume 1*. Cambridge: Cambridge University Press

Bachman, L.F. 1990. *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L.F. 1991. What does language testing have to offer?. *TESOL quarterly*, *25*(4), pp. 671-704.

Bachman, L.F. and Palmer, A. S. 1996. *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford: Oxford University Press.

Baddeley, A. 2003. Working memory and language: An overview. *Journal of communication disorders*, *36*(3), pp. 189-208.

Bailey, K.M. 1996. Working for washback: A review of the washback concept in language testing. *Language testing*, *13*(3), pp. 257-279.

Bailey, R. 1999. The abdication of reason: post-modern attacks upon science and reason. In: Pratt, J. ed. *Improving education: Realist approaches to method and research*. London: Cassell

Baldwin, L. 2018. Monitoring and evaluating progress. In *Leading English in the Primary School*. Oxford: Routledge, pp. 158-174.

Ball, M.J. and Müller, N. 2002. *Mutation in Welsh*. Oxford: Routledge.

Ball, S. J. 2003. The Teacher's Soul and the Terrors of Performativity. *Journal of Education Policy, 18*(2), 215–228.

Becker, A., Matsugu, S. and Al-Surmi, M. 2017. Balancing practicality and construct representativeness for IEP speaking tests. *Asian-Pacific Journal of Second and Foreign Language Education*, *2*, pp. 1-16.

Beerepoot, M.T. 2023. Formative and summative automated assessment with multiple-choice question banks. *Journal of Chemical Education*, *100*(8), pp. 2947-2955.

Benigno, V. and De Jong, J. 2019. Linking vocabulary to the CEFR and the Global Scale of English: A psychometric model. *Developments in language education. A memorial volume in honour of Sauli Takala*, pp. 8-29.

Bennett, R.E. 2011. Formative assessment: A critical review. *Assessment in education: principles, policy & practice*, *18*(1), pp. 5-25.

Berman, A.I., Haertel, E.H. and Pellegrino, J.W. 2020. Comparability of Large-Scale Educational Assessments: Issues and Recommendations. Washington: *National Academy of Education*.

Bialystok, E. 2001. *Bilingualism in development: Language, literacy, and cognition*. Cambridge: Cambridge University Press.

Bialystok, E., Craik, F.I. and Luk, G. 2012. Bilingualism: consequences for mind and brain. *Trends in cognitive sciences*, *16*(4), pp. 240-250.

Biesta, G.J. 2004. Education, accountability, and the ethical demand: Can the democratic potential of accountability be regained?. *Educational theory*, *54*(3), pp. 233-250.

Biggs, J. 1996. Enhancing teaching through constructive alignment. *Higher Education* 32, no.3: 347-364.

Binnick, R.I. 2011. *The Oxford handbook of tense and aspect*. Oxford: Oxford University Press.

Black, P. 2003. The Nature and Value of Formative Assessment for Learning. *Improving Schools*, *6*(3), pp. 7-22.

Black, P. and Wiliam, D. 1998. *Inside the black box: Raising standards through classroom assessment*. London: Kings College London.

Black, P. and Wiliam, D. 2009. Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of personnel evaluation in education)*, *21*, pp. 5-31.

Boaler, J. 2014. Research suggests that timed tests cause math anxiety. *Teaching children mathematics*, *20*(8), pp. 469-474.

Boers, F. and Lindstromberg, S. 2012. Experimental and intervention studies on formulaic sequences in a second language. *Annual Review of Applied Linguistics*, *32*, pp. 83-110.

Bolster, A. 2009. Continuity or a fresh start? A case study of motivation in MFL at transition, KS2–3. *Language Learning Journal*, *37*(2), pp. 233-254.

Bolt, D.M., Cohen, A.S. and Wollack, J.A. 2002. Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*(4), pp. 331-348.

Bourdieu, P. 1991. *Language and symbolic power*. Havard: Harvard University Press.

Bourne, J. 2001. Discourses and identities in a multi-lingual primary classroom. *Oxford Review of Education*, *27*(1), pp. 103-114.

Braund, M. 2009. Progression and continuity in learning science at transfer from primary and secondary school. *Perspectives on Education (Primary Secondary Transfer in Science)*, *2*, pp. 22-38.

Brighouse, H. 2006. *On education*. Oxford: Routledge.

Brown, G. and Yule. G.1983. *Discourse analysis*. Cambridge: Cambridge University Press.

Brown, H.D. and Abeywickrama, P. 2019. *Language assessment: Principles and classroom practices*. London: Pearson.

Brown, J. D. 2005. *Testing in language programs: A comprehensive guide to English language assessment*. London: McGraw-Hill.

Brown, J.D. and Hudson, T. 1998. The alternatives in language assessment. *TESOL quarterly*, *32*(4), pp. 653-675.

Brown, R. and Gilman, A. 1960. *The pronouns of power and solidarity*. Indianapolis: Bobbs-Merrill.

Brown, T., Isbel, S., Logan, A. and Etherington, J. 2020. Predictors of academic integrity in undergraduate and graduate-entry masters occupational therapy students. *Hong Kong Journal of Occupational Therapy*, *33*(2), pp. 42-54.

Burgoyne, K., Kelly, J.M., Whiteley, H.E. and Spooner, A. 2009. The comprehension skills of children learning English as an additional language. *British Journal of Educational Psychology*, *79*(4), pp. 735-747.

Campbell, T. 2015. Stereotyped at seven? Biases in teacher judgement of pupils' ability and attainment. *Journal of Social Policy*, *44*(3), pp. 517-547.

Canale, M. and Swain, M. 1980. Theoretical bases of communicative approaches to second language teaching and testing. *Applied linguistics*, *1*(1), pp. 1-47.

Cao, J. and Stokes, S.L. 2008. Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika*, *73*, pp. 209-230.

Cardiff Metropolitan University. 2019. *Undertaking Professional Enquiry: An Introduction for Lead Enquirers*. Cardiff: Cardiff Met. Available at: https://hwb.gov.wales/storage/aeb2810d-f670-4718-87a1-299696ce5156/guide-to-undertaking-professional-enquiry.pdf [Accessed 20th February 2025]

Carless, D. 2004. Issues in teachers' reinterpretation of a task-based innovation in primary schools. *Tesol quarterly*, *38*(4), pp. 639-662.

Carless, D. 2005. Prospects for the implementation of assessment for learning. *Assessment in Education: Principles, Policy & Practice*, *12*(1), pp. 39-54.

Cauley, K.M. and McMillan, J.H. 2010. Formative assessment techniques to support student motivation and achievement. *The clearing house: A journal of educational strategies, issues and ideas*, *83*(1), pp. 1-6.

Celce-Murcia, M., Dörnyei, Z. and Thurrell, S. 1995. Communicative competence: A pedagogically motivated model with content specifications. *Issues in Applied linguistics*, *6*(2), pp. 5-35.

Chen, X., Aryadoust, V., & Zhang, W. 2024. A systematic review of differential item functioning in second language assessment. *Language Testing*, 0(0). https://doi.org/10.1177/02655322241290188

Chen, X., He, J., Swanson, E., Cai, Z. and Fan, X. 2021. Big five personality traits and second language learning: A meta-analysis of 40 years' research. *Educational Psychology Review*, pp. 1-37.

Cheng, L. 2005. *Changing language teaching through language testing: A washback study* (Vol. 21). Cambridge: Cambridge University Press.

Chiang, S.Y. and Mi, H.F. 2011. Reformulation: a verbal display of interlanguage awareness in instructional interactions. *Language Awareness*, *20*(2), pp. 135-149.

Chiu, C.W. and Chen, T.P. 2023. Speech rate and young EFL learners' listening comprehension. *English Language Teaching*, *16*(7), pp. 74-80.

Chomsky, N. 1965/2015. *Aspects of the Theory of Syntax* (No. 11). Massachusetts: MIT press.

Chowdhury, F. 2018. Grade inflation: Causes, consequences and cure. *Journal of Education and Learning*, *7*(6), pp. 86-92.

Cizek, G.J. 1999. *Cheating on tests: How to do it, detect it, and prevent it*. Oxford: Routledge.

Cizek, G.J. and Bunch, M.B. 2007. *Standard setting: A guide to establishing and evaluating performance standards on tests*. London: SAGE Publications Ltd.

Clapham, C. 1996. *The development of IELTS* (Vol. 4). Cambridge: Cambridge University Press.

Coffman, K.B. and Klinowski, D. 2020. The impact of penalties for wrong answers on the gender gap in test scores. *Proceedings of the National Academy of Sciences*, *117*(16), pp. 8794-8803.

Cohen, A. 1994. *Assessing language ability in the classroom (2nd ed.).* Boston: Heinle & Heinle Publishers.

Cohen, A.D. 2011. Second language learner strategies. In *Handbook of research in second language teaching and learning* (pp. 681-698). Routledge.

Cohen, A.S. and Wollack, J.A. 2006. Test Administration, Security, Scoring. *Educational measurement*, p. 355.

Comrie, B. 1976. *Aspect: An introduction to the study of verbal aspect and related problems* (Vol. 2). Cambridge: Cambridge university press.

Conwy County Borough Council. 2023. *Continwwm Iaith Conwy: Ysgolion Categori T2*. Tîm Athrawon Ymgynghorol y Gymraeg Conwy.

Cortina, J.M. 1993. What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology*, *78*(1), p. 98.

Council of Europe. 2001. *The Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Strasbourg: Council of Europe Publishing

Council of Europe. 2025. *Global scale - Table 1 (CEFR 3.3): Common Reference levels*. Available at: https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale?utm_source=chatgpt.com [Accessed 12th February 2025]

Cox, T.L., Bown, J. and Bell, T.R. 2019. In Advanced L2 Reading Proficiency Assessments, Should the Question Language Be in the L1 or the L2?: Does It Make a Difference?. *Foreign language proficiency in higher education*, pp. 117-136.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. 1972. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York: Wiley.

Cronbach, L.J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), pp. 297-334.

CSC Communications. 2024. Continuum of Welsh Language Patterns. Available at: https://www.cscjes.org.uk/repository/resource/f9c95586-92e5-49dc-bb57-91f829263c6f/overview [Accessed 18th February 2025]

Cumming, A. 2001. ESL/EFL instructors' practices for writing assessment: Specific purposes or general purposes?. *Language testing*, *18*(2), pp. 207-224.

Cumming, A. 2012. Validation of Language Assessments. In Chapelle, C.A. ed. *The Encyclopedia of Applied Linguistics*. Hoboken, N.J.: John Wiley and Sons.

Cummins, J. 2000. *Language, power, and pedagogy: Bilingual children in the crossfire*. Clevedon: Multilingual Matters.

Cutler, A. and Clifton, C. 1999. Comprehending spoken language: A blueprint of the listener. *The neurocognition of language*, pp.123-166.

Davis, D., Shaver, P.R. and Vernon, M.L. 2003. Physical, emotional, and behavioral reactions to breaking up: The roles of gender, age, emotional involvement, and attachment style. *Personality and Social Psychology Bulletin*, *29*(7), pp. 871-884.

Dawson, P., Bearman, M., Dollinger, M. and Boud, D. 2024. Validity matters more than cheating. *Assessment & Evaluation in Higher Education*, *49*(7), pp. 1005-1016.

Day, C. and Gu, Q. 2010. *The new lives of teachers*. Oxford: Routledge.

Dell, G.S. 1986. A spreading-activation theory of retrieval in sentence production. Psychological Review, 93(3), p. 283.

DeVellis, R.F. and Thorpe, C.T. 2021. *Scale development: Theory and applications*. London: Sage publications.

Dewey, J. and Tufts, J.H. 1908. *Ethics*. London: G. Bell

Dinçer, B.H., Antonova-Unlu, E. and Kumcu, A. 2022. Assessing the use of multiple-choice translation items in English proficiency tests: The case of the national English proficiency test in Turkey. *Applied Linguistics Review*, *13*(4), pp. 461-475.

Djiwandono, P.I. 2006. Cultural bias in language testing. *TEFLIN Journal*, *17*(1), pp. 81-89.

Dodeen, H. 2008. Assessing test-taking strategies of university students: developing a scale and estimating its psychometric indices. *Assessment & Evaluation in Higher Education*, *33*(4), pp. 409-419.

Donaldson, G. 2015. *Successful Futures: Independent Review of Curriculum and Assessment Arrangements in Wales*. Available at: https://gov.wales/docs/dcells/publications/150225-successful-futures-en.pdf [Accessed 10th January 2025]

Dörnyei, Z. 2014. *The psychology of the language learner: Individual differences in second language acquisition*. Oxford: Routledge.

Dörnyei, Z. and Henry, A. 2022. Accounting for long-term motivation and sustained motivated learning: Motivational currents, self-concordant vision, and persistence in language learning. In *Advances in motivation science* (Vol. 9, pp. 89-134). London: Elsevier.

Douglas, D. 2000. *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.

Douglas, D. 2010. *Understanding Language Testing*. London: Hodder Education.

Douglas, M. 2011. Spoken language assessment considerations for children with hearing impairment when the home language is not English. *Perspectives on Hearing and Hearing Disorders in Childhood*, *21*(1), pp. 4-19.

Drackert, A. 2016. *Validating language proficiency assessments in second language acquisition research: Applying an argument-based approach*. Oxford: Peter Lang Publishing

Dror, I.E., Basola, B. and Busemeyer, J.R. 1999. Decision making under time pressure: An independent test of sequential sampling models. *Memory & cognition*, *27*(4), pp. 713-725.

D'Sa, J.L. and Visbal-Dionaldo, M.L. 2017. Analysis of multiple choice questions: item difficulty, discrimination index and distractor efficiency. *International Journal of Nursing Education*, *9*(3).

Duff, K., Callister, C., Dennett, K. and Tometich, D. 2012. Practice effects: a unique cognitive variable. *The Clinical Neuropsychologist*, *26*(7), pp. 1117-1127.

Earl, L. and Katz, S. 2002. Leading schools in a data-rich world. In: Leatherwood, K. and Hallinger, P. eds. *Second international handbook of educational leadership and administration*. Dordrecht: Springer Netherlands, pp. 1003-1022.

Elliot, M. 2013. Test taker characteristics. In: Geranpayeh, A. and Taylor, L. eds. *Examining Listening: Research Practices in assessing second language listening. Studies in language testing 35*. Cambridge: Cambridge University Press, pp. 36-77.

Ellis, N. C. and Ferreira-Junior, F. 2009. Construction learning as a function of frequency, frequency distribution, and function. *The Modern Language Journal*, 93, 370–385.

Ellis, R. 2005. Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in second language acquisition*, *27*(2), pp. 141-172.

Ellis, R. and Wulff, S. 2019. Second language acquisition. In Dąbrowska, E. and Divjak, D. eds. *Cognitive Linguistics - A Survey of Linguistic Subfields. p.182-207*

Ercikan, K., Asil, M. and Grover, R. 2018. Digital divide: A critical context for digitally based assessments. *Education Policy Analysis Archives*, *26*, pp. 1-24.

Estyn. 2024. The Welsh language in education and training: 2023-24. Available at: https://annual-report.estyn.gov.wales/2023/the-welsh-language-in-education-and-training/ [Accessed 4th February 2025]

Fandiño, Y.J. 2010. Research as a means of empowering teachers in the 21st century. *Educación y educadores*, *13*(1), pp. 109-124.

Farhan, R. 2019. The Benefits of Bilingualism. *Learning to Teach Language Arts, Mathematics, Science, and Social Studies Through Research and Practice*, *8*(1).

Faulconer, E., Griffith, J. and Gruss, A. 2022. The impact of positive feedback on student outcomes and perceptions. *Assessment & Evaluation in Higher Education*, *47*(2), pp. 259-268.

Fernandes, D.C., Nagtegaal, M., Noordzij, G. and Tio, R.A. 2018. Cumulative assessment: Does it improve students' knowledge acquisition and retention?. *Scientia Medica*, *28*(4), p.11.

Field, J. 2010. Listening in the language classroom. *ELT journal*, *64*(3), pp. 331-333.

Field, J. 2013. Cognitive validity. *Examining listening: Research and practice in assessing second language listening*, *35*, pp. 77-151.

Figueras, N., North, B., Takala, S., Verhelst, N. and Van Avermaet, P. 2005. Relating examinations to the Common European Framework: A manual. *Language Testing*, *22*(3), pp.261-279.

Fishman, J.A., Extra, G. and Gorter, D. eds. 2008. *Multilingual Europe: facts and policies*. New York: Mouton de Gruyter.

Fitzpatrick, T., Morris, S., Clark, T., Mitchell, R., Needs, J., Tanguay, E. and Tovey, B. 2018. *Rapid Evidence Assessment: Effective Second Language Teaching Approaches and Methods.* Cardiff: Welsh Government, GSR report number 31/2018. Available at: https://www.gov.wales/sites/default/files/statistics-and-research/2019-06/180607-effective-second-language-treaching-approaches-methods-en.pdf [Accessed 4th February 2025]

Fleckenstein, J., Leucht, M., & Köller, O. 2018. Teachers' Judgement Accuracy Concerning CEFR Levels of Prospective University Students. *Language Assessment Quarterly, 15*(1), 90–101.

Fletcher, M.I. 2018. *Supporting continuity of learning through assessment information sharing during transition: a comparison of early childhood and new entrant teachers' beliefs,*

experiences and practices: a thesis in partial fulfilment of the requirements for the degree of Master of Education at Massey University, New Zealand (Doctoral dissertation, Massey University).

Foley, J. 2019. Issues on assessment using CEFR in the region. *LEARN Journal: Language Education and Acquisition Research Network*, *12*(2), pp. 28-48.

Fotovatnia, Z. and Dorri, A. 2013. Repair Strategies in EFL Classroom Talk. *Theory & Practice in Language Studies (TPLS)*, *3*(6).

Freire, P. 1996. Pedagogy of the oppressed (revised). *New York: Continuum*, *356*, pp. 357-358.

Fulcher, G. 2004. Deluded by artifices? The common European framework and harmonization. *Language Assessment Quarterly: An International Journal*, *1*(4), pp. 253-266.

Fulcher, G. 2010. The reification of the Common European Framework of Reference (CEFR) and effect-driven testing. *Advances in research on language acquisition and Teaching: Selected Papers*, pp. 15-26.

Fulcher, G. 2013. *Practical language testing*. Oxford: Routledge.

Fullan, M. 2009. *The challenge of change: Start school improvement now!*. London: Corwin Press.

Gedera, D. 2023. A holistic approach to authentic assessment. *Asian Journal of Assessment in Teaching and Learning*, *13*(2), pp. 23-34.

Genon, L.J.D. and Torres, C.B.P. 2020. Constructive Alignment of Assessment Practices in English Language Classrooms. *English Language Teaching Educational Journal*, *3*(3), pp. 211-228.

Geranpayeh, A. and Kunnan, A.J. 2007. Differential item functioning in terms of age in the certificate in advanced English examination. *Language Assessment Quarterly*, *4*(2), pp. 190-222.

Ghapanchi, Z., Khajavy, G.H. and Asadpour, S.F. 2011. L2 Motivation and Personality as Predictors of the Second Language Proficiency: Role of the Big Five Traits and L2 Motivational Self System. *Canadian Social Science*, *7*(6), p.148.

Gierl, M.J., Bulut, O., Guo, Q. and Zhang, X. 2017. Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of educational research*, *87*(6), pp. 1082-1116.

Gilavand, A. 2016. Investigating the impact of environmental factors on learning and academic achievement of elementary students. *Health Sciences*, *5*(7S), pp. 360-9.

Girolamo, T., Ghali, S., Campos, I. and Ford, A. 2022. Interpretation and use of standardized language assessments for diverse school-age individuals. *Perspectives of the ASHA special interest groups*, *7*(4), pp. 981-994.

Glaser, R. and Silver, E. 1994. Chapter 9: Assessment, Testing, and Instruction: Retrospect and Prospect. *Review of Research in Education - REV RES EDUC.* 20. 393-419.

Glaser, R., Chudowsky, N. and Pellegrino, J.W. 2001. *Knowing what students know: The science and design of educational assessment*. Washington: National Academies Press.

Glušac, T. and Milić, M. 2022. Quality of written instructions in teacher-made tests of English as a foreign language. *English Teaching & Learning*, *46*(1), pp. 39-57.

Gnambs, T. 2014. A meta-analysis of dependability coefficients (test–retest reliabilities) for measures of the Big Five. *Journal of Research in Personality*, *52*, pp. 20-28.

Gneezy, U., List, J.A., Livingston, J.A., Qin, X., Sadoff, S. and Xu, Y. 2019. Measuring success in education: The role of effort on the test itself. *American Economic Review: Insights*, *1*(3), pp. 291-308.

Golvardi Yazdi, M.S., Haghighat Shoar, S.M., Sobhani, G., Vafi Sani, F., Khoshkholgh, R., Mousavi Bazaz, N. and Mansourzadeh, A. 2021. Factors affecting students' guesswork in multiple choice questions and corrective strategies. *Medical Education Bulletin*, *2*(4), pp. 297-305.

Gordon, A. 2020. Tests as Drivers of Change in Education: Contextualising Washback, and the possibility of Wash-forward. *VNU Journal of Foreign Studies*, *36*(4).

Green, A. 2007. *IELTS washback in context: Preparation for academic writing in higher education* (Vol. 25). Cambridge: Cambridge University Press.

Green, A. 2018. Linking tests of English for academic purposes to the CEFR: The score user's perspective. *Language Assessment Quarterly*, *15*(1), pp. 59-74.

Grosjean, F. 2010. *Bilingual: Life and reality*. Havard: Harvard University Press.

Gruffudd, H. 2000. Planning for use of Welsh by young people. In C. H. Williams. ed. *Language revitalization: Policy and planning in Wales*. Cardiff: University of Wales Press, pp. 173–208.

Gulliksen, H. 1950. *Theory of Mental Tests*. New Jersey: Wiley & Sons.

Gyllstad, H., Vilkaitė, L. and Schmitt, N. 2015. Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL-International Journal of Applied Linguistics*, *166*(2), pp. 278-306.

Haladyna, T.M. 2004. *Developing and validating multiple-choice test items*. Oxford: Routledge.

Hanif, H. 2020. The role of L1 in an EFL classroom. *The Language Scholar*, *8*(2), pp. 54-62.

Harding, S., Morris, R., Gunnell, D., Ford, T., Hollingworth, W., Tilling, K., Evans, R., Bell, S., Grey, J., Brockman, R. and Campbell, R. 2019. Is teachers' mental health and wellbeing associated with students' mental health and wellbeing?. *Journal of affective disorders*, *242*, pp. 180-187.

Hargreaves, D.H. 2011. Leading a self-improving school system. *Nottingham: National College for School Leadership*.

Harlen, W. 2007. *Assessment of Learning*. London: SAGE Publications.

Harlen, W. and James, M. 1997. Assessment and learning: differences and relationships between formative and summative assessment. *Assessment in education: Principles, policy & practice*, *4*(3), pp. 365-379.

Harlen, W., Crick, R.D., Broadfoot, P., Daugherty, R., Gardner, J., James, M. and Stobart, G. 2002. A systematic review of the impact of summative assessment and tests on students' motivation for learning. *EPPI-Centre, University of London*.

Harris, L.R. and Brown, G.T. 2018. *Using self-assessment to improve student learning*. Oxford: Routledge.

Hasrol, S.B., Zakaria, A. and Aryadoust, V. 2022. A systematic review of authenticity in second language assessment. *Research Methods in Applied Linguistics*, *1*(3), p.100023.

Hasselgreen, A. 2013. Adapting the CEFR for the classroom assessment of young learners' writing. *Canadian modern language review*, *69*(4), pp. 415-435.

Hattie, J. 2008. *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Oxford: Routledge.

Heacox, D. 2012. *Differentiating instruction in the regular classroom: How to reach and teach all learners (Updated anniversary edition)*. Minneapolis: Free Spirit Publishing.

Heaton, J.B. 1988. *Writing English language tests*. London: Longman.

Hembree, R. 1988. Correlates, causes, effects, and treatment of test anxiety. *Review of educational research*, *58*(1), pp. 47-77.

Henning, G. 1992. Dimensionality and construct validity of language tests. *Language Testing*, *9*(1), pp. 1-11.

Henriksen, B. 1999. Three dimensions of vocabulary development. *Studies in second language acquisition*, *21*(2), pp. 303-317.

Heritage, M. 2007. Formative assessment: What do teachers need to know and do?. *Phi Delta Kappan*, *89*(2), pp. 140-145.

Heyworth, F. 2013. Applications of quality management in language education. *Language Teaching*, *46*(3), pp. 281-315.

Hidri, S. 2018. Discrete point and integrative testing. In Liontas, J. I. ed. *The TESOL encyclopedia of English language teaching*. Hoboken, N.J.: John Wiley & Sons. *https://doi.org/10.1002/9781118784235*.

Hill, C.J., Scher, L., Haimson, J. and Granito, K. 2023. Conducting Implementation Research in Impact Studies of Education Interventions: A Guide for Researchers. Toolkit. NCEE 2023-005. *National Center for Education Evaluation and Regional Assistance*.

Hingorjo, M.R. and Jaleel, F. 2012. Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. *JPMA-Journal of the Pakistan Medical Association*, *62*(2), p. 142.

Holland, P. W., and Thayer, D. T. 1988. Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun. eds. *Test Validity*, pp. 129–145. Hillsdale, NJ: Erlbaum.

Huddleston, R. and Pullum, G. 2002. The Cambridge grammar of the English language. Cambridge: Cambridge University Press

Hui, B. and Godfroid, A. 2021. Testing the role of processing speed and automaticity in second language listening. *Applied Psycholinguistics*, *42*(5), pp. 1089-1115.

Hulstijn, J.H. 2007. The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal*, *91*(4), pp. 663-667.

Hurley, S. 2008. The shared circuits model (SCM): How control, mirroring, and simulation can enable imitation, deliberation, and mindreading. *Behavioral and brain sciences*, *31*(1), pp. 1-22.

Isaacs, T. 2016. Assessing speaking. *Handbook of second language assessment*, *12*, pp. 131-146.

Isoré, M. 2009. Teacher evaluation: Current practices in OECD countries and a literature review. *OECD Education Working Papers, No. 23, OECD Publishing*.

Jacob, B.A. and Levitt, S.D. 2003. Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics*, *118*(3), pp. 843-877.

Kadir, J.S., Zaim, M. and Refnaldi, R. 2019, February. Developing instruments for evaluating validity, practicality, and effectiveness of the authentic assessment for speaking skill at junior high school. In *Sixth of International Conference on English Language and Teaching (ICOELT 2018)* (pp. 98-105). Dordrecht: Atlantis Press.

Kahn-Horwitz, J. and Goldstein, Z. 2024. English foreign language reading and spelling diagnostic assessments informing teaching and learning of young learners. *Language Testing*, *41*(1), pp. 60-88.

Kamhi-Stein, L.D. 2000. Adapting US-based TESOL education to meet the needs of nonnative English speakers. *TESOL journal*, *9*(3), pp. 10-14.

Kane, M. T. 2006. Validation. In: Brennan R. L. ed. *Educational Measurement* (4th ed.). Westport: Praeger, pp. 17–64

Kasper, G. and Rose, K.R. 2002. Pragmatic development in a second language. *Language Learning, 52(Suppl1), 1–352*.

Kelchtermans, G. 2005. Teachers' emotions in educational reforms: Self-understanding, vulnerable commitment and micropolitical literacy. *Teaching and teacher education, 21(8),* pp. 995-1006.

Khalifa, H. and Weir, C.J. 2009. *Examining reading*. Cambridge: Cambridge University Press.

Kim, J. 2023. Memorizing and fabricating? Uncovering high-stakes writing test preparation. *Language Education & Assessment*. Available at: http://dx.doi.org/10.2139/ssrn.4353938 [Accessed 6th February 2025]

Kintsch, W. 1988. The role of knowledge in discourse comprehension: a construction-integration model. *Psychological review*, *95*(2), pp. 163-182

Kolen, M. J. and Brennan, R.L. 2014. *Test equating, Scaling, and Linking-Methods and practices*. New York: Springer-Verlag

Konrad, E., Holzknecht, F., Schwarz, V. and Spöttl, C. 2018. Assessing writing at lower levels: Research findings, task development locally and internationally, and the opportunities presented by the extended CEFR descriptors. AR-G, 2018(4)

Koretz, D.M. 2008. *Measuring up: What education testing really tells us*. Havard: Harvard University Press.

Kormos, J. 2014. *Speech production and second language acquisition*. Oxford: Routledge.

Kremmel, B., Eberharter, K., Konrad, E., Guggenbichler, E., Moser-Frötscher, D., Ebner, V. and Spöttl, C. 2023. The CEFR Companion Volume: Opportunities and challenges for language assessment. *Didaktik slawischer Sprachen*, pp. 65-83.

Kuder, G. F., and Richardson, M. W. 1937. The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151–160.

Kuiken, F. 2023. Linguistic complexity in second language acquisition. *Linguistics Vanguard*, *9*(s1), pp. 83-93.

Kunnan, A. J. 2004. Test fairness. In M. Milanovic & C. Weir. eds. *Studies in Language Testing 18: European Language Testing in a Global Context*. Cambridge: Cambridge University Press.

Kwok, O.M., Lai, M.H.C., Tong, F., Lara-Alecio, R., Irby, B., Yoon, M. and Yeh, Y.C. 2018. Analyzing complex longitudinal data in educational research: A demonstration with project English Language and Literacy Acquisition (ELLA) data using xxM. *Frontiers in Psychology*, *9*, p. 790.

Lakin, J. 2014. Test directions as a critical component of test design: Best practices and the impact of examinee characteristics. *Educational Assessment, 19*(1), pp. 17–34.

Lam, P. and McNaught, C. 2008. A three-layered cyclic model of e-learning development and evaluation. *Journal of Interactive Learning Research*, *19*(2), pp. 313-329.

Lazaraton, A. 2005. Non-native speakers as language assessors: recent research and implications for assessment practice. In: Taylor, L. and Weir, C. J. eds. *Studies in Language Testing 27: Multilingualism and Assessment*. Cambridge: Cambridge University Press

Lee, S.D., Kuncel, N.R. and Gau, J. 2020. Personality, attitude, and demographic correlates of academic dishonesty: A meta-analysis. *Psychological Bulletin*, *146*(11), p. 1042-1058.

Leung, C. and Lewkowicz, J. 2006. Expanding horizons and unresolved conundrums: Language testing and assessment. *Tesol Quarterly*, *40*(1), pp. 211-234.

Levelt, W.J. 1989. *Speaking: from intention to articulation*. Cambridge, Mass.: MIT press.

Lewis, C. 2025. Email of EAS Welsh Language Continuum Content – KS2. 19 March.

Lewis, H. 2021. The Governance of Language Revitalisation: The Case of Wales. In: Lewis, H., Mcleod, W. eds. *Language Revitalisation and Social Transformation*. London: Palgrave Macmillan, pp. 277-310.

Lewkowicz, J.A. 2000. Authenticity in language testing: some outstanding questions. *Language testing*, *17*(1), pp. 43-64.

Little, D. 2005. The Common European Framework and the European Language Portfolio: Involving learners and their judgements in the assessment process. *Language testing*, *22*(3), pp. 321-336.

Little, D. 2006. The Common European Framework of Reference for Languages: Content, purpose, origin, reception and impact. *Language Teaching*, *39*(3), pp. 167-190.

Littlewood, W. 2004. The task-based approach: Some questions and suggestions. *ELT journal*, *58*(4), pp. 319-326.

Liu, J., Cohen, S.B., Lapata, M. and Bos, J. 2021. Universal discourse representation structure parsing. *Computational Linguistics*, *47*(2), pp. 445-476.

Llurda, E. 2005. *Non-native language teachers: Perceptions, challenges and contributions to the profession* (Vol. 5). New York: Springer Science & Business Media.

Lord, F.M. 1980/2012. *Applications of item response theory to practical testing problems*. Oxford: Routledge.

Lord, F.M. and Novick, M.R. 2008. *Statistical theories of mental test scores*. North Carolina: IAP.

Lovell, A.E. 2023. Towards the language continuum: Definitions and implications for Welsh learners in English-medium education. *Wales Journal of Education*, *25*(1).

Lundy, L. 2007. 'Voice' is not enough: conceptualising Article 12 of the United Nations Convention on the Rights of the Child. *British educational research journal*, *33*(6), pp. 927-942.

Luoma, S. 2009. *Assessing speaking.* Cambridge: Cambridge University Press.

MacDonald, M.C. 2013. How language production shapes language form and comprehension. *Frontiers in psychology*, *4*, p. 226.

Magwene, P. M. 2023. Simulating sample distribution. Available at: https://bio723-class.github.io/Bio723-book/simulating-sampling-distributions.html [Accessed 26th February 2025]

Major, R.C., Fitzmaurice, S.M., Bunta, F. and Balasubramanian, C. 2005. Testing the effects of regional, ethnic, and international dialects of English on listening comprehension. *Language learning*, *55*(1), pp. 37-69.

Marcenaro-Gutierrez, O. and Vignoles, A. 2015. A comparison of teacher and test-based assessment for Spanish primary and secondary students. *Educational Research*, *57*(1), pp. 1-21.

Martone, A. and Sireci, S.G. 2009. Evaluating alignment between curriculum, assessment, and instruction. *Review of educational research*, *79*(4), pp. 1332-1361.

May, S. 2013. *Language and minority rights: Ethnicity, nationalism and the politics of language*. Oxford: Routledge.

McCrae, R.R., Kurtz, J.E., Yamagata, S. and Terracciano, A. 2011. Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and social psychology review*, *15*(1), pp. 28-50.

McDonald, R.P. 2013. *Test theory: A unified treatment*. New York: Psychology Press.

McLaughlin, D.B. 1983. Statistical analysis of uncertainty propagation and model accuracy. In *Uncertainty and Forecasting of Water Quality* (pp. 305-319). Berlin, Heidelberg: Springer Berlin Heidelberg.

McNamara, T.F. 1996. *Measuring second language performance: a new era in language testing*. New York: Longman.

McNamara, T.F. 2000. *Language testing*. Oxford: Oxford University Press.

Meara, P. 2009. Connected words: Word associations and second language vocabulary acquisition. Amsterdam: John Benjamins.

Meara, P. and Milton, J. 2003. *X_Lex, the Swansea Levels Test*. Newbury: Express.

Meara, PM and B Buxton 1987. An alternative to multiple choice vocabulary tests. Language Testing 4(1987), pp. 142-154.

Mella, M.S. and Gutiérrez, C.S. 2023. Language immersion effects in the use of tú and usted by L1-French and L1-European Portuguese learners of Spanish. *Studies in Second Language Acquisition*, *45*(5), pp. 1162-1185.

Messer, M., Brown, N.C., Kölling, M. and Shi, M. 2024. Automated grading and feedback tools for programming education: A systematic review. *ACM Transactions on Computing Education*, *24*(1), pp. 1-43.

Messick, S. 1990. *Validity of test interpretation and use.* Princeton: Education Testing Service.

Messick, S. 1996. Validity and washback in language testing. *Language testing*, *13*(3), pp. 241-256.

Milton, J. 2010. The development of vocabulary breadth across the CEFR levels. *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*, *1*, pp. 211-232.

Milton, J. and Alexiou, T. 2009. Vocabulary size and the common European framework of reference for languages. In *Vocabulary studies in first and second language acquisition: The interface between theory and application* (pp. 194-211). London: Palgrave Macmillan UK.

Murphy, R. and Wyness, G. 2020. Minority Report: the impact of predicted grades on university admissions of disadvantaged groups. *Education Economics*, *28*(4), pp. 333-350.

Nagai, N., Birch, G.C., Bower, J.V. and Schmidt, M.G. 2020. *CEFR-informed learning, teaching and assessment*. Singapore: Springer.

NAHT Cymru. 2024. Pupils with additional needs being 'let down' says NAHT Cymru amid cash call. *NAHT website* 21 May 2024. Available at: https://www.naht.org.uk/About-Us/NAHT-Cymru/ArtMID/606/ArticleID/2426/Pupils-with-additional-needs-being-%E2%80%98let-down%E2%80%99-says-NAHT-Cymru-amid-cash-call [Accessed 20th February 2025]

Nation, I.S.P. 2001. *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I.S.P. 2006. Language education – Vocabulary. In Brown, K. ed. *Encyclopaedia of language and linguistics*. Oxford: Elsevier, Vol. 6, 2nd ed., pp. 494–499.

Nation, I.S.P. 2009. *Teaching ESL / EFL reading and writing*. New York: Routledge.

Nation, I.S.P. 2013. *Teaching & learning vocabulary*. Boston: Heinle Cengage Learning.

National Assembly for Wales. 2001. *The Learning Country: A Comprehensive Education and Lifelong Learning Programme to 2010 in Wales.* Available at: https://www.education-uk.org/documents/pdfs/2001-learning-country-wales.pdf [Accessed 12th March 2025]

Naumann, A., Rieser, S., Musow, S., Hochweber, J. and Hartig, J. 2019. Sensitivity of test items to teaching quality. *Learning and Instruction*, *60*, pp. 41-53.

Nese, J.F., Lai, C.F. and Anderson, D. 2013. *A primer on longitudinal data analysis in education*. University of Oregon: Behavioral Research and Teaching.

Newton, P.E. 2007. Clarifying the purposes of educational assessment. *Assessment in education*, *14*(2), pp. 149-170.

Nicol, D. and Milligan, C. 2006. Rethinking technology-supported assessment practices in relation to the seven principles of good feedback practice. In *Innovative assessment in higher education*. Oxford: Routledge, pp. 84-98.

Norris, J.M. and Ortega, L. 2000. Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language learning*, *50*(3), pp. 417-528.

Norris, J.M. and Ortega, L. 2009. Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied linguistics*, *30*(4), pp. 555-578.

North, B. 2014. *The CEFR in practice* (Vol. 4). Cambridge: Cambridge University Press.

Norton, B. and Toohey, K. 2001. Changing perspectives on good language learners. *TESOL quarterly*, *35*(2), pp. 307-322.

O'Sullivan, B. 2000. *Towards a model of performance in oral language testing*. Doctoral dissertation, University of Reading.

O'Sullivan, B. and Green, A. 2011. Test Taker Characteristics. In: Taylor, L. ed. *Examining speaking: Research and practice in assessing second language speaking*, 30, pp. 36-64

Oberle, E. and Schonert-Reichl, K.A. 2016. Stress contagion in the classroom? The link between classroom teacher burnout and morning cortisol in elementary school students. *Social science & medicine*, *159*, pp. 30-37.

OECD (Organisation for Economic Cooperation and Development). 2012. Equity and quality in education: Supporting disadvantaged students and schools. *OECD Publishing*.

Özcan, M., Yeniçeri, N. and Çekiç, E.G. 2019. The impact of gender and academic achievement on the violation of academic integrity for medical faculty students, a descriptive cross-sectional survey study. *BMC Medical Education*, *19*, pp. 1-8.

Pajares, F. and Schunk, D. 2001. The development of academic self-efficacy. *Development of achievement motivation. United States*, *7*, pp. 1-27.

Papageorgiou, S. 2010. Linking international examinations to the CEFR: The Trinity College London experience. *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual*, pp. 145-158.

Parry, N.M. and Thomas, E.M., 2024. Legitimising the 'bilingual': Identity issues among L2 Welsh-speaking teenagers in English-medium schools in Wales. In *The Minority Language as a Second Language*. Oxford: Routledge, pp. 60-88

Pearson, P.D., Hiebert, E.H. and Kamil, M.L. 2007. Vocabulary assessment: What we know and what we need to learn. *Reading research quarterly*, *42*(2), pp. 282-296.

Pedhazur, E.J. and Schmelkin, L.P. 1991. *Measurement, design, and analysis: An integrated approach*. New York: Psychology Press.

Pekrun, R., Goetz, T. and Perry, R.P. 2005. Achievement emotions questionnaire (AEQ). User's manual. *Unpublished Manuscript, University of Munich, Munich*.

Peña, E.D. and Quinn, R. 1997. Task familiarity: Effects on the test performance of Puerto Rican and African American children. *Language, Speech, and Hearing Services in Schools*, *28*(4), pp. 323-332.

Phalke, S.S., Shrivastava, A. and Sahgal, P. 2023. Identification of digital font size and font type to enhance the attention span of children living with ADHD in a typical learning environment. *International Journal of Visual Design*, *17*(1).

Pickering, M.J. and Garrod, S. 2013. An integrated theory of language production and comprehension. *Behavioral and brain sciences*, *36*(4), pp. 329-347.

Pools, E. and Monseur, C. 2021. Student test-taking effort in low-stakes assessments: Evidence from the English version of the PISA 2015 science test. *Large-Scale Assessments in Education*, *9*, pp. 1-31.

Porter, T.M. 1996. *Trust in numbers: the pursuit of objectivity in science & public life*. Princeton: Princeton University Press.

Poupore, G. 2013. Task motivation in process: A complex systems perspective. *Canadian Modern Language Review*, *69*(1), pp. 91-116.

Purpura, J. 2004. *Assessing grammar*. Cambridge: Cambridge University Press.

Putwain, D.W., Woods, K.A. and Symes, W. 2010. Personal and situational predictors of test anxiety of students in post-compulsory education. *British Journal of Educational Psychology*, *80*(1), pp. 137-160.

Rahman, K.A., Seraj, P.M.I., Hasan, M.K., Namaziandost, E. and Tilwani, S.A. 2021. Washback of assessment on English teaching-learning practice at secondary schools. *Language Testing in Asia*, *11*(1), p.12.

Ramos, R. and Dario, F. 2015. Incidental vocabulary learning in second language acquisition: A literature review. *Profile Issues in Teachers Professional Development*, *17*(1), pp. 157-166.

Rasch, G. 1960/1993. *Probabilistic Models for Some Intelligence and Attainment Tests*. California: MESA Press.

Raudenbush, S.W., Hernandez, M., Goldin-Meadow, S., Carrazza, C., Foley, A., Leslie, D., Sorkin, J.E. and Levine, S.C. 2020. Longitudinally adaptive assessment and instruction increase numerical skills of preschool children. *Proceedings of the National Academy of Sciences*, *117*(45), pp. 27945-27953.

Rauss, K. and Pourtois, G. 2013. What is bottom-up and what is top-down in predictive coding?. *Frontiers in psychology*, *4*, p. 276.

Read, J. 2000. Assessing Vocabulary. Cambridge: Cambridge University Press.

Read, J. 2019. Key issues in measuring vocabulary knowledge. In *The Routledge handbook of vocabulary studies*. Oxford: Routledge, pp. 545-560.

Rea-Dickins, P. 2001. Mirror, mirror on the wall: Identifying processes of classroom assessment. *Language testing*, *18*(4), pp. 429-462.

Rea-Dickins, P. 2004. Understanding teachers as agents of assessment. *Language Testing*, *21*(3), pp. 249-258.

Realyvásquez-Vargas, A., Maldonado-Macías, A.A., Arredondo-Soto, K.C., Baez-Lopez, Y., Carrillo-Gutiérrez, T. and Hernández-Escobedo, G. 2020. The impact of environmental factors on academic performance of university students taking online classes during the COVID-19 Pandemic in Mexico. *Sustainability*, *12*(21), p. 9194.

Reves, T. and Medgyes, P. 1994. The non-native English speaking EFL/ESL teacher's self-image: An international survey. *System*, *22*(3), pp. 353-367.

Rhys, M. and Smith, K. 2022. 'Everything we do revolves around the exam': What are students' perceptions and experiences of learning Welsh as a second language in Wales?. *Wales Journal of Education*, *24*(1).

Robinson, K., Schmidt, T. and Teti, D.M. 2005. Issues in the use of longitudinal and cross-sectional designs. *Handbook of research methods in developmental science*, pp. 1-20.

Roever, C. and McNamara, T. 2006. Language testing: The social dimension. *International Journal of Applied Linguistics*, *16*(2), pp. 242-258.

Rouffet, C., van Beuningen, C. and de Graaff, R. 2023. Constructive alignment in foreign language curricula: an exploration of teaching and assessment practices in Dutch secondary education. *The Language Learning Journal*, *51*(3), pp. 344-358.

Rubio, V.J., Hernández, J.M., Zaldívar, F., Márquez, O. and Santacreu, J. 2010. Can We Predict Risk-Taking Behavior?. *European Journal of Psychological Assessment*, 26 (2)

Russell, P.O. 2024. Automagic Pilot-Summary Report: A process evaluation of the Automagic Welsh language intervention pilot conducted September 2023 to July 2024. Available at: https://orca.cardiff.ac.uk/id/eprint/174145/1/Automagic%20Pilot%20-%20Summary%20Report%20.pdf [Accessed 16th Mary 2025]

Russell, P.O. 2025. The Assessment of Welsh as a Second Language - A report on current Welsh language assessment practices in mainstream English medium primary schools. Available at: The Assessment of Welsh as a Second Langugae - A report on current Welsh language assessment practices in mainstream English medium schools. Russell 2025.pdf [Accessed 29th June 2025]

Russell, M., Madaus, G. and Higgins, J. 2009. *The paradoxes of high stakes testing: How they affect students, their parents, teachers, principals, schools, and society*. Charlotte, NC: IAP Publishing.

Sahlberg, P. 2021. *Finnish lessons 3.0: What can the world learn from educational change in Finland?*. New York: Teachers College Press.

Saunders, C. and Kulchitsky, J. 2021. Enhancing self-administered questionnaire response quality using code of conduct reminders. *International Journal of Market Research*, *63*(6), pp. 715-737.

Scarino, A. 2013. Language assessment literacy as self-awareness: Understanding the role of interpretation in assessment and in teacher learning. *Language testing*, *30*(3), pp. 309-327.

Schmidt, W.H. and Prawat, R.S. 2006. Curriculum coherence and national control of education: issue or non-issue?. Journal of curriculum studies, 38(6), pp. 641-658.

Schmitt, N. 1996. Uses and abuses of coefficient alpha. *Psychological assessment*, *8*(4), pp. 350-353.

Schmitt, N. 2000. *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schunk, D.H. 1991. Self-efficacy and academic motivation. *Educational psychologist*, *26*(3-4), pp. 207-231.

Schunk, D.H. and Mullen, C.A. 2012. Self-efficacy as an engaged learner. In: Christenson, S. L., Reschly, A. L., Wylie, C. eds. *Handbook of research on student engagement*. Boston, MA: Springer US, pp. 219-235

Segalowitz, N. 2010. *Cognitive bases of second language fluency*. Oxford: Routledge.

Senedd Commission. 2024a. *Welsh Language and Education (Wales) Bill: Stage 1 Report*. Available at: https://senedd.wales/media/0lkocuv4/cr-ld16872-e.pdf [Accessed 11th February 2025]

Senedd Cymru. 2024a. *Children, Young People and Education Committee: 26/09/2024*. Available at: https://record.senedd.wales/Committee/14917#A89901 [Accessed 12th February 2025]

Senedd Cymru. 2024b. *Children, Young People and Education Committee: 09/10/2024*. Available at: https://record.senedd.wales/Committee/14918#A89921 [Accessed 12th February 2025]

Senedd Cymru. 2024c. *WLE 07. Response from: Welsh Language Commissioner*. Available at: https://business.senedd.wales/documents/s154162/WLE%207%20Welsh%20Language%20Commissioner.pdf [Accessed 14th March 2025]

Senedd Cymru. 2024d. *WLE 10. Response from: Association of School and College Leaders (ASCL) Cymru*. Available at: https://business.senedd.wales/documents/s154165/WLE%2010%20Association%20of%20School%20and%20College%20Leaders%20ASCL%20Cymru.pdf [Accessed 14th March 2025]

Senedd Cymru. 2024e. *WLE 12. Response from: Association of Directors of Education (ADEW) and Welsh Local Government Association (WLGA)*. Available at: https://business.senedd.wales/documents/s154167/WLE%2012%20Association%20of%20Directors%20of%20Education%20ADEW%20and%20Welsh%20Local%20Government%20Association%20WLGA.pdf [Accessed 14th March 2025]

Senedd Cymru. 2024f. *Children, Young People and Education Committee: 02/10/2024*. Available at: https://record.senedd.wales/Committee/14117#A89883 [Accessed 14th March 2025]

Shaw, S.D. and Weir, C.J. 2007. *Examining writing: Research and practice in assessing second language writing* (Vol. 26). Cambridge: Cambridge University Press.

Shepard, L.A. 2000. The role of assessment in a learning culture. *Educational researcher*, *29*(7), pp. 4-14.

Shohamy, E. 2020. *The power of tests: A critical perspective on the uses of language tests*. Oxford: Routledge.

Shute, V.J. 2008. Focus on formative feedback. *Review of educational research*, *78*(1), pp. 153-189.

Slabakova, R. 2016. *Second language acquisition*. Oxford: Oxford University Press.

Smith, A.P. 2013. Time of day and performance. In: Smith, A. P. and Jones, D. M. eds. *Handbook of human performance*, *3*, pp. 217-236.

Solano-Flores, G. and Trumbull, E. 2003. Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher*, *32*(2), pp. 3-13.

Sperber, D. and Wilson, D. 1986. *Relevance: Communication and cognition* (Vol. 142). Cambridge, MA: Harvard University Press.

Spolsky, B. 1985. The limits of authenticity in language testing. *Language Testing*, *2*(1), pp. 31-40.

Stæhr, L. S. 2008. Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal, 36, 139–152*.

Steinert, S., Krupp, L., Avila, K.E., Janssen, A.S., Ruf, V., Dzsotjan, D., Schryver, C.D., Karolus, J., Ruzika, S., Joisten, K. and Lukowicz, P. 2024. Lessons learned from designing an open-source automated feedback system for stem education. *Education and Information Technologies*, pp. 1-42.

Stoeckel, T., Bennett, P. and Mclean, S. 2016. Is" I Don't Know" a Viable Answer Choice on the Vocabulary Size Test?. *TESOL Quarterly*, *50*(4), pp. 965-975.

Streiner, D.L. 2003. Starting at the beginning: an introduction to coefficient alpha and internal consistency. *Journal of personality assessment*, *80*(1), pp. 99-103.

Suskie, L. 2008. Using assessment results to inform teaching practice and promote lasting learning. In *Assessment, learning and judgement in higher education*. Dordrecht: Springer Netherlands. pp. 1-20

Swann, J. 2003. How Science can Contribute to the Improvement of Educational Practice. *Oxford Review of Education 29(2),* pp. 253–268.

Tarrant, M., Ware, J. and Mohammed, A. M. 2009. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC medical education*, *9*, pp. 1-8.

Taylor, C.S. 2013. *Validity and validation*. New York: Oxford University Press.

Taylor, D. 1994. Inauthentic authenticity or authentic inauthenticity. *TESL-EJ*, *1*(2), pp. 1-11.

Taylor, L. 2013. Introduction. In: Geranpayeh, A. and Taylor, L. eds. *Examining Listening: Research Practices in assessing second language listening. Studies in language testing 35*. Cambridge: Cambrighe University Press

Taylor, L. and Galaczi, E. 2011. Scoring validity. In: Taylor, L. ed. *Examining speaking: Research and practice in assessing second language speaking*, 30, pp. 171-233.

Teng, M.F. and Wu, J.G. 2024. An investigation of learners' perceived progress during online education: Do self-efficacy belief, language learning motivation, and metacognitive strategies matter?. *The Asia-Pacific Education Researcher*, *33*(2), pp. 283-295.

Thomas, H., Duggan, B., McAlister-Wilson, S., Roberts, L,. Sinnema, C., ColeJones, N. and Glover, A. 2023. *Research on the early implementation of Curriculum for Wales: Wave 2 report*. Cardiff: Welsh Government, GSR report number 88/2023

Tight, M. 2024. Challenging cheating in higher education: a review of research and practice. *Assessment & Evaluation in Higher Education*, *49*(7), pp. 911-923.

Tobin, S. and Grondin, S. 2012. Time perception is enhanced by task duration knowledge: Evidence from experienced swimmers. *Memory & cognition*, *40*, pp. 1339-1351.

Tomlinson, C.A. 2001. *How to differentiate instruction in mixed-ability classrooms*. Alexandira, VA: ASCD.

Tomlinson, C.A. and Moon, T.R. 2013. *Assessment and student success in a differentiated classroom*. Alexandira, VA: ASCD.

Torrance, H. 2012. Formative assessment at the crossroads: Conformative, deformative and transformative assessment. *Oxford Review of Education*, *38*(3), pp. 323-342.

Towne, L. and Shavelson, R.J. eds. 2002. *Scientific research in education*. Washington: National Academies Press.

Treffers-Daller, J., Parslow, P. and Williams, S. 2018. Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*, *39*(3), pp. 302-327.

Tsagari, D., Cheng, L. 2017. Washback, Impact, and Consequences Revisited. In: Shohamy, E., Or, I., and May, S. eds. *Language Testing and Assessment. Encyclopedia of Language and Education*. London: Springer.

Tsagari, D., Vogt, K., Froelich, V., Csépes, I., Fekete, A., Green, A., Hamp-Lyons, L., Sifakis, N. and Kordia, S. 2018. Handbook of assessment for language teachers. Oslo: *Erasmus+*.

Tsoy, E., Zygouris, S. and Possin, K. L. 2021. Current state of self-administered brief computerized cognitive assessments for detection of cognitive disorders in older adults: a systematic review. *The journal of prevention of Alzheimer's disease*, *8*(3), pp. 267-276.

United Kingdom Government. 2010. *Equality Act 2010*. Available at: https://www.legislation.gov.uk/ukpga/2010/15/contents [Accessed 20th February 2025]

United Nations. 1989. The United Nations Convention on the Rights of the Child. Available at: https://downloads.unicef.org.uk/wp-content/uploads/2010/05/UNCRC_united_nations_convention_on_the_rights_of_the_child.pdf [Accessed 11th March 2025]

Ushioda, E. and Dörnyei, Z. 2009. Motivation, language identities and the L2 self: A theoretical overview. *Motivation, language identity and the L2 self*, *2*, pp. 1-8.

Van de Pol, J., Volman, M. and Beishuizen, J. 2010. Scaffolding in teacher–student interaction: A decade of research. *Educational psychology review*, *22*, pp. 271-296.

Van Der Linden, W. J. 2005. A comparison of item-selection methods for adaptive tests with content constraints. *Journal of Educational Measurement*, *42*(3), pp. 283-302.

Van Lier, L. 1989. Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly*, *23*(3), pp. 489-508.

van Zeeland, H. 2013. L2 vocabulary knowledge in and out of context: Is it the same for reading and listening?. *Australian review of applied linguistics*, *36*(1), pp. 52-70.

Vandergrift, L. and Goh, C. 2009. Teaching and testing listening comprehension. *The handbook of language teaching*, pp. 395-411.

Viriya, C. and Sapsirin, S. 2014. Gender differences in language learning style and language learning strategies. *Indonesian Journal of Applied Linguistics*, 3(2), pp. 77-88.

Vogt, K. and Tsagari, D. 2014. Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly*, *11*(4), pp. 374-402.

Watts, T. W., Bailey, D.H. and Li, C. 2019. Aiming further: Addressing the need for high-quality longitudinal research in education. *Journal of Research on Educational Effectiveness*, *12*(4), pp. 648-658.

Webb, S. and Nation P. 2017. *How Vocabulary is Learned*. Oxford: Oxford University Press.

Wei, X. 2024. Text-to-Speech technology and math performance: A comparative study of students with disabilities, English Language Learners, and their general education peers. *Educational Researcher*, *53*(5), pp. 285-295.

Weigle, S. C. 2009. *Assessing writing*. Cambridge: Cambridge University Press.

Weir, C. J. 2005. *Language testing and validation: An evidence-based approach*. Hampshire: Palgrave Macmillan.

Welsh Government. 2012. Welsh Language Strategy: Evidence Review. Available at: https://www.gov.wales/sites/default/files/statistics-and-research/2019-08/120301welshlanguageen.pdf [Accessed 20 March 2025]

Welsh Government. 2017. *Cymraeg 2050: A million Welsh Speakers*. Available at: https://www.gov.wales/sites/default/files/publications/2018-12/cymraeg-2050-welsh-language-strategy.pdf [Accessed: 15 June 2024].

Welsh Government. 2018a. Additional Learning Needs and Education Tribunal (Wales) Act 2018. Cardiff: Welsh Government. Available at: https://www.legislation.gov.uk/anaw/2018/2/contents [Accessed 20th February 2025]

Welsh Government. 2018b. *The National Approach to Professional Learning*. Cardiff: Welsh Government.

Welsh Government. 2019. *Statutory assessment arrangements for the Foundation Phase and end of Key Stages 2 and 3*. Available at: https://hwb.gov.wales/api/storage/53889241-676d-4e36-8e69-b9129f8ccf51/statutory-assessment-arrangements-for-the-foundation-phase-and-end-of-key-stages-2-and-3-190909.pdf [Accessed 5th February 2025]

Welsh Government. 2021a. Additional Learning Needs and Education Tribunal (Wales) Act. Available at: https://www.gov.wales/additional-learning-needs-and-education-tribunal-wales-act [Accessed 24th February 2025]

Welsh Government. 2021b. Guidance on school categories according to Welsh-medium provision. Available at: https://www.gov.wales/sites/default/files/publications/2021-12/guidance-on-school-categories-according-to-welsh-medium-provision.pdf [Accessed 20 March 2025]

Welsh Government. 2022a. *Transition from primary to secondary school guidance: Guidance and information for schools to support transition of learners from Year 6 to Year 7.* Available at: https://hwb.gov.wales/curriculum-for-wales/assessment-arrangements/transition-from-primary-to-secondary-school-guidance/#core-content-of-transition-plans [Accessed 12th February 2025]

Welsh Government. 2022b. *The Transition from Primary to Secondary School (Wales) Regulations 2022*. Available at: https://www.legislation.gov.uk/wsi/2022/566/contents/made [Accessed 19th January 2025]

Welsh Government. 2023. *School Workforce Census results: as at November 2023*. Available at: https://www.gov.wales/school-workforce-census-results-november-2023-html?utm_source=chatgpt.com#149252 [Accessed 5th February 2025]

Welsh Government. 2024a. *Supporting learner progression assessment guidance*. Available at: https://hwb.gov.wales/curriculum-for-wales/assessment-arrangements/supporting-learner-progression-assessment-guidance [Accessed 13th December 2024]

Welsh Government. 2024b. *Welsh Language and Education (Wales) Bill: Bill Summary*. Available at: https://business.senedd.wales/documents/s153667/Bill%20Summary%20-%2018%20September%202024.pdf [Accessed 8th February 2025]

Welsh Government. 2024c. *Welsh Language and Education (Wales) Bill: Integrated Impact Assessment.* Available at: https://www.gov.wales/sites/default/files/publications/2024-07/wl-education-bill-ia.pdf [Accessed 12th February 2025]

Welsh Government. 2024d. *Supporting learner progression: assessment guidance*. Available at: https://hwb.gov.wales/curriculum-for-wales/assessment-arrangements/supporting-learner-progression-assessment-guidance/ [Accessed 13th January 2025]

Welsh Government. 2024e. *Language, literacy and communication: Designing your curriculum*. Available at: https://hwb.gov.wales/curriculum-for-wales/languages-literacy-and-communication/designing-your-curriculum [Accessed 18th February 2025]

Whitty, G. 2006. Education(al) research and education policy making: is conflict inevitable?. *British Educational Research Journal 32(2*), pp. 159–176.

Widdowson, H. 2003. *Defining issues in English language teaching*. Oxford: Oxford University Press.

Wiliam, D. 2013. Assessment: The bridge between teaching and learning. *Voices from the Middle*, *21*(2), pp. 15-20.

Williamson, B. 2017. *Big data in education: The digital future of learning, policy and practice*. London: Sage Publications

Winter, P.C. 2010. *Evaluating the Comparability of Scores from Achievement Test Variations*. North Carolina: Council of Chief State School Officers.

Wise, S.L. 2017. Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*, *36*(4), pp. 52-61.

Wise, S.L. and DeMars, C.E. 2006. An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*(1), pp. 19-38.

Wise, S.L. and Kingsbury, G.G. 2022. Performance decline as an indicator of generalized test-taking disengagement. *Applied Measurement in Education*, *35*(4), pp. 272-286.

Wise, S.L. and Kong, X. 2005. Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), pp. 163-183.

WIZERD. 2023. *The WISERD Education Multi-Cohort Study: Key findings from 2012-2022*. Available at: https://wiserd.ac.uk/wp-content/uploads/The-WISERD-Education-Multi-Cohort-Study_E4.pdf [Accessed 4th February 2025]

Yoliando, F.T. 2020. A comparative study of dyslexia style guides in improving readability for people with dyslexia. *International conference of innovation in media and visual design (IMDES 2020)* (pp. 32-37). Dordrecht: Atlantis Press.

Yu, Q. 2022. A review of foreign language learners' emotions. *Frontiers in Psychology*, *12*, 827104.

Yule, G. 2022. *The Study of Language*. Cambridge: Cambridge University Press.

Zabihi, R. 2018. The Role of Cognitive and Affective Factors in Measures of L2 Writing. *Written Communication*, *35*(1), 32-57.

Zimmerman, B.J. 2013. Theories of self-regulated learning and academic achievement: An overview and analysis. *Self-regulated learning and academic achievement*, pp. 1-36.

Zumbo, B.D. and Rupp, A.A. 2004. Responsible Modeling of Measurement Data for Appropriate Inferences: Important Advances. In: Kaplan, D. ed. *Reliability and Validity Theory.* London: Sage Publications, pp. 73-92.

11. Appendix

Item 1 – Examples of the WLC content developed by CSC (CSC Communication 2024), GWE (Conwy County Borough Council 2023),

Example 1 from the CSC WLC Content

Describing and identifying people: Talking about illnesses

Cam Cynnydd 1	Cam Cynnydd 2	Cam Cynnydd 3
Beth sy'n bod? *What's the matter?* **bola tost** *a tummy ache* **Mae bola tost gyda fi.** *I have a tummy ache.* **Dim byd** *Nothing* **Trueni, o diar.** *What a pity, oh dear.*	**Beth sy'n bod?** *What's the matter?* **Mae pen tost a bola tost gyda fi.** *I have a headache and a tummy ache.* **Does dim braich dost gyda fi.** *I don't have a bad arm.* **Dw i'n drist achos mae pen tost gyda fi.** *I'm sad because I have a headache.* **Mae coes dost gyda fi ond does dim llwnc tost gyda fi.** *I have a bad leg but I don't have a sore throat.* **Oes cefn tost gyda ti?** *Do you have a bad back?* **Oes/Nac oes** *Yes, I have / No, I haven't* **Oes, mae pen tost gyda fi.** *Yes, I have a headache.* **Nac oes, does dim bys tost gyda fi.** *No, I don't have a bad finger.*	**Beth sy'n bod?** *What's the matter?* **Mae annwyd arna i.** *I have a cold.* **Mae peswch arna i ond does dim pen tost gyda fi.** *I have a cough but I don't have a headache.* **Oes peswch arnat ti?** *Do you have a cough?* **Oes/Nac oes** *Yes, I have / No, I haven't* **Oes, mae peswch arna i.** *Yes, I have a cough.* **Nac oes, does dim annwyd arna i.** *No, I don't have a cold.* **Nac oes, does dim annwyd arna i, ond mae cefn tost gyda fi.** *No, I don't have a cold, but I have a bad back.*

Example 2 from the GWE WLC Content

Cam Cynnydd 1	Cam Cynnydd 2	Cam Cynnydd 3
Cyfarch a theimladau Bore da Prynhawn da P'nawn da Nos da Hwyl / Hwyl fawr Ta ta Ta ta tan toc Welai di 'fory / Welai chi 'fory. Pwy wyt ti? ..... ydw i. Sut wyt ti heddiw? Rydw / Dw i'n ..... dda iawn diolch/ drist/wedi blino/sâl Beth sy'n bod? Dw i wedi brifo. Wyt ti'n..... drist/wedi blino/sâl? Ydw / Nac ydw Ydy Tedi yn drist? Ydy / Nac ydy **Lliwiau a rhifau** Pa liw ydy hwn? Coch / glas / melyn / oren / gwyrdd / pinc / gwyn /piws /brown /du / llwyd Pa rif ydy hwn?	**Cyfarch a theimladau** **Patrymau cyfarwydd i'r dysgwyr barhau i'w defnyddio.** **Gweler CC 1** **Patrymau i'w cyflwyno yn CC2:** Rydw i wedi cael llond bol. Rydw i bron â llwgu. Rydw i wedi blino'n lân. Rydw i'n hapus fel y gôg. Dydw i ddim yn teimlo'n dda. Wyt ti wedi cael llond bol? ayyb Ydw, rydw i wedi cael llond bol. Nac ydw, dydw i ddim yn hapus fel y gôg. Sut ydych chi heddiw? Rydyn ni'n ..... **Lliwiau a rhifau** **Patrymau cyfarwydd i'r dysgwyr barhau i'w defnyddio.**	**Cyfarch a theimladau** **Patrymau cyfarwydd i'r dysgwyr barhau i'w defnyddio.** **Gweler CC 2** **Patrymau i'w cyflwyno yn CC3:** Rydw i'n teimlo'n gyffrous/nerfus/siomedig/hyderus/swil Rydw i'n edrych ymlaen ..... Dydw i ddim yn edrych ymlaen.... Wyt ti'n teimlo'n nerfus? Ydw, rydw i'n teimlo'n nerfus. *teimlo'n gyffrous I feel excited **Lliwiau a rhifau** **Patrymau cyfarwydd i'r dysgwyr barhau i'w defnyddio.** **Gweler CC 2** **Patrymau i'w cyflwyno yn CC3:**

Example 3 from the EAS WLC Content

Gwybodaeth bersonol / Providing personal information 1st person present tense Person 1af amser presennol Cyfarchion a theimladau / Greetings and feelings			
Cam Cynnydd 1	Cam Cynnydd 2	Cam Cynnydd 3	
• **Bore da** – Good morning • **Prynhawn da** – Good afternoon • **Pwy wyt ti?** Who are you? **... ydw i** I'm ... • **Sut wyt ti? (Shw mae)** How are you? **Da iawn diolch/** Very well thank you **Gweddol -** OK **Wedi blino -** tired **Bendigedig –** fantastic / **Ofnadwy -** Awful • **Sut wyt ti'n teimlo?** How are you feeling? **Hapus -** happy **Trist** – Sad **Dw i'n hapus / drist** I'm happy / sad • **Faint ydy d'oed di?** How old are you? **Dw i'n ... oed** – I'm ... years old	• **Nos da** – Good night • **Hwyl fawr** – Good bye • **Beth yw d'enw di? -** What is your name? **... ydw i** – I'm ... • **Beth yw enw d'ysgol di?** **Dw i'n mynd i Ysgol...** What's the name of your school? I go to ... school • **Sut wyt ti? (Shw mae)** How are you? **Da iawn diolch/** Very well thank you **Gweddol** – Ok / **Wedi blino** – tired / **Bendigedig** – brilliant / **Ofnadwy -** awful • **A ti? –** And you? • **Sut wyt ti'n teimlo?** How are you feeling? **Dw i'n teimlo'n hapus / drist / sâl / nerfus / ofnus / gyffrous** I'm feeling happy / sad / nervous / fearful / excited • **Faint ydy d'oed di?** How old are you? **Dw i'n ... oed** – I'm ... years old • **Pryd mae dy ben-blwydd di?** When is your birthday? Month and date e.g. **Mai 19 / Tachwedd 6**	• **Noswaith dda** – Good evening • **Welai di nes ymlaen** - see you later • **Beth yw d'enw di? -** What is your name? **... ydw i** – I'm ... • **Beth yw enw d'ysgol di?** **Dw i'n mynd i Ysgol...** What's the name of your school? I go to ... school • **Sut wyt ti? (Shw mae)** How are you? **Dw i'n dda iawn diolch/** Very well thank you / **Gweddol** – Ok / **Wedi blino** – tired / **Bendigedig** – brilliant / **Ofnadwy** – awful • **Beth amdanat ti? –** How about you? • **Sut wyt ti'n teimlo?** How are you feeling? **Dw i'n teimlo'n hapus / drist / sâl / nerfus / ofnus / gyffrous** I'm feeling happy / sad / nervous / fearful / excited • **Faint ydy d'oed di?** How old are you? **Dw i'n ... oed** – I'm ... years old • **Pryd mae dy ben-blwydd di?** When is your birthday? **Mae fy mhen-blwydd ym mis ...** My birthday is in the month of ...	**Year 7 patterns** • **Beth ydy dy enw di?** • What is your name? **Fy enw i ydy...** My name is ... • **Faint ydy dy oed di? Dw i'n...oed** • How old ar you? I'm ...years old • **Pryd mae dy ben-blwydd di?** • When is your birthday? **Mae fy mhen-blwydd i ym mis...**

Item 2 – Example Teacher Rubric

-GALW-

Rubric for Teachers

Selecting the quiz content

To help ensure that the quiz only measures content that the learners have covered in class, you will need to select the content included in the quiz. This is supposed to be cumulative, so ensure you include everything they have covered (even in previous year groups). Language needs to be developed from strong foundations, so if your class is performing poorly on patterns they've covered in previous years, it is important to return to those earlier foundational patterns and not feel you must push on with new material.

You can curate the quiz content on your teacher dashboard, where you will also find an instructional video explaining the other features and how you can use the data you collect.

Setting up the Environment

- Try to ensure that the learners have a comfortably, calm and quiet environment in which they can complete the quiz. They should be spaced apart sufficiently to avoid them distracting each other. Ideally learners should have headphones that allow them to hear the audio clearly, but if this is not possible, ensure the environment allows them to hear their own device.

- If a learner is distressed or uncooperative, it is permissible to delay the assessment until they are more composed. This will avoid misleading performances effecting your class data.
- If the learner has an ALN that prevents them from accessing the assessment, additional support should be provided.
- EAL learners who have poor levels of English should not use the GALW, as it will be difficult to distinguish whether their performance is a reflection of their Welsh or English ability.
- Avoid scheduling use of the GALW too close to break/lunch times, as this can cause learners to rush and provide you with misleading scores.

Structure of the Quiz

- The first time your class uses the GALW, you will need to demonstrate the assessment and play the instructional video on the landing page. Invite questions to ensure that learners understand how to complete the quiz.
- You should explain that the quiz is 20 or 30 questions long, and will take approximately 10 minutes. You can explain that their responses will help you understand what they've learned and what you need to revise in their Welsh lessons. You can explain it is important they try their best, so you don't end up repeating things in lessons they already know.
- The GALW can be used around other classroom tasks, but learners should not break away to do other activities during the assessment.
- You should not need to monitor the quiz once learners are familiar with the functionality, but you should be available to assist if learners encounter a technical problem.
- The data collected from the quiz becomes more accurate the more often your learners use the GALW. Whilst it can be used as a summative tool, it is advised you integrate it into your teaching provision regularly (e.g. at the end of each content block) to maximise its efficacy and accuracy.

Feedback and Results

- You can explain to learners that the quiz will tell then what they did well and what they need to focus on learning.
- You will have access to the learners' actual scores through your teacher dashboard. You can use this information to find out:
    - Which learners are falling behind and require extra support with their Welsh
    - What language patterns have your learners retained and which need revision
    - How well do your learners understand the chunks of language within each phrase, allowing them to use them flexibly with our vocabulary
    - Which learners are making progress, stagnating, or regressing over time