

**REPUBLIC OF TURKEY
ÇUKUROVA UNIVERSITY
INSTITUTE OF SOCIAL SCIENCES
DEPARTMENT OF ENGLISH LANGUAGE TEACHING**

**A CASE STUDY ON HOW TO EXPLORE AND PROFESSIONALLY
DEVELOP IRAQI TEST WRITERS' LANGUAGE ASSESSMENT LITERACY
THROUGH LESSON STUDY MODEL**

Attaa GALALI

Ph.D DISSERTATION

ADANA / 2025

**REPUBLIC OF TURKEY
ÇUKUROVA UNIVERSITY
INSTITUTE OF SOCIAL SCIENCES
DEPARTMENT OF ENGLISH LANGUAGE TEACHING**

**A CASE STUDY ON HOW TO EXPLORE AND PROFESSIONALLY
DEVELOP IRAQI TEST WRITERS' LANGUAGE ASSESSMENT LITERACY
THROUGH LESSON STUDY MODEL**

Attaa GALALI

Supervisor: Prof. Dr. Zuhall OKAN

Member of Examining Committee: Prof. Dr. Cem CAN

Member of Examining Committee: Prof. Dr. Glden İLIN

Member of Examining Committee: Doç. Dr. Fadime YALÇIN ARSLAN

Member of Examining Committee: Dr. Öğr. Üyesi Nebahat BADEM

Ph.D DISSERTATION

ADANA / 2025

ETİK BEYANI

Çukurova Üniversitesi Sosyal Bilimler Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmada;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmada yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde ve ortaya çıkan sonuçlarda herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim. / / 2025

Attaa GALALI

ÖZET

DERS İNCELEMESİ MODELİ YOLUYLA IRAKLI TEST HAZIRLAYICILARININ DİL DEĞERLENDİRME OKURYAZARLIĞININ KEŞFİ VE MESLEKİ GELİŞİMİNE YÖNELİK BİR DURUM ÇALIŞMASI

Attaa GALALI

Doktora Tezi, İngiliz Dili Eğitimi Anabilim Dalı

Danışman: Prof. Dr. Zuhul OKAN

Eylül 2025, 191 sayfa

Bu nitel durum çalışması, Iraklı İngilizce öğretim görevlilerinin, Ders İncelemesi (Dİ) olarak bilinen iş birliğine dayalı mesleki gelişim modeli aracılığıyla Dil Değerlendirme Okuryazarlığını (DDO) nasıl geliştirdiklerini araştırmaktadır. DDO'nin önemi dünya genelinde giderek daha fazla kabul görse de, Irak'ta yapılandırılmış eğitim olanaklarının sınırlı olması, geleneksel ve daha az etkili değerlendirme uygulamalarının sürmesine yol açmaktadır.

Çalışmada Dİ, ortak planlama, sınıf temelli inceleme ve yansıtıcı diyalogu teşvik eden uygulama orally bir yaklaşım olarak ele alınmaktadır. On üniversite düzeyinde Yabancı Dil Olarak İngilizce (YDİ) eğitmeni, planlama, uygulama ve yansıtma aşamalarını içeren Dİ sürecine katılmıştır. Katılımcılar birlikte bir değerlendirme görevi tasarlamış, uygulamış ve gelen öğrenci ve meslektaş geri bildirimleri doğrultusunda revize etmişlerdir.

Veriler, yarı yapılandırılmış mülakatlar, odak grup görüşmeleri, yansıtıcı sınıf gözlemleri, gözlemler ve belge analizi yoluyla toplanmıştır. Tematik analiz sonucunda DDO gelişiminin dört alanda belirginleştiği görülmüştür: (1) geçerlilik, güvenilirlik ve uyum bilgisi; (2) otantik ve iletişimsel değerlendirme tasarımı; (3) test uygulamalarına yönelik eleştirel farkındalık; (4) yerel bağlama duyarlılık.

Dİ süreci, mesleki öz yeterlik duygusunu artırmış ve meslektaşlar arası etkileşimi teşvik ederek fikir alışverişine olanak sağlamıştır. Katılımcıların büyük çoğunluğu, izole ve prosedürel yaklaşımlardan daha yansıtıcı ve pedagojik temelli yöntemlere geçtiklerini belirtmiştir. Bulgular, Dİ'nin sınırlı kaynaklara sahip ortamlarda DDO gelişimi için uygulanabilir ve sürdürülebilir bir model sunduğunu göstermektedir. Modelin öğretmen

eđitimi programlarına entegrasyonu ve uzun vadeli etkileri üzerine ileri arařtırmalar da önerilmektedir.

Anahtar kelimeler: Dil deęerlendirme okuryazarlıęı, ders incelemesi ve öęretmen mesleki gelişimi.



ABSTRACT**A CASE STUDY ON HOW TO EXPLORE AND PROFESSIONALLY
DEVELOP IRAQI TEST WRITERS' LANGUAGE ASSESSMENT LITERACY
THROUGH LESSON STUDY MODEL****Attaa GALALI****PhD Thesis, Department of English Language Teaching****Supervisor: Prof. Dr. Zuhail OKAN****September 2025, 191 pages**

This qualitative case study explores how Iraqi English as a Foreign Language (EFL) test writers develop their Language Assessment Literacy (LAL) through a collaborative professional development model known as Lesson Study (LS). Although the importance of LAL is increasingly recognised worldwide, structured training opportunities in Iraq remain scarce, often resulting in the continued use of conventional and less effective assessment practises.

The study examines LS as a context-sensitive, practise-based framework that fosters collective lesson planning, classroom-based inquiry, and reflective dialogue. Ten university-level EFL instructors engaged in three main stages of LS: planning, implementation, and reflection. The participants jointly designed a classroom-oriented assessment task, implemented it in real teaching environments, and refined it based on feedback from students, peers, and collaborative reflection.

Data were collected through semi-structured interviews, focus groups, reflective journals, classroom observations, and document analysis. Thematic analysis revealed notable development in participants' LAL across four domains: (1) technical competence, including enhanced understanding of validity, reliability, and alignment; (2) practical skills in designing more authentic, communicative assessments; (3) critical awareness in challenging previously held assumptions about testing; and (4) contextual sensitivity in adapting assessment practises to local constraints.

The LS process encouraged professional agency and fostered collegial engagement, enabling the exchange of ideas and enhancing participants' confidence in their assessment practises. Many reported a clear transition from isolated, procedural

approaches to more reflective, pedagogically informed methods. The findings suggest that Lesson Study offers a feasible and sustainable model for LAL development, particularly in resource-constrained environments. The study recommends integrating LS into formal teacher development programmes and encourages further research into its broader application and long-term outcomes.

Keywords: Language assessment literacy, lesson study & teacher professional development.



ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my academic supervisor, Prof. Dr. Zuhale OKAN, for her invaluable guidance, support, and encouragement throughout the course of this study. Her insight, constructive feedback, and continuous motivation have played a crucial role in shaping the direction and completion of this thesis. I am sincerely thankful for her patience, professionalism, and unwavering support at every stage of the research.

I would also like to extend my sincere thanks to the esteemed members of my thesis jury: Prof. Dr. Cem CAN and Assoc. Prof. Dr. Fadime YALÇIN ARSLAN for their thoughtful comments, insightful suggestions, and valuable time during the evaluation of this work.

My heartfelt appreciation goes to the teachers who participated in this study. Their openness, time, and contributions were essential for the success of this research. Without their willingness to engage in the process and share their experiences, this study would not have been possible.

Lastly, I wish to express my deepest gratitude to my family. I am forever thankful to my parents for their love, encouragement, and for being a constant source of strength and belief in my abilities. A special thanks goes to my dear wife for her endless support, patience, and understanding throughout this journey. Her unwavering encouragement and sacrifices have made the completion of this thesis possible.

Attaa GALALI

Adana / 2025

TABLE OF CONTENTS

	Page
ÖZET	iv
ABSTRACT.....	vi
ACKNOWLEDGEMENTS	viii
ABBREVIATIONS.....	xiii
LIST OF TABLES	xiv
LIST OF FIGURES	xvi
LIST OF APPENDICES	xvii

CHAPTER I INTRODUCTION

1.1. Introduction.....	1
1.2. Background of the Study	1
1.3. Statement of the Problem.....	4
1.4. Aim of the Study.....	7
1.5. Significance of the Study.....	7
1.6. Assumptions.....	8
1.7. Limitations	8
1.8. Chapter Summary	9

CHAPTER II LITERATURE REVIEW

2.1. Introduction.....	10
2.2. Theoretical Framework.....	10
2.3. A Brief History of Approaches and Theories of Language Testing and Assessment	12
2.4. Basic terms: Assessment, Measurement, Testing and Evaluation.....	14
2.5. Principles of language assessment.....	16
2.5.1. Validity	16
2.5.2. Reliability	18
2.5.3. Practicality	19

2.5.4. Authenticity	20
2.5.5. Washback.....	20
2.6. Assessment Purposes	21
2.7. Definitions and significance of Language Assessment Literacy (LAL).....	22
2.8. A Review of Studies on Teachers' Language Assessment Literacy	25
2.9. Training for Language Assessment Literacy (LAL).....	28
2.9.1. Research on LAL training	29
2.10. An Overview of Lesson Study Model	31
2.11. Benefits of lesson study implementation	33
2.12. Difficulties in Implementing Lesson Study	35
2.13. Chapter Summary	36

CHAPTER III METHODOLOGY

3.1. Introduction.....	37
3.2. Research Context	38
3.3. Research Design	41
3.4. Participants.....	43
3.4.1. Teachers	43
3.4.2. Researcher.....	44
3.4.3. Experts	44
3.4.4. Qualitative Data Analysis Expert	45
3.5. Data Collection Tools	45
3.5.1. Language Assessment Literacy Questionnaire (LALQ).....	45
3.5.2. Semi-Structured Interviews	48
3.5.3. Document Analysis.....	51
3.5.4. Reflective Journal	51
3.6. Data Collection Procedure	53
3.6.1. Participants' Selection.....	53
3.6.2. Pre-application of the Questionnaire	53
3.6.3. Theoretical Phase.....	54
3.6.3.1. Session One: Introduction to Lesson Study Model	54

3.6.3.2. Session Two: Understanding Collaborative Inquiry in Lesson Study.....	54
3.6.3.3. Session Three: Implementing Lesson Study for Professional Learning Communities	55
3.6.3.4. Session Four: Reflective Practise and Continuous Improvement in Lesson Study	55
3.6.4 Practical Phase	55
3.6.4.1 Administration of the Pre-Interview Questions	56
3.6.4.2. Introducing the Reflective Journal to the Participants	56
3.6.4.3. Practical Implementation of the Lesson Study Model	56
3.6.4.4. Post-application of the Questionnaire	66
3.6.4.5. Administration of the Post-Interview Questions	66
3.7. Data Analysis	67
3.8. Chapter Summary	68

CHAPTER IV

FINDINGS

4.1. Introduction.....	69
4.2. Addressing the First Research Question	69
4.3 Addressing the Second Research Question	91
4.4 Addressing the Third Research Question	104
4.4.1. Theme 1: Test Design Improvement	104
4.4.2. Theme 2: Collaboration Enhancement	106
4.4.3. Theme 3: Student Engagement Improvement	108
4.4.4. Theme 4: Improved Reflection	110
4.4.5. Theme 5: Sustainability in Test Development.....	111
4.5. Addressing the Fourth Research Question.....	115
4.5.1. Theme 1: Pedagogical Knowledge Improvement.....	115
4.5.2. Theme 2: Professional Growth	116
4.5.3. Theme 3: Peer Learning.....	118
4.5.4. Theme 4: Enhanced Assessment Literacy	118
4.5.5. Theme 5: Professional Networking	119
4.5.6. Theme 6: Empowerment.....	120

4.7. Chapter Summary	127
----------------------------	-----

CHAPTER V

DISCUSSION

5.1. Introduction.....	128
5.1.1. Discussing the First Research Question	129
5.1.2. Discussing the Second Research Question	133
5.1.3 Discussing the Third Research Question.....	136
5.1.4. Discussing the Fourth Research Question	139
5.2. Contribution to Literature	143
5.3. Chapter Summary	144

CHAPTER VI

CONCLUSION AND RECOMMENDATIONS

6.1. Introduction.....	146
6.2. Summary of the Study	146
6.3. An Overview of the Study	147
6.3.1. Research Question 1	148
6.3.2. Research Question 2	149
6.3.3. Research Question 3	150
6.3.4. Research Question 4	151
6.4 Implications of the Study.....	151
6.4.1. Implications for Teacher Professional Development	152
6.4.2. Implications for Curriculum Designers and Policy Makers	152
6.4.3. Implications for Language Assessment Practices.....	153
6.4.4. Implications for Institutional Leadership and Resource Allocation	153
6.5. Recommendations for Future Research.....	153
6.6. Researcher's Reflections.....	154
6.7. Chapter Summary	155
REFERENCES.....	157
APPENDICES.....	178
CURRICULUM VITAE.....	191

ABBREVIATIONS

AA: Alternative Assessment

Abbreviation: Full Form

CBT: Computer-Based Testing

EFL: English as a Foreign Language

LA: Language Assessment

LAL: Language Assessment Literacy

LS: Lesson Study

TALQ: Teacher Assessment Literacy Questionnaire



LIST OF TABLES

	Page
Table 1. Demographic Features of the Participants	43
Table 2. Reliability Coefficients for Questionnaire Sections.....	48
Table 3. The academic and professional needs of teachers	69
Table 4. First Case's Training in English Language Testing and Evaluation.....	72
Table 5. Second Case's Training in English Language Testing and Evaluation	73
Table 6. Third Case's Training in English Language Testing and Evaluation	74
Table 7. Fourth Case's Training in English Language Testing and Evaluation	75
Table 8. Fifth Case's Training in English Language Testing and Evaluation	76
Table 9. Sixth Case's Training in English Language Testing and Evaluation.....	77
Table 10. Seventh Case's Training in English Language Testing and Evaluation	78
Table 11. Eighth Case's Training in English Language Testing and Evaluation	79
Table 12. Ninth Case's Training in English Language Testing and Evaluation.....	80
Table 13. Tenth Case's Training in English Language Testing and Evaluation.....	81
Table 14. Summary of the Ten Cases for the Training in English Language Testing and Evaluation.....	82
Table 15. Summary of the Five PhD Holders in Terms of Training in English Language Testing and Evaluation	83
Table 16. Summary of the Five MA Holders in Terms of Training in English Language Testing and Evaluation	84
Table 17. Summary of the Ten Cases for the Academic and Professional Needs of Teachers as Test Writers in Terms of Language Assessment Literacy	85
Table 18. Summary of the Five PhD Holders for the Academic and Professional Needs of Teachers as Test Writers in Terms of Language Assessment Literacy.....	87
Table 19. Summary of the Five MA Holders for the Academic and Professional Needs of Teachers as Test Writers in Terms of Language Assessment Literacy.....	89
Table 20. Results for the Ten Cases in Terms of Listening Assessment Knowledge....	92
Table 21. Results for the PhD Cases in Terms of Listening Assessment Knowledge...	93
Table 22. Results for the MA Cases in Terms of Listening Assessment Knowledge ...	94
Table 23. Results for the Ten Cases in Terms of Reading Assessment Knowledge	95
Table 24. Results for the PhD Cases in Terms of Reading Assessment Knowledge.....	96
Table 25. Results for the MA Cases in Terms of Reading Assessment Knowledge	97

Table 26. Results for the Ten Cases in Terms of Speaking Assessment Knowledge	98
Table 27. Results for the PhD Cases in Terms of Speaking Assessment Knowledge ...	99
Table 28. Results for the MA Cases in Terms of Speaking Assessment Knowledge ..	100
Table 29. Results for the Ten Cases in Terms of Writing Assessment Knowledge	101
Table 30. Results for the PhD Cases in Terms of Writing Assessment Knowledge ...	102
Table 31. Results for the MA Cases in Terms of Writing Assessment Knowledge	103
Table 32. Summary of the Benefits of the 'Lesson Study' Model.....	104
Table 33. Test Design Improvement: Related Sub-Themes and Codes.....	105
Table 34. Collaboration Enhancement: Related Sub-Themes and Codes.....	106
Table 35. Student Engagement Improvement: Related Sub-Themes and Codes.....	109
Table 36. Improved Reflection: Related Sub-Themes and Codes	110
Table 37. Sustainability in Test Development: Related Sub-Themes and Codes.....	112
Table 38. Summary of the Results for the Ten Participants for the Benefits of the 'Lesson Study' Model.....	113
Table 39. Summary of the Results for the Five PhD Participants for the Benefits of the 'Lesson Study' Model.....	113
Table 40. Summary of the Results for the Five MA Participants for the Benefits of the 'Lesson Study' Model.....	114
Table 41. Pedagogical Knowledge Improvement: Related Sub-Themes and Codes...	115
Table 42. Professional Growth: Related Sub-Themes and Codes	117
Table 43. Peer Learning: Related Sub-Themes and Codes.....	118
Table 44. Enhanced Assessment Literacy: Related Sub-Themes and Codes	119
Table 45. Professional Networking: Related Sub-Themes and Codes	120
Table 46. Empowerment: Related Sub-Themes and Codes.....	121
Table 47. Summary of the Results of the Ten Participants for the Contributions of Lesson Study Model Towards Test Writers' Professional Development....	122
Table 48. Summary of the Results of the Five PhD Participants for the Contributions of Lesson Study Model Towards Test Writers' Professional Development....	123
Table 49. Summary of the Results of the Five MA Participants for the Contributions of Lesson Study Model Towards Test Writers' Professional Development....	124

LIST OF FIGURES

	Page
Figure 1. Lesson Study Process (Dudley, 2014).....	61
Figure 2. The most common terms among participants' reflections and responses over Lesson Study cycles	108
Figure 3. The density map representing the importance of terms according to their frequency	111
Figure 4. The network map	126



LIST OF APPENDICES

	Page
APPENDIX A. Language Assessment Literacy Questionnaire (LALQ)	178
APPENDIX B. Semi Structured Interviews (SSI)	187
APPENDIX C. Reflective Journal on the implementation of Lesson Study Model as a tool to Professionally Develop Test Writers' Language Assessment Literacy	189



CHAPTER I

INTRODUCTION

1.1. Introduction

This chapter introduces the Lesson Study Model (LS) as a tool for investigating language assessment literacy (LAL) development among Iraqi test writers during exploration and professional development. It also deals with the study's background information, statement of problems and objectives, overall significance, and assumptions. The study's limitations are also presented at the end of the chapter.

1.2. Background of the Study

Testing and assessment are integral parts of any curriculum. A teacher's ability to interact effectively with different teaching-learning-assessment practises is crucial since it enables them to use assessment for support (Berry et al., 2019). Despite these facts, Fulcher (2012) raises concerns over the sufficiency and quality of teacher training concerning evaluation. Fulcher highlights that many teacher education programs fail to address the complex conceptual and practical knowledge required for effective assessment. He argues that true assessment literacy goes beyond learning how to use tests; it involves understanding the broader implications of assessment for learning, fairness, and decision-making in the classroom.

Tests are increasingly important and widely used across educational systems. Studies indicate that teachers spend a considerable amount of time on them (Popham, 2009), making it clear that assessments cannot be ignored. In other words, teachers need strong foundations in practicing assessment; they should have skills necessary for assessment literacy at large, besides treating these as its essential ingredients. Both formative and summative assessments play key roles in measuring a teacher's instructional effectiveness as well as learners' achievements (Evans, 2013).

The concept of 'Assessment Literacy' (AL), initially introduced by Stiggins (1991) which refers to understanding about what needs to be done when assessing somebody else's performance or product; knowledge/skills required by those engaged in such practises, either teachers or administrators. According to Taylor (2013), involvement in language assessment (LA) should not only be limited to language teachers themselves

but extend further than this point too because even other stakeholders like school heads, parents, and politicians need awareness on LA so that they can appreciate its implications better especially while making decisions based upon scores.

The language testing domain has gone through significant changes throughout its history. In the past, "testing" was commonly used and favoured as the term that brought about different changes in the LA landscape. Researchers have identified many historical phases of language testing. Spolsky (1978) stated that three significant shifts in language testing and assessment history were noted, and these include: the pre-scientific era, the scientific era characterized by psychometric-structuralist approaches, and the Integrative Sociolinguistic era. These changes are also reflected in the evolution of teaching methodology.

The earliest of these, the pre-scientific era existed with its sharing informal and unsystematic appraisal of language from the late 19th century to the early 20th century. At that time, teachers would utilize their intuition. They may well have just thought an opinion or charged their charges to compose an essay or responded to some oral questions. Test methods were never really considered. Unauthorized was never introduced unlike standardized testing to insure consistency or fairness. Despite all its various weaknesses, there was the realization that language learning had to be evaluated, even though the tools and techniques were somewhat primitive in terms of evaluation.

With the growth of structuralist linguistics and behaviourist psychology, language testing entered the scientific or psychometric-structuralist period in the 1940s through the 1960s. This shift in paradigm entailed a methodological change by emphasizing objectivity in the measurement of language ability through standardized tests. Language was conceived as a system of discrete components-grammar, vocabulary, and phonology-that could be tested individually by using such separate techniques as multiple-choice questions. Test development was seen as a procedure to be carried out under the psychometric principles of reliability, validity, and item analysis so that the test would be statistically rigorous. On the one hand, it helped it to be formalized and put to large-scale use, but in time, language testing was criticized for ignoring communicative competence and for being narrowly focused on isolated linguistic components.

Addressing what was considered the shortcomings of the scientific era. This period shifted into what Spolsky (1978) called the integrative sociolinguistic era during 1970s and 1980s. This era was influenced by the spread of sociolinguistic theories and the development of communicative approaches to language teaching. Assessment began

to be construed as the evaluation of language used for meaningful communication. Test instruments were then designed with language tasks that closely simulated authentic language use within real-world contexts, the so-called "communicative tasks" such as interviews or group discussion, and integrated writing assignments. Assessment took on a more holistic view, concerned with discourse-level abilities and communicative effectiveness instead of isolated grammatical knowledge. With further pushes toward formative assessment and learning-oriented evaluation, the integrative sociolinguistic era helped secure language testing in the field of language teaching and learner development.

Following the historical developments in language testing, recent debates have also emerged. Additionally, Brown (2004) has raised some concerns regarding current debates within the LA field. He mentions two main ones: computer-based testing (CBT) and alternative assessment (AA). It is clear that traditional ways of doing LA, like pen-and-paper tests that largely depend on numerical scores, are inadequate indicators of student progress.

Brown (2004) highlights two major developments in the field of language assessment: CBT and AA. He argues that while CBT offers benefits such as efficiency and immediate feedback, it also raises issues related to fairness, technological access, and the authenticity of language tasks. Similarly, Brown (2004) emphasizes that AA shifts focus from discrete-point testing to more authentic and process-oriented evaluations. These include portfolios, projects, and self-assessments, which require teachers to adopt more reflective and analytical roles in assessment design. According to Brown (2004), both CBT and AA demand a higher level of AL, as teachers must carefully select, implement, and interpret various assessment tools to align with learners' needs and instructional goals.

Besides, Mertler (2003) argues that alternative methods of evaluating students' knowledge and skills are different from traditional tests because they require learners to demonstrate abilities that cannot be measured by means of multiple-choice items. In addition, it also assesses not only the product but also the process through such forms as projects, posters, portfolios, self-evaluation sheets etc. At the same time, formal examinations are mainly used for this purpose in conventional assessment systems. Such an approach allows teachers to incorporate various types of evaluation into their lessons and have precise results at the same time. Furthermore, according to Yildirim (2013), individualization is one of its main characteristics – every student needs a unique approach; therefore, active learning becomes possible when children know what is

expected from them in terms of outcome or output produced during instruction phases; finally, inclusiveness means that all youngsters have equal opportunities for success regardless of their background or abilities.

The importance of LAL in second language education cannot be undervalued. Lado (1961) stressed the essentiality for teachers to possess adequate comprehension and hands-on experience in assessment to evaluate their language learning accurately. It allows teachers to understand, interpret, and apply knowledge regarding student achievement, as highlighted by Scarino (2013). Scarino further stated that possessing a strong level of LAL is critical as it allows English language teachers to establish a connection between language teaching methodologies and assessment practises, enabling them to choose and create suitable assessments from various options.

On the other hand, teacher development models vary across different contexts worldwide, and this diversity is reflected in the practises of teacher training. The professional development activities for teachers within these models encompass a range of practises, including pre-service Bachelor's-level education, participation in workshops and professional meetings, and informal endeavours such as exploring field-specific publications (Jones & Dexter, 2014). Additionally, in-service training and short-term courses are common opportunities for introducing teachers to the latest advancements in their respective fields. However, it is often observed that these activities tend to overlook the unique needs and contexts of individual participants.

Relatively, Lesson Study (LS) is an approach of professional development that promotes reflection skills of language teachers on their instructional methods through a periodic procedure of planning a lesson in collaboration, then observing the planned lesson and analysing students' comprehension and learning, initially originating in Japan (Gutierrez 2016). It enhances essential reflective abilities, which are significant for the professional development of teachers, while it also embodies the principles of social constructivism. This pedagogical approach emphasizes the significance of fostering a collaborative environment where students can actively exchange, explore, and develop innovative ideas based on classroom practises, thereby fostering both learning and professional advancement (Fernandez and Yoshida 2004).

1.3. Statement of the Problem

Language assessment literacy is increasingly recognized as a vital teaching competency. In the field of English as a Foreign Language (EFL) instruction, educators

must possess advanced literacy skills and engage in continuous professional development. Teachers must be literate in language assessment in order to boost the learning process (DeLuca & Klinger, 2010). They also mentioned that EFL teachers' illiteracy in LA might affect the quality of teaching as they are closely interrelated. To them, teachers have limited knowledge of assessment literacy since they had no or ineffective training in LA.

Starkey et al. (2009) suggest that professional development programmes hold equal significance for both novice and experienced teachers, as highlighted by their research findings. In addition, Rogers et al. (2007) stressed the fact that the role of veteran teachers in participating in professional development programmes for self-renewal is similar to novice teachers, since both are required to develop their skills and acquire new teaching strategies. This is the issue that Johnson (2009) calls "Teachers as Learners of Teaching."

According to Stiggins (1991), LAL is the essential knowledge and abilities that educators and other involved parties (stakeholders) require to manage the assessment process in language teaching effectively. Relatively, teachers need to take some roles in assessment activities. Inbar-Lourie (2017) stated that teachers need to be aware of the terms and concepts of the uses of assessment as teaching and assessment are two interrelated and complementary activities. In addition, according to Fulcher (2012), three factors affect AL; two external factors and one internal factor. The two external factors influencing AL are, first, the widespread use of large-scale assessments across educational systems, and second, the increasing role of language testing in immigration policies and decision-making. These external pressures have heightened the demand for valid and reliable language assessments, placing additional expectations on language teachers to understand and align with national and international standards. The internal factor, on the other hand, stems from the evolving role of assessment within the classroom itself. Assessment is no longer seen as a separate or final step in instruction but as an integral part of the teaching and learning process — what Fulcher (2012) refers to as “assessment for learning.” This shift emphasizes the use of assessment to support student progress and instructional improvement, reinforcing the need for teachers to be both knowledgeable and reflective in their assessment practises.

Therefore, many studies have been carried out to stress the significance of language teachers' assessment literacy and its impact on the teaching process. The reason for such focus is grounded in the widely acknowledged notion that teachers' knowledge

and perspectives on assessment greatly influence their instructional strategies (Popham, 2009). Despite the significance of assessment within the educational domain, particularly in the context of language learning, instructors often do not receive sufficient training in language testing and assessment. Therefore, multiple studies consistently indicated inadequate levels of teacher assessment and language assessment literacy (e.g., Alkharusi, 2011; Mertler, 2004; Plake, Impara, & Fager, 1993). Further, there has been limited exploration in the literature concerning the effects of lesson study on LAL.

In the Iraqi context, particularly at Salahaddin University, language courses at Salahaddin University in Iraq are provided to more than 500 learners holding MA and PhD degrees (i.e., future teachers in higher education settings). They must complete the B2 level in order to be able to apply for the Academic Title (s), which is a preliminary requirement to be eligible to teach at the tertiary level at Iraqi universities. The whole course is taught and covered by only one teacher. The teachers who teach at language centres at Salahaddin University might not be graduates of ELT or related departments. These teachers are employed based on their significant fields of study rather than their background in language teaching or applied linguistics. It has been noted that there is an academic gap in the assessment practises in these centres. Therefore, the study will examine the problems in detail, and it will also propose implications for the teachers themselves on how to improve themselves professionally.

In summary, the problem is twofold: at the global level, teachers' limited assessment literacy and the lack of systematic professional development opportunities remain pressing concerns, while at the local level, Iraqi higher education language centres face acute challenges due to limited training, heavy teaching loads, and non-specialist staff. Although previous studies have underscored the importance of language assessment literacy, few have investigated how it can be systematically developed in practice. In particular, the potential of the Lesson Study Model to foster assessment literacy among test writers has remained underexplored. These problems, together with the gaps identified in the literature (see Chapter Two), shaped the focus of this study. The present research therefore seeks not only to respond to these challenges but also to contribute original insights to the literature, as later outlined in the Discussion chapter.

Accordingly, this study seeks to respond to these problems and the corresponding gaps in the literature by investigating how Iraqi test writers' language assessment literacy can be developed through the Lesson Study Model. In doing so, the research not only addresses the lack of empirical studies in the Iraqi context but also demonstrates the

processes through which collaborative professional development can enhance teachers' assessment practices. The original contributions of this work are presented in detail in Chapter Five (Discussion).

1.4. Aim of the Study

This study aims to evaluate and explore EFL teachers' LAL level (their Language Assessment Knowledge). In addition, the possible effectiveness of the 'Lesson Study Model' as a professional development activity for teachers is examined. Therefore, for these purposes, the study aims to find out answers to the following research questions:

1. What are the academic and professional needs of teachers who teach at the higher education level at language centres as test writers in terms of language assessment literacy as they perceive?
2. What is the level of language assessment knowledge of the teachers?
3. To what extent do the teachers, as test writers, benefit from the 'lesson study' model to design language tests?
4. How does this model contribute to test writers' professional development?

1.5. Significance of the Study

LAL holds a central role in the instructional process, as it aids in measuring the degree to which academic objectives are achieved. It enables the identification of whether students are acquiring the expected knowledge and skills. Consequently, assessment emerges as a vital tool in motivating students when implemented systematically and appropriately. Moreover, it facilitates lecturers in evaluating the effectiveness of the taught course and determining the need for any necessary adjustments. According to Mihai (2010), assessment encompasses a diverse set of methods utilized to gather data regarding students' knowledge and, most importantly, their ability to apply and demonstrate that knowledge. Hence, teachers take the responsibility of assessing learners' performance and academic accomplishments.

The current study will examine the teachers as test writers at language centres in higher education settings in Iraq. Those who have MA and PhD degrees must be proficient enough in English to be able to teach at the tertiary level in any major. They must score a minimum of 5 in IELTS or any equivalent standardized test. Otherwise, they must attend language courses provided by public universities in Iraq. They must reach the

B2 level to be eligible to teach at universities. Therefore, the study will deal with those teachers as test writers who teach future teachers.

The study will examine the teachers' knowledge level in terms of their assessment literacy level. (i.e., whether they are assessment literate) and shed light on different factors that affect assessment literacy, including years of experience and their field of specialty.

The current research study will propose the areas in which teachers need to be professionally developed in terms of language assessment literacy. This will help teacher trainers and stakeholders shape the training programmes accordingly.

Further, the study will look at possible benefits, as shown in the literature, of the 'lesson study' model as a reflective activity to improve their ability to design language tests that support instructional goals. This could be gained through collaborative work with more knowledgeable colleagues within a sociocultural theory framework.

Finally, the study deals with insights of teachers as test writers, whether the mentioned model can aid and guide them to design more valid, reliable, and practical language tests. The results can be used as guidelines for other teachers and similar educational institutions to shape their programmes accordingly.

1.6. Assumptions

It is assumed that the teachers, as test writers at language centres, are not comprehensively assessment literate. This might be due to their previous education. Some of them are graduates of departments rather than the English Language Department, such as language and literature, and linguistics. However, they are tasked with teaching all four language skills and developing standardized language assessments. This limited literacy is evident in their test designs, especially in the omission of listening and speaking skills from final exams. On the other hand, the Lesson Study Model is assumed to help teachers design better communicative and collaborative language assessments.

1.7. Limitations

Although the present study provides meaningful insights into the literature on language assessment literacy through the Lesson Study Model, certain limitations should also be considered.

The present study was conducted with a small, purposely chosen sample of ten EFL teachers from a single university language centre. Though this allowed for rich context-based data collection, it has its limitations regarding generalisation. The

experiences of teachers in other Iraqi institutions or cultural setups may be considerably different.

Collective reflective journals and interviews served as key sources of data collection. While these methods are valid for capturing internal changes in participants' growth and attitudes, they may still be influenced by social desirability bias and selective recall. To mitigate this risk, triangulation was employed. Nevertheless, future studies may benefit from incorporating additional methods, such as direct classroom observation and test-development analysis, to provide more objective insight into practice.

1.8. Chapter Summary

This chapter introduced the study by presenting the background to LAL and its importance in EFL teaching. The problem statement highlighted both the global concern over teachers' limited training in assessment and the local challenges in Iraqi higher education, particularly at Salahaddin University language centres, where teachers often lack ELT backgrounds and work with large classes and limited resources. The research aims, objectives, and questions were defined to address these issues by exploring how Iraqi test writers' LAL can be developed through the LS Model. The chapter also explained the significance of the study in terms of its contributions to theory, practice, and teacher professional development in Iraq, and it outlined the scope, limitations, and organisation of the thesis. By framing both the international and local relevance of LAL and specifying the research focus, this chapter laid the foundation for the study. The next chapter reviews the literature on LAL, teacher professional development, and LS Model.

CHAPTER II

LITERATURE REVIEW

2.1. Introduction

This chapter reviews the literature that supports the theoretical orientation and practical aims of this study. It begins by addressing language assessment changes over time and looks at how these changes impact current thoughts about teaching and learning. The language assessment literacy issue is stressed, especially in relation to teachers and professional development mechanisms that enhance it.

To build up the conceptual framework for the research, an overview of ~~some~~ relevant theoretical perspectives is provided: reflective practice and metacognition. These two link together to help explain how teachers come to know their assessment practice and develop it further.

Then, the chapter deals with the historical background of language testing, describing the path of the field from its early psychometric traditions to more communicative and integrative forms of assessment that are a better fit with current pedagogical values. This sequence is followed by an explication of some key terms: Testing, evaluation, measurement, and assessment are discussed and differentiated to avoid ambiguity in their usage throughout this study.

Subsequent sections examine the multifaceted nature of LAL, offering various definitions and highlighting its significance for language teachers. This includes a review of empirical studies that examine the current state of teachers' assessment literacy, their training needs, and the challenges they face in implementing sound assessment practises.

Finally, the chapter introduces the Lesson Study Model as a collaborative professional development framework, examining its origins, core features, benefits, and implementation challenges. By integrating insights from both theoretical and empirical sources, this chapter establishes the scholarly groundwork for the present investigation and identifies the research gaps that this study aims to address.

2.2. Theoretical Framework

The evolution of the professional language assessment competency is a remote matter with its realization being reflective and metacognitive. The central premise of this

conception is that teachers ought not to be limited to socializing technical skills alone but must intent structured self-awareness to analyse and change their instructional decisions. Schön (1983, 1987) highlights two types of reflection: reflection-in-action and reflection-on-action. Reflection-in-action consists of a teacher thinking on his/her feet when confronted with a problem during the delivery of instruction; reflection-on-action is where the teacher has time to think about the incident some other time after the event and uses that insight to change the way they teach. This basis was further supplemented by Dewey (1933), who argued that reflection had to be intentional and organized for it to become a process by which one grows professionally.

Building on this premise, Brown's (1987) metacognitive model highlighted self-monitoring and decision making as aspects to consider in instructional contexts. According to this model, teaching is effective not just by virtue of content knowledge but as a consequence of the extent to which the teacher is able to plan, monitor, and evaluate their own thinking and behaviour under classroom conditions. Similarly, Reynolds and Suter (2012) conceptualized professional development in the Reflective Teaching Model as a cyclical process through which teachers inquire, experiment, and analyse their experiences to develop better pedagogical skills. Collectively, these theoretical inputs support the argument that reflection is not simply an exercise of looking back. Instead, it is an active and essential dimension in instructional design and assessment planning. Therefore, this study adopts these reflective and metacognitive perspectives as a theoretical foundation, mainly because they align closely with the core principles of the Lesson Study Model. Lesson Study, as a collaborative professional development approach, requires teachers to critically examine their classroom practices, including their assessment strategies, in a systematic and evidence-based manner, through cycles of planning, observation, reflection, and revision. Teachers refine their instructional techniques and develop more profound insights into assessment principles and learner needs. Given the study's focus on exploring the professional development of Iraqi EFL test writers in relation to language assessment literacy, these theories provide a meaningful framework to interpret how reflective inquiry can foster lasting pedagogical and assessment-oriented development.

2.3. A Brief History of Approaches and Theories of Language Testing and Assessment

Evaluation plays a crucial role in language education and instruction, deeply intertwined with the teaching process, providing support, guidance, and facilitation. It is widely accepted that language assessment (LA) is an inherent part of the holistic process rather than an external entity (Coombe et al., 2012). However, as Viengsang (2016) points out, language testing has historically been viewed as separate from teaching. Heaton (2011) also highlights that studies illuminate the importance of LA for students' motivation and academic achievements. Because tests and assessments are dynamic, they have evolved alongside changes in applied linguistics and language teaching methodologies. Therefore, LA has shifted from simple word tests into more advanced and complicated communicative language tests throughout history (Bachman, 1990).

Spolsky (1978) elucidated three distinct eras of historical changes in language testing and assessment: the prescientific era, the scientific (psychometric structuralist) era, and the Integrative Sociolinguistic era. During the prescientific era, the teaching process relied more on teachers' intuitions rather than a well-established method of instruction. The testing approaches in this era did not align with any principles of Language Assessment (LA), as no established principles for LA had yet emerged. Assessments during this time primarily relied on teachers' intuition, primarily evaluating 'knowledge'. This emphasis reflected the understanding of language instruction at the time—language education was predominantly considered to encompass grammatical, lexical, and phonetic knowledge exclusively. Consequently, the tests administered during this period concentrated on assessing grammatical rules, morphological changes in words, and the application of vocabulary. Scoring for these types of tests was notably empirical and subjective, as there was an absence of a standardized scoring rubric.

The second period was the scientific period (i.e., psychometric-structuralist period) which was influenced by structuralist linguistics. Language teaching in the 1940s and early 1950s was largely influenced by both linguistics and psychology. Linguistics offered the description of the nature of language while psychology emphasized the process of learning a language. In accordance with the developments of linguistics and psychology, language testing has been influenced by the theories of language learning and the psychological considerations as well.

During this period, language was defined as a system of symbols, tracing back to the invention of semiotics theory, which addressed the interrelations among independent

components of language, such as grammar, vocabulary, and pronunciation (Saussure et al., 1961). Additionally, Behaviourism (Skinner, 1974) facilitated the development of the Audio-lingual method. Lado's (1961) framework for language emerged from this method and posited that language proficiency encompasses the comprehensive knowledge of the four language skills. Although widely criticized, this framework laid the foundation for future developments in language testing. It is imperative to recognize its contributions to the field of language testing, as it established the foundation for all subsequent testing activities.

This led to the establishment of various psychometric characteristics of language assessment (LA), such as reliability and validity. Consequently, discrete point testing emerged, which aimed to evaluate a singular item at each testing point (for instance, grammar tests should be isolated from other components of language). To achieve reliability objectives, improvements in the objectivity of test scoring were sought through the development of multiple-choice items. To establish validity, test items needed to accurately represent the content of the materials being assessed. Assessments during this period encompassed numerous items, as each item was designed to evaluate a specific element of language. Language assessments were devised in a practical manner, utilizing a multiple-choice format to facilitate scoring, which enhanced the practicality of the tests. Lado's framework and discrete point testing faced scrutiny and criticism for fragmenting elements of language. Carroll (1961), as cited in Farhady (2018), posited that all components of a language ought to be assessed in an integrated manner rather than disaggregating language components into discrete elements.

Farhady (2018) further concluded that Carroll's (1968) suggestion gained popularity during the 1960s and '70s, coinciding with the integrative sociolinguistic era described by Spolsky (1978). He suggested that alternative assessment formats should be considered since discrete point items did not address the test taker's overall language proficiency. Therefore, Oller (1983), a key figure in promoting integrative assessment and a significant source of research in this area for nearly a decade, proposed a different hypothesis regarding language testing, which he termed the "unitary factor hypothesis." He believed that proficiency in a language represents a unified set of language competencies. Consequently, the general ability of a test taker should be evaluated. To support this, he employed statistical methods such as factor analysis to interpret test results. He concluded that the hypothesis was supported by findings related to mental and psychological factors, such as aptitude and attitude. Oller's hypothesis and claims were,

however, questioned regarding the accuracy of the interpretation of statistical analyses. Farhady (1983) and Bachman & Palmer (1982) rejected his hypothesis after conducting research that yielded different and contradictory results from Oller's. According to these researchers, language ability consists of various factors rather than a single one.

The communicative era is a phenomenon that dates to the late 1970s and early 1980s. Canale and Swain (1980) proposed a framework asserting that the ability to communicate in real-world contexts should be considered when assessing language proficiency. According to their framework, language ability consists of three competencies: linguistic, discourse, and strategic competencies. While Canale and Swain's framework focused on using and testing language in real contexts, Bachman's model represents language ability in a broader domain, arguing that it also pertains to other contexts beyond the academic realm. Consequently, he incorporated additional knowledge bases into his model, including textual and functional knowledge.

On the other hand, Fulcher (2000) argued that communicative language testing emerged in reaction to earlier models, such as Lado's concerns about focusing on designing multiple-choice items to enhance reliability. He stated that this framework seemed to reject the significance of reliability and validity in language testing. Fulcher further elaborates on the role of communicative language testing, emphasizing the importance of task definition, context, and the relationship between the situational use of the target language and test tasks, particularly in testing language for specific purposes.

2.4. Basic terms: Assessment, Measurement, Testing and Evaluation

Ewell and Cumming (2017) explain that the term "assessment" is derived from the Latin root meaning "to sit beside," which they interpret as evaluating a person's performance in complex tasks through close observation. The concept of sitting beside signifies the act of observing closely to make an estimation. Historically, it pertained to serving as an assistant to a judge by remaining beside them.

Assessment originated as a tool for tax collection, intended to provide impartial support rather than enforce hierarchical control over learning (Clapham, 2000). Today, however, assessments are used in diverse fields—such as finance, medicine, and law—which have diverged significantly from this original function. It was intended to serve as an impartial aid rather than establishing a hierarchical relationship where one individual dictates the learning of another (Clapham, 2000). However, assessments are now extensively utilized across various domains such as finance, property law, career

advancement, and medicine, among others, which possess no direct correlation to their original purpose. Wherever assessment is used, it requires gathering evidence through observation to make informed judgments. While terminology may differ across professions, the underlying purpose remains the same. Different professions may employ distinct terminologies associated with assessment, yet they all encompass the same fundamental concept. For instance, an educational psychologist would utilize the term 'assessment' when referring to the employment of psycho-educational tests, observational data, interview data, etc., aimed at accruing evidence related to learning, social skills, behaviour, communication, or adaptive functioning in students.

Brown (2004) describes assessment as a continuous process that requires feedback through multiple methods, assessing students' performance in relation to set objectives while also considering their prior progress.

According to Brown and Abeywickrama (2010), tests are just one component within the broader framework of assessment, used to measure achievement under specific conditions. Unlike evaluations, test results are usually communicated directly to learners.

Cumming and Ewell (2017) describe "assessment" as an evolving concept shaped by the epistemological needs of various fields. In education, assessment may overlap with evaluation, measurement, or testing methods, but each term serves a distinct purpose. Therefore, it is essential for practitioners to define and distinguish these terms clearly. These terms can be used interchangeably in some cases because they may share common processes and methodologies, yet they serve different purposes. According to other scholars, however, the varied usage and definitions of these terms among multiple stakeholders in the field make it essential for us as practitioners to establish clear distinctions between them.

Measurement in an educational context is the systematic practise of assigning numbers to qualities or dimensions to enhance understanding and ensure accurate representation. In educational research, measurement refers to the use of psychometrics to assess how much of a characteristic, trait, or attribute an individual possesses. Classical Test Theory (CTT) compares the measurement of tangible objects with objective tools (like rulers and scales) to the measurement of non-observable constructs that lack a physical standard for comparison. These theoretical concepts represent levels or amounts of the manifestation of specific characteristics exhibited by individuals.

Although measurement and assessment are often used interchangeably, they are not synonymous. Assessment involves a broader process of gathering and interpreting

information, not limited to quantitative outcomes. Therefore, it's important to distinguish between numerical data and more holistic judgments.

On the contrary, evaluation involves making decisions for specific purposes in particular situations based on gathered data that has been quantified. Evaluation aligns more closely with the concept of assessment than with measurement. The term "evaluation" is etymologically linked to assessing value, worth, or quality, whereas "assessment" is rooted in assigning a value. In this context, assigning value during an assessment plays a neutral role since it helps us make decisions about our preferences without claiming whether something has any true merit. A value judgment can be complete only when some values are objectively assigned, and others are judged subjectively against them. Assessments aim to collect data fairly on immeasurable aspects to provide an authentic, unbiased opinion on worthiness, while evaluations utilize these data inputs for informed judgments regarding the same worthiness. Consequently, different definitions of evaluation emphasize using available or gathered data to make worthwhile judgments in educational activities, whereas assessments involve sense-making. In educational settings, evaluation specifically refers to using evidence obtained through investigations into the effectiveness of programmes or institutions, while assessment implies establishing something as a fact. It may therefore employ a cost-benefit analysis approach within productivity dimensions in higher education contexts.

2.5. Principles of language assessment

According to Brown (2004), for a test to fulfil the criteria of good testing, teachers as test writers must understand the principles governing language assessment, as these shape the properties that tests should possess. Thus, the five key principles that underpin effective tests include validity, reliability, practicality, authenticity, and washback. Each principle will be briefly explained below.

2.5.1. Validity

Within educational assessment, the notion of validity is significant in ensuring that an evaluation measures what it purports to measure (Stobard, 2012). A test becomes valid if and only if it can accurately assess a specific skill, knowledge, or ability as well as its components, which are targeted by the test (Brown, 2004). Furthermore, claims regarding the validity of tests should be backed by evidence (Messick, 1989). For example, content-based evidence known as content validity refers to how appropriate the

contents selected for inclusion in the item pool match up against the behaviours being considered. Suppose we take an example where people are assessed on their conversational speaking skills by engaging them in real-life discussions. In this case, this method of assessment would have high content validity because everything being tested is derived from what is expected in daily conversations among individuals. This contrasts with asking someone to answer questions based on given options, as multiple-choice items may lack alignment with the real-life communication tasks the test aims to assess (Hughes et al., 2020).

(Coombe et al., 2012) stated that content validity, often referred to as "congruence," pertains to the alignment between a test's content and the construct it is intended to measure.

Criterion-related validity, also referred to as criterion-referenced evidence, assesses whether a test correlates with a standard. In this situation, it may be necessary to compare test results with those from an independent and reliable measure. For example, a teacher-developed assessment on present modals can be compared with a textbook-based examination that also covers this topic. Criterion-related validity is established when there is a strong association between these two sets of scores. There are two types of criterion-related validity: concurrent validity and predictive validity (Fulcher & Davidson, 2007). Concurrent validity involves administering tests around the same time, while predictive validity assesses how well examinees will perform in the future by administering tests at different times.

Concurrent validity reflects current proficiency, while predictive validity estimates future performance. This is evident when strong performance on the final examination of an English course aligns with real-life language proficiency. Conversely, if students who score highly on a proficiency test are found to be capable of undertaking a university-level course, then predictive validity has been established.

The third category of evidence is construct-related evidence, also known as construct validity. It examines whether tests accurately measure what they are intended to measure. In linguistic terms, constructs refer to aspects such as fluency and communicative competence, among others, used within this field. Although teachers may find it challenging to measure such qualities directly, Brown & Abeywickrama (2010) assert that construct validity can be attained through a careful examination of the alignment between tests and their underlying theoretical principles.

Consequential validity examines all outcomes that result from the use of a test. Literature emphasizes the necessity of considering an assessment's consequences. McNamara (2000) asserts that consequentiality should never be overlooked in any educational setting, as teachers inevitably interpret scores based on this concept with a view toward improving learning rather than making judgments alone. This could also yield positive results for stakeholders benefiting from curriculum development, making it essential for every stage of test construction, administration, scoring, interpretation, reporting, and so on.

Gronlund (1998) defines face validity as the perceived fairness or relevance of whether something measures what it is supposed to and how useful that information may be in promoting student achievement motivation, driven by its impact on individuals' actions toward the tasks set before them. When these criteria are absent, both parties involved—the teacher and the student—may exhibit low interest due to perceived invalidity. Such feelings can lower self-efficacy, leading to decreased effort levels and poor grades obtained. However, if learners feel that criterion-referenced assessments accurately describe their own abilities, then intrinsic motivation will increase even further, ultimately leading to mastery of related subjects. This will ensure heightened personal satisfaction and pride displayed by learners when they accomplish such feats.

2.5.2. Reliability

Another essential part of an assessment is reliability, which refers to how consistent or dependable tests are. It indicates whether similar results will be achieved each time the same test is administered under varied conditions, such as different places or scorers. Hughes (2020) argues that assessments need to provide stable outcomes, regardless of who scores them or when they are administered. However, too much delay between two administrations could result in students gaining new knowledge and skills, thereby affecting the reliability of the test. On the other hand, short intervals might help learners recall items, thus influencing the reliability of these exams. Therefore, teachers should keep this in mind while administering tests.

According to Heaton (1975), there are four primary groups of factors that affect reliability: student-related reliability, rater reliability, test administration reliability, and test reliability. First, factors such as temporary illness, anxiety, fatigue, and motivation can hinder students from performing to their true ability. Additionally, Mousavi (2002) has noted that “test-wiseness” and strategies for effective test-taking among students may

create a gap between the observed score and the actual score of the test-taker. Second, subjectivity, bias, and human error are some of the rater-related factors that could influence scoring.

It is important to recognize that both inter-rater reliability, involving multiple scorers, and intra-rater reliability, involving a single scorer, can influence this matter. Inter-rater reliability occurs when two or more scorers provide inconsistent scores for the same test due to factors such as inattentiveness, inexperience, bias, or neglecting scoring criteria. Conversely, intra-rater reliability is common among classroom teachers and can be affected by factors like unclear scoring criteria, fatigue, carelessness, and bias. For example, when a teacher is assigned to read 50 tests within a week, scoring and reliability issues may occur. In these situations, it can be helpful to read through half of the tests before assigning any scores and then review all the tests again, ensuring fair judgment and scoring, as suggested by Brown (2004). Furthermore, establishing scoring criteria beforehand and using rubrics can help achieve rater reliability.

Moreover, test administration reliability is affected by various factors, including temperature and lighting variations within the same room, the conditions of desks and chairs, noise levels, and differences in photocopying quality. It is crucial to take these elements into account and recognize their potential impact on students' performance when conducting a test. Finally, test reliability relates to the inherent qualities of the test itself. Certain features of the test, such as its length, structure, and item construction, can affect its reliability. For instance, if a test is excessively long, students may experience fatigue, which can negatively affect performance. Likewise, poorly constructed and ambiguous test items can adversely impact performance. In conclusion, these four factors significantly influence the consistency and dependability of tests, underscoring the importance of considering them when interpreting test reliability by instructors and test administrators (Cohen, 2001).

2.5.3. Practicality

When conducting a test, various practical considerations must be thoroughly evaluated. As stated by Harris (1969), these considerations include factors such as cost-effectiveness, ease of administration and scoring, and simplicity of interpretation. Further elaborating on this notion, Farhady (2012) argues that practicality encompasses all aspects related to the processes of a test, including development, administration, and scoring. For example, while assigning students an essay task may be a valid approach to

assessing their writing skills, it becomes impractical when handling a large group of learners. Highlighting the importance of practicality, Brown (2004) emphasizes that an effective test should demonstrate practical qualities, such as affordability, adherence to time constraints, ease of administration, and an efficient scoring process. Additionally, Farhady (2012) cautions that both low and high practicality factors need careful consideration during the test's development and scoring phases. For example, developing a multiple-choice test can be challenging, but scoring it is relatively straightforward. In contrast, designing an essay task may be more straightforward than developing a high-quality multiple-choice test, yet it requires more time for scoring, and it often requires more time and introduces greater scoring subjectivity. Therefore, the practicality of the procedures involved in developing and scoring different types of tests can vary significantly. Finally, practicality must be balanced with the other principles for optimal test design.

2.5.4. Authenticity

The authenticity of a language test task is determined by how closely its features align with those of a real-life task in the target language. This concept is also related to generalizability, which refers to the extent to which we can apply our understanding of test scores beyond the testing environment (Bachman & Palmer, 1996). According to Brown (2004), key features of authentic tests include the use of natural language, real-world tasks, and contextually meaningful content.

2.5.5. Washback

There are various definitions of washback found in the literature. According to Cohen (1994), washback refers to the influence that an evaluation tool has on teaching practises and beliefs. He further explains that it describes "the extent to which teachers and learners are encouraged or discouraged by its introduction and use to engage in certain kinds of activities." Hughes (2020) describes washback as the effects of testing on both teaching and learning, which can include both positive and negative consequences. Messick (1996) identifies two types of washback: positive and negative. Positive washback occurs when language tests directly assess communicative skills through genuine tasks that involve all four skills: listening, reading, writing, and speaking. Conversely, negative washback happens when learners focus predominantly on preparing for test formats like multiple-choice items, which do not reflect real writing tasks, which

do not require actual writing during real tests. This approach creates an imbalance between what should be taught and learned (Hughes, 2020).

Positive or negative washback was classified by Bailey (1996) based on how it either helps or hinders learners in achieving their educational goals. While terminology may vary, the core concept of washback remains consistent. In contrast, Wall (1993) offered a different perspective on the causes of washback, stating that it occurs when exams compel individuals to engage in activities they would not typically pursue. From this viewpoint, certain areas of study may receive more attention from students who believe these subjects are linked to the examination (Wall, 2012).

Therefore, washback can significantly influence curriculum design, instructional priorities, and learner engagement, as educators often align their teaching methods and content with the perceived demands and format of high-stakes assessments.

2.6. Assessment Purposes

It is crucial for language teachers to understand why tests and assessments are utilized in their classrooms. This understanding will assist everyone involved in setting clear goals to choose appropriate assessment methods and provide suitable resources for testing. Within a specific timeframe, teachers should assess how much students have progressed in terms of knowledge acquisition, as well as how these knowledge gains were achieved (Bachman, 1996).

According to Cohen (1994), there are three main functions served by assessments, each reflecting a different stakeholder's agenda. Institutional programme managers may want to use evaluations as tools for gaining insight into various aspects of curriculum implementation among different teachers within an institution. Administrative appraisals play a key role in determining learners' placement, certification status, and their progression into higher levels within the institution, thus ensuring enrolment in appropriate language courses.

Secondly, assessments serve instructional purposes that enable teachers to diagnose and highlight learning outcomes among students. Here, Cohen (2001) refers to this diagnostic function, where evidence of progress can also facilitate a comprehensive evaluation of the effectiveness of the curriculum.

Finally, assessments can also serve as research tools that facilitate the examination and evaluation of course-specific proficiency levels demonstrated by second language learners over time. Therefore, for such research to be conducted effectively, teachers must

have the necessary skills for test development, adaptation, and administration to achieve this objective.

Brown and Race (2012) emphasized the objectives of assessment for instruction, which consist of classroom-based tests designed to measure students' growth and achievement. They also observed that teachers frequently concentrate on evaluating learners' progress within a specific course. Brown outlined three primary functions for classroom-based assessment: diagnostic, which identifies learners' strengths and weaknesses at the outset; progress, which monitors learning over time; and achievement, which measures whether learners have met specific course objectives. These functions are particularly relevant to the present study because they reflect the range of purposes that test writers at Salahaddin University need to address in their practice, yet which are often approached without sufficient assessment literacy or systematic training.

2.7. Definitions and significance of Language Assessment Literacy (LAL)

Assessment, as defined by American Federation of Teachers (AFT), National Education Association (NEA), and National Council on Measurement in Education (NCME) (1990), refers to any method of gathering student information. This data serves multiple purposes: guiding educational decisions, providing feedback on student progress, evaluating instructional and curriculum effectiveness, and informing policy. While these organizations offered a broad definition of assessment, Stiggins (1991) introduced the concept of *assessment literacy* to highlight the specific knowledge educators and stakeholders need to apply assessments effectively. Many researchers have explored various aspects or components linked to assessment literacy in response to Stiggins' concept development regarding what it means to be knowledgeable about assessments. Scholars have defined language assessment literacy in various ways. Most definitions emphasize that stakeholders, especially language teachers—must understand testing principles, measurement techniques, and their practical applications, and the practical application of testing (Inbar-Lourie et al., 2008; Stiggins et al., 2001; Taylor et al., 2009).

Assessment literacy refers to educators' understanding of how tests effectively reveal what students know, enabling them to design assessments aligned with learning goals. This includes recognizing when to use specific assessment formats, as certain methods, such as open-ended tasks, may more accurately demonstrate student comprehension than others like multiple-choice tests (Stiggins, 2002). Paterno (2001)

adds another dimension, stating that teachers must also be familiar with best practises, which include understanding key terms such as reliability and validity, developing alternative methods of assessing learning besides traditional tests among other considerations.

Additionally, Volante and Fazio (2007) define assessment literacy as the ability to recognize when and where to use specific types of tests within the context of their administration to ensure they effectively serve educational purposes.

Davies (2008) argued that LAL consists of three components: skills, knowledge, and principles. Similarly, Fulcher (2012) proposed a three-dimensional view of assessment literacy. The first component is having the necessary skills, knowledge, and abilities to construct, maintain, and evaluate a test, whether it is a classroom or standardized assessment. Secondly, understanding what goes into testing should involve awareness of the underlying concepts, including ethics. Thirdly, one should be able to see how various aspects fit together historically, socially, politically, and philosophically. This broader awareness enables individuals to understand why we do what we do and encourages critical engagement with current practises, which can impact various levels, from schools to society as a whole.

Newfields (2006) presents a context-dependent view of assessment literacy, suggesting that its meaning varies depending on whether the stakeholder is a university student, a secondary school teacher, or a professional examiner. Newfields views assessment literacy as something that can mean different things to different people. For students, being assessment literate means having enough information about exams to navigate them successfully and achieve high marks when needed. For teachers, this involves marking ethically and accurately while considering fairness towards all students, regardless of their background or abilities. Professional test developers should base their practises on established principles of effective test design, implementation, and evaluation. However, his observation that many educators lack the necessary knowledge to conduct effective assessments echoes earlier concerns raised in the field.

Historically, Lado (1961) emphasized that teachers need both theoretical knowledge and practical experience to evaluate language learners effectively. Nunan (1988) later observed a gap in teachers' knowledge about testing and the ineffective application of testing methods and techniques in second language classrooms.

Stiggins (1991) noted that many educational systems have identified inadequate teacher preparation in assessment as a major concern. This has prompted calls from

international organizations like UNESCO for improved teacher training in assessment literacy skills—something that governments at all levels should heed.

Scarino (2013) rightly points out that language assessment literacy is critical in second language education. Falsgraf (2005) adds that this enables teachers to understand, evaluate, and use student performance data to improve instruction, while Siegel and Wissehr (2011) note another benefit: it helps teachers know different types of tests to choose those best suited for specific learning objectives.

Mertler (2004) highlights the systemic importance of teacher competence in assessment, arguing that ‘assessment quality will always be limited by the weakest link in the chain. Popham (2006) reinforces the need for assessment literacy by asserting that ‘educators must be more knowledgeable about measurement than the people they are trying to measure,’ highlighting the importance of teacher expertise in educational assessment. Additionally, Gottheiner and Siegel (2012) argue that assessment-literate teachers can accurately interpret test results and engage students in meaningful discussions about their comprehension. These discussions then inform the setting of targeted language learning goals. Finally, Eames (2003) concludes that instructors’ understanding of evaluation plays a crucial role in promoting student achievement and professional development. This, in turn, contributes to the overall reputation and effectiveness of educational institutions.

Teachers’ language assessment literacy requirements, in many cases, are related to their skills to assess receptive skills, which constitute some of the more difficult dimensions of language testing. Farrall (2012) drew attention to the complex nature of the relationship between language and literacy and cognition in the measurement of reading and the implications for teachers to be aware of that context to ensure fair and equitable evaluations of learning. Similarly, Mahmood et al. (2021) noted the substantial lack of reading assessment knowledge and skills among many Pakistani ESL teachers in their study, demonstrating deficiencies in practice. In addition, Rasyid, Rauf, Junaid, Nur, & Syukur (2023) stated that listening comprehension tests in EFL classrooms has still remained a difficult task. These results show that teachers are not trained and do not as always have the resources to adhere to best-marking practices of the receptive skills.

In addition to receptive skills, teachers also need to be able to be open to and engage in comprehensive and inclusive forms of language assessment. Razavipour (2013) highlighted the need to develop test literacy in high-stakes situations, stating that integrative assessment provides teachers with a more accurate indication of students’

abilities. Berger (2023) further stressed the importance of assessment literacy in classroom practices and maintained that teachers need to learn to design tests that incorporate different skills rather than assess them independently. North's (2005) research on the CEFR also indicated the need for transparency and diversity in testing tasks, with Stabler-Havener (2018) demonstrating that teachers must be familiar with a wide variety of assessment types to accurately assess performance. Consistent with this, Kavakli and Arslan (2019) showed how teachers evolve over time from learning fundamental assessment ideas to implementing them. In combination, these studies indicate that whilst teachers are increasingly supportive of integrative and inclusive approaches, they lack belief and competence in respect of their ability to act effectively on them.

2.8. A Review of Studies on Teachers' Language Assessment Literacy

Numerous studies have been conducted on teachers' assessment literacy, which pertains to their understanding and beliefs about testing (Coombe et al., 2012; Wang et al., 2008). This is significant because it is widely accepted that what teachers know and believe influences their practises (Alkharusi et al., 2012; Popham, 2009; Stiggins, 1995).

However, several studies have consistently shown low levels of language testing competence among teachers (Alkharusi, 2011; Mertler, 2004; Plake, Impara, & Fager, 1993). Interestingly, there appears to be a lack of recent empirical studies that either confirm or refute these findings, highlighting a gap in the current literature.

For example, Alkharusi et al. (2012) utilized the Teacher Assessment Literacy Questionnaire (TALQ) to evaluate the assessment knowledge of 167 teachers and found the results to be unsatisfactory. Similarly, Mohamed (2016) identified moderate-to-low levels of assessment literacy among secondary school teachers in Malaysia. Xu and Brown conducted a large-scale study (2017) that evaluated the level of assessment knowledge among 891 EFL teachers in Chinese universities using an adapted version of the TALQ and reported insufficient levels of assessment literacy. Among other recommendations, the researchers also advocated for continuous professional development programmes on assessment within language teaching institutions.

Similar concerns about teachers' perceptions of assessment have been raised elsewhere. In a study involving 512 primary school teachers, Birgin and Baki (2009) found that many teachers considered themselves competent in traditional methods of assessment but lacked skills in alternative methods. Sheehan and Munro (2017)

established through classroom observations, interviews, and focus groups with EFL teachers that those with minimal training in assessment lack confidence in producing assessment materials. According to them, there is a need for classroom-based training on assessment, though the researchers observed that such teachers were able to implement various kinds of assessments successfully even without formal training. Some studies also indicate that self-confidence and knowledge of assessment among educators can improve if they receive education or training on this subject; for instance, DeLuca & Klinger (2010) and Mertler (2009).

Several studies have indicated a gap between what teachers do in assessing their students' learning in school and what is recommended by test experts concerned with the measurements used during exam periods. For example, Alsarimi's (2001) research revealed that most science instructors preferred traditional evaluation methods over alternative forms among 246 randomly selected respondents from secondary schools in Oman's central region. This preference negatively affected learners' self-efficacy beliefs and created the impression among them that these assessments did not align with the instructional goals set by the instructors earlier.

Additionally, research shows a misalignment between the classroom practises employed by educators during testing periods and the techniques recommended by experts responsible for establishing national examination standards worldwide. An example can be cited where Alsarimi (2001) discovered that an overwhelming majority of science tutors sampled preferred conventional assessment strategies instead of considering alternative approaches, thus diminishing students' confidence and raising questions about whether such tests adequately align with the previously planned teaching content for the academic year.

Another recent study conducted by Xu et al. (2017) aimed to validate the 'Teacher Assessment Literacy Questionnaire (TALQ)' by examining how various demographic factors impact assessment performance among English as a Foreign Language (EFL) teachers in China. The researchers also intended to assess the level of assessment literacy among these teachers. TALQ was adapted for this study and included a paper-based questionnaire containing 24 assessment literacy scenarios and 7 demographic variables: age, qualifications, experience, and training. A total of 891 Chinese EFL teachers from 7 regions completed the questionnaire. However, the findings suggested that the specific contextualization required some revisions within the adapted instrument for use with university teachers in China. Furthermore, certain areas exhibited basic levels of

participants' assessment literacy, indicating significant room for improvement. Although demographic variables had minimal impact on their performance based on the data collected during the study period, the researchers recommended placing greater emphasis on policy awareness creation campaigns targeting such groups, as it has been found that the lack of awareness is a major barrier to achieving high standards in the teaching profession.

Wach (2012) conducted a study that focused on probing language assistants regarding EFL teachers' beliefs and practises related to their native or non-native speaker status. The study aimed to explore how school type and English language proficiency shape teachers' beliefs and roles in Poland. In total, this study involved 87 participants (native and non-native English-speaking teachers), including 42 university teachers and 45 school teachers. Participants had been teaching for as little as six months to as long as thirty years. The researcher used open-ended questionnaires to collect data, which revealed that university teachers demonstrated greater awareness of assessment functions, areas, and features than their counterparts in schools. Most respondents centered their focus on grammar and vocabulary assessment, viewing it as separate from teaching; however, some also considered other language skills. Interestingly, there was no significant difference in the opinions expressed by teachers based on whether they were native or non-native speakers. According to these findings, it is recommended that different types of assessments be taught during training sessions for EFL educators so they can appreciate evaluation across various language areas and skills.

Research on technology in language testing has expanded considerably over the past two decades. Chalhoub-Deville (2001) first highlighted the importance of investigating attitudes toward the use of technology in assessment as part of efforts to improve testing methodologies. Following this, a growing body of work has focused on computer-assisted language testing and computer-based tests (CBT). Chapelle (2008), for example, defines CBT as a process in which language performance is assessed directly through computers, emphasising both technological and pedagogical considerations.

Much of this research has concentrated on learners rather than teachers. Jeong (2014), for instance, conducted a study comparing the outcomes of CBTs with those of traditional paper-based tests, focusing primarily on student performance. In contrast, Brown (2013) observed that very little research has examined the knowledge and competencies teachers need in order to integrate computer technology effectively into classroom-based language assessment.

Therefore, the term "digital literacy assessment" originated from a study conducted by Eyal in 2012. She discovered that twenty-first-century teachers are concerned about adopting technology that can support assessment for learning; they also promote peer assessment and the development of reflective skills. According to her findings, a teacher proficient in digital literacy assessment knows how to effectively use various applications and technological systems to enhance student progress and adjust evaluation approaches.

Furthermore, teachers must equip their students with evaluative skills that foster reflective learning, as emphasized by Eyal (2012). The proposed levels suggest that when learners are more engaged during evaluations, teachers become more competent in assessing them; this is what the author stated. Among other points mentioned were alternative methods for measuring knowledge, such as tests or quizzes. However, these methods were not all utilized equally, depending on the level of digital literacy assessment among educators.

2.9. Training for Language Assessment Literacy (LAL)

Sato et al. (2008) found that teachers are the most significant factor influencing student achievement. This emphasizes that quality instruction cannot be achieved without effective classroom assessments, as highlighted by Eckhout et al. (2005). Stiggins (1999) further asserted that no matter how much effort teachers put into delivering content, it becomes ineffective if they do not employ appropriate assessments to gauge student progress during lessons.

Assessment is a complex process in educational settings (Plake, 1993). Brookhart (1998, 1999) asserts that testing significantly impacts various aspects of a teacher's role, making its importance indispensable. Therefore, it is essential that educators develop a comprehensive understanding of assessment methodologies so they can determine the most appropriate method to use according to their teaching objectives (Siegel & Wissehr, 2011).

In foreign language education LA serves multiple functions. The primary one is that it helps in monitoring the progress made by learners hence enabling teachers to plan their lessons accordingly. Secondly, students find tests motivating because they push them towards achieving higher levels of competence within the target language; this was

also mentioned earlier on. Data can be collected at various points using LA which then provides insights into both student and teacher development.

Good LA should not only measure accurately but also create positive attitudes thereby minimizing negative perceptions towards testing through its impact on motivation. Teachers can create a conducive environment for learning through proper evaluation methods thus reducing anxiety related with exams among other negative emotions associated with it.

Research has shown that many teachers struggle with assessment (Campbell et al., 2002). This issue often arises when teachers and administrators lack the knowledge to properly evaluate educational outcomes (Stiggins, 2001). Their inability to effectively use a variety of assessment types contributes to assessment illiteracy (Mertler & Campbell, 2005). As a result, students may suffer from under-assessment or lack of assessment altogether, which highlights the importance of teachers being well-versed in effective assessment practises.

In conclusion, incorporating Language Assessment Literacy (LAL) training into teacher education programmes is crucial for equipping prospective teachers with the knowledge and skills needed to apply effective assessment strategies, ultimately enhancing learning outcomes and supporting language acquisition in their classrooms.

2.9.1. Research on LAL training

A shift has occurred in the perception of LA from a mere testing concept to a broader assessment perspective (O'Loughlin, 2006). This change resulted from evolving foreign language teaching methods that have influenced the understanding and application of LA. It is now recognized as not only a technical and psychometric field but also a social practise encompassing political and ethical dimensions.

Previous studies on teachers' Language Assessment Literacy (LAL) indicated that many teachers were eager to receive training in areas such as test design, evaluation, and item development. For example, Vogt and Tsagari (2014) found that among the topics for which in-service language teachers expressed a need for further training were test design skills and test evaluation skills. Fulcher (2012) observed that language teachers are interested in learning how to develop items specifically for multiple-choice questions. Kunnan's (2000) study found that language teachers showed little interest in ethics or fairness in LA, despite these being key points of debate in professional development.

Huhta et al.'s survey (2005) aimed to identify assessment training needs among European teachers. However, it faced challenges distinguishing between 'teachers,' 'trainers,' and 'experts,' and some countries, like Finland, were overrepresented. Nevertheless, the research identified several common areas where more training was needed: assessment portfolio development; creating classroom tests; peer assessment; analysing test results; self-assessment; providing feedback on learners' progress; understanding the principles of language assessment; using assessment statistics; preparing test item banks; statistical evaluations of items; conducting interviews; and scoring procedures, among others.

Xu and Brown's study (2016) analysed 100 sources, including both empirical and non-empirical studies, with the aim of providing an extensive framework targeting the essential requirements for assessment-literate language teachers. They argued that assessment literacy requires broad knowledge of assessment purposes, content, methods, grading, feedback, self- and peer assessment, and ethics. Teachers should also reflect on their personal beliefs regarding evaluation, as these beliefs emerge through classroom practises, including teaching, assessment, and participation in CPD activities. According to them, a failure by teachers to understand these aspects would negatively impact students' performance in various aspects of language learning.

Walters (2010) and Arias et al. (2012) conducted studies that not only identified needs but also aimed to enhance language teachers' Language Assessment Literacy (LAL). They facilitated the creation of test specifications that enabled a critical understanding of standards-based testing among teachers.

López and Bernal (2009) reported that language teachers have varying assessment practises. It was found that teachers who received training in language assessment applied assessment methods designed to enhance both teaching and learning processes. In contrast, those without training viewed it merely as a means of assigning grades. As a result, teachers without assessment training often equate evaluation solely with grading, a view considered overly narrow by researchers.

Brown (2013) observed that studies on what language teachers need, believe and can do in relation to the implementation of computer-based language testing (CBT) are rare. Recent research, however, indicates that this landscape is changing. For example, studies on the digital assessment literacy of EFL teachers at the tertiary level find that they consider that they are proficient in traditional assessment but are less confident in their digital assessment expertise even if they acknowledge the need to use it in a modern

context (Oman, 2023). More general studies indicate that, in self-assessments, higher-education academics report as being of low or low-medium levels of digital competence, with the assessment process in the field evaluation processes being the most problematic ones (Zhang, 2023). These studies also infer that teachers' technology attitudes, skills and access determine their digital literacy as a whole.

2.10. An Overview of Lesson Study Model

The Japanese term “jugyokenkyu” directly translates into English as “lesson study.” “Jugyo” means “lesson,” while “kenkyu” stands for either “study” or “research” (Gutierrez 2016). Lesson study is unique because it involves groups or teams of teachers collaborating on action-research activities aimed at improving the overall quality of instruction (Ono & Ferreira, 2010). They also observe that while lesson study has been practised since the 19th century, it gained significant popularity in Japan after the late 1960s when Tokyo Normal School and its affiliated school, founded in the early 1870s, were credited with being its birthplace.

Lesson study was introduced to Western educators by Stigler and Hiebert (1999). The authors argue that lesson study was originally designed for teaching mathematics; therefore, they emphasize the high teaching standards observed in math education in Japan. They also point out that through lesson study, Japanese teachers can actively participate in developing and implementing a learner-centered curriculum specifically designed to enhance students’ conceptual understanding of mathematics.

According to Lewis (2002), the Lesson Study Model involves collaboration among teachers, prospective teachers, and other specialists aimed at preparing for and engaging in teaching activities as effectively as possible. It also encompasses the assessment and application of this approach, including its outcomes. Cerbin and Kopp (2006) define lesson study as teachers working in small teams to jointly design, deliver, observe, evaluate, and refine specific class lessons known as research lessons. Exemplifying a professional development model, it is widely practised among Japanese educators, where virtually all teachers engage in lesson study teams at some point during the academic year.

Furthermore, during the implementation of lesson study, educators with a shared goal typically meet frequently to plan lessons designed to enhance teaching skills or deepen their understanding of subject matter content (Cerbin and Kopp, 2006). Lesson study reflects social constructivism, which asserts that learning and professional growth

are fostered through the collaborative sharing, exploration, and formulation of new ideas grounded in classroom experiences (Dadds, 1997; Kriewaldt, 2012). In today's educational climate, where there is pressure on teachers to meet standards, lesson study has emerged as an effective form of professional development because it provides a platform for constructive critique of teaching practises through the sharing and enhancement of lessons (Stigler & Hiebert, 1999).

Lewis (2000) identified three characteristics of lesson study, or "lesson research," as he referred to it:

- Research lessons are observed by other teachers.
- Research lessons are recorded.
- Research lessons are discussed.

(Fernandez & Chokshi, 2002, pp. 4-6)

The Lesson Study Model also fosters essential reflection skills in teachers for professional development through knowledge sharing and both individual and group discussions.

The Lesson Study Model has provided various forms of support to teachers in different areas, according to Fernandez and Yoshida (2004). These areas include general, professional, and content knowledge. It also involves understanding students' conceptions and engaging in practises such as planning lessons, implementing those plans, observing peers, and reflecting on both peer and self-performance.

A three-cycle process characterizes how the Lesson Study Model operates (Lewis & Hurd, 2011). In this model, teachers initially collaborate to plan the lesson. To collectively explore the desired student objectives, they form small groups. Next, they move into the planning stage, where a team develops a plan for investigating students' learning during the lesson. Afterward, one teacher takes responsibility for facilitating the planned lessons, during which one member of the group is selected to act as the model teacher, while the other members actively participate by attending class and collecting data through observation methods.

Finally, the teachers evaluate the effectiveness of the implemented lesson plan. During this stage, impressions about the teaching practise are shared by the designated model teacher while the other group members, acting as observers, provide their

comments based on the instructional practise, considering the learners' learning experiences.

2.11. Benefits of lesson study implementation

At the beginning of the 20th century, the concept of research lessons or Lesson Study was recognized worldwide and has since been accepted as an effective means of enhancing education by developing a professional knowledge base (Lewis, 2009). This recognition was initially made by Yoshida (1999). According to Hashimoto et al. (2003), the research lesson model has become a continuous practise in Japan to the extent that it is now considered a norm among Japanese teachers and administrators alike. Stigler and Hiebert (1999), along with other researchers from the United States, have shown significant interest in lesson study as a means for teachers' professional development. The motivation behind teacher professional development stems from the need to improve teaching practises, skills, and beliefs. These needs are supported by various factors, such as changes in curriculum, advances in pedagogy, and the integration of new classroom technologies (Doig, Groves & Fujii, 2011).

Lesson study is an effective professional learning method for in-service and pre-service teachers, as well as experienced practitioners (Miller & Xu, 2017; Chokshi & Fernandez, 2004). Researchers have extensively examined the relationship between lesson study and teachers' pedagogical content knowledge (Cerbin & Kopp, 2006). The concept of pedagogical content knowledge was first introduced by Shulman (1986), who argued that it describes effective ways to represent subject matter, facilitating comprehension by others. It encompasses useful forms of representation, potential analogies, selected illustrations, examples, explanations, and feasible demonstrations or performances. Lewis (2002) found that engaging in real teaching episodes with colleagues—by observing each other's instructional strategies and receiving feedback on one's lessons—leads to significant professional growth among teachers, a process known as “learning from reflection.” According to McDowell (2010), these experiences profoundly influence their teaching philosophy. Furthermore, McDowell (2010) noted that teachers who engage in more critical reflection develop a deeper understanding of pedagogical content knowledge.

Doig et al. (2011) also argued that the Lesson Study Model does not focus on perfecting a single lesson; rather, it values the generation of teachers' ideas and the exploration of different teaching methods. They further stated that conducting research

lessons prompts participants and observers to reflect deeply on both specific and general aspects of teaching, leading to a deeper understanding of the subject being taught. According to Lewis & Tsuchida (1997), through research lessons, a teacher can view teaching from multiple angles, as immediate decision-making and colleagues' feedback can provide a holistic view of their teaching profile that may not have been apparent before. Moreover, they regarded Lesson Study as a powerful method for disseminating new content and approaches, particularly during curriculum reforms, since teachers are not only exposed to its implementation but also engage in conversations about the reasons behind these changes. Yoshida (2012) argues that Japanese teachers benefit from the designated time for reflection built into the structure of Lesson Study, which supports ongoing professional growth by providing opportunities to observe various types of lessons both within and outside their schools.

The findings of McMahon and Hines (2008) and Rekalidou, Karadimitriou & Moumoulidou (2014) suggest that lesson study provides a constructive platform for teachers' collaborative reflection aimed at improving their teaching practises. Experienced teachers grow professionally through active involvement in lesson study. According to Saito, Hadiprawiro, and Empedhe (2008), lesson study is fundamentally about reflecting on students' learning. Chassels and Melville's results indicated that through lesson study, teachers can create professional learning communities where they deepen their understanding of the curriculum, while also engaging in critical habits of observation, analysis, and reflection.

Furthermore, in addition to being an intensive examination of a lesson with its various collaborative elements, targeted feedback, reflective practises, etc., this creative approach played a crucial role in bridging the gap between theory and practise, thereby fostering personal growth and significantly contributing to professional development (Dania, Voutsina, & Moumouri, 2013). According to Zhou and Xu (2017), research lessons serve as effective means to promote the exchange of innovative teaching approaches, the development of learning communities, and the recognition of the benefits of collaboration. They also emphasized that while engaging in lesson study, teachers have opportunities to enhance their professional identities.

Conversely, there has been limited research on how lesson study (LS) impacts learning activities (LA). In a study conducted by Fernandez (2010), active learning—characterized by deep discussions, planning, and practise—can be facilitated when an informed consultant supports continuous collaboration among teachers along with testing

and analytical capabilities. These components collectively constitute what he termed micro-teaching within the lesson study framework. In addition, Cinkara (2020) experimented with a model called “the quiz study model” to assess its effectiveness in enhancing teachers' professional development concerning assessment practises. He discovered that it can be a valuable tool for fulfilling assessment-related professional development requirements. The quiz study model suggests applicability not only in similar teaching environments but also in teacher education contexts, where pre-service educators can adopt it as an alternative professional development model for themselves.

2.12. Difficulties in Implementing Lesson Study

There are several challenges that may arise during the implementation of lesson study, despite its many benefits. According to Harle (2008), there is concern about whether lesson study can be adapted to different educational systems due to cultural diversity. The entire process of lesson study can take several weeks or even months because of its extensive nature. Chassels and Melville (2009) conducted research that revealed it requires significant time, thus presenting time-related challenges during implementation because it demands more time than expected.

Similarly, Carrier's (2011) findings indicated that student learning was not sufficiently emphasized in lesson study, as one key area was overlooked. For teachers to conduct regular lesson studies effectively, they need both collaborative planning time and adequate periods to implement what they have planned.

Moreover, some teachers may be reluctant to spend their free periods collaborating on aspects like implementing reflections later while planning their teaching and discussing these areas together since lesson study relies on collaboration (Lewis, 2002). Additionally, Rekalidou et al.'s (2014) research evaluated how effective this model could be when colleagues work together towards achieving common goals within schools; however, some staff members experienced challenges while making significant efforts, leading to tensions that could have been avoided if not for factors arising from individuals' attempts to meet the demands set forth by programmes over specific periods. Furthermore, according to Chassels' and Melville's (2009) findings about the advantages and disadvantages of this strategy, similarities can be observed between preservice teacher education and in-service training, particularly when used as a professional development approach among teachers. They found that the main issues primarily

cantered around time constraints within schools and administrative structures that hindered effective teacher collaboration.

Giannakidou, Gioftsali & Tzioras (2013) concluded that the learner participants in their study did not have enough time to reflect critically on their experiences because the process was implemented too quickly. Additionally, teachers' lack of reflective thinking about their actions, compounded by their inexperience, increased their stress during implementation.

2.13. Chapter Summary

This chapter reviewed the theoretical and empirical literature relevant to the study. It examined the concept of LAL, its definitions, frameworks, and the competencies it requires of teachers, as well as its role in supporting effective teaching and learning. The chapter also explored professional development models for teachers, with a particular focus on collaborative approaches, before turning to the Lesson Study Model (LSM) as a form of sustainable, practice-based professional development. In addition, the review discussed principles of language assessment, common challenges faced by teachers in assessment practice, and previous studies on teachers' levels of LAL worldwide. Special attention was given to the limited research in the Iraqi context, where teachers' LAL has not been systematically examined, and to the absence of studies applying LSM to language test writing. The chapter concluded by identifying key gaps in the literature, namely the lack of empirical work in the Iraqi higher education context, the scarcity of studies exploring the processes through which LAL develops, and the neglect of contextual realities influencing assessment practice. These gaps provided the rationale for the current study. The next chapter describes the methodology adopted to investigate these issues.

CHAPTER III

METHODOLOGY

3.1. Introduction

This investigation aimed to illuminate the academic and professional needs of teachers as test writers regarding language assessment literacy. Moreover, the study sought to determine the extent to which higher education instructors at Language Centres possessed language assessment skills. Additionally, the research examined how the participating teachers utilized the 'lesson study' model to design language tests. It should also be noted that this study aimed to demonstrate the role of the 'lesson study' model in enhancing the professional development of test writers. Another goal of this research was to uncover test writers' attitudes toward the Lesson Study Model in relation to language assessment literacy and professional development. Ultimately, this research project focused on identifying the potential contributions of the lesson study process to language assessment literacy among test writers and how it supports their professional development.

This section elaborates on the research methodology. Therefore, the 'Context' of the research is described, focusing on the educational landscape and the importance of language assessment literacy within EFL teaching. Based on this foundational knowledge, the 'Research Design' is detailed, with the rationale explained thoroughly. Additionally, the alternative methodologies considered are discussed. This is followed by the introduction of the study's 'Participants,' the criteria used for their selection, and how the sample size was determined. Finally, the measures taken to ensure unbiased representation are described in detail.

The 'Instruments' section provides detailed information about the tools and measures used to collect data on EFL teachers' language assessment literacy. This part explains how each instrument was developed and adapted, highlighting the relevance of each one and offering a thorough description of their roles in the study.

Additionally, the 'Data Collection Procedure' delivers comprehensive information on the systematic processes used to gather data from participants. This section also

addresses ethical considerations, validation procedures, and the implementation of instruments.

'The 'Data Analysis' section discusses the methods and techniques used for processing, organizing, and interpreting the collected data. In this regard, emphasis is placed on the alignment between the selected data analysis methods and the research aims, focusing on the rigor and precision of the approach. Indeed, this section can be regarded as the cornerstone of the study, providing the necessary groundwork. This aids in fully understanding the methodology and the ensuing findings of the study.

3.2. Research Context

This investigation was conducted at the language centre affiliated with Salahaddin University in Erbil, Iraq. This Language Centre is dedicated to enhancing English proficiency among various learners. Established in 2005 as the first public language centre in Iraq, it plays a crucial role in providing English language courses to civil servants, university academics, graduates from Iraqi universities, and private sector staff.

Seeking to enhance L2 skills for academic and professional advancement, this centre is on a mission to reinforce language learning and cultural exchange by offering English courses to numerous students aiming to improve their English language proficiency. The growing number of students at this centre is a testament to its popularity, along with the increasing demand for effective L2 instruction in the region. This centre enrolls students from diverse professional and cultural backgrounds, all seeking to master English as a requirement for their personal and professional development.

A glance at the demographic details of the applicants enrolled in this centre reveals a wide spectrum of students belonging to various age groups, educational backgrounds, and occupations. This diversity enriches the learning experience by enhancing cultural interactions and paving the way for a dynamic exchange of views and perspectives. As a result of this level of diversity, the centre fosters a vibrant learning community where learners can benefit from each other's experiences and ideas.

Committed to providing high-quality instruction, the Language centre utilizes a diverse array of instructional resources and materials to meet the needs of all learners. These resources include the following:

- Textbooks
- Multimedia tools
- Online platforms
- Access to language labs

As a result, the centre creates a dynamic learning environment where learners are encouraged to engage and develop their skills actively. Additionally, the centre utilizes various assessment methods to track learners' progress and proficiency in English, ensuring effective measurement of learning outcomes.

This centre makes thoughtful efforts to design and structure the curriculum to provide learners at various proficiency levels with opportunities for comprehensive learning. From the very beginning, with a focus on basic language skills, to advanced stages that include developing communication strategies, the courses and programmes are tailored to meet the diverse needs and goals of students. Consequently, this structured approach facilitates effective L2 acquisition, creating a supportive learning environment that promotes continuous improvement and growth.

Aiming to provide effective instruction, the Language Centre places a strong emphasis on selecting and training qualified instructors with the necessary expertise. As a result, these instructors participate in well-structured training programmes that equip them with essential pedagogical skills and knowledge, enabling them to facilitate meaningful learning experiences. This focus on teacher development ensures that instructors meet the critical requirements for delivering engaging and effective lessons that address learners' needs.

With a commitment to continuous improvement, the Language Centre gathers feedback from both students and teachers to enhance L2 instruction and programme delivery. Such feedback mechanisms allow stakeholders to gain valuable insights that inform curriculum development, teaching methodologies, and programme effectiveness. The integration of feedback into decision-making processes ensures that the centre's programmes remain responsive to the evolving needs and expectations of students.

Despite its accomplishments, the Language Centre faces several challenges and barriers regarding L2 assessment and test writing. Some of these challenges include limited resources, varying levels of learner proficiency, and the need for alignment

between assessment practises and international standards. To address these challenges, the centre adopts a proactive strategy that emphasizes providing ongoing support and training for instructors, fostering collaboration with external partners, and staying informed about current trends and best practises in language assessment.

To ensure teachers are aware of the latest trends and effective practises in L2 assessment, the Language Centre offers a variety of professional development opportunities. These opportunities include hands-on workshops, meetings, and collaborative projects, enabling instructors to enhance their skills, expand their knowledge, and stay abreast of the latest developments in language assessment. By investing in instructor professional development, the centre ensures that educators have the knowledge necessary to teach effectively.

Additionally, the Language Centre actively seeks collaborations and partnerships with external organizations and institutions to improve its language assessment practises and outcomes. By engaging with stakeholders such as government agencies, universities, and private sector organizations, the centre promotes a culture of collaboration and knowledge exchange that benefits both students and teachers. These partnerships allow the centre to utilize external expertise, resources, and best practises to enhance its language assessment methods and encourage continuous improvement. In this research project, the cooperation among the Language Centre authorities reflects the collaborative atmosphere, which contributes to enhancing professional development opportunities for instructors.

As for quality assurance and constant improvement, the Language Centre has developed effective mechanisms such as regular assessments, evaluations, and feedback. This makes it possible for the centre to pinpoint the areas for improvement and the implementation of purposeful interventions. Placing enormous emphasis on quality assurance, the centre makes sure that its language assessment practises satisfy the excellence standards, contributing to materializing learning outcomes for students.

Having a forward-looking approach, the Language Centre has set high goals and standards regarding language assessment literacy, as well as professional development for teachers. Some of these goals are as follows:

- An increase in the opportunities for professional development
- Enhanced collaboration with external partners
- The integration of innovative technologies into L2 assessment practises
- The reinforced culture of constant improvement

Thanks to these aims and priorities, the centre seeks to reinforce its reputation as a pioneer English Learning institution in the region. All in all, the centre at Salahaddin University-Erbil contributes greatly to the promotion of English language proficiency among various learners.

Due to its inclusive courses and programmes, loyal instructors, creative approaches to assessment, as well as its commitment to constant improvement, the centre tries to stick to its mission of improving English learning and cultural exchange. Placing emphasis on collaboration, innovation, and the quality of teaching, the centre serves as the leading English learning institute in Iraq.

3.3. Research Design

This research used a case study approach to examine the professional growth of Iraqi test writers within the Lesson Study Model. The focus and purposes of the study objectives can best be addressed using case study methodology because it offers in-depth, context-sensitive analysis of complicated issues within real-life situations (Yin, 2009). Unlike some designs that prioritize generalization, case studies concentrate on comprehension of a bounded system in this particular instance, the experiences of test writers from Iraq, through abundant rich data (Duff, 2018). This approach aligns with the study's aim to uncover how and why teachers' language assessment literacy (LAL) evolves within a collaborative professional framework.

The shift to a case study design was driven by the study's contextual constraints and epistemological goals. With a small, purposive sample of 10 participants, statistical generalization was neither feasible nor appropriate (Gerring, 2004). Instead, the case study framework allowed for analytical generalization, wherein findings are theoretically transferable to similar contexts (Tsang, 2014). By focusing on a single, bounded case, Iraqi test writers engaged in Lesson Study, the design facilitates a granular examination of their needs, knowledge levels, and developmental trajectories (Hancock et al., 2021).

As Duff (2014) pointed out, applied linguistics case studies tend to capture the ‘messy reality’ of language practises and therefore are suitable for investigating the relationships between personal and organizational cultural influences on LAL development in an LAL context.

This study also strengthens credibility through triangulation by drawing from various data sources, including interviews, reflective journals, and questionnaires (Yin, 2009). While the original mixed-method plan sought breadth through quantitative data, the case study prioritizes depth, leveraging qualitative tools to unpack participants’ lived experiences. For instance, reflective journals provided longitudinal insights into test writers’ evolving attitudes, while interviews illuminated the socio-cultural barriers shaping their engagement with Lesson Study (Thomas, 2011). This aligns with Duff’s (2018) assertion that case studies in language education thrive when examining how individuals or groups navigate complex, situated learning processes.

Furthermore, the exploratory nature of the research questions necessitated a flexible, iterative design. Case studies permit emergent themes to guide analysis, which is critical when investigating under-researched populations like Iraqi test writers (Hancock et al., 2021). Questions about *how* Lesson Study influences professional development (RQ3) and *why* attitudes shift (RQ4) require a design that privileges context over quantification. As Gerring (2004) argues, case studies are particularly suited to causal inference in complex, real-world scenarios where variables cannot be isolated. By treating the Lesson Study intervention as an embedded case, the study disentangles the mechanisms through which collaborative inquiry fosters LAL growth.

Critically, the design addresses concerns about validity and rigor. Yin (2009) emphasizes construct validity through operationalizing theoretical concepts (e.g., LAL) via multiple evidence sources. Thematic analysis of interviews and journals, coupled with questionnaire-derived profiles of assessment knowledge, ensures that findings are grounded in participants’ realities rather than abstract metrics. Additionally, the iterative comparison of data streams mitigates researcher bias, a key advantage in studies involving reflexive practises like Lesson Study (Duff, 2014).

3.4. Participants

The sample of this study was made up of the teachers, the researcher, field experts, and a qualitative data analysis expert. What follows is a description of the participants and their involvement in the research process.

3.4.1. Teachers

The participants of the present study comprised 10 Iraqi EFL teachers teaching at different levels of language proficiency in the Language Centre of Salahaddin University. All the participants were native speakers of Arabic and 6 were male while 4 were female teachers. Their age ranged between 31 to 45. Their teaching experience fell within the range of a few months to 15 years. Table 1 displays the demographic information of the participants.

Table 1.

Demographic Features of the Participants

Pseudonym	Gender	Age	Academic Degree	Experience	Major
Ms. Ranya	Female	31	MA	4	TEFL
Ms. Dalia	Female	39	MA	2	TEFL
Dr. Tara	Female	44	PhD	6	TEFL
Dr. Suzan	Female	40	PhD	13	TEFL
Dr. Yousif	Male	41	PhD	10	TEFL
Dr. Kamal	Male	42	PhD	8	TEFL
Mr. Rebaz	Male	32	MA	7	TEFL
Mr. Haval	Male	35	MA	6	TEFL
Dr. Farman	Male	36	PhD	14	TEFL
Mr. Salar	Male	38	MA	2	TEFL

Maximum variation sampling was considered to select the participants in order to consider the variations existing in real settings. Maximum variation sampling can help yield results which highlight the shared patterns while cutting across individual participants to capitalize on the extant heterogeneity in the sample (Patton, 2002). Accordingly, in recruiting the participants, attempts were made to include both male and female teachers. Moreover, variations in terms of academic degrees, and age to the extent possible were considered to accommodate a wide range of participants.

The teachers were selected based on convenience non-random sampling procedures as the researcher could not choose teachers based on pure random sampling. Therefore, the researcher negotiated with 18 teachers to take part in the study and was

able to win the consent of 10 to partake in the present research project. The participants were selected from those teachers teaching at the Language Centre of Salahaddin University and EFL teachers teaching in public state schools were excluded due to two reasons. Firstly, it was not manageable for the researcher to choose participants from private institutes and public schools. Moreover, delimiting the participants to a specific educational context would yield more consistent results as teachers' professional development is context-sensitive (Sandholtz & Ringstaff, 2016).

3.4.2. Researcher

The researcher was involved in different stages of the implementation of the study including participants selection, instruments selection and administration, delivering the intervention of the Lesson Study Model, and data analysis. As for the participant selection, the researcher was involved in negotiation with the University administrators and participants. With regards to instruments selection, the researcher drawing on the extant literature tried to choose the best instruments for the data collection. Moreover, the researcher did his best to administer the data collection tools so that there were no ambiguities for the participants to collect rigorous and rich data. With respect to the intervention of the Lesson Study Model, the researcher made attempts to carry out both the theoretical and practical aspects of the model based on the available literature as effectively and efficiently as possible. Thus, the researcher had a key role in this study since he was present for all the intervention sessions and made attempts to carry out the intervention to perfection. In so doing, the researcher planned each intervention session based on the available theoretical frameworks and presented the relevant contents to the participants. The researcher also monitored the participants while they were involved in the lesson study plan interventions. As for the data analysis section, the researcher organised the quantitative data and analysed it. Concerning the qualitative data analysis, the researcher was one of the analysers of the data.

3.4.3. Experts

Three experts holding PhD in TEFL assisted the researcher in developing the semi-structured interview questions. These experts had about 15 years of teaching experience at university and were also experienced in the field of educational research. The experts were teaching at Salahaddin University-Erbil in Iraq.

3.4.4. Qualitative Data Analysis Expert

In this study, a co-researcher assisted the researcher with collecting and analysing the qualitative data. The research assistant holds a PhD in TEFL with more than 10 years of teaching experience and qualitative data analysis.

3.5. Data Collection Tools

This investigation made use of multiple data collection tools, including a five-section questionnaire and two sets of semi-structured interviews. The research was conducted during the 2022–2023 academic year at Salahaddin University. The process started in October 2022 with the collection of participants' existing test papers and the first planning meetings. The initial Lesson Study Model (LSM) cycle, which involved the stages of planning, observation, and reflection, was carried out between October and December 2022. The second cycle was organised from January to February 2023, during which the revised test tasks were implemented and evaluated. Classroom observations were aligned with the regular course assessments, and interviews together with reflective writings were collected after each cycle. Altogether, the data collection covered a period of about five months. The data collection tools are described as follows:

3.5.1. Language Assessment Literacy Questionnaire (LALQ)

As an effective data collection instrument, questionnaires are frequently used in educational research. Such an instrument is aimed at collecting data due to some advantages; questionnaires yield many benefits in terms of data collection in educational research, such as high efficiency, standard format, respondent's anonymity, provision of quantitative data, and cost-effectiveness. These advantages render questionnaires an important instrument for collecting data and gaining information on participants' perspectives. In fact, questionnaires serve as an efficient tool for gathering data from a large number of people in a relatively short period of time.

According to Dillman et al. (2014), questionnaires enable researchers to gather responses from a large sample, offering a broader perspective on the research topic. Additionally, they provide standardized measures for data collection, ensuring consistent question presentation and uniform response recording. This standardization enhances both the reliability and validity of the data. (Fowler, 2013).

Also, the use of questionnaires allows for the anonymity of the participants, enabling them to provide honest responses without any concern for judgment or bias. This paves the way for more accurate responses, especially in terms of sensitive topics (Tourangeau et al., 2000). Moreover, questionnaires yield quantitative data that researchers can easily analyse using statistical methods. Therefore, researchers can spot different patterns, trends, and relationships within the data, giving important insights into the research topic (Babbie, 2020). Last but not the least, questionnaires serve as cost-effective data collection method compared to other techniques such as interviews. The administration and analysis of questionnaire requires fewer resources, rendering them an attractive option (Bryman, 2016).

This study administered a questionnaire made up of five sections (appendix A). Section one elicited information on the participants' backgrounds (e.g., gender, the educational qualifications, specializations, experience, and the possible membership at a testing office). Section two was aimed at collecting data about self-perceived assessment skills among the teachers. This section contained four items that sought to elicit information on the extent to which the participants used the assessment issue stated by each of the statements regarding the four language skills of reading, listening, writing, and speaking.

The participants were required to indicate their level of assessment knowledge by ticking one of the choices next to each statement according to the five-point scale: 1 not competent, 2 a little competent, 3 somewhat competent, 4 competent, and 5 very competent. It should be noted that the four items in this section were only part of the whole questionnaire aiming to assess language assessment literacy. Moreover, the second section was not aimed to measure language assessment literacy. It was aimed at assessing how skilled the respondents thought themselves to be in using the assessment issue described by each of the statements. The content validity of all the items in the questionnaire was subject to expert opinion prior to administering the questionnaire. To this aim, the questionnaire items were discussed in a panel of experts consisting of two PhD holders and it was decided that the items measured what they were supposed to measure. Moreover, Cronbach's Alpha was also applied, and the results indicated that the questionnaire and its different sub-sections had adequate levels of internal consistency.

The third section was given to the participants to gather data regarding training in English language testing and evaluation. In this section, the participants were required to

specify the amount of training they received in the domains of language testing and evaluation during their university education. This section comprised 10 items and the respondents were required to indicate the amount of training they had received using a four-point Likert scale consisting of none (1), little (2), sufficient/intermediate (3), and extensive/advanced (4). The Likert-scale provided options from none (i.e., no training at all) to extensive/advanced training and therefore provided coverage for the possible range and magnitude of training received by all the participants. The fourth section was employed to collect data regarding language assessment knowledge. This section was administered to assess the level of language assessment knowledge of the teachers who teach at higher education level at Language Centres. The fourth section had four sub-sections tapping into assessment knowledge of the four language skills of reading, listening, writing, and speaking. Each section contained 10 items with three options consisting of true, false, and don't know. The difference between the second and fourth sections was that while in the second section the respondents provided their self-perceptions of assessment skills via four items, the fourth section aimed to measure their assessment knowledge in a detailed manner with 40 items. The last section tapped into the academic and professional needs of teachers as test writers in terms of language assessment literacy. This section consisted of 26 items on a 5-point Likert scale format comprising not at all essential (1), little essential (2), somewhat essential (3), essential (4), and very essential (5). All sections of the questionnaire were administered to the participants prior to the implementation of the Lesson Study Model. After the Lesson Study Model, all the sections were filled out by the participants once again except for sections one and three. Section one was not administered again as the researcher had already gathered the demographic information of the participants. As for section three, it was not administered since it dealt with the amount of training the participants had received in the domains of language testing and evaluation during their university education. The following table demonstrates the results of Cronbach's Alpha for different sections and sub-sections of the questionnaire.

Table 2.

Reliability Coefficients for Questionnaire Sections

Section/Sub-section	Number of Items	Cronbach's Alpha (α)
Self-Perceived Assessment Skills	4	0.82
Training in Testing/Evaluation	10	0.89
Language Assessment Knowledge	40	0.90
→ Sub-section 4.1: Reading	10	0.87
→ Sub-section 4.2: Listening	10	0.92
→ Sub-section 4.3: Writing	10	0.93
→ Sub-section 4.4: Speaking	10	0.88
Section 5: Academic/Professional Needs	26	0.93

As indicated in the above table, all Cronbach's Alpha values for the sections and sub-sections of the questionnaire surpass the acceptable threshold of .70. Accordingly, the questionnaire and all its sections and sub-sections possess satisfactory levels of reliability (Hulin et al., 2001; Tavakol & Dennick, 2011).

3.5.2. Semi-Structured Interviews

Semi-structured interviews serve as a widely used data collection instrument in the context of educational research, given their multiple advantages, such as their capability of in-depth exploration, high flexibility, engagement power, contextual understanding, as well as data triangulation. Such a type of data collection instrument makes it easy for the researcher to deeply explore participants' attitudes, past experiences, and beliefs. Using open-ended questions, researchers are able to gain nuanced insights that may not be obtained through other data collection methods (Rubin & Rubin, 2011).

Moreover, semi-structured interviews allow for a balance between structure and flexibility, enabling researchers to use consistently a set of questions while probing the unplanned avenues that may occur during the interview at the same time (Denzin & Lincoln, 2000). Such a level of flexibility lets researchers gain thorough insights into participants' answers, which makes it possible to gather rich, detailed data. Also, such a type of interview fosters participant engagement by paving the way for the creation of a conversational and interactive platform which can be used for data collection. Being so, participants find it easier to share their thoughts and experiences using a semi-structured interview format, yielding more transparent and authentic responses (Seidman, 2006).

In fact, interviews allow researchers to gain a contextual awareness of participants' experiences embedded in cultural, social, and professional environments. Probing

participants' perspectives, researchers are able to lay bare the complex nature of their experiences and interpretations (Kvale & Brinkmann, 2009). As another advantage, semi-structured interviews serve as a data triangulation technique whereby researchers can collect multiple perspectives on a given topic. Carrying out interviews with several participants, researchers make a comparison across several viewpoints. This increases the credibility and validity of the study findings (Creswell & Creswell, 2017).

This study developed and used two semi-structured interviews, namely, a set of pre-interview questions prior to the implementation of the Lesson Study Model and a set of post-interview questions after the intervention. To pose the interview questions, the researcher reviewed the literature related to Lesson Study Model, the research findings on language assessment literacy, as well as professional development. This was followed by the preparation of the list of tentative questions, which was studied by the panel of experts. Following that, minor modifications were made to the questions.

Then, the list of questions was piloted on two of the study participants to identify any ambiguities with respect to grammar and wording. Then, minor modifications were made to the questions to enhance their clarity and readability. The final list of pre-interview questions was as follows:

1. How would you describe your understanding of the concept of quiz study in the context of language assessment?
2. Have you had any prior experience with collaborative activities focused on developing language assessments?
3. What are your expectations for engaging in the quiz study model as a professional development activity for language assessment?
4. Are there specific goals you hope to achieve through participating in the quiz study process?
5. From your perspective, what potential benefits do you anticipate for your language assessment literacy from participating in a collaborative quiz study?
6. How do you think quiz study can contribute to your proficiency in designing effective language assessments?
7. How comfortable are you with collaborating with colleagues on the design and evaluation of language assessments?

8. What skills or qualities do you think are important for effective collaboration within a quiz study group focused on language assessment?
9. Briefly describe your experience in designing language assessments for your EFL classes.
10. What challenges, if any, have you faced in the past when designing language assessments?

The final list of post-interview questions was as follows:

1. Reflect on your overall experience with the quiz study model in the context of language assessment. What aspects of the process did you find most valuable?
2. How has participating in quiz study influenced your perceptions of collaborative professional development in the field of language assessment?
3. In what ways, if any, did the quiz study model impact your approach to designing language assessments?
4. Can you provide specific examples of changes or improvements you made to your assessments based on insights gained from the quiz study?
5. Evaluate the effectiveness of communication and collaboration within your quiz study group, particularly concerning the design and evaluation of language assessments.
6. How did the collaborative aspects of quiz study contribute to the development and refinement of the language assessments?
7. How did the feedback received from your colleagues during the post-implementation meeting influence your understanding of the language assessments you designed?
8. Reflect on the role of reflective practises in quiz study. How did reflection contribute to your professional growth in language assessment?
9. Do you see elements of the quiz study experience that you can apply to future language assessment planning and design?
10. How might the quiz study model continue to impact your language assessment practises beyond this study?

3.5.3 Document Analysis

As part of the data collection process, document analysis was conducted. The documents analysed were mainly teacher-made quizzes, classroom tests, and assessment tasks that participating teachers designed and used in their classes before and during the Lesson Study cycles. These documents were collected from the English Language Centre at Salahaddin University. The rationale for selecting them was that they directly reflected the teachers' existing assessment literacy in practice and served as authentic evidence of their test-writing habits. In addition, revised versions of the same documents, developed collaboratively during the Lesson Study sessions, were also included in the analysis to trace changes and improvements in test design. Thus, the documents provided a concrete basis for comparing participants' assessment practices before and after their involvement in the study.

3.5.4 Reflective Journal

Reflective journals are a valuable tool for collecting data in educational research, particularly in studies focused on professional development and teacher training. The literature review indicates that the use of reflective journals yields rich information through descriptions, resulting in the following outcomes:

- Enhancing critical thinking
- Fostering self-reflection,
- Reinforcing professional development
- Improving data authenticity.

As argued by Kolb (2014), reflective journals pave the way for participants to give a detailed reflective description of their experiences in their own words. As a result, researchers can get access to rich, nuanced data that may not be obtained through other data collection methods. Also, using reflective journaling, participants are provided with the opportunity to self-reflect on their teaching practises, perceptions, and experiences, leading to an increase in self-awareness and personal growth (Boud et al., 2013).

Moreover, reflective journals enable individuals to contemplate deeply over their past experiences, actions, and beliefs. Reflecting on their own thoughts and writing them,

participants start engaging in critical thinking processes, which results in a deeper understanding of their own practises (Moon, 2006). Furthermore, reflective journaling serves as an effective tool for professional development given that it makes it possible for the participants to track their growth over time. Documenting their thoughts and reflections, these individuals can pinpoint the areas that need to be improved and set goals for their own development (Zeichner & Liston, 2013).

Also, reflective journals open up a direct window through which participants' thoughts and experiences can be seen. This provides authentic data that is immune to external factors, enhancing the credibility and validity of the research findings (Moon, 2006). Therefore, reflective journals can be used as an effective data collection tool in educational research, especially in case of the studies that deal with professional development and teacher training. Given that such journals promote critical thinking, encourage self-reflection, obtain rich descriptions, support professional development, and enhance data authenticity, they enable researchers to gain invaluable insights into participants' experiences and perspectives. The present study's use of reflective journal drew on Schön's (1983, 1987) two types of reflection - reflection-in-action & reflection-on-action. The participants had to reflect prior to the experience, during the experience, and following the experience.

As for before the experience, they were asked to think about the things that could have happened, the things that they felt might be a challenge, and the things that they could do to prepare for the challenges and experiences. With regard to the during the experience phase, they were required to reflect upon what was happening at the moment as they made a particular decision, whether it was working out as expected, and were they dealing with the challenges well, and was there anything they should do, say or think to make the experience successful. Concerning after the experience, the participants were required to describe their thoughts immediately after, and/or later when they had more emotional distance from the event, whether they would do anything differently before or during a similar event, and what were the takeaways from the experience/lesson. They were also required to write their reflections on the process regarding how the process contributed to their language assessment literacy level as a test-writers and how its implementation assisted them to improve professionally.

3.6. Data Collection Procedure

This study was conducted in four phases including participants' selection, administration of the questionnaire and the interviews prior to the Lesson Study Model, application of the Lesson Study Model comprising the theoretical and practical phases, and post-intervention application of the questionnaire and the interviews, which are fully elaborated on below:

3.6.1. Participants' Selection

To recruit the participants for the study, the researcher initially contacted the head of Language Centre, Salahaddin university to obtain their consent for the research study at Language Centre. In so doing, the title of the research along with the aims of the study and how the study was going to unfold were explained. Upon receiving the consent from the head of Language Centre, the researcher contacted 17 teachers at the Language Centre. Out of the 17 participants, 10 consented to partake in the study. The prospective participants were sent an email in which they were briefed on the purposes of the study and the procedures. Moreover, they were informed that the collected data would be kept confidential with the researcher and used only for research purposes. Moreover, the participants were informed that their participation in the study was voluntary, and they could withdraw from the study at any stages of data collection. The participants were also provided with the researcher's contact information to ask any questions or concerns about the study.

3.6.2. Pre-application of the Questionnaire

Upon receiving the consent form from the participants, the researcher gave them a questionnaire consisting of 5 sections to fill out. The respondents were asked to send any queries they had regarding different sections of the questionnaire and how to fill it out. Following the questionnaires completion, the participants delivered the questionnaires to the researcher via email.

3.6.3. Theoretical Phase

In the theoretical phase which consisted of four 90-minute sessions, the researcher delineated the theoretical underpinnings of the quiz study based on the Lesson Study Model.

3.6.3.1. Session One: Introduction to Lesson Study Model

The first session kicked off with the introduction of the teachers to the foundational principles and theoretical foundations of the Lesson Study Model. At the very beginning, a thorough history and origins of Lesson Study were presented, with a focus on its roots in Japanese educational practises and its evolving growth as a collaborative professional development approach.

Based on the research findings of Stigler and Hiebert (1999), the teachers were provided with ample insights into the main components of Lesson Study, such as collaborative lesson planning, observation, and reflection. Using interactive forum and case studies, the teachers managed to probe the theoretical framework of Lesson Study as a formative assessment tool that enhances teacher learning and student achievement.

3.6.3.2. Session Two: Understanding Collaborative Inquiry in Lesson Study

Drawing on the underlying knowledge gained in the first session, session two aimed to improve the teachers' understanding of collaborative inquiry within the Lesson Study Model. In accordance with the research findings of Lewis (2002) and Fernandez and Yoshida (2004), the teachers participated in the activities that focused on the contribution of collaborative inquiry to fostering collaboration among teachers, knowledge sharing, and pedagogical improvement. Engaging in discussions and case analyses, the teachers proceeded to explore the way in which collaborative inquiry processes (e.g., lesson planning activities and post-lesson discussions) play a role in improving teaching practises and student learning outcomes within the Lesson Study framework.

3.6.3.3. Session Three: Implementing Lesson Study for Professional Learning Communities

Session three unfolded as follows: the teachers tried to probe the notion of Professional Learning Communities (PLCs) with regard to the implementation of Lesson Study. Using the work of Hord (1997) and DuFour and Eaker (1998), the teachers sought to find out how PLCs can provide a propping structure for collaborative lesson planning, peer observation, and reflective practise. The teachers examined the contribution of PLCs to improving constant improvement through role-playing exercises and scenario analyses. They shared accountability in implementing Lesson Study. The session focused on the important role of trust-building and creating norms within PLCs to streamline effective collaboration and professional growth.

3.6.3.4. Session Four: Reflective Practise and Continuous Improvement in Lesson Study

The last session dealt with reinforcing a culture of reflective practise and constant improvement within the Lesson Study Model. Based on the research of Schön (1987) and Zeichner and Liston (2013), the teachers participated in those reflective activities that pushed them to have critical evaluation of their experiences with Lesson Study. Accordingly, they proceeded to identify areas that needed to be improved and set goals for constant professional development. Indeed, due to guided reflection guidelines and peer feedback sessions, the teachers managed to make action plans thereby making reflective practises as a component of their daily teaching routines. They also leveraged Lesson Study as a vehicle for constant improvement. The session focused on the important role of self-assessment, as well as collaborative reflection in improving teaching effectiveness and student learning outcomes within the Lesson Study framework.

3.6.4 Practical Phase

In practice, each Lesson Study cycle consisted of three main stages: planning, observation, and reflection. During the planning sessions, participants analysed their existing test papers and collaboratively designed new tasks. In the observation stage, these tasks were trialled in real classrooms, where both student performance and teacher

administration were systematically noted. In the reflection sessions, the group evaluated the strengths and weaknesses of the test items and revised them based on the classroom evidence. The cycle was conducted twice, which allowed participants not only to redesign tasks but also to test their effectiveness under authentic classroom conditions.

3.6.4.1 Administration of the Pre-Interview Questions

The first stage of the practical phase involved conducting the pre-interviews. In so doing, the researcher carried out interviews with the participants individually in a face-to-face manner. Each interview lasted about 40 minutes and the interviewees' responses were recorded after seeking their permission. The teachers were given the choice to select either Arabic or English as the language during the time of the interviews and they all opted for English.

3.6.4.2. Introducing the Reflective Journal to the Participants

The participants were given the guidelines for the reflective journal. In so doing, a session of 1 hour was held in which all the participants were present and they were informed about the before the experience, during the experience, after the experience parts of the journal. The participants were requested to document their experiences and reflections on the entire process, contributing valuable insights to the study through writing reflective journals.

3.6.4.3. Practical Implementation of the Lesson Study Model

The first stage of the practical implementation of the Lesson Study Model was a *collaborative meeting*. A joint meeting was organised to define the content and specifications of the quiz. During the initial session of the lesson study plan, the participants engaged in a collaborative process to define the content and specifications of the quiz for EFL learners. The steps involved in this process comprised:

- **Brainstorming and Idea Generation:** using this technique, the participants engaged in collective activities aimed at generating ideas and making suggestions for the quiz content in accordance with their understanding of EFL learners' needs. Accordingly, they shared their viewpoints and insights to build a diverse pool of potential quiz topics.

- **Discussion and Consensus Building:** After undergoing the idea generation phase, participants participated in reflective discussions where they evaluated and refined the proposed quiz topics. Accordingly, they reflected on the relevance, difficulty level, as well as compliance with learning objectives. The aim was to make sure that the quiz served as effective assessment of EFL learners' language proficiency.
- **Defining Quiz Specifications:** Following the consensus over the quiz content, participants defined the specifications and details, such as the number of questions, questions formats (e.g., multiple-choice, fill-in-the-blank), scoring scheme, and testing duration. This step was aimed at developing clear-cut guidelines for designing and implementing the quiz.
- **Feedback and Iteration:** The participants provided each other with different types of feedback. They shared their ideas to improve the quality and validity of the quiz. This iterative approach paved the way for constant improvement and refinement given the constructive input provided by all the participants.

Stage two which covers the practical implementation of the Lesson Study Model was comprised of *designing and reviewing the quiz*. The second session involved the participants' engagement in a structured process aimed at designing and reviewing a quiz for EFL learners. The steps followed for the design of the quiz individually were as follows:

- **Understanding the Objectives:** Initially participants were introduced to the learning objectives, as well as assessment criteria to make sure that the quiz content is aligned with the EFL learners' intended outcomes.
- **Creating Quiz Items:** Each participant was encouraged to develop quiz items that represent the language skills and competencies embedded in the lesson. They considered factors, including question format, difficulty level, and relevance to the learning material.
- **Reviewing and Refining Items:** Following the design of quiz items, participants proceeded to review their questions critically to assess clarity, the level of accuracy, and compliance with EFL learners' needs. Accordingly,

they revised the items as needed to improve the quality and appropriateness of the quiz.

This was followed by reviewing the quiz items, with the participants taking the following steps:

- **Sharing and Presenting Quiz Items:** Each one of the participants shared their quiz items with the group, elaborating on the logic behind their choices. They highlighted the specific considerations that should be considered during the design process.
- **Providing Feedback:** All members of each group shared constructive feedback by ascertaining each other's quiz items given the predetermined criteria such as relevance to learning objectives, the level of language complexity, and fairness. The participants made suggestions for enhancement, diagnosing the possible issues or inconsistencies.
- **Collaborative Revision:** Engaging in team discussion, participants made some revisions to the quiz items so that the feedback received from peers could be addressed. They tried to assure the coherence and balance of the final version of the quiz, as well as its alignment with the stated assessment goals for EFL learners.

Quiz administration was covered in stage three of the practical implementation of the Lesson Study Model. In this session, the students were provided with the quiz during regular class time. The aim was to gather real-time data. As the quiz administration unfolded, the participants used a structured process thereby administering the quiz to EFL learners during regular class time to collect real-time data. The following steps were followed in this process:

- **Setting Up the Quiz:** the quiz materials and instructions were prepared by the participants, making sure that it was now ripe time to distribute the quiz.

- **Briefing the Students:** Prior to the administration of the quiz, participants gave some instructions to the students on the objectives of the assessment, and how they could complete it, along with any relevant guidelines or expectations.
- **Distributing the Quiz:** The quizzes were distributed among the students by the participants so that each one of the learners received a copy of the assessment.
- **Monitoring the Administration:** Participants engaged in the observation of the students who were completing the quiz, raising any difficulties or complexities arising during the assessment process.
- **Peer Observation:** Participants observed their counterparts, tracking the way in which they administered the quiz and participated in the interaction with the students to gain insights and feedback for improvement.
- **Collecting Real-Time Data:** Participants gathered real-time data on how learners performed and responded during the quiz administration, recording observations and writing down any emerging trends or patterns.
- **Filling Out Reflective Journals:** In this part, the participants proceeded to complete reflective journals in which they could write down their thoughts, feelings, and insights about the quiz administration process. This included the perceived challenges, successes, and the areas that needed to be improved.

Feedback constituted stage four in the operationalization of the Lesson Study Model, with the participants encouraged to provide feedback, both from their own viewpoint and that of their students, by holding a post-implementation meeting. In these meetings, participants tried to offer feedback in a principled way as educators and they also tried to imagine what feedback the students would give. The following steps were followed to provide feedback during this meeting:

- **Reflection on the Quiz Administration:** All the participants began to reflect on the quiz administration process, which involved taking into account their perceptions, challenges, and successes related to the implementation. To this end, they shared their thoughts on the effectiveness of the quiz design, instructions, and overall student engagement.

- **Peer Discussion and Feedback:** This part involved the participants' engagement in collective discussions where they shared their perceptions, experiences and insights regarding the quiz administration. They provided feedback to each other given their judgments and observations, raising the areas that needed to be improved.
- **Student Feedback Analysis:** This part involved the participants' reflection on the feedback provided by their students regarding the quiz experience. The factors considered were as follows: learners' responses, performance, and any comments or suggestions given by the learners.
- **Comparison and Synthesis:** The participants proceeded to compare their own perspectives with those of other students in terms of the feedback provided. The aim was to identify shared themes, discrepancies, or areas of alignment.

They synthesized this information to gain a comprehensive understanding of the quiz's impact on student learning.

Action Planning: Based on the feedback gathered, participants collaboratively developed action plans for future improvements in quiz design, administration, and student expectations. They set goals for enhancing the assessment process and promoting student learning outcomes.

The practical implementation of the lesson study plan was repeated three times in line with the lesson study cycle proposed by Dudley (2014). The model is illustrated in Figure 1.

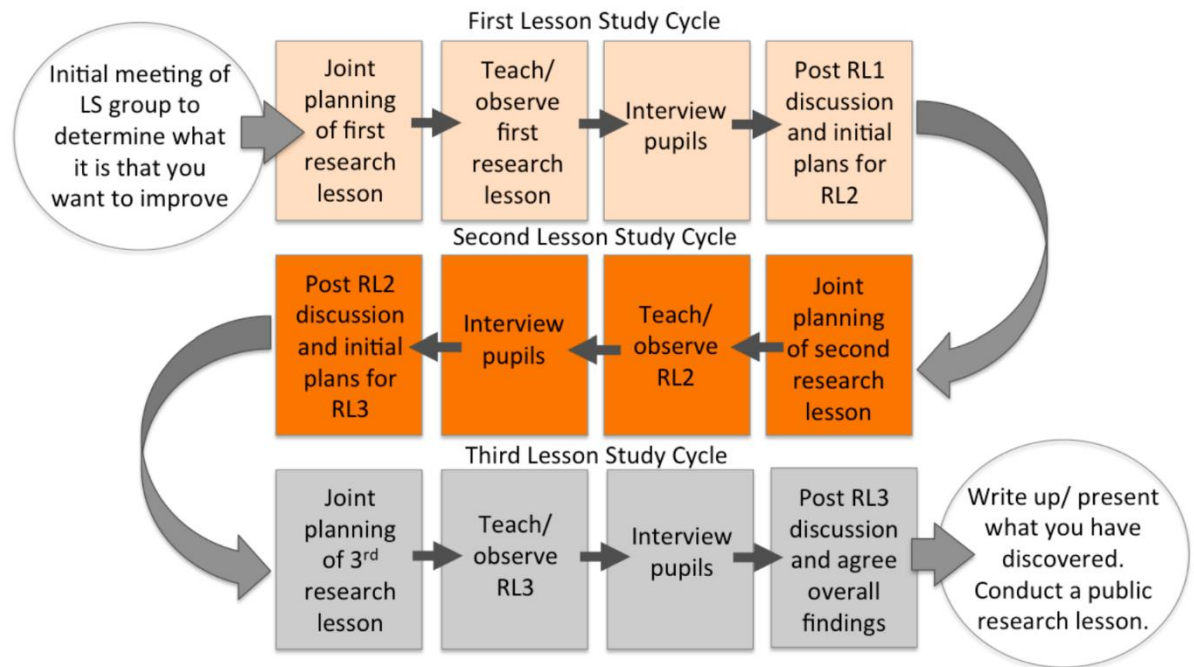


Figure 1. Lesson Study Process

Source: Dudley, 2014

As mentioned, the lesson study plan was repeated three times since three cycles had roughly similar stages. However, for the sake of more clarification, a breakdown of the three cycles with the stages within each cycle is provided. This Lesson Study Model has three cycles. The first lesson study cycle has the following stages:

- Joint planning of first research lesson
- Teach/observe first research lesson (RL)
- Interview pupils
- Post RL1 discussion and initial plans for RL2

The second lesson study cycle has the following stages:

- Post RL2 discussion and initial plans for RL3
- Interview pupils

- Teach observe RL2
- Joint planning of second research plan

The third lesson study cycle has the following stages:

- Joint planning of 3rd research plan
- Teach/observe RL3
- Interview pupils
- Post RL3 discussion and agree on overall findings

Cycle 1: Initial planning and first research lesson (RL)

Objective: To collaboratively identify assessment needs, design the first quiz, and establish baseline insights into teachers' Language Assessment Literacy (LAL).

Stage 1: Joint planning of RL1: A collaborative meeting was organised to define the quiz's content and specifications. Participants engaged in:

- Brainstorming and idea generation: Participants proposed quiz topics aligned with EFL learners' needs, sharing perspectives on grammar, vocabulary, and communicative competence.
- Discussion and consensus building: Debates arose over prioritizing formative vs. summative assessment goals. For instance, one group emphasized real-world application, while another stressed grammatical accuracy. Through facilitated dialogue, they agreed on a hybrid approach.
- Defining quiz specifications: The group finalized 25 items (15 multiple choice questions (MCQs), 5 fill-in-the-blanks, 5 short-answer questions) with a 40-minute timeframe.
- Feedback and iteration: A preliminary rubric was critiqued for vague descriptors; participants rephrased criteria to enhance clarity.

Two participants with prior testing experience led a mini-workshop on aligning items to CEFR levels, resolving disagreements about difficulty calibration. A ‘content map’ was created to visualize skill coverage.

Stage 2: Teach/observe RL1: The quiz was administered in class. Steps included:

- Setting up and briefing: Teachers distributed printed quizzes but faced delays due to inconsistent digital instructions.
- Monitoring and peer observation: Observers noted students struggling with ambiguous fill-in-the-blank prompts (e.g., tense markers not contextualized).
- Reflective journals: One teacher wrote, “I realized my MCQs had ‘trick’ options that confused even proficient learners.”

An observer noted that 30% of students skipped short-answer questions, prompting a post-hoc discussion about task intimidation. Audio recordings captured teacher-student interactions during clarification requests.

Stage 3: Interview Pupils: Post-quiz, 10 students were interviewed. Key feedback included: “The MCQ section felt repetitive, but the short answers. Let me explain my ideas. I didn’t understand what ‘contextually appropriate’ meant in Question 12.

Interviews revealed disparities between teachers’ assumptions and learner experiences. For example, participants had assumed visuals in MCQs would aid comprehension, but students found them distracting.

Stage 4: Post-RL1 discussion & RL2 planning

Participants analysed outcomes:

- Revised ambiguous items (e.g., added tense markers to blanks).
- Agreed to reduce MCQs from 15 to 12 and include 3 scenario-based tasks.
- Planned a peer-review protocol for RL2 to standardize rubric application.

A ‘problem bank’ was created to document recurring issues (e.g., vague rubrics, time mismanagement), with solutions prioritized for RL2.

Cycle 2: Iterative refinement and second research lesson (RL2)

Objective: To address RL1 gaps, refine test design, and evaluate improvements in LAL.

Stage 1: Joint planning of RL2

Building on RL1 insights, the group:

- Replaced 3 MCQs with listening comprehension tasks using audio clips.
- Designed a standardized observation checklist for peer reviewers.
- A subcommittee piloted the revised quiz with 5 students, discovering that audio instructions improved clarity. This led to integrating QR codes for audio support in RL2.

Stage 2: Teach/observe RL2

RL2 administration featured:

- Stricter timekeeping: A countdown timer projected on-screen.
- Peer reviewers scored short answers live using the revised rubric.

Discrepancies in peer scoring emerged (e.g., one teacher awarded full marks for creativity, another deducted for grammar). This highlighted the need for rubric calibration.

Stage 3: Interview pupils

Student feedback from RL2 noted:

- “The audio examples helped me understand the context better.”
- “I still ran out of time for the last scenario task.”

Three interviewees suggested splitting the quiz into two shorter assessments, a proposal debated in post-RL2 discussions.

Stage 4: Post-RL2 discussion & RL3 planning

Key decisions included:

- Splitting the quiz into a 20-minute grammar/vocabulary test and a 25-minute communicative task.
- Conducting a rubric-norming workshop before RL3.

Participants reviewed video clips of RL2 administrations, identifying nonverbal cues of student confusion (e.g., prolonged pauses on specific items).

Cycle 3: Finalization and third research lesson (RL3)

Objective: To consolidate findings, evaluate professional growth, and formalize test-writing guidelines.

Stage 1: Joint Planning of RL3

The group:

- Finalized two separate quizzes with distinct rubrics.
- Trained peers on rubric norming using sample student responses.

A ‘think-aloud’ session was conducted, where teachers verbalized their scoring rationales to align interpretations of ‘task completion’ criteria.

Stage 2: Teach/observe RL3

Administration highlights:

- Students completed the segmented quizzes with fewer time complaints.

- Peer scorers achieved 85% inter-rater reliability post-training.

Teachers noted increased confidence in justifying scores, e.g., “I now reference specific rubric descriptors instead of general impressions.”

Stage 3: Interview pupils

RL3 feedback indicated:

- Preference for segmented quizzes: “It felt less overwhelming.”
- Requests for delayed grammar/vocabulary testing to prioritize fluency tasks.

Five students critiqued the listening tasks’ cultural relevance (e.g., audio scenarios assumed familiarity with Western contexts), prompting a discussion on localization.

Stage 4: Post-RL3 discussion and findings consolidation

The group identified and wrote down on LAL growth areas: rubric design, time management, and cultural sensitivity. As Mr Rebaz remarked, “I now see tests as dialogues, not monologues.”

The three cycles revealed progressive refinement of test-design practises, collaborative problem-solving, and heightened metacognitive awareness among participants. Teachers transitioned from viewing assessments as isolated products to iterative, learner-centred processes.

3.6.4.4. Post-application of the Questionnaire

After the completion of the lesson study plan cycles, the participants were requested to fill out 3 sections of the questionnaire administered during the pre-application phase.

3.6.4.5. Administration of the Post-Interview Questions

At the end of the study, of the researcher carried out post-interviews with the participants individually in a face-to-face manner. Upon finishing the interviews, the

researcher transcribed the interview questions verbatim, and the collected data became ready for qualitative data analysis.

3.7. Data Analysis

In this study, for quantitative analysis, only the frequency and percentages were calculated and reported. Furthermore, qualitative analysis was also carried out on the interview contents and reflective journal contents. Throughout the entirety of the qualitative data analysis process, the application of the six delineated stages of thematic analysis as proposed by Braun and Clarke (2006) was rigorously adhered to. These stages encompass the sequential progression of 1) achieving a comprehensive understanding of the data, 2) formulating initial codes, 3) identifying overarching themes, 4) reviewing and refining these identified themes, 5) clearly defining and labelling the themes, and ultimately 6) compiling the findings into a cohesive report. In order to ensure the dependability and robustness of the analytical procedures, an assistant was engaged to provide support in the execution of the thematic analysis methodology, as advocated by Hsieh and Shannon (2005). Initially, a meticulous process of multiple readings of both the interview transcripts and reflective journal entries was undertaken to foster a deep familiarity with the dataset. Concurrently, preliminary impressions were noted down and recorded for further reference. Subsequently, the data underwent a systematic process of coding, categorization, and condensation into meaningful units. The co-researcher independently conducted a parallel analysis of the data following a similar approach. A meticulous examination was conducted to identify and address any discrepancies in findings between the principal researcher and their assistant. To guarantee the accuracy of the analysis, an objective measure of inter-rater reliability was applied using Holsti's (1969) coefficient. The resulting values exceeded the benchmark of 0.7, signifying a high degree of consistency between the two coders. Moreover, to enhance the trustworthiness of the analytical results, we implemented a member checking procedure, as recommended by Nassaji (2020). This involved engaging in discussions with three participants to validate and confirm that our interpretations and conclusions accurately reflected their experiences and perspectives.

3.8. Chapter Summary

This chapter described the methodology of the study. It justified the use of a qualitative case study approach as the most appropriate strategy for investigating the development of Iraqi test writers' LAL through LSM. The chapter presented the research setting, in terms both context – the language centre of Salahaddin University – and participants, with the latter being individuals with MA and PhD holders who designed and conducted language tests. Multiple sources of evidence were documents, observation, semi-structured interviews and personal writings. The procedures for implementing the LSM cycles—planning, observation, and reflection—were explained in detail, along with the process of classroom trials of test tasks. The methods of data analysis were described including thematic analysis of qualitative data and visualisations such as networks and density maps. Ethical issues, validity checks, and limitations of the approach were also outlined. This chapter laid the foundations by developing an explicit & systematic methodology, which was used to inform the results that are presented in the following chapter.

CHAPTER IV

FINDINGS

4.1. Introduction

In this chapter, the results of both quantitative and qualitative data analyses to address the research questions are presented. In so doing, initially in each section, one of the research questions is stated and the pertinent data analysis performed to respond to that research question is elaborated followed by the presentation of the results.

4.2. Addressing the First Research Question

The first research question of the current study was:

RQ1: What are the academic and professional needs of teachers who teach at the higher education level at language centres as test writers in terms of language assessment literacy as they perceive?

Table 3

The academic and professional needs of teachers

Skills	Not Competent (1)	A Little Competent (2)	Somewhat Competent (3)	Competent (4)	Very Competent (5)	Total
Reading	0 (0%)	3 (30%)	4 (40%)	3 (30%)	0 (0%)	10 (100%)
Listening	1 (10%)	4 (40%)	3 (30%)	2 (20%)	0 (0%)	10 (100%)
Writing	0 (0%)	0 (0%)	2 (20%)	3 (30%)	5 (50%)	10 (100%)
Speaking	0 (0%)	0 (0%)	3 (30%)	2 (20%)	5 (50%)	10 (100%)

As illustrated in Table 3, 50% of the participants regarded themselves to be very competent in assessing writing and 50% also in assessing speaking with none of them being very competent in assessing reading or listening. And yet another requirement for communication testing literacy can be deduced. The extent of assessment literacy for productive skills is sufficient among the subjects to assess speaking and writing competences, when it comes to the respective skills of how to or not to design, conduct, and rate speaking and writing performance.

By contrast, it is interesting to note that none ~~one~~ of the participants considered themselves to be particularly good at ~~the~~ receptive skills (i.e. reading and listening). This striking gap in perceived skills could indicate an under emphasis on these skills in their routine teaching or perhaps limited access to specific training on the evaluation of receptive language skills. This pattern emphasizes an area of focus for professional development.

In addition, 30% of the participants perceived themselves to be little competent for the assessment of the reading, while the percentage has been higher, 40%, for the listening. This would indicate a considerable degree of insecurity in building or judging reading and listening tests. It is also interesting to note that 10% of the respondents affiliate themselves with the lowest response category, not at all competent, in assessing listening skills, highlighting listening as the most problematic aspect of assessment literacy.

Participants also indicated that they were not completely unskilled in some basic skills: 40% felt somewhat competent in reading, and 30% felt slightly competent in listening, suggesting that these participants may feel the need to develop their existing abilities further. This moderate confidence may indicate a general belief in pedagogical transfer stemming from their teaching background. However, they may lack expertise in applying assessment principles, such as developing valid test items for receptive skills or interpreting diagnostic data from comprehension tasks.

There was convergence between the two groups of teachers: both groups had the same percentage (30%) rating themselves as competent in reading, but a slightly lower percentage (20%) did so for listening. This confirms that constructing listening tests is seen as more difficult and perhaps less systematically approached by these teachers. This may be due to logistical constraints, such as limited questioning time and the lack of space for complex listening activities, which require additional materials. It may also stem from uncertainty about how to evaluate tasks that support the development of listening comprehension.

To sum up, the table clearly shows the highest perceived assessment competence in writing and speaking, with 50% or more of participants reporting that they felt very competent. This number increases further when those who felt competent or reasonably competent are included. In contrast, reading and listening received no ratings in the “very

competent” category, and many participants reported feeling only marginally competent in these areas.

This distribution suggests that teachers feel more confident evaluating students' productive skills than receptive ones. This tendency may be influenced by various interconnected factors, such as their teaching practices, preferred assessment activities, the greater variety of task formats available for productive skills, and their familiarity with producing or assessing spoken and written outputs. It is also possible that current institutional assessment patterns prioritise speaking and writing, thereby discouraging teachers from investing in the development of their knowledge and skills in assessing receptive abilities.

One thing worth mentioning is that even though some teachers had more experience or higher degrees than others, I chose not to compare them based on these differences. This was a conscious decision, in line with the focus of the study and based on my supervisor's guidance. What mattered more was looking at how teachers view their own assessment skills in real, everyday teaching contexts — not reducing them to academic titles or years of experience.

There may be an evident discrepancy in the perception of the receptive skills and teachers may need to appropriate external support. High-quality professional development experiences—for example, workshops, collaborative planning, or lesson study projects—can contribute to meeting this requirement. Through these practices, instructors will become more confident about the way they assess their students and will also help make the assessment process a fairer and more equitable testing that reflects students' L2 skills on use.

Table 4 shows the results for the training in English language testing and evaluation for the first case.

Table 4.

First Case's Training in English Language Testing and Evaluation

Concepts and Topics - None (No learning), Little (Very shallow), Sufficient/Intermediate (Enough learning to be used in practise), Extensive/Advanced (In-depth and comprehensive enough to work as a test and assessment expert).	Please indicate the extent of instruction or training you received at university in each of the following topics related to English language testing and evaluation.			
	None	Minimal	None	Extensive
1. Designing classroom-based assessments				❖
2. Developing various test types (e.g., diagnostic, placement, achievement, norm-referenced, integrative, direct/indirect, objective/subjective)			❖	
3. Providing test-based feedback to learners				❖
4. Applying peer and self-assessment techniques				❖
5. Implementing informal, ongoing, or non-traditional assessment approaches				❖
6. Utilising the European Language Portfolio or similar tools				❖
7. Ensuring the reliability of assessments				❖
8. Establishing test validity				❖
9. Assessing the four English language skills (listening, speaking, reading, writing)			❖	
10. Creating and applying scoring rubrics or criteria				❖
Total Choices	0	0	2	8
Total Score	0	0	6	32
Total Grand Score			38	

As shown in Table 4, the first case, who holds a PhD degree and has 15 years of teaching experience, has obtained a total grand score of 38 for training in English language testing and evaluation.

Table 5 shows the results for the training in English language testing and evaluation for the second case.

Table 5.

Second Case's Training in English Language Testing and Evaluation

Concepts and Topics - None (No learning), Little (Very shallow), Sufficient/Intermediate (Enough learning to be used in practise), Extensive/Advanced (In-depth and comprehensive enough to work as a test and assessment expert).	Please indicate the extent of instruction or training you received at university in each of the following topics related to English language testing and evaluation.			
	None	Minimal	Moderate	Extensive
1. Designing classroom-based assessments				❖
2. Developing various test types (e.g., diagnostic, placement, achievement, norm-referenced, integrative, direct/indirect, objective/subjective)			❖	
3. Providing test-based feedback to learners				❖
4. Applying peer and self-assessment techniques			❖	
5. Implementing informal, ongoing, or non-traditional assessment approaches				❖
6. Utilising the European Language Portfolio or similar tools				❖
7. Ensuring the reliability of assessments				❖
8. Establishing test validity				❖
9. Assessing the four English language skills (listening, speaking, reading, writing)			❖	
10. Creating and applying scoring rubrics or criteria				❖
Total Choices	0	0	3	7
Total Score	0	0	9	28
Total Grand Score			37	

As seen in Table 5, the second case, who holds a PhD degree and has 12 years of teaching experience, has obtained a total grand score of 37 for training in English language testing and evaluation. The one-point lower score of case #2 in comparison to the first case can possibly be rooted in the amount of teaching experience.

Table 6 displays the results for the training in English language testing and evaluation for the third case.

Table 6.

Third Case's Training in English Language Testing and Evaluation

Concepts and Topics - None (No learning), Little (Very shallow), Sufficient/Intermediate (Enough learning to be used in practise), Extensive/Advanced (In-depth and comprehensive enough to work as a test and assessment expert).	Please indicate the extent of instruction or training you received at university in each of the following topics related to English language testing and evaluation.			
	None	Minimal	Moderate	Extensive
1. Designing classroom-based assessments				❖
2. Developing various test types (e.g., diagnostic, placement, achievement, norm-referenced, integrative, direct/indirect, objective/subjective)			❖	
3. Providing test-based feedback to learners				❖
4. Applying peer and self-assessment techniques			❖	
5. Implementing informal, ongoing, or non-traditional assessment approaches				❖
6. Utilising the European Language Portfolio or similar tools				❖
7. Ensuring the reliability of assessments				❖
8. Establishing test validity			❖	
9. Assessing the four English language skills (listening, speaking, reading, writing)			❖	
10. Creating and applying scoring rubrics or criteria				❖
Total Choices	0	0	4	6
Total Score	0	0	12	24
Total Grand Score			36	

As revealed in Table 6, the third case, who holds a PhD degree and has 10 years of teaching experience, has obtained a total grand score of 36 for training in English language testing and evaluation. The lower score of case #3 in comparison to the first and second cases can possibly be attributed to teaching experience as an important factor influencing such results.

Table 7 depicts the results for the training in English language testing and evaluation for the fourth case.

Table 7.

Fourth Case's Training in English Language Testing and Evaluation

Concepts and Topics - None (No learning), Little (Very shallow), Sufficient/Intermediate (Enough learning to be used in practise), Extensive/Advanced (In-depth and comprehensive enough to work as a test and assessment expert).	Please indicate the extent of instruction or training you received at university in each of the following topics related to English language testing and evaluation.			
	None	Minimal	Moderate	Extensive
1. Designing classroom-based assessments				❖
2. Developing various test types (e.g., diagnostic, placement, achievement, norm-referenced, integrative, direct/indirect, objective/subjective)			❖	
3. Providing test-based feedback to learners			❖	
4. Applying peer and self-assessment techniques			❖	
5. Implementing informal, ongoing, or non-traditional assessment approaches			❖	
6. Utilising the European Language Portfolio or similar tools				❖
7. Ensuring the reliability of assessments				❖
8. Establishing test validity				❖
9. Assessing the four English language skills (listening, speaking, reading, writing)			❖	
10. Creating and applying scoring rubrics or criteria				❖
Total Choices	0	0	5	5
Total Score	0	0	15	20
Total Grand Score			35	

As depicted in Table 7, the fourth case, who holds a PhD degree and has 9 years of teaching experience, has obtained a total grand score of 35 for training in English language testing and evaluation, this lower score in comparison with the previous cases who are more experienced is further evidence for the influence of teaching experience as an important factor in relation to training in English language testing and evaluation.

Table 8 illustrates the results for the training in English language testing and evaluation for the fifth case.

Table 8.

Fifth Case's Training in English Language Testing and Evaluation

Concepts and Topics - None (No learning), Little (Very shallow), Sufficient/Intermediate (Enough learning to be used in practise), Extensive/Advanced (In-depth and comprehensive enough to work as a test and assessment expert).	Please indicate the extent of instruction or training you received at university in each of the following topics related to English language testing and evaluation.			
	None	Minimal	Moderate	Extensive
1. Designing classroom-based assessments				❖
2. Developing various test types (e.g., diagnostic, placement, achievement, norm-referenced, integrative, direct/indirect, objective/subjective)			❖	
3. Providing test-based feedback to learners			❖	
4. Applying peer and self-assessment techniques			❖	
5. Implementing informal, ongoing, or non-traditional assessment approaches			❖	
6. Utilising the European Language Portfolio or similar tools			❖	
7. Ensuring the reliability of assessments			❖	
8. Establishing test validity				❖
9. Assessing the four English language skills (listening, speaking, reading, writing)			❖	
10. Creating and applying scoring rubrics or criteria				❖
Total Choices	0	0	7	3
Total Score	0	0	21	12
Total Grand Score			33	

As illustrated in Table 8, the fifth case, who holds a PhD degree and has 8 years of teaching experience, has obtained a total grand score of 33 for training in English language testing and evaluation, which is further evidence for the influence of teaching experience as an important factor regarding training in English language testing and evaluation.

Table 9 demonstrates the results for the training in English language testing and evaluation for the sixth case.

Table 9.

Sixth Case's Training in English Language Testing and Evaluation

Concepts and Topics - None (No learning), Little (Very shallow), Sufficient/Intermediate (Enough learning to be used in practise), Extensive/Advanced (In-depth and comprehensive enough to work as a test and assessment expert).	Please indicate the extent of instruction or training you received at university in each of the following topics related to English language testing and evaluation.			
	None	Minimal	Moderate	Extensive
1. Designing classroom-based assessments			❖	
2. Developing various test types (e.g., diagnostic, placement, achievement, norm-referenced, integrative, direct/indirect, objective/subjective)		❖		
3. Providing test-based feedback to learners			❖	
4. Applying peer and self-assessment techniques			❖	
5. Implementing informal, ongoing, or non-traditional assessment approaches			❖	
6. Utilising the European Language Portfolio or similar tools			❖	
7. Ensuring the reliability of assessments			❖	
8. Establishing test validity		❖		
9. Assessing the four English language skills (listening, speaking, reading, writing)			❖	
10. Creating and applying scoring rubrics or criteria			❖	
Total Choices	0	2	8	0
Total Score	0	4	24	0
Total Grand Score			28	

As shown in Table 9, the sixth case, which involves a participant with an MA degree and 7 years of teaching experience, received a total score of 28 in training related to English language testing and evaluation. Notably, there is a marked decline of 5 points from case #5 to case #6. Given that there is only a one-year difference in teaching experience between the two cases, this significant drop can primarily be attributed to the difference in educational qualifications, although teaching experience may also play a role in this context.

Table 10 portrays the results for the training in English language testing and evaluation for the seventh case.

Table 10.

Seventh Case's Training in English Language Testing and Evaluation

Concepts and Topics - None (No learning), Little (Very shallow), Sufficient/Intermediate (Enough learning to be used in practise), Extensive/Advanced (In-depth and comprehensive enough to work as a test and assessment expert).	Please indicate the extent of instruction or training you received at university in each of the following topics related to English language testing and evaluation.			
	None	Minimal	Moderate	Extensive
1. Designing classroom-based assessments		❖		
2. Developing various test types (e.g., diagnostic, placement, achievement, norm-referenced, integrative, direct/indirect, objective/subjective)		❖		
3. Providing test-based feedback to learners		❖		
4. Applying peer and self-assessment techniques			❖	
5. Implementing informal, ongoing, or non-traditional assessment approaches			❖	
6. Utilising the European Language Portfolio or similar tools			❖	
7. Ensuring the reliability of assessments			❖	
8. Establishing test validity		❖		
9. Assessing the four English language skills (listening, speaking, reading, writing)		❖		
10. Creating and applying scoring rubrics or criteria			❖	
Total Choices	0	5	5	0
Total Score	0	10	15	0
Total Grand Score			25	

In Table 10, the seventh case—a participant with an MA degree and seven years of teaching experience—has achieved a total grand score of 25 in training related to English language testing and evaluation. Notably, this significant decline is evident when comparing this case to the last PhD holder, which reinforces the critical role that educational degrees play in training for English language testing and evaluation.

Table 11 presents the results for the training in English language testing and evaluation for the eighth case.

Table 11.

Eighth Case's Training in English Language Testing and Evaluation

Concepts and Topics - None (No learning), Little (Very shallow), Sufficient/Intermediate (Enough learning to be used in practise), Extensive/Advanced (In-depth and comprehensive enough to work as a test and assessment expert).	Please indicate the extent of instruction or training you received at university in each of the following topics related to English language testing and evaluation.			
	None	Minimal	Moderate	Extensive
1. Designing classroom-based assessments				❖
2. Developing various test types (e.g., diagnostic, placement, achievement, norm-referenced, integrative, direct/indirect, objective/subjective)		❖		
3. Providing test-based feedback to learners				❖
4. Applying peer and self-assessment techniques		❖		
5. Implementing informal, ongoing, or non-traditional assessment approaches			❖	
6. Utilising the European Language Portfolio or similar tools			❖	
7. Ensuring the reliability of assessments			❖	
8. Establishing test validity		❖		
9. Assessing the four English language skills (listening, speaking, reading, writing)		❖		
10. Creating and applying scoring rubrics or criteria			❖	
Total Choices	0	4	4	2
Total Score	0	8	12	8
Total Grand Score			28	

As Table 11 displays, the eighth case, who holds an MA degree and has 4 years of teaching experience, has obtained a total grand score of 28 for training in English language testing and evaluation. As it can be seen, the sudden decline is still persistent in comparison with the PhD cases.

Table 12 illustrates the results for the training in English language testing and evaluation for the ninth case.

Table 12.

Ninth Case's Training in English Language Testing and Evaluation

Concepts and Topics - None (No learning), Little (Very shallow), Sufficient/Intermediate (Enough learning to be used in practise), Extensive/Advanced (In-depth and comprehensive enough to work as a test and assessment expert).	Please indicate the extent of instruction or training you received at university in each of the following topics related to English language testing and evaluation.			
	None	Minimal	Moderate	Extensive
1. Designing classroom-based assessments		❖		
2. Developing various test types (e.g., diagnostic, placement, achievement, norm-referenced, integrative, direct/indirect, objective/subjective)		❖		
3. Providing test-based feedback to learners				❖
4. Applying peer and self-assessment techniques			❖	
5. Implementing informal, ongoing, or non-traditional assessment approaches			❖	
6. Utilising the European Language Portfolio or similar tools	❖			
7. Ensuring the reliability of assessments		❖		
8. Establishing test validity		❖		
9. Assessing the four English language skills (listening, speaking, reading, writing)		❖		
10. Creating and applying scoring rubrics or criteria		❖		
Total Choices	1	6	2	1
Total Score	1	12	8	4
Total Grand Score			25	

As Table 12 shows, the ninth case (MA holder with three years of teaching experience) had little training exposure to the area of language test and evaluation. The participant reported some instruction in those six areas (low instructional frequency), specifically: designing classroom assessments, developing types of tests, ensuring reliability, establishing validity, language skills, and rubrics. Except peer/self-assessment

and informal/non-traditional assessment, which were moderate, examples of extensive training were found for test-based feedback. It should be noted that teachers did not receive guidance about the use of the European Language Portfolio. The total grand score of 25 suggests that the participant's preparation in assessment is uneven and generally limited, with stronger grounding in feedback provision but weaker foundations in technical aspects such as reliability and validity.

Table 13 depicts the results for the training in English language testing and evaluation for the tenth case.

Table 13.

Tenth Case's Training in English Language Testing and Evaluation

Concepts and Topics - None (No learning), Little (Very shallow), Sufficient/Intermediate (Enough learning to be used in practise), Extensive/Advanced (In-depth and comprehensive enough to work as a test and assessment expert).	Please indicate the extent of instruction or training you received at university in each of the following topics related to English language testing and evaluation.			
	None	Minimal	Moderate	Extensive
1. Designing classroom-based assessments			❖	
2. Developing various test types (e.g., diagnostic, placement, achievement, norm-referenced, integrative, direct/indirect, objective/subjective)	❖			
3. Providing test-based feedback to learners				❖
4. Applying peer and self-assessment techniques			❖	
5. Implementing informal, ongoing, or non-traditional assessment approaches				❖
6. Utilising the European Language Portfolio or similar tools		❖		
7. Ensuring the reliability of assessments		❖		
8. Establishing test validity	❖			
9. Assessing the four English language skills (listening, speaking, reading, writing)		❖		
10. Creating and applying scoring rubrics or criteria	❖			
Total Choices	3	3	2	2
Total Score	3	6	6	8
Total Grand Score			23	

As Table 13 indicates, the ninth case, who holds an MA degree and has one year of teaching experience, has obtained a total grand score of 23 for training in English language testing and evaluation. Similar to the previous MA cases, this case has also gained a lower score in comparison with all the PhD-holding cases.

Table 14 provides the summary of the results for the 10 cases in terms of training in English language testing and evaluation.

Table 14.

Summary of the Ten Cases for the Training in English Language Testing and Evaluation

Concepts and Topics - None (No learning), Little (Very shallow), Sufficient/Intermediate (Enough learning to be used in practise), Extensive/Advanced (In-depth and comprehensive enough to work as a test and assessment expert).	Please indicate the extent of instruction or training you received at university in each of the following topics related to English language testing and evaluation.			
	None	Minimal	Moderate	Extensive
1. Designing classroom-based assessments		2	2	6
2. Developing various test types (e.g., diagnostic, placement, achievement, norm-referenced, integrative, direct/indirect, objective/subjective)	1	4	5	
3. Providing test-based feedback to learners		1	3	6
4. Applying peer and self-assessment techniques		1	8	1
5. Implementing informal, ongoing, or non-traditional assessment approaches			6	4
6. Utilising the European Language Portfolio or similar tools	1	1	4	4
7. Ensuring the reliability of assessments		2	4	4
8. Establishing test validity	1	4	1	4
9. Assessing the four English language skills (listening, speaking, reading, writing)		4	6	
10. Creating and applying scoring rubrics or criteria	1	1	3	5
Total Choices	4	20	42	34
Total Score	4	40	126	136
Total Grand Score			306	

As depicted in Table 14, the total grand score of all the cases is 306. Since the maximum obtainable score for each case is 40 (10 items * 4), and thus the maximum obtainable score for all the cases is 400, the score of 306 out of 400 is fairly higher than the average of 200 (400/2). Thus, the overall training in English language testing and evaluation for all the cases can be considered to be at a satisfactory level.

Table 15 summarizes the results for the five PhD holders in terms of training in English language testing and evaluation.

Table 15.

Summary of the Five PhD Holders in Terms of Training in English Language Testing and Evaluation

Concepts and Topics - None (No learning), Little (Very shallow), Sufficient/Intermediate (Enough learning to be used in practise), Extensive/Advanced (In-depth and comprehensive enough to work as a test and assessment expert).	Please indicate the extent of instruction or training you received at university in each of the following topics related to English language testing and evaluation.			
	None	Minimal	Moderate	Extensive
1. Designing classroom-based assessments				5
2. Developing various test types (e.g., diagnostic, placement, achievement, norm-referenced, integrative, direct/indirect, objective/subjective)			5	
3. Providing test-based feedback to learners			2	3
4. Applying peer and self-assessment techniques			4	1
5. Implementing informal, ongoing, or non-traditional assessment approaches			2	3
6. Utilising the European Language Portfolio or similar tools			1	4
7. Ensuring the reliability of assessments			1	4
8. Establishing test validity			1	4
9. Assessing the four English language skills (listening, speaking, reading, writing)			5	
10. Creating and applying scoring rubrics or criteria				5
Total Choices			21	29
Total Score			63	116
Total Grand Score			179	

As shown in Table 15, the total grand score of all the PhD cases is 179. Since the maximum obtainable score for each case is 40 (10 items * 4), and thus the maximum obtainable score for all the PhD cases is 200, hence the score of 179 out of 200 is quite higher than the average of 100 (200/2). Thus, the overall training in English language testing and evaluation for the PhD cases can be considered to be at a highly satisfactory level.

Table 16 summarizes the results for the five MA holders in terms of training in English language testing and evaluation.

Table 16.

Summary of the Five MA Holders in Terms of Training in English Language Testing and Evaluation

Concepts and Topics - None (No learning), Little (Very shallow), Sufficient/Intermediate (Enough learning to be used in practise), Extensive/Advanced (In-depth and comprehensive enough to work as a test and assessment expert).	Please indicate the extent of instruction or training you received at university in each of the following topics related to English language testing and evaluation.			
	None	Minimal	Moderate	Extensive
1. Designing classroom-based assessments		2	2	1
2. Developing various test types (e.g., diagnostic, placement, achievement, norm-referenced, integrative, direct/indirect, objective/subjective)	1	4		
3. Providing test-based feedback to learners		1	1	3
4. Applying peer and self-assessment techniques		1	4	
5. Implementing informal, ongoing, or non-traditional assessment approaches			4	1
6. Utilising the European Language Portfolio or similar tools	1	1	3	
7. Ensuring the reliability of assessments		2	3	
8. Establishing test validity	1	4		
9. Assessing the four English language skills (listening, speaking, reading, writing)		4	1	
10. Creating and applying scoring rubrics or criteria	1	1	3	
Total Choices	4	20	21	5
Total Score	4	40	63	20
Total Grand Score			127	

As presented in Table 16, the total grand score of all the MA cases is 127. Since the maximum obtainable score for each case is 40 (10 items * 4), and thus the maximum obtainable score for all the MA cases is 200, hence the score of 127 out of 200 is quite close to the average of 100 (200/2). Thus, the overall training in English language testing and evaluation for the MA cases can be considered to be roughly satisfactory.

As comparison of the obtained scores for the PhD and MA cases presented in the preceding two tables (Tables 15 and 16), shows that the MA holders have gained considerably lower than their PhD counterparts, which is an indication that MA holders are in need of more training in English language testing and evaluation.

Table 17 provides the summary of the results for the 10 cases in terms of the academic and professional needs of teachers as test writers in terms of language assessment literacy.

Table 17.

Summary of the Ten Cases for the Academic and Professional Needs of Teachers as Test Writers in Terms of Language Assessment Literacy

Topics	Not at all essential	Little essential	Somewhat essential	Essential	Very essential	Total
1. Evolution and background of language assessment	7	2	1			14
2. Selection and determination of test content	3	5	2			19
3. Appropriate techniques for evaluating language proficiency	2	2	4	2		26
4. Designing and constructing language evaluations	4	1	4	1		22
5. Formulating detailed test guidelines		3	6	1		28
6. Evaluating test efficiency through item performance analysis	5		3	2		22
7. Categories and functions of test elements and formats			3	6	1	38
8. Organizing and executing language evaluations	6	4				14

9.	Analysing examinee results and achievement levels	2	3		4	1	29
10.	Evaluating comprehension abilities in reading and listening				3	7	47
11.	Measuring expressive abilities in speaking and writing	3	5	2			19
12.	Testing combined language skills in integrated tasks		1	2	4	3	39
13.	Assessing specific language aspects: grammar, vocabulary, and pronunciation	2	4	4			22
14.	Evaluating learners with diverse educational needs	1	3	6			25
15.	Using portfolio-based methods for student assessment	2	2	1	1	4	33
16.	Implementing learner-driven self-evaluations	3	1	1	3	2	30
17.	Facilitating peer-led evaluation processes	2	1	2	2	3	33
18.	Statistical methods applied in test analysis	4	1		2	3	29
19.	Foundational concepts in language evaluation	1	3	2	4		29
20.	Applying detailed scoring rubrics for speaking and writing	3	2		4	1	28
21.	Using overall impression scales for productive skill rating	2	8				18
22.	Supporting students in preparation for formal exams (e.g., FCE)	4	5	1			17
23.	Moral and professional responsibilities in test design	4	2	4			20
24.	Communicating student outcomes and giving feedback	4	4	2			18
25.	Familiarity with CEFR proficiency descriptors	3	4	3			20
26.	Using assessment outcomes to enhance teaching strategies	4	1	3	2		23
Grand Total		71	134	168	164	125	662

As indicated in the above table, the item stipulating Assess receptive skills (reading and listening) has obtained the highest score (47), followed by Assess integrated skills (39) and Types and purposes of items, tasks, and tests (38). Due to the high scores for these items, the primary academic and professional needs of teachers as test writers in terms of language assessment literacy are those related to assessing receptive skills, integrative testing, and test, task, and item types. Moreover, assess students through portfolios (33), Assess students through peer assessment (33), and Assess students through self-assessment rate have also been stated as the needs by the respondents.

Table 18 provides the summary of the results for the five PhD holders in terms of the academic and professional needs of teachers as test writers in terms of language assessment literacy.

Table 18.

Summary of the Five PhD Holders for the Academic and Professional Needs of Teachers as Test Writers in Terms of Language Assessment Literacy

Topics	Not at all essential	Little essential	Somewhat essential	Essential	Very essential	Total
1. Evolution and background of language assessment	5					5
2. Selection and determination of test content	3	2				7
3. Appropriate techniques for evaluating language proficiency	2	2	1			9
4. Designing and constructing language evaluations	4	1				6
5. Formulating detailed test guidelines		3	2			12
6. Evaluating test efficiency through item performance analysis	5					5
7. Categories and functions of test elements and formats			3	2		17
8. Organizing and executing language evaluations	4	1				6

9. Analysing examinee results and achievement levels				4	1	21
10. Evaluating comprehension abilities in reading and listening				2	3	23
11. Measuring expressive abilities in speaking and writing	3	2				7
12. Testing combined language skills in integrated tasks		1	2	2		16
13. Assessing specific language aspects: grammar, vocabulary, and pronunciation	2	2	1			9
14. Evaluating learners with diverse educational needs	1	1	3			12
15. Using portfolio-based methods for student assessment	2	2	1			9
16. Implementing learner-driven self-evaluations	3	1	1			8
17. Facilitating peer-led evaluation processes	2	1	2			10
18. Statistical methods applied in test analysis	4	1				6
19. Foundational concepts in language evaluation	1	3	1			10
20. Applying detailed scoring rubrics for speaking and writing	3	2				7
21. Using overall impression scales for productive skill rating	1	4				9
22. Supporting students in preparation for formal exams (e.g., FCE)	4	1				6
23. Moral and professional responsibilities in test design	4	1				6
24. Communicating student outcomes and giving feedback	4	1				6
25. Familiarity with CEFR proficiency descriptors	3	2				7
26. Using assessment outcomes to enhance teaching strategies	4	1				6
Grand Total	64	70	51	40	20	245

Based on the information in the above table, the PhD holding cases have obtained an overall score of 245 in relation to their academic and professional needs as test writers in terms of language assessment literacy. Their primary assessment literacy needs in a descending order are Assess receptive skills (reading and listening) (23), Interpreting test-takers' scores and performances (21), Types and purposes of items, tasks and tests (17), and assess integrated skills (16).

Table 19 provides the summary of the results for the five MA holders in terms of the academic and professional needs of teachers as test writers in terms of language assessment literacy.

Table 19.

Summary of the Five MA Holders for the Academic and Professional Needs of Teachers as Test Writers in Terms of Language Assessment Literacy

Topics	Not at all essential	Little essential	Somewhat essential	Essential	Very essential	Total
1. Evolution and background of language assessment	2	2	1			9
2. Selection and determination of test content		3	2			12
3. Appropriate techniques for evaluating language proficiency			3	2		17
4. Designing and constructing language evaluations			4	1		16
5. Formulating detailed test guidelines			4	1		16
6. Evaluating test efficiency through item performance analysis			3	2		17
7. Categories and functions of test elements and formats				4	1	21
8. Organizing and executing language evaluations	2	3				8
9. Analysing examinee results and achievement levels	2	3				8

10. Evaluating comprehension abilities in reading and listening				1	4	24
11. Measuring expressive abilities in speaking and writing		3	2			12
12. Testing combined language skills in integrated tasks				2	3	23
13. Assessing specific language aspects: grammar, vocabulary, and pronunciation		2	3			13
14. Evaluating learners with diverse educational needs		2	3			13
15. Using portfolio-based methods for student assessment				1	4	24
16. Implementing learner-driven self-evaluations				3	2	22
17. Facilitating peer-led evaluation processes				2	3	23
18. Statistical methods applied in test analysis				2	3	23
19. Foundational concepts in language evaluation			1	4		19
20. Applying detailed scoring rubrics for speaking and writing				4	1	21
21. Using overall impression scales for productive skill rating	1	4				9
22. Supporting students in preparation for formal exams (e.g., FCE)		4	1			11
23. Moral and professional responsibilities in test design		1	4			14
24. Communicating student outcomes and giving feedback		3	2			12
25. Familiarity with CEFR proficiency descriptors		2	3			13
26. Using assessment outcomes to enhance teaching strategies			3	2		17
Grand Total	7	64	117	124	105	417

As the information in the above table shows, the MA holding cases have gained an overall score of 417 in relation to their academic and professional needs as test writers

in terms of language assessment literacy. Their primary assessment literacy needs in a descending order are Assess receptive skills (reading and listening) (24), Assess students through portfolios (24), Assess integrated skills (23), Assess students through peer assessment (23), and Language tests and statistical approaches (23), followed by Using analytic scales for productive skills (21), and Types and purposes of items, tasks and tests (21). Moreover, the MA cases have obtained a substantially higher score (417) as compared with their PhD counterparts (245) in terms of their academic and professional needs as test writers regarding language assessment literacy.

4.3 Addressing the Second Research Question

The second research question of the current study was:

***RQ2:** What is the level of language assessment knowledge of the teachers who teach at higher education level at Language Centres?*

To address this research question, Language Assessment Knowledge Scale scores were computed. The scale consisted of four sections for the listening, reading, speaking, and writing assessment. Table 20 shows the results for the ten cases in terms of listening assessment knowledge.

Table 20.

Results for the Ten Cases in Terms of Listening Assessment Knowledge

Assessing Listening Items	True	False	Do not Know
1. Using reading passages as listening materials can present challenges.	3	2	5
2. Authentic listening materials are essential for evaluating learners' real-world comprehension skills	7	1	2
3. Note-taking can be applied to any type of listening text.	3	1	6
4. Listening tests should avoid inference questions that rely on test-taker intelligence.	9	1	0
5. Errors in grammar or spelling may impact scores in listening evaluations.	1	8	1
6. Listening cloze tasks promote targeted or selective listening.	6	3	1
7. Tasks involving phonemic contrasts, such as minimal pairs, are considered integrative assessments.	6	2	2
8. Evaluating note-taking responses is generally straightforward.	3	6	1
9. Discrete-point listening tests typically assess surface-level or literal understanding.	7	2	1
10. Selective listening requires learners to identify specific pieces of information.	8	0	2
Total Correct Score	58		
Total "Do not Know" Responses	18		

As indicated in the above table, the cases all together have obtained an overall correct score of 58 out of a maximum of 100 ($10 \times 10 = 100$). The score of 58 out of the maximum of 100 stands quite close to the mean score of 50 ($100/2 = 50$). The obtained score is an indication that the ten cases have scored quite low in terms of listening assessment knowledge. Moreover, they have responded with "do not know" 18 times. Table 21 shows the results for the PhD cases in terms of listening assessment knowledge.

Table 21.

Results for the PhD Cases in Terms of Listening Assessment Knowledge

Assessing Listening Items	True	False	Do not Know
1. Using reading passages as listening materials can present challenges.	2	1	2
2. Authentic listening materials are essential for evaluating learners' real-world comprehension skills	4	0	1
3. Note-taking can be applied to any type of listening text.	2	0	3
4. Listening tests should avoid inference questions that rely on test-taker intelligence.	5	0	0
5. Errors in grammar or spelling may impact scores in listening evaluations.	0	5	0
6. Listening cloze tasks promote targeted or selective listening.	4	1	0
7. Tasks involving phonemic contrasts, such as minimal pairs, are considered integrative assessments.	4	0	1
8. Evaluating note-taking responses is generally straightforward.	1	4	0
9. Discrete-point listening tests typically assess surface-level or literal understanding.	4	1	0
10. Selective listening requires learners to identify specific pieces of information.	5	0	0
Total Correct Score	34		
Total "Do not Know" Responses	4		

As seen in the above table, the PhD cases all together have obtained an overall correct score of 34 out of a maximum of 50 ($10 \times 5 = 50$). The score of 34 out of the maximum of 50 stands fairly above the mean score of 25 ($50/2 = 25$). The obtained score of 34 is an indication that the PhD cases have scored fairly satisfactory in terms of listening assessment knowledge. Moreover, they have responded with "do not know" 4 times.

Table 22 displays the results for the MA cases in terms of listening assessment knowledge.

Table 22.

Results for the MA Cases in Terms of Listening Assessment Knowledge

Assessing Listening Items	True	False	Do not Know
1. Using reading passages as listening materials can present challenges.	1	1	3
2. Authentic listening materials are essential for evaluating learners' real-world comprehension skills	3	1	1
3. Note-taking can be applied to any type of listening text.	1	1	3
4. Listening tests should avoid inference questions that rely on test-taker intelligence.	4	1	0
5. Errors in grammar or spelling may impact scores in listening evaluations.	1	3	1
6. Listening cloze tasks promote targeted or selective listening.	2	2	1
7. Tasks involving phonemic contrasts, such as minimal pairs, are considered integrative assessments.	2	2	1
8. Evaluating note-taking responses is generally straightforward.	2	2	1
9. Discrete-point listening tests typically assess surface-level or literal understanding.	3	1	1
10. Selective listening requires learners to identify specific pieces of information.	3	0	2
Total Correct Score	24		
Total "Do not Know" Responses	14		

Based on the information in the above table, the MA cases all together have gained an overall correct score of 24 out of a maximum of 50 ($10 \times 5 = 50$). The score of 24 out of the maximum of 50 stands fairly lower the mean score of 25 ($50/2 = 25$). The obtained score of 24 is an indication that the MA cases have scored fairly unsatisfactory in terms of listening assessment knowledge. Furthermore, they have responded with "do not know" 14 times.

Table 23 shows the results for the ten cases in terms of reading assessment knowledge.

Table 23.

Results for the Ten Cases in Terms of Reading Assessment Knowledge

Items for Assessing Reading	True	False	Do not know
1. Summarizing a reading text is a valid method for evaluating reading skills.	7	2	1
2. The top-down model emphasizes the reader's overall understanding of a text during evaluation.	6	2	2
3. Cloze procedures are suitable for identifying a reader's grasp of a text's central message.	3	7	0
4. Using a well-known passage in a reading test is not necessarily problematic.	1	5	4
5. A single reading passage can adequately serve in a comprehension test.	3	7	0
6. Distractors that are grammatically incorrect in multiple-choice reading items may reduce question quality.	8	2	0
7. Spelling errors are considered in scoring reading assessments.	0	6	4
8. Evaluating the complexity of vocabulary plays a key role in assessing reading abilities.	8	2	0
9. Incorporating "not mentioned" options in true/false questions offers distinct benefits.	4	3	3
10. The reliability of a reading text increases with the number of items.	4	5	1
Total Correct Score	64		
Total "Do not know" Responses	15		

As indicated in the above table, the cases all together have obtained an overall correct score of 64 out of a maximum of 100 ($10 \times 10 = 100$). The score of 64 out of the maximum of 100 stands quite above the mean score of 50 ($100/2 = 50$). The obtained score of 64 out of 100 is an indication that the ten cases have scored fairly high in terms of reading assessment knowledge. Moreover, they have responded with "do not know" 15 times.

Table 24 shows the results for the PhD cases in terms of reading assessment knowledge.

Table 24.

Results for the PhD Cases in Terms of Reading Assessment Knowledge

Items for Assessing Reading	True	False	Do not know
1. Summarizing a reading text is a valid method for evaluating reading skills.	4	1	0
2. The top-down model emphasizes the reader's overall understanding of a text during evaluation.	4	0	1
3. Cloze procedures are suitable for identifying a reader's grasp of a text's central message.	1	4	0
4. Using a well-known passage in a reading test is not necessarily problematic.	0	3	2
5. A single reading passage can adequately serve in a comprehension test.	0	5	0
6. Distractors that are grammatically incorrect in multiple-choice reading items may reduce question quality.	5	0	0
7. Spelling errors are considered in scoring reading assessments.	0	4	1
8. Evaluating the complexity of vocabulary plays a key role in assessing reading abilities.	5	0	0
9. Incorporating "not mentioned" options in true/false questions offers distinct benefits.	2	2	1
10. The reliability of a reading text increases with the number of items.	3	2	
Total Correct Score	39		
Total "Do not know" Responses	5		

As seen in the above table, the PhD cases all together have gained an overall correct score of 39 out of a maximum of 50 ($10 \times 5 = 50$). The score of 39 out of the maximum of 50 stands quite above the mean score of 25 ($50/2 = 25$). The obtained score of 39 is an indication that the PhD cases have scored quite satisfactory in terms of reading assessment knowledge. Moreover, they have responded with "do not know" 5 times.

Table 25 displays the results for the MA cases in terms of reading assessment knowledge.

Table 25.

Results for the MA Cases in Terms of Reading Assessment Knowledge

Items for Assessing Reading	True	False	Do not know
1. Summarizing a reading text is a valid method for evaluating reading skills.	3	1	1
2. The top-down model emphasizes the reader's overall understanding of a text during evaluation.	2	2	1
3. Cloze procedures are suitable for identifying a reader's grasp of a text's central message.	2	3	0
4. Using a well-known passage in a reading test is not necessarily problematic.	1	2	2
5. A single reading passage can adequately serve in a comprehension test.	1	4	0
6. Distractors that are grammatically incorrect in multiple-choice reading items may reduce question quality.	3	2	0
7. Spelling errors are considered in scoring reading assessments.	0	2	3
8. Evaluating the complexity of vocabulary plays a key role in assessing reading abilities.	3	2	0
9. Incorporating "not mentioned" options in true/false questions offers distinct benefits.	2	1	2
10. The reliability of a reading text increases with the number of items.	1	3	1
Total Correct Score	25		
Total "Do not know" Responses	10		

Based on the information in the above table, the MA cases all together have gained an overall correct score of 25 out of a maximum of 50 ($10 \times 5 = 50$). The score of 25 out of the maximum of 50 equals the mean score of 25 ($50/2 = 25$). The obtained score of 25 is an indication that the MA cases have scored unsatisfactory in terms of reading assessment knowledge. Furthermore, they have responded with "do not know" 10 times.

Table 26 exhibits the results for the ten cases in terms of speaking assessment knowledge.

Table 26.

Results for the Ten Cases in Terms of Speaking Assessment Knowledge

Items for Assessing Speaking	True	False	Do not Know
1. Role plays are employed when the emphasis is on evaluating discourse skills.	7	3	0
2. Assessing speaking skills may require more than one task.	6	2	2
3. Interlocutors showing interest through signals can be problematic.	6	2	2
4. Ending a task when a learner can't reach the criterion level is appropriate.	2	6	2
5. Simultaneously using holistic and analytic scales is problematic.	8	0	2
6. Oral reading is an appropriate method for evaluating speaking proficiency.	3	7	0
7. Checklists can be used to assess in-class oral presentations.	8	1	1
8. Interactive tasks with more than two learners can pose difficulties.	7	2	1
9. Assessing speaking skills can involve learners participating in discussions.	7	2	1
10. Speaking exams assess both production and comprehension together.	6	1	3
Total Correct Scores	68		
Total "Do not know" Responses	15		

As indicated in the above table, the cases all together have obtained an overall correct score of 68 out of a maximum of 100 for speaking assessment knowledge ($10 \times 10 = 100$). The score of 68 out of the maximum of 100 stands fairly above the mean score of 50 ($100/2 = 50$). The obtained score is an indication that the ten cases have scored quite high in terms of speaking assessment knowledge. Moreover, they have responded with "do not know" 15 times.

Table 27 shows the results for the PhD cases in terms of speaking assessment knowledge.

Table 27.

Results for the PhD Cases in Terms of Speaking Assessment Knowledge

Items for Assessing Speaking	True	False	Do not Know
1. Role plays are employed when the emphasis is on evaluating discourse skills.	4	1	0
2. Assessing speaking skills may require more than one task.	4	0	1
3. Interlocutors showing interest through signals can be problematic.	3	1	1
4. Ending a task when a learner can't reach the criterion level is appropriate.	1	4	0
5. Simultaneously using holistic and analytic scales is problematic.	5	0	0
6. Oral reading is an appropriate method for evaluating speaking proficiency.	1	4	0
7. Checklists can be used to assess in-class oral presentations.	5	0	0
8. Interactive tasks with more than two learners can pose difficulties.	4	1	0
9. Assessing speaking skills can involve learners participating in discussions.	4	0	1
10. Speaking exams assess both production and comprehension together.	4	0	1
Total Correct Scores	41		
Total "Do not know" Responses	4		

As seen in the above table, the PhD cases all together have gained an overall correct score of 41 out of a maximum of 50 in terms of speaking assessment knowledge ($10 \times 5 = 50$). The score of 41 out of the maximum of 50 stands quite above the mean score of 25 ($50/2 = 25$). The obtained score of 41 is an indication that the PhD cases have scored quite satisfactory in terms of reading assessment knowledge. Moreover, they have responded with "do not know" 4 times.

Table 28 demonstrates the results for the MA cases in terms of writing assessment knowledge.

Table 28.

Results for the MA Cases in Terms of Speaking Assessment Knowledge

Items for Assessing Speaking	True	False	Do not Know
1. Role plays are employed when the emphasis is on evaluating discourse skills.	3	2	0
2. Assessing speaking skills may require more than one task.	2	2	1
3. Interlocutors showing interest through signals can be problematic.	3	1	1
4. Ending a task when a learner can't reach the criterion level is appropriate.	1	2	2
5. Simultaneously using holistic and analytic scales is problematic.	3	0	2
6. Oral reading is an appropriate method for evaluating speaking proficiency.	2	3	0
7. Checklists can be used to assess in-class oral presentations.	3	1	1
8. Interactive tasks with more than two learners can pose difficulties.	3	1	1
9. Assessing speaking skills can involve learners participating in discussions.	3	1	1
10. Speaking exams assess both production and comprehension together.	2	1	2
Total Correct Scores	27		
Total "Do not know" Responses	11		

Based on the information in the above table, the MA cases all together have gained an overall correct score of 27 out of a maximum of 50 in terms of speaking assessment knowledge ($10 \times 5 = 50$). The score of 27 out of the maximum of 50 is just above the mean score of 25 ($50/2 = 25$). The obtained score of 27 is an indication that the MA cases have scored fairly satisfactory in terms of speaking assessment knowledge. Furthermore, they have responded with "do not know" 11 times.

Table 29 exhibits the results for the ten cases in terms of writing assessment knowledge.

Table 29.

Results for the Ten Cases in Terms of Writing Assessment Knowledge

Items for Assessing Writing	True	False	Do not Know
1. Providing two prompts and asking students to choose one promotes scoring consistency.	0	8	2
2. Analytic scoring reveals learners' strengths and weaknesses.	8	1	1
3. Scoring scales and scores remain consistent across different levels.	0	7	3
4. In the final stages of writing assessment, mechanical issues like punctuation and spelling should be considered in grading.	6	2	2
5. Students are typically required to respond to at least two writing tasks in the exam.	2	6	2
6. Assessing writing skills by giving learners a reading text to use as a basis for their written response is a valid method.	7	2	1
7. Asking learners to present and justify their viewpoints in written form is an effective way to measure writing proficiency.	2	6	2
8. Using visual prompts to support writing can be problematic.	1	7	2
9. Holistic scoring is used to assess overall writing ability at the end of a term.	9	0	1
10. When it comes to assessing written work, using analytic scoring tends to provide more consistent results than holistic methods.	7	2	1
Total Correct Scores	71		
Total "Do not know" Responses	17		

As indicated in the above table, the cases all together have obtained an overall correct score of 68 out of a maximum of 100 for writing assessment knowledge ($10 \times 10 = 100$). The score of 71 out of the maximum of 100 stands quite above the mean score of 50 ($100/2 = 50$). This obtained score is an indication that the ten cases have scored very high in terms of writing assessment knowledge. Moreover, they have responded with "do not know" 15 times.

Table 30 shows the results for the PhD cases in terms of writing assessment knowledge.

Table 30.

Results for the PhD Cases in Terms of Writing Assessment Knowledge

Items for Assessing Writing	True	False	Do not Know
11. Providing two prompts and asking students to choose one promotes scoring consistency.	0	5	0
12. Analytic scoring reveals learners' strengths and weaknesses.	5	0	0
13. Scoring scales and scores remain consistent across different levels.	0	4	1
14. In the final stages of writing assessment, mechanical issues like punctuation and spelling should be considered in grading.	4	1	0
15. Students are typically required to respond to at least two writing tasks in the exam.	0	4	1
16. Assessing writing skills by giving learners a reading text to use as a basis for their written response is a valid method.	4	1	0
17. Asking learners to present and justify their viewpoints in written form is an effective way to measure writing proficiency.	1	3	1
18. Using visual prompts to support writing can be problematic.	0	4	1
19. Holistic scoring is used to assess overall writing ability at the end of a term.	5	0	0
20. When it comes to assessing written work, using analytic scoring tends to provide more consistent results than holistic methods.	4	1	0
Total Correct Scores	42		
Total "Do not know" Responses	4		

As seen in the above table, the PhD cases all together have gained an overall correct score of 42 out of a maximum of 50 in terms of writing assessment knowledge ($10 \times 5 = 50$). The score of 42 out of the maximum of 50 stands quite above the mean score of 25 ($50/2 = 25$). The obtained score of 42 is an indication that the PhD cases have scored very satisfactory in terms of reading assessment knowledge. Moreover, they have responded with "do not know" 4 times.

Table 31 demonstrates the results for the MA cases in terms of writing assessment knowledge.

Table 31.

Results for the MA Cases in Terms of Writing Assessment Knowledge

Items for Assessing Writing	True	False	Do not Know
1. Providing two prompts and asking students to choose one promotes scoring consistency.	0	3	2
2. Analytic scoring reveals learners' strengths and weaknesses.	3	1	1
3. Scoring scales and scores remain consistent across different levels.	0	3	2
4. In the final stages of writing assessment, mechanical issues like punctuation and spelling should be considered in grading.	2	1	2
5. Students are typically required to respond to at least two writing tasks in the exam.	2	2	1
6. Assessing writing skills by giving learners a reading text to use as a basis for their written response is a valid method.	3	1	1
7. Asking learners to present and justify their viewpoints in written form is an effective way to measure writing proficiency.	1	3	1
8. Using visual prompts to support writing can be problematic.	1	3	1
9. Holistic scoring is used to assess overall writing ability at the end of a term.	4	0	1
10. When it comes to assessing written work, using analytic scoring tends to provide more consistent results than holistic methods.	3	1	1
Total Correct Scores	29		
Total "Do not know" Responses	13		

Based on the information in the above table, the MA cases all together have gained an overall correct score of 29 out of a maximum of 50 in terms of writing assessment knowledge ($10 \times 5 = 50$). The score of 29 out of the maximum of 50 is fairly above the mean score of 25 ($50/2 = 25$). The obtained score of 29 is an indication that the MA cases have scored fairly satisfactory in terms of speaking assessment knowledge. Furthermore, they have responded with "do not know" 13 times.

Overall, based on the information available in Tables 20 to 31, it can be inferred that the level of language assessment knowledge of the teachers who teach at higher education level at Language Centres in an ascending order belongs to listening, reading, speaking, and writing. Furthermore, PhD holders enjoy a higher level of language

assessment knowledge in comparison with their MA counterparts in all four language skills.

4.4 Addressing the Third Research Question

The third research question of the present study was:

RQ3: To what extent do the teachers as test writers benefit from the 'lesson study' model to design language tests?

Interview and reflective journal data were thematically analysed, revealing five major themes (table 32). Each is presented below with a table of sub-themes and codes, followed by detailed explanations that integrate participants' voices. This approach demonstrates not only what benefits emerged, but also how these were experienced and interpreted by the participants themselves.

Table 32.

Summary of the Benefits of the 'Lesson Study' Model

No	Theme	Number of Mentions (Frequency out of 10)	Percentage (Out of 10)
1	Test design improvement	8	80%
2	Engagement improvement	6	60%
3	Collaboration enhancement	7	70%
4	Improved reflection	6	60%
5	Sustainability in test development	5	50%
Total	Five Themes	32	

4.4.1 Theme 1: Test Design Improvement

During classroom observations of the first trial, several students struggled with the clarity of instructions and the amount of time required for the writing task. Some students left sections blank or asked repeated questions about the task format. After revisions, the second trial showed improvement: students completed more items within the allocated time, and fewer clarification requests were made. This indicates that the changes in test design enhanced both clarity and feasibility.

Table 33 presents the sub-themes and codes related to the “Test Design Improvement” theme. This theme captures how teachers enhanced the alignment of test content with learning objectives, diversified assessment methods, and improved validity and reliability checks through collaborative lesson study cycles.

Table 33.

Test Design Improvement: Related Sub-Themes and Codes

Sub-Themes	Codes
Alignment with learning objectives	Curriculum mapping for test alignment, standardised assessment criteria, blueprint use
Diverse assessment methods	Task variety (open vs. closed questions), rubric development, formative feedback
Validity & reliability checks	Bias identification, pilot testing, alignment analysis (test vs. outcomes)

As shown in Table 33, the theme “*Test Design Improvement*” captures how the lesson study approach shifted teachers’ thinking about assessment. One noticeable change, under the sub-category of aligning tests with intended learning outcomes, was that instructors began moving away from creating tests based purely on intuition. Instead, they started designing items that more deliberately connected to external course objectives. For example, Mr. Rebaz mentioned that the lesson study “really helped me understand the nuisances around test development such as use of test specifications, blueprints and how each item relates to outcomes.” The approach also encouraged a move away from focusing too heavily on simpler elements — like isolated grammar questions — and toward more balanced coverage of language skills overall.

For various assessment forms, participants mentioned learning to include different genres of tasks and create elaborated rubrics. Mr. Rebaz also concluded that, “I didn’t know how to create test items before,” adding that the process helped them integrate short answer tasks alongside scenario-based writing and speaking prompts, all supported by clear scoring guides. Moreover, Mr. Salar noted that, “engagement of learners was significantly better after those changes — they weren’t just trying to guess what was

next.” Formative applications of assessments, such as oral practice embedded in speaking tests, were also developed, transforming tests from purely evaluative occasions into instructionally oriented events.

In validity and reliability checks, teachers have started creating pilot exams and looking at their tests for biases. Dr Kamal noted that early pilot testing revealed “a 12% variation in scores simply by the change in the order of the questions,” which prompted several rounds of resequencing and standardisation. Dr. Suzan reported that collaborative reviews had frequently flagged references that not all students had the cultural background to understand in reading passages, leading to changes that made such tests fairer for all students.

To sum up, these experiences demonstrate that the lesson study not only introduced new tools into the teacher’s toolkit but also promoted a mindset of purposeful, research-informed design. Teachers transitioned from developing examinations as a means of “covering” syllabus points to being strategic designers of assessments that were closely aligned with learning trajectories and student contexts.

4.4.2 Theme 2: Collaboration Enhancement

Table 34 summarizes the results regarding *the Collaboration Enhancement theme*, illustrating how the lesson study approach restructured professional conversations among teachers about test construction.

Table 34.

Collaboration Enhancement: Related Sub-Themes and Codes

Sub-Themes	Codes
Peer feedback mechanisms	Structured reviews of test drafts, cross-disciplinary critiques, iterative refinement
Shared ownership in test development	Rotating facilitator roles, collective responsibility for outcomes, informal accountability
Professional communities beyond cycles	Continuing groups for test sharing, resource pooling across institutions

Further, regarding peer feedback, the difference was clear: instead of working in isolation and rushing to prepare test papers on their own, teachers began organising themselves into critique circles. Dr. Yousif captured the influence, “The biggest thing was the collaboration – I know I learned a lot from people’s comments. They pointed out things that I had never noticed.” They would discuss how people would push back against the kind of unspoken assumptions we make — why, for example, would a listening script have idioms from some weird country somewhere, right, or why would we make it so vague, the rubric? These critiques influenced tasks; Mr Salar, for example, noted that they “argued about what was an appropriate level of cohesion for a good 20 minutes and then we revised together”.

The second sub-theme shared ownership derived from practices such as the rotation of roles, sometimes offering facilitation support, sometimes monitoring the process, spreading responsibility and avoiding domination. Teachers reported establishing shared timelines to ensure the development of rubric drafts or blueprint revisions; Mr. Rebaz explained that doing so created “a real shared commitment. We weren’t doing it because leadership told us to — there was also this sense of accountability toward one another.”

To add another level of analysis that supports the thematic findings, word cloud visualisation was utilised to represent the most important terms from participants’ interview transcripts and reflective journals. This image provides a comprehensive overview of the occurrence of emergent concepts and of how they related to one another. As shown in Figure 2, the main concepts in the word cloud are “feedback,” “formative,” “peer assessment,” and “collaboration.” These were consistently referenced in participants’ reflections on their development in relation to assessment. The frequency of these terms indicates a common focus on the principles of formative assessment and collaborative learning as integral elements of their developing language assessment literacy. This visual mapping confirms the earlier findings and supports the thematic output from the data.



Figure 2. The most common terms among participants' reflections and responses over Lesson Study cycles

These findings show how lesson study made better tests, but even more so, a stronger professional culture around test creation — one that was more critical, supportive and openly shared.

4.4.3 Theme 3: Student Engagement Improvement

During the initial administration, quite a large number of students seemed academically passive as they demonstrated minimal output throughout the speaking tasks. Some of the students appeared disinterested; they stood or sat quietly while awaiting their turn. Students, however, seemed more interactive and more present during the second observations upon revisions. Especially, they requested clarification less frequently, interacted during the speaking activity more, and even looked motivated to do well. This change indicated that the new test items were easier and more interesting.

Table 35 displays the results related to the *Student Engagement Improvement* theme, illustrating how the Lesson Study Model reshaped teachers' perspectives to view students as active contributors in assessment design and implementation.

Table 35.

Student Engagement Improvement: Related Sub-Themes and Codes

Sub-Themes	Codes
Active participation strategies	Piloting test tasks with students, integrating their feedback, designing authentic scenarios
Transparency in assessment criteria	Co-created rubrics, pre-assessment clarification sessions, opportunities for student self-checking

As seen in Table 35, the theme *Student Engagement Improvement* depicts that the Lesson Study Model facilitated a shift of teachers' orientation of students being perceived as passive subjects of assessments to being active participants.

Recurring active participant strategies involved piloting new speaking or writing tasks with students and revising according to learner feedback. Workshops "got us doing instead of just theorizing," Dr Suzan recalled "Student reactions surprised us; they found unclear instructions we had overlooked.". He also stated that "Through incorporating this input, tasks were more transparent and had purpose, and students took more ownership over the process."

To make assessment-related criteria transparent, teachers held sessions to unpack the test rubric, and occasionally, co-constructed it with students. Mr. Salar commented, "There was a special value that "...I could utilize through the tests and the items in which I was able to increase the level of engagement for my learners once again." Some participants shared that students used the rubrics to self-assess drafts before submission, which led to profound self-critiques and a more collaborative final marking discussion.

All these practices transformed tests into a component of learning. Students had a better idea of what was expected of them. They considered assessments as opportunities to exhibit their growth rather than difficulties, which helped teachers continue to adjust their instruments.

4.4.4 Theme 4: Improved Reflection

Teachers were also noticed, jotting down in test administration the students' reactions and problems. These comments were then employed in the reflection sessions. Mr Rebaz reflected after the second trial and wrote "I had understood how some of the items that I thought were simple were confusing for the students. Their struggle to unravel led me to reconsider my assumptions."

Table 36 shows how the theme *Improved Reflection emerged* as a core aspect of the changing teacher practices. In reflective practices, teachers reported how lesson study promoted the systematization of the collection of written challenges on the test. Dr. Kamal reflected, "I was able to think more systematically and put what I reflected on into practice in the classroom," that is, reflections from a single test cycle often informed immediate classroom changes.

Table 36.

Improved Reflection: Related Sub-Themes and Codes

Sub-Themes	Codes
Reflective practice routines	Detailed journals on test design, group discussions of recurring issues, targeted action plans
Critical evaluation of validity	Identifying subtle biases, pilot checks, explicit mapping to objectives

Moreover, concerning validity, the participants tracked from wide notions of 'fairness' to the target check — mapping of item to outcome, removal of culturally load.

The importance of validity, rubrics, and practical strategies was also reinforced through a density map generated using VOSviewer. The density map represented the importance of terms according to their frequency and centrality within the dataset. For instance, "feedback", "rubric", "reflection", and "validity" emerged as bright yellow and green zones, hence, they were both ubiquitous and prominent in the themes. Three clusters in the spatial map were detected: One was labelled "validity increase" and "feedback system" which may represent wider institution concerns about quality of the

assessment; the second, central, cluster was identified around “feedback” and “rubric” focusing on practical aspects of assessment strategies; and a third cluster to the far right side of the map was highlighted with “reflection”, “validity”, “technical growth” and “assessment” items that reflected a reflective and developmental disposition of the data.

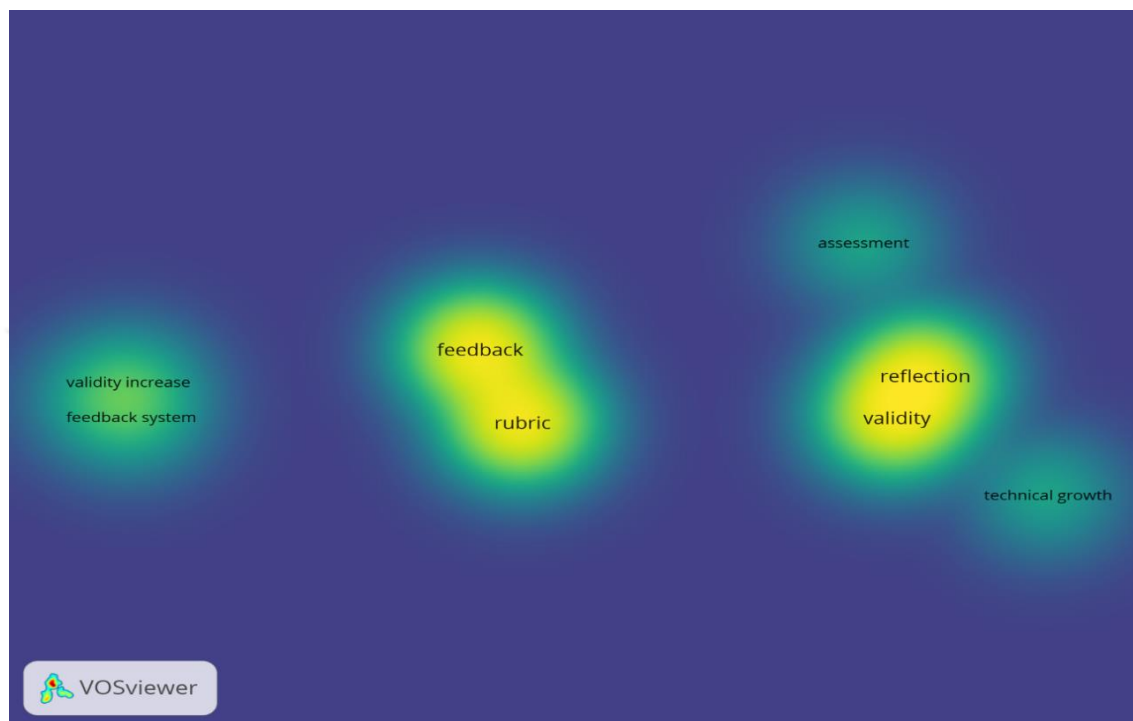


Figure 3. The density map representing the importance of terms according to their frequency

4.4.5 Theme 5: Sustainability in Test Development

Sustainability in Test Development highlights the apparent far-reaching effects of the lesson study cycles.

By institutional logic, teachers reported that leadership started allocating time to test design meetings, legitimizing that previously been inserted into personal hours. Mr Rebaz pointed out, "It would die in committee without leadership support. Initial discussions about slim budgets for bringing in outside assessment experts indicated that lesson study was becoming institutionalized within broader institutional agendas.

In the context of professional networks and continued collaboration, teachers felt that the camaraderie fostered by their lesson study experiences persisted even after formal

meetings concluded. They maintained connections through shared folders and messaging groups, exchanged examples of rubric applications, and visited colleagues at other schools to explore innovative practices, such as portfolio assessments. As Ms. Ranya noted, this collaborative model "has a rolling influence on the test development capacity of teachers," indicating the early stages of a culture that values iterative and collective improvement.

Table 37 illustrates how the theme Sustainability in Test Development became fundamental to sustaining the changes in teachers' practices. There was an emphasis by teachers on the need for institutional support; which has been exemplified by adding the time schedule for lesson observation; including budget for the assessment training, and lesson study in the organizational plan, which facilitated continuation. Participants also emphasized the importance of professional networks and continuation practices such as creating WhatsApp groups, engaging in informal mentorship, and exchanging ideas with educators in other institutions. These practices nurtured continued sharing and professional growth, outside the territory of the cycles of Lesson Study.

Table 37.

Sustainability in Test Development: Related Sub-Themes and Codes

Sub-Themes	Codes
Institutional support mechanisms	Dedicated time slots, initial moves toward budgets for assessment training, embedding LS in policy
Professional networks & continuation	WhatsApp groups for resources, informal mentorship, cross-institution idea sharing

Table 38 shows the summary of the results for all the ten participants in relation to the extent the teachers as test writers benefit from the 'lesson study' model to design language tests. Give some figures from the table

Table 38.

Summary of the Results for the Ten Participants for the Benefits of the 'Lesson Study' Model

No	Theme	Number of Mentions (Frequency out of 10)	Percentage (Out of 10)
1	Test design improvement	8	80%
2	Engagement improvement	6	60%
3	Collaboration enhancement	7	70%
4	Improved reflection	6	60%
5	Sustainability in test development	5	50%
Total	Five Themes	32	

As shown in Table 38, the most frequently reported benefit of the LS model was test design improvement, mentioned by 8 out of 10 participants (80%). Collaboration enhancement was also emphasised strongly, with (70%) referring to it as a major outcome. Moreover, reflection and engagement improvement were both noted by 6 participants (60%), indicating that teachers valued the reflective and participatory dimensions of the process as well as the technical aspects of test writing. Finally, half of the participants (50%) highlighted that LS approach encouraged them to maintain consistent practices beyond the immediate cycles.

Table 39 demonstrates the summary of the results for the five PhD participants in relation to the extent the teachers as test writers benefit from the 'lesson study' model to design language tests.

Table 39.

Summary of the Results for the Five PhD Participants for the Benefits of the 'Lesson Study' Model

No	Theme	Number of Mentions (Frequency out of 5)	Percentage (Out of 5)
1	Test design improvement	5	100%
2	Engagement improvement	4	80%
3	Collaboration enhancement	4	80%
4	Improved reflection	3	60%
5	Sustainability in test development	3	60%
Total	Five Themes	19	

Table 40 displays the summary of the results for the five MA participants in relation to the extent the teachers as test writers benefit from the 'lesson study' model to design language tests. Same above

Table 40.

Summary of the Results for the Five MA Participants for the Benefits of the 'Lesson Study' Model

No	Theme	Number of Mentions (Frequency out of 5)	Percentage (Out of 5)
1	Test design improvement	3	60%
2	Engagement improvement	2	40%
3	Collaboration enhancement	3	60%
4	Improved reflection	3	60%
5	Sustainability in test development	2	40%
Total	Five Themes	13	

Table 40 shows that three out of five participants (60%) reported improvements in test design, collaboration, and reflection as outcomes of the lesson study model. Engagement improvement and sustainability in test development were each mentioned by two participants (40%). These results highlight that the main benefits identified by this group were concentrated around test design, collaborative practices, and reflective thinking, while engagement and sustainability were recognised but to a lesser extent.

As it can be noticed in Tables 39 and 40, the total number of theme instances for the PhD and MA holders regarding the benefits of the 'Lesson Study' model turned out to be 19 and 13, respectively, which is an indication that PhD holders have referred to the emergent themes more than their MA counterparts.

4.5 Addressing the Fourth Research Question

The fourth research question of the present study was:

RQ4: How does this model contribute to test writers' professional development?

To explore this, interview data and reflective journals were thematically analysed. The findings revealed six major themes, each discussed below with its own table of sub-themes and codes, followed by detailed explanations that weave in participants' experiences. This structure highlights how the Lesson Study Model contributed to test writers' evolving professional knowledge, skills, networks, and identities.

4.5.1 Theme 1: Pedagogical Knowledge Improvement

Table 41 details the sub-themes and codes related to *Pedagogical Knowledge Improvement*. This theme explores how the Lesson Study Model deepened teachers' understanding of assessment theory and its practical classroom application.

Table 41.

Pedagogical Knowledge Improvement: Related Sub-Themes and Codes

Sub-Themes	Codes
Practical application of assessment theories	Integration of formative vs. summative assessments, alignment with Bloom's Taxonomy, backward design principles
Adaptation to student-centred assessment	Designing tasks for diverse learning styles, incorporating student feedback, differentiated strategies

The prominent theme was *Improvement of Pedagogical Knowledge* (Table 41), which described how it enhanced teachers' knowledge of assessment and their application of it.

In the real-world application of assessment theories, Ms. Dalia remarked that the process had helped her improve her understanding of assessment principles and the related classroom skills. For some, it marked a subtle shift in test development: moving from simply matching test items to the lower levels of Bloom's taxonomy to making intentional efforts to target the highest levels of thinking skills. Mr. Rebaz explained "just not designing tests or blue printing them and drafting balanced item sets where it's not just a matter of filling in points on a piece of paper, it's matching each item back to an outcome".

In the subtheme re-adaptation to student-centred assessment, the teachers turned to tasks that accounted for various learning styles from fixed formats. Dr. Suzan explained, "Student feedback suggested our rubrics were unclear so we amended them -- they responded better." Mr. Salar stated, "With the tests and test items I created, I can increase involvement of my learners more." So, teachers started giving options — write an essay or create a presentation or assemble a portfolio, for example — that play to varying strengths, and rethought tests as opportunities to show what they know.

These practices demonstrate the extent to which the Lesson Study Model encouraged teachers to put theory into practice, using the principles of curriculum alignment and student viewpoints to form more responsive assessment practices.

4.5.2. Theme 2: Professional Growth

Table 42 summarises the *Professional Growth* theme, highlighting how teachers developed reflective practices and greater autonomy in test design as a result of their participation in the lesson study process.

Table 42.

Professional Growth: Related Sub-Themes and Codes

Sub-Themes	Codes
Reflective practice and self-evaluation	Journals on past assessments, goals for skill gaps (e.g. rubrics), tracking growth in discussions
Increased autonomy in test design	Confidence in original tasks, less reliance on textbook tests, experimenting with formats

The theme *Professional Growth*, described in Table 42, reflected professional development. Through the lesson study, teachers developed their reflective ability and independence.

Regarding the reflective practice, Dr. Farman commented, "the Lesson Study Model helped me with continuous development of my test writing skills (e.g., reflecting on my designed tests' feedback) and teaching (e.g., [elaboration on] feedback reflected in my teaching)." Teachers shifted from generic impressions to targeted criticisms, like probing journal tests for why specific tasks continually tripped students up and then word modification or structure in the next cycle. Dr. Tara said that monitoring these improvements throughout classes had turned "what used to be a bit of a fog about how writing gets scored into laser-focused questions around bias."

The sub-theme of enhanced autonomy illustrated how lesson study raised confidence. Dr. Yousif acknowledged that "At the beginning, I stuck to textbooks and didn't want to risk 'bad' items." Later, after getting peer feedback and revising them together, they started creating their scenarios for speaking exams and began to pilot and refine them. As they summarized it: "Risky it but it worked out — students were so engaged and the scores better aligned with ability." This move from adopting (being passive) to designing something (being active) demonstrates how lesson study improves skill and agency.

4.5.3 Theme 3: Peer Learning

Table 43 provides a summary of the main components of the *Peer Learning* theme. It reflects how collaborative practices among outside schools deepened teachers' assessment practices and professional thinking.

Table 43.

Peer Learning: Related Sub-Themes and Codes

Sub-Themes	Codes
Collaborative problem-solving	Brainstorming tricky test items, peer critiques, bias reduction strategies
Cross-institutional knowledge exchange	Adopting best practices from other universities, comparative test reviews, workshops by peers

As *peer learning* became a central focus of professional development, Table 43 outlines how this process unfolded. Mr. Haval remarked that his peers gave him “multiple viewpoints and actions on test development and administration, which built who I am as a person.” Other participants noted that peer sessions encouraged them to question their assumptions — for example, recognising cultural bias in a test prompt or realising how unclear their grading language had become.

4.5.4 Theme 4: Enhanced Assessment Literacy

The *Enhanced Assessment Literacy* theme (Table 44) illustrates how teachers moved from theory to practice in gaining proficiency in reliability, validity, and data-informed test redesign.

Table 44.

Enhanced Assessment Literacy: Related Sub-Themes and Codes

Sub-Themes	Codes
Mastery of validity and reliability	Piloting tests for consistency, item difficulty stats, CEFR alignment
Feedback loops for redesign	Post-test analysis, student data use, iterative refinement of flawed tasks

The *Enhanced Assessment Literacy theme*, depicted in Table 44, presents how participants advanced to greater support from basic knowledge of testing principles to operational expertise. Regarding test validity and reliability, Dr. Suzan described piloting a test was “Actually variation was eye-opening — we thought we were consistent until data showed otherwise.” This resulted in the standardisation of the examiner scripts and the re-scoring of standardization, which immediately increased reliability.

The participants used basic statistics in the same way: they looked at which items were too easy or too hard and then reworked them. They also started to explore why some objectives constantly underperformed and developed more focused test content.

Under feedback cycles, Dr. Suzan said, “Every cycle is not just correcting — it’s rebuilding trust in our tools.” The participants analyzed patterns in post-test results, determined mixed-up or misaligned prompts, and revised them, turning assessment into a live, evidence-based process rather than a static one.

4.5.5. Theme 5: Professional Networking

Table 45 illustratively summarizes the *Professional Networking* theme and the lesson study process that led to continued collaborative efforts, mentoring and resource sharing between test writers.

Table 45.

Professional Networking: Related Sub-Themes and Codes

Sub-Themes	Codes
Building communities of practice	Regular LS meetings, online sharing platforms, joint assessment manuals
Mentorship and role modelling	Shadowing expert test writers, peer coaching on rubrics, feedback from senior assessors

As shown in Table 45, the theme of *Professional Networking* captures how lesson study drew teacher teams into active and ongoing social interaction. Reflecting on the growing sense of community, Dr. Tara noted, “The meetings, sharing rubrics instantly and resolving testing issues kept us on the same page.” Through co-authoring internal guides on test specifications and common pitfalls, teachers began embedding new assessment norms into the institution’s shared practices and memory.

She also reflected on the value of shadowing more experienced colleagues, stating, “Just watching it balance strict criteria with student rapport was a masterclass — it’s one thing to read about fairness, another to see it done live.” Sometimes critical feedback from senior assessors could be challenging, it often prompted teachers to revise vague rubrics and minimise subjectivity in their grading practices. This mentorship evidence can boost development beyond what an independent test writer alone can create.

4.5.6 Theme 6: Empowerment

Table 46 sums up the *Empowerment* theme, reflecting how teachers gained confidence, took on leadership roles, and embraced a professional identity as assessment experts through the lesson study experience.

Table 46.

Empowerment: Related Sub-Themes and Codes

Sub-Themes	Codes
Increased agency in decision-making	Advocacy for curriculum-aligned assessments, leadership on test committees, proposing new policies
Recognition of professional identity	Calling themselves “assessment experts,” contributions to publications, conference participation

Table 46 shows that *Empowerment* as a theme enables teachers to move beyond technical skills to a sense of professional agency. In their reported decision-making, Dr. Yousif explained, “I felt confident to lead our testing committee pushing for anonymised grading and unified rubrics. Mr. Rebaz continued, “I was like, ‘Oh, I’m realising now that it’s not just right, it’s actually a responsibility to raise those gaps that we found’ — I think that gave me the confidence to do it.”

In professional identity, Mr. Rebaz described, “I used to tell them I taught English. Now, people ask, and I introduce myself as a specialist in assessment — and owning it matters.” This illustrates how the lesson study not only equipped the participants with new approaches but positioned them to influence the assessment culture of their schools or colleges.

Table 47 shows the summary of the results for all the ten participants in relation to the contributions of Lesson Study Model towards test writers’ professional development.

Table 47.

Summary of the Results of the Ten Participants for the Contributions of Lesson Study Model Towards Test Writers' Professional Development

No	Theme	Number of Mentions (Frequency out of 10)	Percentage (Out of 10)
1	Pedagogical knowledge improvement	9	90%
2	Professional growth	6	60%
3	Peer learning	7	70%
4	Enhanced assessment literacy	7	70%
5	Professional networking	8	80%
6	Empowerment	6	60%
Total	Six Themes	43	

Table 47 presents an overview of findings from the ten participants in relation to how the Lesson Study Model contributed to test writers' professional development. The results show that improvement in pedagogical knowledge (having the most scores) was the most reported theme based on the data by 9 out of 10 respondents (90%). Then came networking online (80%), proving what a difference it makes to collaborate and connect with others through this process. Further, 70% of participants recognised both peer learning and increased assessment literacy (dual benefit of shared practice and greater testing expertise). Professional development and self-empowerment were also recognised by 60% of participants, reflecting the model's holistic personal and professional impact. To sum up, the themes did indeed emerge as a rich and nuanced portraiture of the impact of the Lesson Study Model on teachers' growth as test writers.

Table 48 demonstrates the summary of the results for the five PhD participants in relation to the contributions of Lesson Study Model towards test writers' professional development.

Table 48.

Summary of the Results of the Five PhD Participants for the Contributions of Lesson Study Model Towards Test Writers' Professional Development

No	Theme	Number of Mentions (Frequency out of 5)	Percentage (Out of 5)
1	Pedagogical knowledge improvement	5	100%
2	Professional growth	4	80%
3	Peer learning	4	80%
4	Enhanced assessment literacy	4	80%
5	Professional networking	5	100%
6	Empowerment	4	80%
Total	Six Themes	26	

Table 48 shows the five PhD participants' perceptions about how the Lesson Study Model contributed to test writers' professional development. The results indicate that both pedagogical knowledge enhancement and networking were recognized as important (100%) to all and hence instructed as a central part of professional development. Professional development, peer learning, increased assessment literacy, and empowerment were also addressed by 80% of participants, demonstrating a strong influence across multiple dimensions. Collectively, these findings imply the Lesson Study Model not only contributed to the knowledge and connections for our PhD participants but significant development, collaboration, and confidence as test writers for them as well.

Table 49 displays the summary of the results for the five MA participants in relation to the contributions of Lesson Study Model towards test writers' professional development.

Table 49.

Summary of the Results of the Five MA Participants for the Contributions of Lesson Study Model Towards Test Writers' Professional Development

No	Theme	Number of Mentions (Frequency out of 5)	Percentage (Out of 5)
1	Pedagogical knowledge improvement	4	80%
2	Professional growth	2	40%
3	Peer learning	3	60%
4	Enhanced assessment literacy	3	60%
5	Professional networking	3	60%
6	Empowerment	2	40%
Total	Six Themes	17	

As it can be noticed in Tables 48 and 49, PhD and MA holders reported these themes in relation to the contribution of the LS Model to their growth as test writers with all boxes filled for each group, but the frequency of mentioned themes varied. The PhD participants reported 26 examples across the emergent themes while the MA participants reported 17, indicating that PhDs were more engaged and in-depth with the academic application of the model. Regarding pedagogical knowledge enhancement, (80% of the MA participants) LS also played a key role in enhancing their knowledge of practice in teaching related to assessment. Peer learning, improvement in assessment literacy, and networking were also mentioned by 50% of the participants and peer learning, improvement in assessment literacy, and networking were mentioned by 60% of MA participants, thus highlighting the social, reflective, and relational aspects of the model in supporting growth around more than technical aspects of assessment skills. Professional development and autonomy were less frequently discussed (40% and 40%, respectively), indicating that many participants did not believe its impact extended to broader career progression and agency building. The PhD holders also repeated, though, these themes more often, especially in topics related to assessment literacy and professional development, which may evidence their higher level of study completion as well as longer teaching experience and exposure to educational measurement discourse. The findings point to the effects of the LS Model in both groups, however, a more comprehensive and stronger impact on the PhD holders in relation to the MA participants were evident in the areas of pedagogical knowledge, collaborative learning and consciousness of the principles of assessment compared to the lesser emphasis on these dimensions of professional development and empowerment in the MA participants.

Before presenting the summary of this chapter, it is useful to illustrate how the different themes identified in the findings were interrelated. To achieve this, a network map was generated from participants' reflections and interview transcripts. This visualisation provided an overview of the relationships among the most frequently mentioned concepts and highlighted how technical aspects of assessment (e.g., feedback, rubrics, validity) were linked to reflective dimensions (e.g., professional growth and critical reflection).

This topological relationship was also illustrated in terms that appeared close together in the network map and were conceptually similar. From the emergent structure, two main thematic clusters developed. The first cluster, illustrated in red lines, related "feedback," "rubric," "feedback system," and "validity increase" and represented the procedural and technical aspects of assessment practice. The second cluster, illustrated in green, included "reflection," "validity," and "technical growth," suggesting an emphasis on reflective practice and professional development. Interestingly, "assessment" occupied a bridging position between clusters, serving as a connective concept that linked technical processes with reflective depth.

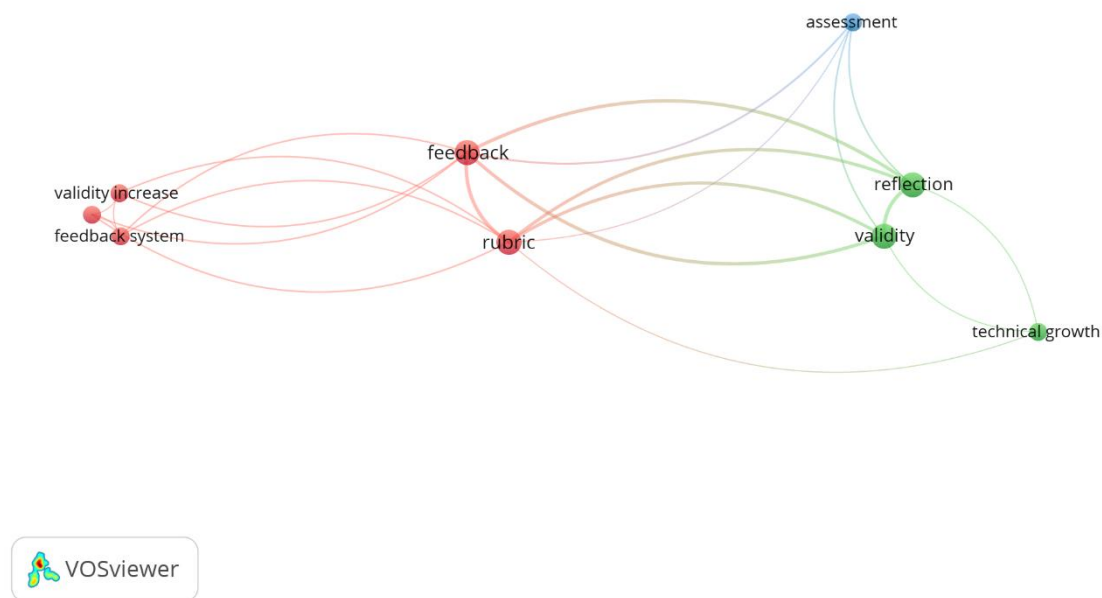


Figure 4. The network map

These graphic representations are supportive of the overall claims of the study: that the lesson study approach summarizes a dual track as a model of professional learning which develops not only technical craft, but reflective capacity as well. Clustering of core terms also suggests that the participant did not think of feedback, rubrics and validity as being stand-alone structures, but instead as interrelated aspects of a process of pedagogical development and professional development. It was the existence of “technical growth” within the reflective cluster, however, that further indicated that teachers did not merely regard professional development as cognitive reflection, but as practical, skill-based movement.

In conclusion, the VOSviewer visualisations provide supportive evidence of the unitary and organic character of learning in the lesson study system. They provide a visually compelling story that captures the essence of the qualitative results, depicting professional development as a dynamic interaction of structured feedback practices, validity-oriented assessment, and reflective inquiry.

4.7 Chapter Summary

This chapter presented the findings of the study in relation to the research questions. It reported on the baseline levels of participants' LAL, showing relative strengths in writing assessment and weaknesses in listening and speaking assessment. The chapter also described how differences in degree (PhD vs. MA) and years of experience shaped participants' assessment knowledge. The implementation of the LSM cycles was analysed, with evidence showing how collaborative planning, classroom observations, and group reflections led to the revision and improvement of test tasks. Observations highlighted initial challenges, such as unclear instructions and poor student engagement, which were addressed in the second cycle through test revisions. The findings also showed increased teacher collaboration, greater awareness of assessment principles, and enhanced reflection on practice. Visual analyses, including network and density maps, illustrated the development of participants' LAL across technical and reflective dimensions. Overall, the findings demonstrated both immediate improvements in test design and longer-term growth in teachers' professional capacities. The next chapter discusses these findings in relation to the literature and highlights the study's original contributions.

CHAPTER V

DISCUSSION

5.1. Introduction

The present chapter presents the findings of the study with regard to the research questions and other studies. It seeks to analyze the findings reported in Chapter Four, and to shed light on their relevance to teacher development, assessment literacy, and Lesson Study Model implementation in the Iraqi EFL educational environment. The section starts with an overview of the main themes which emerged from the analysis of the data, followed by the relationship between these findings and relevant theoretical viewpoints and research literature. Special focus is placed on the effect of the Lesson Study Model on test writers' pedagogical content knowledge, professional learning, collegial work, and evaluative work. Lastly, the chapter discusses the implications of these findings for policy, practice, and future research.

This study aimed to explore the academic and professional needs of teachers as test writers regarding their Language Assessment Literacy (LAL). Additionally, it sought to uncover the level of language assessment knowledge among teachers at higher education institutions within Language Centres. The research also intended to examine how teachers, as test writers, benefited from the Lesson Study Model in designing language tests. Furthermore, the study investigated how this model contributed to the professional development of test writers.

The results of this study regarding the academic and professional needs of teachers as test writers indicated that the assessment literacy competence of the participants, in descending order, was highest for writing, then speaking, followed by reading, and finally listening. In other words, the participants demonstrated the highest level of assessment literacy in writing, followed by speaking as the productive language skills. However, their assessment literacy levels for reading and listening, which are receptive skills, were lower. Furthermore, the results showed that PhD holders achieved substantially higher scores in assessment literacy competence across all four language skills compared to their MA-holding counterparts. It was also found that teaching experience significantly influenced self-perceived assessment skills; teachers with more extensive teaching experience reported higher scores in their self-perceived assessment skills. The results

indicated that overall training in English language testing and evaluation for the PhD and MA groups was rated as highly satisfactory and roughly satisfactory, respectively, suggesting that MA holders need more training in English language testing and evaluation. Additionally, teachers' primary academic and professional needs as test writers in terms of language assessment literacy relate to assessing receptive skills, integrative testing, and understanding test, task, and item types. The respondents also expressed the need to assess students through portfolios, peer assessment, and self-assessment. Moreover, the results illustrated that MA holders exhibited greater academic and professional needs as test writers concerning language assessment literacy.

The results of this study regarding the level of language assessment knowledge among teachers at higher education Language Centres indicated that their knowledge levels, in ascending order, belong to listening, reading, speaking, and writing. Furthermore, PhD holders demonstrated a higher level of language assessment knowledge compared to their MA counterparts across all four language skills.

The current study's results regarding how teachers as test writers benefited from the Lesson Study Model in designing language tests revealed five themes: improvements in test design, enhanced collaboration, increased student engagement, better reflection, and sustainability in test development.

The results concerning the contributions of the Lesson Study Model to test writers' professional development showed six themes: improvement in pedagogical knowledge, professional growth, peer learning, enhanced assessment literacy, professional networking, and empowerment.

5.1.1. Discussing the First Research Question

RQ1: What are the academic and professional needs of teachers who teach at the higher education level at language centres as test writers in terms of language assessment literacy as they perceive?

The study's results indicate that self-perceived assessment literacy in productive skills (writing and speaking) is regarded as higher than in receptive skills (reading and listening), which corroborates what the related literature states about assessment practices. Bachman and Palmer (1996) and Fulcher (2012) argue that traditional testing of language abilities has emphasized discrete-point assessments of grammar and

vocabulary, which readily lend themselves to scoring productive tasks. In contrast, writing and speaking assessments tend to be evaluated against criteria and rubrics with which teachers have become socialized to be familiar and confident (Brown, 2004). Conversely, the assessment of receptive skills, especially in standardized contexts, may rely on indirect assessments, such as multiple-choice questions, which teachers might view as not genuinely reflecting the so-called "real" use of the language. As a result, their confidence in designing or rating such assessments may have been diminished (Spolsky, 1978). The differences may also stem from curricular and institutional priorities. Language programmes typically emphasize productive skills since their examples are more evident within academic or professional contexts, while receptive skills are often viewed as receiving only secondary consideration (Heaton, 2011).

In the findings of this investigation, self-perceived assessment literacy among PhD holders is found to be higher than that of MA holders across all four skills. Such observations raise whether there are any links between academic qualifications and perceived competency in assessment practice. Previous studies addressed this issue, indicating a positive role for higher academic degrees in variables related to some teachers, such as self-efficacy, pedagogical thinking, and reflective practices. For example, Akbari and Moradkhani (2010) demonstrated that Iranian EFL teachers with advanced academic degrees reported higher levels of self-efficacy than their peers without such degrees. This suggests that higher education may allow educators better to question their mastery of assessment theories and practices.

Additionally, Akbari and Moradkhani (2012) concluded from their study that teachers with academic degrees related to ELT demonstrate more effective pedagogical thought processes compared to those without such degrees. Therefore, the extensive knowledge these teachers acquire during their PhD programmes will enhance their assessment literacy. The results of this study align with those of Mahmoodarabi and Khodabakhsh (2015), who pointed out a significant influence of EFL teachers' academic degrees on their perceptions of critical pedagogy. A profound understanding of educational theories gained through advanced academic degrees enables PhD holders to approach assessment research more nuancedly, resulting in a higher level of perceived assessment literacy than that of teachers who hold only a master's degree. Exposure to advanced training in research and theoretical frameworks, along with the advanced studies of PhD holders, likely equips them to integrate multiple skills with sophisticated

assessment methods. On the other hand, MA programmes tend to be more employment-oriented (Obeidat & Shehadeh, 2020). They may not sufficiently address more complex aspects of receptive skills, such as analysis of inferential comprehension or discourse coherence (Bachman & Palmer, 1996). Teachers face challenges in designing tasks to evaluate higher cognitive processes, as noted by Alkharusi (2011). Therefore, tailored teacher-training programmes in receptive skill assessment, focusing on task design, authenticity, and critical reflection, may improve this situation, as suggested by Xu and Brown (2016).

Regarding the two reasons for the disparity in self-perceived assessment literacy, the first is the greater sophistication of the knowledge acquired during PhD studies; the second is the emphasis on research and critical analysis in doctoral programmes. Generally, PhD programmes involve extensive coursework, research work, and familiarity with literature. This results in a deeper understanding of the theories behind language acquisition, assessment strategies, and their pedagogical implications. Such programmes equip doctoral candidates with a wealth of tools necessary for critically evaluating and implementing a wide variety of effective assessment strategies across all four language skills. Consequently, PhD holders tend to have a higher level of self-perceived competence regarding the assessment of student performance, leading to better self-perceived assessment literacy scores. In contrast, master's programmes often focus more on practical teaching skills (Obeidat & Shehadeh, 2020; Tezel, 2006) compared to PhD programmes, typically lacking an exploration of theoretical frameworks or advanced assessment techniques. As Seifalian and Derakhshan (2018) pointed out, this may contribute to a sense of reduced self-efficacy among their participants.

PhD programmes emphasize research and critical thinking, enabling candidates to engage with newly explored educational issues, such as assessment literacy. This emphasis on research fosters an inquiry-driven environment where PhD candidates learn to critically analyse current assessment practices and adapt them based on research evidence. This skill is crucial for developing assessments that cater to diverse learner needs. Soodmand Afshar and Farahani (2018) noted that reflective teaching practices can significantly improve pedagogy; in fact, the reflective skills gained through the PhD programme may enhance self-perceived effectiveness in assessment literacy among PhD graduates.

The conclusion of this study on the relationship between teaching experience and self-perceived assessment literacy will contribute to the existing literature regarding teacher efficacy and professional development. Indeed, participants found that teachers with more experience had higher scores in self-assessment of their assessment skills. This is further supported by previous empirical studies that highlight the significant role of experience in shaping educators' perceptions and competencies. Research indicates that experienced teachers possess a better awareness of classroom dynamics and the evolution of assessments compared to their inexperienced counterparts. For example, the study by Carter et al. (1988) demonstrated that experienced teachers are more adept at perceiving and processing visual information in the classroom than novice teachers. This skill is vital for effective testing and assessment. Thus, experienced teachers have a comprehensive understanding of learner behaviours and learning needs, allowing them to analyse these factors in relation to their assessment strategies. Consequently, this adaptability is essential for accurately assessing students, reflecting the teachers' profound grasp of pedagogical principles.

Moreover, Gatbonton's (2008) study showed that experienced ESL teachers possess significantly deeper pedagogical expertise compared to novice teachers. This level of knowledge enhances classroom practises, particularly in assessment. With increased teaching experience, teachers are able to critically reflect on their practises, leading to more effective assessment techniques that account for students' diverse learning profiles. Similarly, Fitriyah et al. (2022) concluded that experienced EFL teachers surpassed their novice peers in classroom-based language assessment literacy.

Such results can be attributed to increased knowledge and skills, along with reflective practise. As teachers gain experience, they accumulate a wealth of knowledge about assessment practises. This knowledge is obtained through theoretical learning, formal training, professional development, and hands-on activities in the classroom. This enables teachers to acquire a variety of strategies for effectively assessing learning. In this context, Akiri and Dori (2022) note that teachers develop professionally by engaging in a continuous learning process reinforced by practise, which enhances their ability to implement various assessment methods tailored to learners' needs. This process boosts their self-efficacy regarding assessment skills. In fact, experienced teachers tend to have greater confidence when facing the complexities of evaluating learner performance. Furthermore, experienced teachers' participation in reflective practise allows them to

critically assess their own teaching methods and assessment strategies. Tschannen-Moran and Hoy (2007) observe that reflective practises significantly enhance teachers' self-efficacy. As teachers reflect on their experiences, they can identify their own weaknesses, enabling them to improve their assessment performance and make necessary adjustments in their assessment approaches. This reflective process nurtures their existing skills and fosters a growth mindset that promotes continuous improvement. Consequently, it is likely that experienced teachers view themselves as more professional in assessment practises due to their ability to critically evaluate their effectiveness over time, especially in contrast to less experienced teachers.

Additionally, Estaji and Ghiasvand (2022) highlight the differences in how novice and experienced teachers perceive classroom supervision and professionalism. This discrepancy may stem from experienced teachers' deeper understanding of assessment expectations, leading to a greater self-perception of their skills. By navigating various educational contexts, experienced teachers gain insights into effective assessment practises that align with institutional standards, thereby enhancing their confidence.

5.1.2. Discussing the Second Research Question

***RQ2:** What is the level of language assessment knowledge of the teachers who teach at higher education level at Language Centres?*

The findings of this investigation indicate that the order of language assessment knowledge among the teachers for listening, reading, speaking, and writing aligns with the hierarchy of language skills and reflects institutional priorities (Genç et al., 2020; Farhady & Tavassoli, 2018). Curricula tend to prioritize productive skills, particularly writing and speaking, due to their prominent roles in academic and professional domains (Heaton, 2011). As a result, significant training for teachers has been conducted on designing and assessing these skills, which has enhanced their confidence in these areas. In contrast, receptive skills are mainly assessed through formats many teachers regard as standardized (e.g., multiple-choice), raising concerns about their authenticity and pedagogical value (Tavassoli & Farhady, 2018). Alkharusi (2011) similarly found that teachers struggle to create tasks that measure higher-order cognitive processes in listening and reading, and this challenge is compounded by inadequate training.

This study highlights that teachers have several academic and professional needs related to language assessment literacy. They primarily require expertise in evaluating receptive skills, integrated testing, and essential assessment techniques such as dynamic assessment, online portfolio assessment, peer assessment, and self-assessment. Research indicates that greater emphasis should be placed on receptive skills, as these areas are particularly challenging to assess effectively. Farrall (2012) examined the complex relationship between language, literacy level, and cognition in the context of reading assessment, arguing that teachers need to understand these relationships for fair learner evaluation. Mahmood et al. (2021) echoed this point by highlighting differences in reading assessment awareness among Pakistani ESL teachers, noting that many lack the skills and knowledge necessary for effective assessment practises. Rasyid et al. (2023) also asserted that assessing listening comprehension skills poses significant challenges in EFL classrooms. Furthermore, it appears that teachers' needs regarding receptive skills assessment arise from inadequate training and resources, often hindering effective assessment strategies. This indicates that teachers must be knowledgeable about integrated testing, highlighting the increasing recognition of holistic assessments over isolated evaluations of language skills. Accordingly, Razavipour (2013) emphasized the need for developing assessment literacy in high-stakes contexts, where teachers can better understand students' capabilities through an integrative approach. Berger (2023) underscored the role of assessment literacy in classroom processes, suggesting that teachers should cultivate the ability to create assessments that integrate various language skills. This reflects that teachers are aware of the importance of integrative testing but lack confidence in implementing such practises effectively. Participants' preferences for diverse tests, tasks, and item types relate to North's (2005) study on the Common European Framework of Reference (CEFR), highlighting the importance of transparency in the variety of language assessment. Additionally, in a separate investigation, Stabler-Havener (2018) examined how teachers conceptualize assessment literacy and concluded that teachers should thoroughly understand various assessment formats to accurately evaluate student performance. The study also explores how Kavakli and Arslan (2019) describe the transition of teachers from learning early assessment concepts to applying them in practise. This current study suggests that while teachers recognize the necessity of diverse assessment types, they require additional support for effective application.

Contrary to contemporary principles and approaches that offer new pathways for student agency and reflective learning, teachers need to understand new assessment methods, such as portfolios and peer assessment. For instance, Wu (2018) explains how teachers' writing assessment literacy can be improved through e-portfolios, while Deneen and Brown (2016) emphasize the impact of teachers' conceptualization of assessment on their assessment literacy. Ng et al. (2020) stress the necessity for teachers to develop a deeper appreciation of the peer-evaluation process to enhance their own assessment literacy, specifically highlighting peer assessment. Additionally, Thong-Iam (2019) acknowledges the significance of self-assessment for enhancing EFL teachers' professional competencies. The present study reveals that teachers increasingly recognize these methods as valuable insights for promoting student engagement and ownership of learning.

These results can be attributed to inadequate professional development and evolving educational standards. The primary reason teachers feel the need to gain language assessment literacy relates to their ineffective professional development programmes. Many teacher training programmes lack comprehensive training on receptive skills assessment or integrative testing methods. Consequently, as Singh et al. (2022) observe, teachers are unprepared to effectively tackle these challenges. Such ineffective programmes can create gaps in teachers' knowledge and skills, causing them to resort to traditional assessment methods that may not align with best practises. Additionally, language instruction is rapidly evolving, with formative assessment techniques such as portfolios, peer assessment, and self-assessment being increasingly emphasized (Yan & Carless, 2022). Given the prevalence of more holistic approaches to student evaluation, teachers find it essential to adapt their assessment practises but lack the necessary training to do so with confidence. This evolving context demands continuous professional development that considers contemporary assessment practises, equipping teachers with the tools they need to meet these new expectations.

5.1.3 Discussing the Third Research Question

***RQ3:** To what extent do the teachers as test writers benefit from the 'lesson study' model to design language tests?*

This study provides valuable insights into the benefits that teachers gain as test writers from the Lesson Study Model applied to language test design. The thematic analysis revealed five key themes: improvements in test design, enhanced collaboration, increased student engagement, heightened reflection, and sustainability in test development.

The findings of this study align with previous empirical research on the role of lesson study in enhancing teacher competencies. For example, Fernandez & Yoshida (2004) highlight that collaborative lesson planning and peer observation form the foundation of lesson study and help develop a shared understanding of pedagogical goals. This aligns with the principal theme of this study, which focuses on enhancing collaboration. Ongoing dialogue among teachers refines instructional strategies, as evidenced by participants' reports of collaboratively improving test designs. Likewise, structured reflection during lesson study fosters metacognitive awareness, according to McDowell (2010), thus reinforcing the theme of heightened reflection. The participants' critical reflection on their assessment practises is in harmony with Schön's (1983) concept of reflection-on-action, where post-lesson analysis allows educators to iteratively refine their methods.

Zhou and Xu's assertion (2017) that lesson study establishes lasting professional learning communities further supports the theme of sustainability in test development. Their research highlights how this extended collaboration fosters long-term pedagogical innovation, mirrored in participants' sentiments about maintaining even improved practices. Moreover, the cyclical nature of lesson study—comprising planning, implementing, observing, and reflecting (Cerbin & Kopp, 2006)—provides a theoretical foundation for enhancing test design. This ongoing process enables teachers to advance systematically to address gaps in assessment literacy, as noted by participants who appreciated the iterative feedback cycles.

The focus on improving student engagement aligns with the social constructivist principles that underpin lesson study (Dadds, 1997). Participants of the current study

transitioned towards implementing communicative language testing that embraced learner-centered approaches and encouraged student interactions, similar to the designs of Canale and Swain (1980). Their emphasis on real-life tasks likely served as an asset due to Bachman and Palmer's (1996) focus on authenticity; previous attention to real-world relevance possibly deepened students' engagement, fostering a stronger connection with the material.

A key justification for these outcomes lies in lesson study's dual focus on structured collaboration and contextual adaptability. Unlike traditional one-off workshops, lesson study's embedded, cyclical format (Lewis & Hurd, 2011) ensures that professional development is context-specific and responsive to institutional challenges, such as the assessment literacy gaps identified among Iraqi test writers. This aligns with Fulcher's (2012) assertion that practical LAL training must address localized needs through participatory models.

The first theme, test design improvement, is supported by the literature, emphasizing the role of structured professional development in advancing assessment practices. According to Cinkara (2020), focusing on collaborative quiz literature aids EFL test writers in enhancing their assessment techniques. This, in turn, leads to higher quality language tests. Similarly, Rock and Wilson (2005) noted that lesson study enhances teachers' ability to plan, observe, and reflect, thereby significantly improving the quality of educational assessments. The teachers in this study found lesson study highly beneficial as they could exchange best practices; additionally, they received constructive feedback on their test designs. This is reinforced by Cinkara's (2020) findings, which support the idea that collaborative efforts lead to improved test quality.

One of the important themes was collaboration. This aligns with previous research emphasizing the significance of collaborative professional development models. Similarly, DuFresne (2007) elaborated on how lesson study fostered a collaborative atmosphere among teachers, enabling them to work toward common goals. This collaborative approach is also supported by Norwich et al. (2018), who demonstrate how inter-professional collaboration plays a crucial role in enhancing teaching practises. In the current study, the teachers felt a stronger sense of community with their peers as they sought a shared opportunity to plan and evaluate language tests. This community further strengthened their individual practises, resulting in a collective understanding of effective assessment strategies.

The term 'improvement of student engagement' is a core area because it highlights the growing recognition of how assessments contribute to learners' motivation and active participation in learning. Hannah (2019) states that the lesson study approach could significantly enhance practice by aligning these assessments more closely with learners' needs and interests to boost engagement. Additionally, Setyaningsih (2018) notes that an optimized lesson study would yield better learning outcomes due to assessments being more relevant to students. The teachers in this study held a positive view of lesson plans since their collaborative nature often encouraged them to consider students' perspectives when designing the tests. This situation resulted in engaging and well-suited assessments to learners' needs.

Improved reflection emerged as a key theme among participants. Abd Rahim and Abd Rahim (2020) noted that lesson study encouraged teachers to critically assess their practises, leading to a deeper understanding of their teaching and assessment methods. Supporting this viewpoint, McDonald (2009) emphasized the vital role of reflective practise in professional development through lesson study. Participants reported that the structured reflection processes inherent in lesson study enabled them to critically evaluate their assessment strategies, resulting in more informed decisions regarding test design. This finding aligns with current literature, highlighting the importance of reflective practises in enhancing teacher effectiveness.

The final theme addressed sustainability in test development, indicating that the benefits of lesson study extend beyond immediate practise enhancements. Vogt and Tzagari (2014) emphasized the ongoing development of assessment literacy among second and foreign language teachers. Such continual progress can be achieved through sustainable professional development models like lesson study. Subadi et al. (2013) argued that lesson study can serve as a long-term developmental framework for teachers. In this study, participants shared their experiences of collaboration through lesson study, which subsequently enabled them to make informed decisions about their test-writing practises over time. This emphasis on sustainability is crucial for fostering a culture of continuous improvement in assessment literacy.

The results can be attributed to two main factors: structured professional development and a focus on student-centred learning. The former equips teachers with a framework that allows for deep engagement with their practise. In contrast to traditional professional development models, which often consist of limited seminars, lesson study

emphasizes improving the interactive nature of assessment development through iterative cycles of planning, teaching, observing, and reflecting. This sustained engagement fosters meaningful dialogue among educators. As noted by Rock and Wilson (2005), such collaboration enhances both individual and collective assessment practises. Due to its systematic approach, lesson study has been shown to contribute to significant improvements in this area.

Furthermore, the Lesson Study Model's emphasis on student engagement aligns well with contemporary educational paradigms that advocate for learner-centred approaches. By considering students' perspectives in assessment design, teachers can develop tests that are academically rigorous and relevant to learners. As highlighted by Hannah (2019), these assessments are engaging for students. In this study, the focus on student needs played a crucial role in enhancing teachers' motivation to refine their assessment practices, ultimately resulting in more meaningful learning experiences for students.

5.1.4. Discussing the Fourth Research Question

RQ4: *How does this model contribute to test writers' professional development?*

This study provided significant insights into the role of the Lesson Study Model in the professional development of test writers. The results from the thematic analysis revealed six key themes:

- Improvement in pedagogical knowledge
- Professional growth
- Peer learning
- Enhanced assessment literacy
- Professional networking
- Empowerment

The findings of this study align with previous empirical research demonstrating the transformative effect of lesson study, particularly regarding the implementation of collaborative and reflective approaches. For instance, the theme of enhanced pedagogical content knowledge resonates with the work of Chong and Kong (2012), which highlights that lesson study fosters a shared understanding of teaching and the acquisition of new knowledge through collaborative planning and peer teaching. This structured collaboration enables educators to systematically refine their methodologies, as reflected in participants' reports of an improved ability to align assessments with pedagogical objectives.

Additionally, as Flanagan et al. (2024) noted, ongoing participation in lesson study facilitates learning through reflection, which is directly correlated with professional development. The increased confidence and competence of study participants as test writers illustrate the cyclical nature of lesson study, characterized by repetitive cycles of planning, implementation, and constructive criticism that contribute to gradual yet significant skill development—a phenomenon also highlighted by Follmer et al. (2024) in their examination of professional learning communities.

Peer learning has a valid reason in the form of social constructivism which underpins lesson study (Rapp, 2015). It allows investigation collectively with access to more pooled expertise of educators—one of those aspects which Bocala (2015) argues is necessary for the formation of an adaptive teaching strategy. This indeed mirrors how participants shared about employing peers to fill in the missing assessment framework gaps, an example of López and Bernal (2009): cross-collaboration moving beyond isolation. The theme of augmented assessment literacy is also validated through the feedback loops that characterize lesson study.

Themes of professional networking and empowerment are likewise nurtured by lesson study's ability to develop sustainable communities of practice. Yoshida (2012) proposed that designated reflection time in the lesson study, permitting autonomy, led teachers to take charge of their development view supported by participants who described increased agency in their decision-making. Further evidence for the flexibility of the model was seen in how participants adapted collaborative strategies to fit Iraq's rather unique educational landscape, thereby addressing the cultural and contextual difficulties identified by Harle (2008). This flexibility not only helps reduce institutional barriers but, as Zhou and Xu (2017) underscored in their view of teacher identity

development, also helps sustain long-lasting professional relationships. Thus, these converging themes highlight the dual nature of lesson study as a reflective and socially embedded practice. By elevating cyclical improvement and collective problem-solving, it rises above the constraints of traditional professional development to provide a holistic entrance for educators to cope with the complexity of language assessment in the changing academic spaces.

The study clearly highlights a range of benefits associated with the Lesson Study Model. The findings corroborate existing empirical research that provides a more nuanced understanding of effective teacher development. McDonald (2009) and Abd Rahim and Abd Rahim (2020) also address the enhancement of pedagogical knowledge, arguing that lesson study facilitates a deeper understanding of teaching practices. Participants in this study experienced remarkable growth in their pedagogical skills, which enhanced their ability to conduct assessments. Cevik and Muldur (2021) supported this notion, suggesting that the collaborative nature of the process allows teachers to refine their teaching methods. Notably, the iterative cycles inherent in lesson study enabled teachers to critically examine their own critically practices, maintaining a sustained focus on improving their pedagogical knowledge.

Another prominent theme was professional development, which aligns with previous studies demonstrating lesson study's contributions to ongoing professional growth. For example, Taylor et al. (2005) found that lesson study provided teachers with opportunities for reflective practice, a crucial component of professional development. Similarly, Podhorsky and Fisher (2007) asserted that the collaborative environment fostered by lesson study encouraged teachers to become proactive in their own professional development. The researchers involved in this study reported a noticeable enhancement in their confidence and competence as test writers. This indicates that collaborative professional development is a significant pathway for personal and professional advancement.

A recurring theme highlighted by the participants was the significance of peer learning. Liu (2020) emphasized that lesson study promotes collaborative inquiry among teachers, enabling them to learn from each other's experiences. Gardner et al. (2012) also pointed out that peer learning can be a defining feature of an effective professional development model. Participants noted that sharing insights and strategies with colleagues enhanced their understanding and cultivated a supportive learning

environment. The value of such collaboration is further supported by Liu (2020), reinforcing the case for fostering peer interactions in the enhancement of teaching practises.

Another key theme identified was the enhancement of assessment literacy. This finding aligns with the research conducted by Cinkara (2020), which concluded that collaborative activities, such as quiz studies, significantly improved teachers' assessment practices. Respondents felt more capable of designing valid and reliable assessments as a result of their participation in lesson study. This outcome is consistent with the findings of Cevik and Muldur (2021), whose study established that lesson study positively influences teachers' assessment literacy. By engaging in the collaborative evaluation and refinement of assessment tools, teachers gain a clearer understanding of assessment principles, ultimately leading to improved student performance.

Another theme highlighted by the participants involves professional networking, which is especially relevant in today's interconnected educational landscape. DuFresne (2007) described how lesson study introduces a new form of collaboration into teaching, facilitating networking and partnership advantages stemming from this shift. Moreover, McDonald (2009) argued that it encourages resource sharing among teachers. This study showed that participation in lesson study increased participants' professional network connections by providing access to a variety of perspectives and resources. Therefore, the collaborative model of lesson study plays an important role in sustaining longer-lasting relationships among teachers compared to traditional formal processes.

Empowerment emerges as a critical theme, exemplifying the transformative potential of lesson study for educators. Lesson study empowers teachers, as Sami et al. (2020) note, by giving them a voice in their professional development and decision-making. According to Taylor et al. (2005), teacher participation in lesson study grants them agency. The teachers involved in this study felt more confident in the skills needed for test-writing and felt empowered to initiate activities pertinent to their professional lives. This degree of empowerment is vital for fostering a culture of continuous improvement and innovation in educational settings.

Consequently, the findings can be understood in terms of collaborative learning environments and reflective practises. Lesson study inherently increases collaboration among teachers, providing significant professional development. This contrasts with

traditional models that rely on one-off training sessions. Continuous conversations, experience sharing, and collaborative problem-solving characterize lesson study. Enhancing collaborative learning among individuals enriches their knowledge within a community of practice and empowers skill enhancement (Rock & Wilson, 2005). Regular peer interactions contributed to the marked achievements across all themes. The defining feature of lesson study's reflective practices may well be their effectiveness in professional development. Planning, observation, and reflection can occur in a continuous cycle, enabling participants to critically analyse their teaching methods and assessment strategies (Abd Rahim & Abd Rahim, 2020). Such focused reflection allows teachers to identify areas of their practice needing improvement and to develop feasible strategies for enhancing their competencies. The literature supports the integration of reflection with professional development; thus, the positive outcomes in this study can be attributed to the structured reflective practices implemented within the lesson study context.

5.2. Contribution to Literature

This research directly addresses the issues and needs presented in the Statement of the Problem (Chapter One) and the Literature Review (Chapter Two). Although there is consistent evidence in the literature that teachers generally lack adequate assessment literacy and that the opportunities for teacher training in this field are inadequate, there is very little empirical evidence about the development of assessment literacy in practice.

In the context of Iraq, there has been very little if any systematic investigation into the language assessment literacy of test writers and practitioners, and how this may be addressed through collaborative forms of professional development. By addressing this problem in the Lesson Study Model, the present study makes some contributions that contribute to the existing literature.

First, this study adds to the literature by reporting on the development of test writers in a higher education language centre in Iraq that is somewhat under-researched to date. This contextual information is an important contribution, since it shows how assessment practices and issues arise from the local situations that the teachers encounter in their professional practice. In this regard, high teaching workloads, large class sizes, and improper teacher training in language assessment. In foregrounding these “local

proprieties,” the study draws attention to the need to approach assessment research with a context-sensitive lens, rather than relying on universalised principles.

Second, the study also contributes methodologically by demonstrating how the LS can be applied toward test design. Rather than thinking of professional development in terms of one-time training, planning, observation and reflection cycles provided teachers an opportunity to (re-)develop their tests collaboratively in a systematic manner. This process links student responses in classroom-trials to actual test-revision clarifies how participation in LS develops teachers’ assessment literacy. This level of detail has been lacking in many of the professional development articles, which have focused, often uncritically, on the outcomes of PD rather than offering a detailed account of how change occurs.

A further contribution is that the growth process of teachers’ assessment literacy is both technical and reflective. Study participants gained practical skills, including writing more clear rubrics and matching tasks to learning outcomes, as well as developing a more reflective stance toward their own practices. This dual-track development process illustrates that professional learning is not simply about acquiring technical skills, but also about developing critical consciousness and self-reflection.

In sum, these contributions move beyond simply relating the findings to existing literature. They show how the present research addressed global concerns about teachers’ assessment literacy, responded to the local challenges of Iraqi higher education, and advanced the field by demonstrating how collaborative professional development through LS can produce both immediate improvements in test quality and longer-term growth in teachers’ reflective capacities. In this way, the study not only fills the identified gaps but also contributes new perspectives on how LAL can be developed, sustained, and adapted across diverse educational contexts.

5.3. Chapter Summary

This chapter discussed the results and compared them to the literature and described original contributions of the study. It demonstrated how the research supported, expanded, and in some instances contradicted what was previously known about LAL and PD. A key contribution was the contextual expansion of LAL research to Iraqi higher education language centres, where limited resources and institutional constraints shape

assessment practice. The study also contributed by demonstrating the processes through which LS Model develops teachers' LAL, linking classroom observations directly to test revisions and professional growth. Equally important was the uncovering of dual-track development, in that teachers enhanced both their technical abilities to design tests and their reflective abilities to critique practice. The concept of "local proprieties" was also advanced to capture the influence of institutional conditions, class size, and available resources on assessment practices. Collectively, these contributions went beyond the description of findings in relation to literature and provided new knowledge about how, for whom, and under what circumstances LSM can facilitate sustainable LAL development. The chapter ended by laying the grounds for implications, recommendations and conclusions in the last chapter.



CHAPTER VI

CONCLUSION AND RECOMMENDATIONS

6.1. Introduction

This research aimed to investigate how Iraqi EFL teachers, specifically those who have served as test writers at the Language Centre of Salahaddin University, can enhance their language assessment literacy through a collaborative professional development model known as Lesson Study (LS). Recognizing that language assessment is a crucial component of effective teaching, this study addresses prevalent concerns regarding the inadequacy of teacher training in this area. It offers assessment writers opportunities to engage in reflective and collaborative practices through LS, evaluating their current levels of assessment literacy while positioning LS as a meaningful developmental model. Adopting a qualitative case study approach, the research employs interviews, questionnaires, and reflective journals to triangulate data and provide a comprehensive, context-sensitive analysis of the teachers' developmental trajectories.

6.2. Summary of the Study

The introductory chapter established foundational concepts for the research and identified the primary issue: the limited language assessment literacy among Iraqi test writers and the necessity for sustainable professional development. It elaborated on the significance of language assessment literacy (LAL) in English as a Foreign Language (EFL) instruction and presented the research questions to guide the inquiry. The chapter also introduced the Lesson Study Model as a potentially effective intervention, outlining the study's parameters, assumptions, and limitations.

The study also thoroughly investigated the theoretical underpinnings and historical development of language assessment. It defined crucial terms such as assessment, testing, measurement, and evaluation and elaborated on important concepts, including validity, reliability, washback, and practicality. Furthermore, it traced the evolution of LAL and reviewed global perspectives and challenges within teacher education. This section also discusses the origin and function of the Lesson Study Model, framing it as a constructivist, practice-based approach to teacher learning.

The qualitative research design was thoroughly articulated and justified with respect to the chosen case study methodology. A detailed description of the research context—expressly, the Language Centre of Salahaddin University—was provided, along with an explanation of the participant selection process. An in-depth discussion covered the data collection tools employed in the study, which included semi-structured interviews, reflective journals, and questionnaires. Following this, the procedures for data analysis and ethical considerations were explored comprehensively to ensure the study was conducted with transparency and rigor.

The findings were then presented according to the thematic areas corresponding to the research questions. The chapter highlighted that teachers possess limited formal knowledge of language assessment principles and primarily implement outdated traditional practices. However, participation in Language Assessment cycles significantly improved awareness, attitudes, and skills. As a result, teachers increasingly became reflective, collaborative, and intentional in designing valid and communicative assessments.

Additionally, the findings were analysed in the context of relevant literature. It was suggested that LS facilitated a transition from teacher-centred assessment models, which were predominantly test-driven, to more robust learner-centred approaches. This practice promoted reflective teaching, encouraged peer learning, and contextualized the advancement of assessment literacy within genuine teaching environments. The discussion also highlighted institutional and cultural barriers that hindered teachers' engagement with LAL and their professional development.

6.3. An Overview of the Study

This section briefly summarizes each study's aims and corresponding research questions to outline the scope of the research.

6.3.1. Research Question 1

What are the academic and professional needs of teachers who teach at the higher education level at language centres as test writers in terms of language assessment literacy as they perceive?

The focus of this question was to clarify the professional gaps of Iraqi EFL teachers as test writers, particularly concerning their understanding and application of language assessment principles. The findings aimed to identify the competencies that are lacking and determine areas requiring targeted development. The results of this research question can be summarized as follows:

- Teachers reported having no formal training in assessment theory; however, some mentioned concepts such as validity, reliability, washback, and assessment for learning (AfL).
- There was limited awareness regarding the alignment of assessments with the goals of communicative language teaching.
- Teachers expressed the need for training in item writing, rubric design, and performance assessment, particularly in relation to speaking and listening skills.
- Others relied on intuition and personal experience rather than a systematic foundation of knowledge in assessment.

These findings highlight significant academic and professional needs, suggesting the urgent requirement for LAL-based training tailored to the Iraqi EFL context, where test writers must meet established educational standards.

6.3.2. Research Question 2

What is the level of language assessment knowledge of the teachers who teach at higher education level at Language Centres?

The purpose of this research question was to assess teachers' existing knowledge and understanding of language assessment concepts and strategies, as well as their readiness and ability to design fair, valid, and effective language tests. The findings of this research can be summarized as follows:

- The overall level of Language Assessment Literacy (LAL) among participants was found to be low to moderate.
- While most teachers were familiar with multiple-choice items and grammar-focused assessments, their understanding of holistic and communicative assessment methods was limited.
- Key concepts such as formative assessment, consequential validity, and authentic testing were rarely comprehended and underutilized.
- There was a significant gap in teachers' ability to apply theoretical concepts to practical assessment design tasks.

This limited knowledge base highlighted the need for organised, long-term development of LAL; otherwise, teachers may continue to rely on outdated or unbalanced assessment practices that fail to align with modern pedagogical goals.

6.3.3. Research Question 3

To what extent do the teachers as test writers benefit from the 'Lesson Study' model to design language tests?

This inquiry aimed to explore the impact of the Lesson Study Model on teachers' abilities to design more informed and pedagogically sound language assessment tasks. It encompassed practical outcomes such as collaboration in lesson design and reflection throughout that process. The findings from this research can be summarized as follows:

- Teachers gained significant insights from the comprehensive experience of Lesson Study cycles, which enhanced their capacity to create more valid, communicative, and learner-centred assessment tasks.
- The Lesson Study framework facilitated a shared understanding of assessment principles among teachers, who collaboratively negotiated test goal setting, test formats, and evaluation criteria.
- In their test design efforts, participants began to emphasize skills that had previously been overlooked, such as speaking and listening.
- The iterative and reflective nature of Lesson Study enabled participants to revise their assessments based on student feedback and peers' observations.

The Lesson Study Model effectively bridged the gap between assessment theory and practise, establishing a meaningful, human context for English as a Foreign Language (EFL) test development.

6.3.4. Research Question 4

How does this model contribute to test writers' professional development?

The aim of this study was to explore how Lesson Study contributes to teachers' professional development, particularly concerning reflection on practise, pedagogical confidence, and collaborative learning behaviours. The findings regarding this research question can be summarized as follows:

- Lesson Study fostered an identity transformation among teachers, shifting them from mere deliverers of content to reflective practitioners and informed designers of assessments.
- The collaborative nature of the process sustained peer learning, facilitated constructive feedback, and promoted a sense of shared responsibility.
- Teacher participants reported an increase in their motivation and confidence in both classroom teaching and assessment development.
- Reflections recorded in journals provided clear evidence of deeper contemplation regarding student needs, fairness in assessment, and alignment of instruction.

In the current study of Lesson Study, we recognize it as an effective tool for developing assessment literacy and a mechanism for holistic, continuous professional development. It empowers teachers to grow within their teaching roles and to become active agents in their own professional development.

6.4 Implications of the Study

The results of this study hold significant implications for various stakeholders in the field of language education and assessments, particularly concerning EFL teaching in higher education institutions in Iraq and similar contexts.

6.4.1. Implications for Teacher Professional Development

The findings of the study indicate that traditional approaches to teacher professional development (e.g., one-time workshops or authoritative training) did not effectively equip teachers with the necessary skills and knowledge for fair and effective language assessment. In contrast, the Lesson Study (LS) model successfully promoted reflective practice, collaborative learning, and a contextualized application of assessment principles. Therefore, it is highly recommended that:

- LS needs be integrated into a comprehensive framework of ongoing professional development for teachers.
- Professional development activities take place in the actual teaching contexts of the teachers to ensure that theory and practice can mutually enlighten each other.
- Activities are organised around peer mentoring and group reflection as a standard method of bridging learning that extends beyond the training.

6.4.2. Implications for Curriculum Designers and Policy Makers

The study shows that many teachers involved in test writing lack a fundamental understanding of essential assessment constructs. This necessitates interventions at the policy level to ensure that:

- LAL becomes a core component of the teacher education curriculum at both undergraduate and graduate levels.
- Assessment literacy frameworks are created at national or institutional levels to guide training content and evaluation.
- Clear standards are set for test development and evaluation in public and university-level Language Centres.

6.4.3. Implications for Language Assessment Practices

Increasing assessment literacy enhances the quality of design and evaluation. The study revealed that after engaging with LS, teachers created rehabilitative, balanced, and non-discriminatory assessments. Therefore, institutions should:

- Encourage test writers to include all four language skills, particularly the under-assessed skills (i.e., speaking and listening).
- Provide access to sample rubrics, assessment banks, and professional guidance to foster diverse test construction in terms of abilities and applications.
- Promote cross-institution sharing of best practises and review processes for tests, ensuring consistency and fairness.

6.4.4. Implications for Institutional Leadership and Resource Allocation

- The successful implementation of Lesson Study requires time, administrative support, and a culture that values teacher collaboration. Institutional leaders should:
- Allocate time in the teachers' duty schedule for collaborative planning, observation, and reflection.
- Recognize the significance of practitioner-led research and support initiatives that arise within the teaching community.
- Form strategic partnerships with external experts to enhance capacity in areas such as LAL and methodologies for reflective teaching.

6.5. Recommendations for Future Research

The current investigation opens various possibilities for further exploration, particularly in language assessment literacy, professional development, and the Lesson Study Model's application in language teaching fields.

Future research should adopt longitudinal designs to evaluate the sustainability and lasting impact of Lesson Study on teachers' assessment literacy. Tracking participants

over one or more academic years could reveal how deeply the changes are embedded and whether the initial improvements are maintained.

It would also be beneficial to conduct research comparing different institutions across diverse contexts and educational sectors in Iraq. Such studies could illuminate context-specific factors that contribute to the success of Lesson Study while offering insights with broader applicability.

While the present study primarily employed qualitative methods, subsequent investigations could incorporate quantitative measures, such as pre- and post-test assessments of language assessment literacy knowledge and evaluation rubrics, to validate the impact. This approach could explore the relationship between improvements in assessment literacy and student achievement.

Additionally, given that institutional support has been highlighted as a challenge, examining how administrative policies, leadership styles, and institutional culture influence the implementation and outcomes of collaborative professional development models like Lesson Study would be worthwhile.

Future investigations could also consider integrating technology in Lesson Studies, such as video reflections, online collaborative tools, or platforms utilizing AI for feedback, to enhance the model's dissemination and accessibility in resource-limited settings.

Given the significant developmental impact of Lesson Study observed in this study, further research could examine its application in pre-service teacher education programmes. This would help ensure that new teachers enter the profession with a stronger foundation in language assessment literacy, and collaborative learning.

6.6. Researcher's Reflections

The interest in conducting this study is not only academic, but also a personal and professional search. My reading as a scholar encountered many problems, and it was in the process of reading that I was enlightened.

Holding a claim on the time of participants emerged as a key consideration in a crowded academic environment. It required persistence and adaptation to organise the Lesson Study meeting and keep teachers continually engaged. But in spite of these

disappointments, teachers' willingness to collaborate has reaffirmed my belief in the potential of collegial professional development.

A second key reflection is related to my dual role as LS cycles leader, researcher. It was difficult for me to juggle these two roles, as I needed to guide without using my influence to influence my thinking. I also learned to listen, to step backward when it was time to do so, and to let teachers' perspectives guide the shape of conversations and the frame of the rewrites themselves.

It also raised my awareness of the practical hurdles of writing tests in Iraqi higher education. Yet, despite the apparent ease with which research appears to generalize principles of assessment, what I elucidated were "local proprieties": a realization that enacted practice was heavily influenced by local conditions -- the imperative of the institution (what it required or what the students brought to it), the numbers of students for whom there was a need to cater, the kinds of resources that were to hand, and so on. It was those reminders which made me aware of the need that the professional development was not only for conveying theory but also for addressing the local situation.

Last but not least, when studying LAL personally. I learned about theory and practice of testing by reading candidates' papers, visiting classroom trials, and reflecting upon them in a critical manner. The process has also taught me a lot about co-learning or co-generative learning of lessons for teachers and of researchers.

6.7. Chapter Summary

This chapter concluded the thesis by summarising the research aims, questions, methodology, and key findings. It highlighted the study's major contributions, namely the contextualised understanding of LAL in Iraq, the operationalisation of LS Model as a professional development model for test writers, and the demonstration of both technical and reflective growth in teachers' assessment literacy. The implications for theory and practice were discussed, including recommendations for integrating LS Model into teacher professional development programmes and for aligning assessment practices more closely with local realities. The chapter also outlined recommendations for policy-makers and institutions to support sustainable LAL development, as well as directions for future research, such as investigating LAL across different contexts and extending the use

of collaborative models. The thesis ended by reaffirming the importance of enhancing teachers' assessment literacy for improving teaching and learning in higher education.



REFERENCES

- Abd Rahim, E. M., & Abd Rahim, M. E. (2020). Professional Development through Lesson Study: A Practical Model to Refine Teaching, Learning and Teacher Practise Knowledge. *Journal of Advanced Research in Social and Behavioural Sciences*, 19(1), 33-39.
- Akbari, R., & Moradkhani, S. (2010). Iranian English teachers' self-efficacy: Do academic degree and experience make a difference? *Pazhuhesh-e Zabanha-ye Khareji*, 56, 25-47. <https://www.sid.ir/paper/83446/en>
- Akbari, R., & Moradkhani, S. (2012). Is a degree relevant? A comparison of pedagogical thought units of teachers with and without ELT-related academic credentials. *Australian Journal of Teacher Education (Online)*, 37(12), 95-111. <https://ro.ecu.edu.au/ajte/vol37/iss12/8>
- Akiri, E., & Dori, Y. J. (2022). Professional growth of novice and experienced STEM teachers. *Journal of Science Education and Technology*, 31(1), 129-142. <https://doi.org/10.1007/s10956-021-09936-x>
- Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied linguistics*, 14(2), 115-129.
- Alkharusi, H. (2011). Self-perceived assessment skills of pre-service and in-service teachers. *Jurnal Pendidikan Malaysia*, 36(2), 9-17.
- Alkharusi, H. (2011). Teachers' classroom assessment skills: Influence of gender, subject area, and grade level. *Journal of Educational and Psychological Studies*, 5(1), 1-9.
- Alsarimi, A. M. (2001). New trends in assessment in the Sultanate of Oman: Goals and characteristics. *Educational Measurement: Issues and Practise*, 20(4), 27-30.
- Anfara Jr, V. A., Lenski, S. J., & Caskey, M. M. (2009). Using the lesson study approach to plan for student learning. *Middle School Journal*, 40(3), 50-57.
- Arias, C. I., Maturana, L. M., & Restrepo, M. I. (2012). Evaluation in foreign language learning: Towards fair and democratic practises. *Lenguaje*, 40(1), 99-126.
- Arias, G. (2014). *Bilingual language assessment: Contemporary practise versus recommended practise* (Master's thesis). Illinois State University. <https://ir.library.illinoisstate.edu/etd/293/>
- Assessment, A. Glossary of Assessment Terms.
- Babbie, E. R. (2020). *The practise of social research*. Cengage AU.

- Babbie, E. R. (2020). *The practise of social research*. Cengage AU.
- BACHMAN, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly*, 16(4), 449. <https://doi.org/10.2307/3586464>
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practise*. Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practise: Designing and developing useful language tests* (Vol. 1). Oxford University Press.
- Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language testing*, 13(3), 257-279.
- Baricaua Gutierrez, S. (2016). Building a classroom-based professional learning community through Lesson Study: Insights from elementary school science teachers. *Professional Development in Education*, 42(5), 801–817. <https://doi.org/10.1080/19415257.2015.1119709>
- Batool, F., & Malik, S. (2022). Lesson Study Model versus Traditional Methods of Teaching: Effects on Prospective Teachers' Instructional Practises and Students' Achievement. *International Research Journal of Education and Innovation*, 3(1), 214-225.
- Berger, A. (2023). A Difficulty-Informed Approach to Developing Language Assessment Literacy for Classroom Purposes. *Chinese Journal of Applied Linguistics*, 46(2), 289-307.
- Berry, V., Sheehan, S., & Munro, S. (2019a). What does language assessment literacy mean to teachers? *ELT Journal*, 73(2), 113–123. <https://doi.org/10.1093/elt/ccy055>
- Berry, V., Sheehan, S., & Munro, S. (2019b). What does language assessment literacy mean to teachers? *ELT Journal*, 73(2), 113–123. <https://doi.org/10.1093/elt/ccy055>
- Birgin, O., & Baki, A. (2009). An investigation of primary school teachers' proficiency perceptions about measurement and assessment methods: The case of Turkey. *Procedia-Social and Behavioural Sciences*, 1(1), 681-685.
- Bocala, C. (2015). From experience to expertise: The development of teachers' learning in lesson study. *Journal of teacher education*, 66(4), 349-362.

- Boud, D., Keogh, R., & Walker, D. (2013). Promoting reflection in learning a model. In *Boundaries of adult learning* (pp. 32-56). Routledge.
- Boud, D., Keogh, R., & Walker, D. (2013). Promoting reflection in learning a model. In *Boundaries of adult learning* (pp. 32-56). Routledge.
- Boyacı, Z., & Erdamar, G. (2023). Lesson Study's Effect on Math Teacher Candidates' Attitudes, Self-Efficacy, and Teaching Anxiety. *Asian Journal of Instruction (E-AJI)*, 11(2), 23-42.
- Bozkurt, E., & Özdemir, İ. E. Y. (2018). A Collaborative Approach in Professional Development: Lesson Study. *Kastamonu Education Journal*, 26(1), 109-116.
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101.
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101.
- Brookhart, S. M. (1998). Teaching about Grading and Communicating Assessment Results.
- Brookhart, S. M. (1999). Teaching about communicating assessment results and grading. *Educational Measurement: Issues and Practise*, 18(1), 5-13.
- Brown, A. L. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms. *Metacognition, motivation, and understanding*, 65-116.
- Brown, C., Taylor, C., & Ponambalum, L. (2016). Using design-based research to improve the lesson study approach to professional development in Camden (London). *London Review of Education*, 14(2).
- BROWN, H. D. (2004). *Language assessment*. LONGMAN.
- Brown, H. D. (2004). *Language assessment: Principles and classroom practises*. Pearson Education.
- Brown, H. D., & Abeywickrama, P. (2004). *Language assessment. Principles and Classroom Practises*. White Plains, NY: Pearson Education, 20.
- Brown, J. D. (2013). Research on computers in language testing: Past, present and future. *Contemporary computer-assisted language learning*, 73-94.
- Brown, S., & Race, P. (2012). Using effective assessment to promote learning. In *University teaching in focus* (pp. 74-91). Routledge.
- Bryman, A. (2016). *Social research methods*. Oxford university press.
- Bryman, A. (2016). *Social research methods*. Oxford university press.

- Campbell, C., Murphy, J. A., & Holt, J. K. (2002, October). Psychometric analysis of an assessment literacy instrument: Applicability to preservice teachers. In *Annual meeting of the mid-western educational research association, Columbus, OH*.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied linguistics*, 1(1), 1-47.
- CANALE, M., & SWAIN, M. (1980). Theoretical bases of communicative approaches to Second language teaching and testing. *Applied Linguistics*, 1(1), 1-47. <https://doi.org/10.1093/applin/i.1.1>
- Carr, M., Jones, C., & Lee, W. (2005). Beyond listening: can assessment practise play a part? In *Beyond listening* (pp. 129-150). Policy Press. <https://bristoluniversitypressdigital.com/abstract/book/9781447342403/ch008.xml>
- Carrier, S. J. (2011). Implementing and integrating effective teaching strategies including features of lesson study in an elementary science methods course. *The Teacher Educator*, 46(2), 145-160.
- Carter, K., Cushing, K., Sabers, D., Stein, P., & Berliner, D. (1988). Expert-novice differences in perceiving and processing visual classroom information. *Journal of Teacher Education*, 39(3), 25-31. <https://doi.org/10.1177/002248718803900306>
- Cerbin, W., & Kopp, B. (2006). Lesson study as a model for building pedagogical knowledge and improving teaching. *International journal of teaching and learning in higher education*, 18(3), 250-257.
- Çetin, K. (2024). *Collaboration and Reflection in English Language Teaching Practicum: Insights from the Practicum Lesson Study Model* [PhD - Doctoral Programme]. Middle East Technical University. <https://open.metu.edu.tr/handle/11511/109530>
- Cevik, A., & Muldur, M. (2021). A model trial for the "Teaching Practise" course within Turkish teaching programmes: Lesson study. *Journal of Language and Linguistic Studies*, 17(S1), 403-422.
- Chalhoub-Deville, M. (2001). Language testing and technology: Past and future. *Language Learning & Technology*, 5(2), 95-98.
- Chassels, C., & Melville, W. (2009). Collaborative, reflective and iterative Japanese lesson study in an initial teacher education programme: Benefits and challenges. *Canadian Journal of Education/Revue canadienne de l'éducation*, 32(4), 734-763.

- Cheng, E. C., & Chan, J. K. (2021). *Developing metacognitive teaching strategies through lesson study*. Springer Singapore.
- Chokshi, S., & Fernandez, C. (2004). Challenges to importing Japanese lesson study: Concerns, misconceptions, and nuances. *Phi Delta Kappan*, 85(7), 520–525. <https://doi.org/10.1177/003172170408500710>
- Chong, W. H., & Kong, C. A. (2012). Teacher collaborative learning and teacher self-efficacy: The case of lesson study. *The journal of experimental education*, 80(3), 263-283.
- Cinkara, E. (2020). Quiz Study as a Professional Development Activity for Tertiary-Level EFL Test Writers. *TESL-EJ*, 24(3), n3.
- Cinkara, E. (2020). Quiz Study as a Professional Development Activity for Tertiary-Level EFL Test Writers. *TESL-EJ*, 24(3), n3.
- Clapham, C. (2000). Assessment and testing. *Annual Review of Applied Linguistics*, 20, 147-161.
- Cohen, A. D. (1994). Assessing language ability in the classroom.
- Cohen, A. D. (2001). Second language assessment. *Teaching English as second*.
- Coombe, C. A., Davidson, P., O'Sullivan, B., & Stoyanoff, S. (2013). *The Cambridge Guide to Second Language Assessment*. Cambridge University Press.
- Coombe, C., Troudi, S., & Al-Hamly, M. (2012). Foreign and second language teacher assessment literacy: Issues, challenges, and recommendations. *The Cambridge guide to second language assessment*, 1, 20-29.
- Coombe, C., Vafadar, H., & Mohebbi, H. (2020). Language assessment literacy: What do we need to learn, unlearn, and relearn? *Language Testing in Asia*, 10(1), 3. <https://doi.org/10.1186/s40468-020-00101-6>
- Crusan, D., Plakans, L., & Gebril, A. (2016). Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practises. *Assessing writing*, 28, 43-56. <https://doi.org/10.1016/j.system.2019.01.006>
- Dadds, M. (1997). Continuing professional development: nurturing the expert within. *British journal of in-service education*, 23(1), 31-38.
- Dania, A., Voutsina, I., & Moumouri, A. (2013). The Application of the Lesson Study Model on the teaching practise of physical education university students. In *Proccedings of the International Scientific Conference eRA-8. Athens, Greece*.
- Davies, A. (2008). Textbook trends in teaching language testing. *Language testing*, 25(3), 327-347.

- Deluca, C., & Klinger, D. A. (2010). Assessment literacy development: Identifying gaps in teacher candidates' learning. *Assessment in Education: Principles, Policy & Practise*, 17(4), 419-438. doi:10.1080/0969594x.2010.516643
- Deneen, C. C., & Brown, G. T. (2016). The impact of conceptions of assessment on assessment literacy in a teacher education programme. *Cogent Education*, 3(1), 1225380.
- Denzin, N. K., & Lincoln, Y. S. (2000). The future of qualitative research. *Handbook of Qualitative Research* (pp.1018-1023). Sage.
- Denzin, N. K., & Lincoln, Y. S. (2000). The future of qualitative research. *Handbook of Qualitative Research* (pp.1018-1023). Sage.
- Dewey, J. (2012). *How we think*. Martino Publishing.
- Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method*. John Wiley & Sons.
- Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method*. John Wiley & Sons.
- Doig, B., Groves, S., & Fujii, T. (2011). The critical role of Task Development in Lesson Study. *Lesson Study Research and Practise in Mathematics Education*, 181–199. https://doi.org/10.1007/978-90-481-9941-9_15
- Doig, B., Groves, S., & Fujii, T. (2011). The critical role of task development in lesson study. *Lesson study research and practise in mathematics education: Learning together*, 181-199.
- Dudley, P. (2014). *Lesson study: A handbook*. LS Developments. <https://lessonstudy.co.uk/wp-content/uploads/2012/03/new-handbook-revisedMay14.pdf>
- Dudley, P. (2014). *Lesson study: A handbook*. LS Developments. <https://lessonstudy.co.uk/wp-content/uploads/2012/03/new-handbook-revisedMay14.pdf>
- Dudley, P., & Vrikki, M. (2019). Teachers' collaborative dialogues in contexts of lesson study. In *The Routledge international handbook of research on dialogic education* (pp. 217-226). Routledge.
- Duff, P. (2018). *Case study research in applied linguistics*. Routledge.
- Duff, P. A. (2014). Case study research on language learning and use. *Annual Review of Applied Linguistics*, 34, 233-255.

- DuFour, R., & Eaker, R. (1998). *Professional learning communities at work: Best practises for enhancing student achievement*. Solution Tree Press.
- DuFour, R., & Eaker, R. (1998). *Professional learning communities at work: Best practises for enhancing student achievement*. Solution Tree Press.
- DuFresne, C. R. (2007). *Using the Lesson Study Model of professional development to enhance teacher collaboration* (Doctoral dissertation, Walden University).
- Eames, C. (2003). Researching in cooperative education: How a practitioner met the challenge. In *Handbook for research in cooperative education and internships* (pp. 71-94). Routledge.
- Estaji, M., & Ghiasvand, F. (2022). Classroom supervision and professionalism: Matches and mismatches in the perceptions of novice and experienced teachers. *Applied Research on English Language*, 11(3), 1-36.
<https://doi.org/10.22108/are.2022.131707.1817>
- Ewell, P. T., & Cumming, T. (2017). History and conceptual basis of assessment in higher education. In *Enhancing Assessment in Higher Education* (pp. 3-26). Routledge.
- Eyal, L. (2012). Digital assessment literacy—The core role of the teacher in a digital environment. *Journal of Educational Technology & Society*, 15(2), 37-49.
- Falbe, K. N., & Seglem, R. (2023). Teaching Is Messy: Using Lesson Study to Reimagine Student-Centred Clinical Experiences. *Education Sciences*, 13(7), 735.
- Falsgraf, C. (2005). Why a national assessment summit. *New visions in action. National Assessment Summit*, 6.
- Farhady, H. (2012). Principles of language assessment. *The Cambridge guide to second language assessment*, 37-46.
- Farhady, H. (2018). History of language testing and assessment. *The TESOL Encyclopedia of English Language Teaching*, 1-7.
<https://doi.org/10.1002/9781118784235.eelt0343>
- Farhady, H., & Tavassoli, K. (2018). Developing a language assessment knowledge test for EFL teachers: A data-driven approach. *Iranian Journal of Language Teaching Research*, 6(3), 79-94.
- Farrall, M. L. (2012). *Reading assessment: Linking language, literacy, and cognition*. John Wiley & Sons.
- Fernandez, C., & Chokshi, S. (2002). A practical guide to translating lesson study for a US setting. *Phi Delta Kappan*, 84(2), 128-134.

- Fernandez, C., & Yoshida, M. (2004). *Lesson study: A Japanese approach to improving mathematics teaching and learning*. Lawrence Erlbaum Associates.
- Fernandez, C., & Yoshida, M. (2004). *Lesson study: A Japanese approach to improving mathematics teaching and learning*. Lawrence Erlbaum Associates.
- Fernandez, M. L. (2010). Investigating how and what prospective teachers learn through microteaching lesson study. *Teaching and teacher education*, 26(2), 351-362.
- Fitriyah, I., Masitoh, F., & Widiati, U. (2022). Classroom-based language assessment literacy and professional development need between novice and experienced EFL teachers. *Indonesian Journal of Applied Linguistics*, 12(1), 124-134. <https://doi.org/10.17509/ijal.v12i1.46539>
- Flanagan, B., Hourigan, M., & Leavy, A. (2024). Implementation of Lesson Study in the context of STEM education in a rural setting in Ireland: challenges and opportunities. *International Journal for Lesson & Learning Studies*, 13(2), 57-70.
- Follmer, D. J., Groth, R., Bergner, J., & Weaver, S. (2024). Theory-based evaluation of lesson study professional development: Challenges, opportunities, and lessons learned. *American Journal of Evaluation*, 45(2), 292-312.
- Fowler Jr, F. J. (2013). *Survey research methods*. Sage publications.
- Fowler Jr, F. J. (2013). *Survey research methods*. Sage publications.
- Fulcher, G. (2000). The 'communicative' legacy in language testing. *System*, 28(4), 483–497. [https://doi.org/10.1016/s0346-251x\(00\)00033-6](https://doi.org/10.1016/s0346-251x(00)00033-6)
- Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly*, 9(2), 113–132. <https://doi.org/10.1080/15434303.2011.642041>
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. London and New York: Routledge.
- Gardner, J., Galanouli, D., Devlin, G., Magee, M., Mc Sweeney, K., McHenry, M., ... & Mitchell, S. (2012). Exploring Japanese lesson study as a model of peer-to-peer professional learning. In *Standing Conference on Teacher Education, North and South* (pp. 3-24).
- Gatbonton, E. (2008). Looking beyond teachers' classroom behaviour: Novice and experienced ESL teachers' pedagogical knowledge. *Language Teaching Research*, 12(2), 161-182. <https://doi.org/10.1177/1362168807086286>

- Genç, E., Çalışkan, H., & Yüksel, D. (2020). Language assessment literacy level of EFL teachers: A focus on writing and speaking assessment knowledge of the teachers. *Sakarya University Journal of Education*, 10(2), 274-291.
- Genç, E., Çalışkan, H., & Yüksel, D. (2020). Language assessment literacy level of EFL teachers: A focus on writing and speaking assessment knowledge of the teachers. *Sakarya University Journal of Education*, 10(2), 274-291. <https://dergipark.org.tr/en/pub/suje/issue/56193/626156>
- Gerring, J. (2004). What is a case study and what is it good for?. *American Political Science Review*, 98(2), 341-354.
- Giannakidou, E., Gioftsali, K., & Tzioras, E. (2013). The reflective action of prospective teachers when implementing an applied version of the Lesson Study Model. *Hellenic Journal of Research in Education*, 1, 30-58.
- Gottheiner, D. M., & Siegel, M. A. (2012). Experienced middle school science teachers' assessment literacy: Investigating knowledge of students' conceptions in genetics and ways to shape instruction. *Journal of Science Teacher Education*, 23, 531-557.
- Gronlund, G. (1998). Portfolios as an assessment tool: Is collection of work enough. *Implementation Programme: Teacher Guide*, 266.
- Gutiérrez Rojas, A. M. (2022). Understanding listening assessment in a Colombian EFL context. <https://manglar.uninorte.edu.co/bitstream/handle/10584/11471/1095793324.pdf?sequence=1&isAllowed=y>
- Hakim, B. (2015). English language teachers' ideology of ELT assessment literacy. *International Journal of Education and Literacy Studies*, 3(4), 42-48.
- Hancock, D. R., Algozzine, B., & Lim, J. H. (2021). *Doing case study research: A practical guide for beginning researchers*. Teachers College Press.
- Hannah, M. (2019). Improving professional practise through" lesson study". In *EDULEARN19 Proceedings* (pp. 1940-1946). IATED.
- Harle, C. B. (2008). *The lesson study professional development process: Exploring the learning experiences of elementary and middle school teachers*. The University of Texas at Austin.
- Harris, D. P. (1969). *Testing English as a Second Language*.
- Heaton, J. B. (2011). *Writing English language tests* (3rd ed.). UK: Longman Group.

- Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Addison-Wesley. <https://doi.org/10.3390/su13147665>
- Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Addison-Wesley. <https://doi.org/10.3390/su13147665>
- Hord, S. M. (1997). *Professional learning communities: Communities of continuous inquiry and improvement*. Southwest Educational Development Laboratory.
- Hord, S. M. (1997). *Professional learning communities: Communities of continuous inquiry and improvement*. Southwest Educational Development Laboratory.
- Howell, J. B., & Saye, J. W. (2016). Using lesson study to develop a shared professional teaching knowledge culture among 4th grade social studies teachers. *The journal of social studies research*, 40(1), 25-37.
- Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9), 1277–1288. <https://doi.org/10.1177/1049732305276687>
- Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9), 1277–1288. <https://doi.org/10.1177/1049732305276687>
- Hughes, A. (2020). *Testing for language teachers*. Cambridge university press.
- Huhta, A., Hirvalä, T., & Banerjee, J. (2005). European survey of language testing and assessment needs. Part 2: Regional findings.
- Hulin, C., Netemeyer, R., & Cudeck, R. (2001). Can a reliability coefficient be too high? *Journal of Consumer Psychology*, 10 (1), 55-58.
- Iksan, Z. H., Zakaria, E., & Daud, M. Y. (2014). Model of lesson study approach during micro teaching. *International education studies*, 7(13), 253-260.
- Impara, J. C., Plake, B. S., & Fager, J. J. (1993). Teachers' assessment background and attitudes toward testing. *Theory into practise*, 32(2), 113-117.
- Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language testing*, 25(3), 385-402.
- Inbar-Lourie, O. (2017). Language assessment literacy. *Language testing and assessment*, 257-270.
- Jeong, H. (2014). A comparative study of scores on computer-based tests and paper-based tests. *Behaviour & Information Technology*, 33(4), 410-422.
- Johnson, K. E. (2009). *Second language teacher education: A sociocultural perspective*. Routledge.

- Kanellopoulou, E. M., & Darra, M. (2019). The Implementation of the Lesson Study in Basic Teacher Education: A Research Review. *Higher Education Studies*, 9(3), 65-78.
- Karimi, M. N., & Asadnia, F. (2022). Variations in novice and experienced L2 teachers' pedagogical cognitions and the associated antecedents in tertiary-level online instructional contexts. *Language Awareness*, 1-20. <https://doi.org/10.1080/09658416.2022.2033251>
- Kavakli, N., & Arslan, S. (2019). Towards a Continuum from Know-How to Show-How for Developing EFL Student-Teachers' Assessment Literacy. *International Online Journal of Education and Teaching*, 6(1), 223-232.
- Kolb, D. A. (2014). *Experiential learning: Experience as the source of learning and development*. FT press.
- Kolb, D. A. (2014). *Experiential learning: Experience as the source of learning and development*. FT press.
- Kunnan, A. J. (Ed.). (2000). *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (No. 9). Cambridge University Press.
- Kvale, S., & Brinkmann, S. (2009). *Interviews: Learning the craft of qualitative research interviewing*. Sage.
- Kvale, S., & Brinkmann, S. (2009). *Interviews: Learning the craft of qualitative research interviewing*. Sage.
- Lado, R. (1961). *Language testing; the construction and use of foreign language tests: A teacher's book*. New York: McGraw-Hill.
- Lam, R. (2019). Teacher assessment literacy: Surveying knowledge, conceptions and practises of classroom-based writing assessment in Hong Kong. *System*, 81, 78-89. <https://doi.org/10.1016/j.system.2019.01.006>
- Lamb, P., & King, G. (2021). Developing the practise of pre-service physical education teachers through a dyad model of lesson study. *European Physical Education Review*, 27(4), 944-960.
- Lewis, C. (1997). Planned educational change in Japan: The shift to student-centred elementary science. *Journal of Educational Policy*, 12(5), 313-331.
- Lewis, C. (2002). *Lesson study: A handbook of teacher-led instructional change*. Research for Better Schools.

- Lewis, C. (2002). *Lesson study: A handbook of teacher-led instructional change*. Research for Better Schools.
- Lewis, C. C., & Hurd, J. (2011). Lesson study step by step: How teacher learning communities improve instruction. (*No Title*).
- Lewis, C. C., & Tsuchida, I. (1999). A lesson is like a swiftly flowing river: How research lessons improve Japanese education. *Improving Schools*, 2(1), 48–56. <https://doi.org/10.1177/136548029900200117>
- Liu, S. H. (2020). Employing Lesson Study Model to Collaboratively Investigate the Use of Student Peer-discussion for Teaching: A Case Study on an Elementary School. *Chung Cheng Educational Studies*, 19(2), 1-42.
- López Mendoza, A. A., & Bernal Arandia, R. (2009). Language testing in Colombia: A call for more teacher education and teacher training in language assessment. *Profile Issues in Teachers Professional Development*, 11(2), 55-70.
- Mahmood, S., Hussain, H. B., & Mahmood, S. (2021). Reading Assessment Awareness: A Comparative Analysis of Public and Private School ESL Teachers in Pakistan. *Pakistan Languages and Humanities Review*, 5(1), 12-22. <https://plhr.org.pk/issues/v5/1/reading-assessment-awareness-a-comparative-analysis-of-public-and-private-school-esl-teachers-in-pakistan.pdf>
- Mahmoodarabi, M., & Khodabakhsh, M. R. (2015). Critical Pedagogy: EFL Teachers' Views, Experience and Academic Degrees. *English Language Teaching*, 8(6), 100-110. <https://eric.ed.gov/?id=EJ1075436>
- Mao, A. (2022). Literature review of language testing theories and approaches. *OALib*, 09(05), 1–5. <https://doi.org/10.4236/oalib.1108741>
- Mao, A. (2022). Literature review of language testing theories and approaches. *Open Access Library Journal*, 9(5), 1-5.
- McDonald, S. E. (2009). *A model of teacher professional development based on the principles of lesson study* (Doctoral dissertation, Queensland University of Technology).
- McDowell, A. V. (2010). *Preservice teachers' use of lesson study in teaching nature of science*. Georgia State University.
- McMahon, M. T., & Hines, E. (2008). Lesson study with preservice teachers. *The Mathematics Teacher*, 102(3), 186-191.
- McNamara, T. F. (2000). *Language testing*. Oxford University Press.

- Mertler, A. C. (2003). Secondary teachers' assessment literacy: *Does classroom experience make a difference? American Secondary Education*, 33(1), 49-64.
- Mertler, C. A. (2004). Secondary teachers' assessment literacy: Does classroom experience make a difference?. *American secondary education*, 49-64.
- Mertler, C. A., & Campbell, C. (2005). Measuring Teachers' Knowledge & Application of Classroom Assessment Concepts: Development of the "Assessment Literacy Inventory". *Online Submission*.
- Messick, S. (1996). Validity and washback in language testing. *Language testing*, 13(3), 241-256.
- Mihai, F. (2010). *Assessing English language learners in the content areas: A research-into-practise guide for educators*. University of Michigan Press.
- Miller, K., & Zhou, X. (2014). Learning from classroom video: What makes it compelling and what makes it hard. In *Video research in the learning sciences* (pp. 321-334). Routledge.
- Mitchell, D. L. (2017). *The influence of lesson study on teacher practise: A case study* (Doctoral dissertation, California State Polytechnic University, Pomona).
- Mohamed, S., Kamis, A., & Ali, N. (2016). Gauging the assessment literacy of Malaysia's home economics teachers: An empirical study. *Malaysian Journal of Society and Space*, 12(3), 130-138.
- Moon, J. A. (2006). *Learning journals: A handbook for reflective practise and professional development*. Routledge.
- Moon, J. A. (2006). *Learning journals: A handbook for reflective practise and professional development*. Routledge.
- Mousavi, S. J., Mobini, B., Mehdian, H., Akbarnia, B., Bouzari, B., Askary-Ashtiani, A., ... & Parnianpour, M. (2010). Reliability and validity of the Persian version of the Scoliosis Research Society-22r questionnaire. *Spine*, 35(7), 784-789.
- Nami, F., Marandi, S. S., & Sotoudehnama, E. (2016). CALL teacher professional growth through lesson study practise: an investigation into EFL teachers' perceptions. *Computer Assisted Language Learning*, 29(4), 658-682.
- Nassaji, H. (2020). Good qualitative research. *Language Teaching Research*, 24(4), 427-431. <https://doi.org/10.1177/1362168820941288>
- Nassaji, H. (2020). Good qualitative research. *Language Teaching Research*, 24(4), 427-431. <https://doi.org/10.1177/1362168820941288>

- Newfields, T. (2006, May). Teacher development and assessment literacy. In *Authentic communication: Proceedings of the 5th Annual jalt Pan-sig Conference* (pp. 48-73).
- Ng, W. S., Xie, H., Wang, F. L., & Li, T. (2020). Peer assessment of peer assessment plan: a deep learning approach of teacher assessment literacy. *International Journal of Innovation and Learning*, 27(4), 450-466.
- North, B. (2005, May). The CEFR levels and descriptor scales. In *Multilingualism and assessment: Achieving transparency, assuring quality, sustaining diversity. Proceedings of the ALTE Berlin Conference* (pp. 21-66).
- Norwich, B., & Ylonen, A. (2015). A design-based trial of Lesson Study for assessment purposes: evaluating a new classroom based dynamic assessment approach. *European Journal of Special Needs Education*, 30(2), 253-273.
- Norwich, B., Fujita, T., Adlam, A., Milton, F., & Edwards-Jones, A. (2018). Lesson study: an inter-professional collaboration approach for Educational Psychologists to improve teaching and learning. *Educational Psychology in Practise*, 34(4), 370-385.
- Nunan, D. (1988). *Syllabus design*. Oxford University Press.
- Obeidat, M. M., & Shehadeh, M. M. (2020). Graduate Students' Perspective about the MA TEFL Programme at Hashemite University in Jordan. *International Journal of Higher Education*, 9(3), 28-38.
- Oller Jr, J. W. (1983). *Issues in language testing research*. Newbury House Publishers, Inc., Rowley, MA 01969.
- Oman, J. A. (2023). English as a foreign language teachers' digital assessment literacy in higher education: A self-perception study. *Cogent Education*, 10(1), 2239535. <https://doi.org/10.1080/2331186X.2023.2239535>
- Ono, Y., & Ferreira, J. (2010). A case study of continuing teacher professional development through lesson study in South Africa. *South African journal of education*, 30(1).
- Ono, Y., & Ferreira, J. (2010a). A case study of continuing teacher professional development through Lesson Study in South Africa. *South African Journal of Education*, 30(1), 59–74. <https://doi.org/10.15700/saje.v30n1a320>
- Ono, Y., & Ferreira, J. (2010b). A case study of continuing teacher professional development through Lesson Study in South Africa. *South African Journal of Education*, 30(1), 59–74. <https://doi.org/10.15700/saje.v30n1a320>

- Öz, S., & Atay, D. (2017). Turkish EFL instructors' in-class language assessment literacy: Perceptions and practises. *ELT Research Journal*, 6(1), 25-44.
- Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.
- Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.
- Podhorsky, C., & Fisher, D. (2007). Lesson study: An opportunity for teacher led professional development. In *Handbook of teacher education* (pp. 445-456). Dordrecht: Springer Netherlands.
- Popham, J. W. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory Into Practise*, 48, 4-11.
- Popham, W. J. (2006, September). Defining and enhancing formative assessment. In *Annual Large-Scale Assessment Conference, Council of Chief State School Officers, San Francisco, CA*.
- Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory into practise*, 48(1), 4-11.
- Rapp, M. J. (2015). Social constructivism as the theoretical underpinning for the employment of collaborative approaches such as lesson study towards teacher development. *International Journal of Social and Cultural Studies*, 8, 17-34.
- Rasyid, M. N. A., Rauf, M., Junaid, S., Nur, S., & Syukur, H. (2023). Assessing Listening Comprehension Skills in Indonesian Islamic Higher Education EFL Classrooms: Current Practises, Challenges and Solutions. *Indonesian TESOL Journal*, 5(2), 321-338. <https://doi.org/10.24256/itj.v5i2.4187>
- Razavipour, K. (2013). Assessing assessment literacy: Insights from a high-stakes test. *Journal of Research in Applied Linguistics*, 4(1), 111-131.
- Rekalidou, G., Karadimitriou, K., & Moumoulidou, M. (2014). Implementation of Lesson Study with students. Collaboration, reflection and feedback. *Hellenic Journal of Research in Education, "Improving the Education of Future Teachers in Crisis of Institutions: Proposals, Applications*, 7-28.
- Rock, T. C., & Wilson, C. (2005). Improving teaching through lesson study. *Teacher education quarterly*, 32(1), 77-92.
- Rogers, W. T., Cheng, L., & Hu, H. (2007). ESL/EFL instructors' beliefs about assessment and evaluation. *Canadian and International Education*, 36(1), 39-61.

- Rubin, H. J., & Rubin, I. S. (2011). *Qualitative interviewing: The art of hearing data*. sage.
- Rubin, H. J., & Rubin, I. S. (2011). *Qualitative interviewing: The art of hearing data*. sage.
- Rushton, I., & Suter, M. (2012). *Reflective practise for teaching in lifelong learning: n/a*. McGraw-Hill Education (UK).
- Şahin, Ş., & Kılıç, A. (2020). Evaluation of the teaching practise course carried out with the Lesson Study Model. *Tuning Journal for Higher Education*, 8(1), 99-127.
- Saito, E., Hawe, P., Hadiprawiroc, S., & Empedhe, S. (2008). Initiating education reform through lesson study at a university in Indonesia. *Educational Action Research*, 16(3), 391-406.
- Sami, R., Piri, M., & Talebi, B. (2020). Modelling the Empowerment of Primary School's Teachers based on Lesson Study Approach. *Educational and Scholastic studies*, 9(3), 223-251.
- Sandholtz, J. H., & Ringstaff, C. (2016). The influence of contextual factors on the sustainability of professional development outcomes. *Journal of Science Teacher Education*, 27(2), 205-226.
- Sandholtz, J. H., & Ringstaff, C. (2016). The influence of contextual factors on the sustainability of professional development outcomes. *Journal of Science Teacher Education*, 27(2), 205-226.
- Sato, M., Wei, R. C., & Darling-Hammond, L. (2008). Improving teachers' assessment practises through professional development: The case of National Board Certification. *American Educational Research Journal*, 45(3), 669-700.
- Saussure, F. de, Bally, C., & Baskin, W. (1961). *Course in General Linguistics; edited by Charles Bally and Albert Sechehaye in collaboration with Albert Reidlinger; translated from the French by Wade Baskin*. P. Owen.
- Scarino, A. (2013). Language assessment literacy as self-awareness: Understanding the role of interpretation in assessment and in teacher learning. *Language Testing*, 30(3), 309-327.
- Schon, D. (1983). donald schon (schön): learning, reflection and change. Accessed April, 11, 2004.
- Schön, D. A. (1987). *Educating the reflective practitioner: Toward a new design for teaching and learning in the professions*. Jossey-Bass.

- Schön, D. A. (1987). *Educating the reflective practitioner: Toward a new design for teaching and learning in the professions*. Jossey-Bass.
- Schon, D. A. (2000). *Reflective practitioner: How professionals think in action*. BasicBooks.
- Seidman, I. (2006). *Interviewing as qualitative research: A guide for researchers in education and the social sciences*. Teachers' college press.
- Seidman, I. (2006). *Interviewing as qualitative research: A guide for researchers in education and the social sciences*. Teachers' college press.
- Seifalian, M., & Derakhshan, A. (2018). The relationship between Iranian EFL teachers' burnout and self-efficacy across English-related vs. non-English-related academic degrees. *International Journal of English Language & Translation Studies*, 6(2), 99-110.
- Setyaningsih, S. (2018, March). Improving the quality of learning in science through optimization of lesson study for learning community. In *IOP Conference Series: Materials Science and Engineering* (Vol. 332, No. 1, p. 012043). IOP Publishing.
- Shahzamani, M., & Tahririan, M. H. (2021). Iranian medical ESP practitioners' reading comprehension assessment literacy: Perceptions and practises. *Iranian Journal of English for Academic Purposes*, 10(1), 1-15.
https://journalscmu.sinaweb.net/article_122276.html?lang=en
- Sheehan, S., & Munro, S. (2017). Assessment: attitudes, practises, needs.
- Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational researcher*, 15(2), 4-14.
- Siegel, M. A., & Wissehr, C. (2011). Preparing for the plunge: Preservice teachers' assessment literacy. *Journal of Science Teacher Education*, 22, 371-391.
- Singh, C. K. S., Singh, H. K. J., Singh, T. S. M., Moneyam, S., Abdullah, N. Y., & Zaini, M. F. (2022). ESL Teachers' Assessment Literacy in Classroom: A review of past studies. *Journal of Language and Linguistic Studies*, 18.
- Skinner, B. F. (1974). *About behaviourism*. Knopf.
- Skinner, B. F. (1974). *About Behaviourism: By BF Skinner*. Knopf.
- Soodmand Afshar, H., & Farahani, M. (2018). Inhibitors to EFL teachers' reflective teaching and EFL learners' reflective thinking and the role of teaching experience and academic degree in reflection perception. *Reflective Practise*, 19(1), 46-67.
<https://doi.org/10.1080/14623943.2017.1351353>

- Spolsky, B. (1978). Linguistics and language testers. In B. Spolsky (Ed.), *Papers in applied linguistics: Advances in language testing*, 2 (pp. v–x). Arlington, VA: The Centre for Applied Linguistics.
- Stabler-Havener, M. L. (2018). Defining, conceptualizing, problematizing, and assessing language teacher assessment literacy. *Studies in Applied Linguistics and TESOL*, 18(1).
- Standards for teacher competence in Educational Assessment of Students. (1990). *Educational Measurement: Issues and Practise*, 9(4), 30–30. <https://doi.org/10.1111/j.1745-3992.1990.tb00391.x>
- Starkey, L., Yates, A., Meyer, L. H., Hall, C., Taylor, M., Stevens, S., & Toia, R. (2009). Professional development design: Embedding educational reform in New Zealand. *Teaching and teacher education*, 25(1), 181-189.
- Stiggins, R. (1991). Assessment Literacy. *The Phi Delta Kappan* 72(7), 534 – 539.
- Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan*, 77(3), 238.
- Stiggins, R. J. (1999). Are You Assessment Literate? *High School Magazine*, 6(5), 20-23.
- Stiggins, R. J. (2001). The unfulfilled promise of classroom assessment. *Educational Measurement: Issues and Practise*, 20(3), 5-15.
- Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan*, 83(10), 758-765.
- Stigler, J. W., & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom*. Free Press.
- Stigler, J. W., & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom*. Free Press.
- Stobart, G. (2006). The validity of formative assessment. *Assessment and learning*, 133-146.
- Subadi, T., Khotimah, R. P., & Sutarni, S. (2013). A lesson study as a development model of professional teachers. *International Journal of Education*, 5(2), 102.
- Sun, Q., & Zhang, L. J. (2022). Understanding novice and experienced teachers' cognitions and practises for sustainable teacher development: The case of form-focused instruction in English language teaching. *Sustainability*, 14(8), 4711. <https://doi.org/10.3390/su14084711>

- Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53. <https://doi.org/10.5116/ijme.4dfb.8dfd>
- Tavassoli, K., & Farhady, H. (2018). Assessment knowledge needs of EFL teachers. *Teaching English Language*, 12(2), 45-65.
- Taylor, A. R., Anderson, S., Meyer, K., Wagner, M. K., & West, C. (2005). Lesson study: A professional development model for mathematics reform. *Rural Educator*, 26(2), 17-22.
- Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, 29, 21-36.
- Taylor, L. (2013). Communicating the theory, practise and principles of language testing to test stakeholders: Some reflections. *Language Testing*, 30(3), 403-412.
- Tezel, K. V. (2006). *A Study of an American University Master'S Programme in TESOL: Multiple Perspectives in Programme Evaluation* (Doctoral dissertation, Middle East Technical University (Turkey)).
- Thomas, G. (2011). A typology for the case study in social science following a review of definition, discourse, and structure. *Qualitative Inquiry*, 17(6), 511-521.
- Thong-Iam, N. (2019). Professional Development Needs of Thai EFL University Teachers: A Focus on the Assessment Literacy and Practise of Student Self-assessment in Classrooms. (*MANUTSAT PARITAT: Journal of Humanities*), 40(2).
- Tourangeau, R., Rips, L. J., & Rasinski, K. (Eds.). (2000). *The psychology of survey response*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511819322>
- Tourangeau, R., Rips, L. J., & Rasinski, K. (Eds.). (2000). *The psychology of survey response*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511819322>
- Tsagari, D., & Vogt, K. (2017). Assessment literacy of foreign language teachers around Europe: Research, challenges and future prospects. *Papers in Language Testing and Assessment*, 6(1), 41-63.
- Tschannen-Moran, M., & Hoy, A. W. (2007). The differential antecedents of self-efficacy beliefs of novice and experienced teachers. *Teaching and Teacher Education*, 23(6), 944-956. <https://doi.org/10.1016/j.tate.2006.05.003>

- Viengsang, R. (2016). EXPLORING PRE-SERVICE ENGLISH TEACHERS' LANGUAGE ASSESSMENT LITERACY. *Modern Journal of Language Teaching Methods*, 6(5), 432.
- Vogt, K., & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly*, 11(4), 374-402.
- Volante, L., & Fazio, X. (2007). Exploring Teacher Candidates' Assessment Literacy: Implications for Teacher Education Reform and Professional Development. *Canadian Journal of Education*, 30(3), 749-770.
- Wach, A. (2012). Classroom-based language efficiency assessment: A challenge for EFL teachers. *Glottodidactica*, 39(1), 81-92.
- Wall, D. (Eds.). (2012). *The Routledge handbook of language testing*. New York, NY: Routledge.
- Walters, F. S. (2010). Cultivating assessment literacy: Standards evaluation through language-test specification reverse engineering. *Language Assessment Quarterly*, 7(4), 317-342.
- Wang, T. H., Wang, K. H., & Huang, S. C. (2008). Designing a web-based assessment environment for improving pre-service teacher assessment literacy. *Computers & Education*, 51(1), 448-462.
- Westheimer, J. (2008). Learning among colleagues: Teacher community and the shared enterprise of education 1. *Handbook of research on teacher education*, 756-783.
- Wong, M. E. (2018). Lesson study for the enhancement of the use of assistive technology in teaching and learning. In *Lesson Study for Inclusive Education*. London: Taylor and Francis.
- Wu, T. M. (2018). *Exploring teachers' writing assessment literacy in multilingual first-year composition: A qualitative study on e-portfolios*. Arizona State University.
- Xu, Y., & Brown, G. T. (2017). University English teacher assessment literacy: A survey-test report from China. *Papers in Language Testing and Assessment*, 6(1), 133-158.
- Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practise: A reconceptualization. *Teaching and Teacher Education*, 58, 149-162. <https://doi.org/10.1016/j.tate.2016.05.010>
- Yan, Z., & Carless, D. (2022). Self-assessment is about more than self: the enabling role of feedback literacy. *Assessment & Evaluation in Higher Education*, 47(7), 1116-1128.

- Yin, R. K. (2009). *Case study research: Design and methods*. Sage.
- Yoshida, M. (2012). Mathematics Lesson Study in the United States. *International Journal for Lesson and Learning Studies*, 1(2), 140–152.
<https://doi.org/10.1108/20468251211224181>
- Yoshida, M. (2012). Mathematics lesson study in the United States: Current status and ideas for conducting high quality and effective lesson study. *International Journal for Lesson and Learning Studies*, 1(2), 140-152.
- Yu, L. (2021). *The impact of lesson study on teacher professional development*. (Doctoral Dissertation). University of Canterbury.
<https://ir.canterbury.ac.nz/server/api/core/bitstreams/fdfc7df6-a6e0-461b-97dd-f819f79a5382/content>
- Zeichner, K. M., & Liston, D. P. (2013). *Reflective teaching: An introduction*. Routledge.
- Zeichner, K. M., & Liston, D. P. (2013). *Reflective teaching: An introduction*. Routledge.
- Zhang, W. (2023). Investigating digital literacy among EFL teachers: A large-scale survey. *Frontiers in Psychology*, 14, 1153339.
<https://doi.org/10.3389/fpsyg.2023.1153339>
- Zhou, G., & Xu, J. (2017). Microteaching Lesson Study: An approach to prepare teacher candidates to teach science through inquiry. *International Journal of Education in Mathematics, Science and Technology*, 5(3), 235–235.
<https://doi.org/10.18404/ijemst.296039>

APPENDICES

APPENDIX A. Language Assessment Literacy Questionnaire (LALQ)

Dear Participant,

Thank you for participating in this PhD case study, "Exploring and Professionally Developing Iraqi Test Writers' Language Assessment Literacy Through Lesson Study Model." Your involvement is crucial in investigating the effectiveness of the Lesson Study Model in enhancing language assessment literacy among test. Your insights will play a vital role in advancing our understanding of professional development strategies for language assessment. Your responses are confidential. Your valuable contribution is highly appreciated.

SECTION 1. PARTICIPANTS' BACKGROUNDS

1. Gender

☐ Male ☐ Female

2. Last Qualification

☐ BA Degree ☐ MA Degree ☐ PhD Degree

Field of Specializations:

☐ Applied Linguistics	☐ TESOL
☐ English Language Teaching	☐ TEFL
☐ English Language & Literature	☐ Other
☐ Linguistics	☐ Please specify _____

3. Teaching Experience

☐ 1-5 years ☐ 6- 10 years ☐ 11- 15 years ☐ 16- 20 years ☐ more than 21years

4. Have you ever served as a member of a testing office?

☐ yes ☐ no

5. Were you trained in a separate testing/assessment programme during pre-service training?

☐ yes ☐ no

6. Have you participated in any professional development programmes/courses/training on language assessment?

☐ yes ☐ no

7. Presently, which language proficiency levels based on the Common European Framework (CEF) are you teaching? [More than one option can be selected.]

☐ A1 ☐ A2 ☐ B1 ☐ B2 ☐ C1 ☐ C2

8. Which of the following statements best represent your perception of overall knowledge, understanding, and application of language testing?

☐ Very unprepared ☐ Somewhat unprepared ☐ Somewhat prepared ☐ Very prepared

SECTION 2. TEACHER SELF PERCEIVED ASSESSMENT SKILLS

Please rate your level of competence in applying each of the assessment practises described below. Indicate your response by selecting one option for each statement, using the following five-point scale:

SKILLS	Not Competent	A Little Competent	Somewhat Competent	Competent	Very Competent
1. Reading					
2. Listening					
3. Writing					
4. Speaking					

SECTION 3. TRAINING IN ENGLISH LANGUAGE TESTING AND EVALUATION

Concepts and Topics - None (No learning), Little (Very shallow), Sufficient/Intermediate (Enough learning to be used in practise), Extensive/Advanced (In-depth and comprehensive enough to work as a test and assessment expert).	Please indicate the extent of instruction or training you received at university in each of the following topics related to English language testing and evaluation.			
	None	Minimal	Moderate	Extensive
1. Designing classroom-based assessments				
2. Developing various test types (e.g., diagnostic, placement, achievement, norm-referenced, integrative, direct/indirect, objective/subjective)				
3. Providing test-based feedback to learners				
4. Applying peer and self-assessment techniques				
5. Implementing informal, ongoing, or non-traditional assessment approaches				
6. Utilising the European Language Portfolio or similar tools				
7. Ensuring the reliability of assessments				
8. Establishing test validity				
9. Assessing the four English language skills (listening, speaking, reading, writing)				
10. Creating and applying scoring rubrics or criteria				

SECTION 4. LANGUAGE ASSESSMENT KNOWLEDGE SCALE

Items	True	False	Don't Know
-------	------	-------	---------------

ASSESSING READING

1. Summarizing a reading text is a valid method for evaluating reading skills.			
2. The top-down model emphasizes the reader's overall understanding of a text during evaluation.			
3. Cloze procedures are suitable for identifying a reader's grasp of a text's central message.			
4. Using a well-known passage in a reading test is not necessarily problematic.			
5. A single reading passage can adequately serve in a comprehension test.			
6. Distractors that are grammatically incorrect in multiple-choice reading items may reduce question quality.			
7. Spelling errors are considered in scoring reading assessments.			
8. Evaluating the complexity of vocabulary plays a key role in assessing reading abilities.			
9. Incorporating "not mentioned" options in true/false questions offers distinct benefits.			
10. The reliability of a reading text increases with the number of items.			

ASSESSING LISTENING

11. Using reading passages as listening materials can present challenges.			
12. Authentic listening materials are essential for evaluating learners' real-world comprehension skills			
13. Note-taking can be applied to any type of listening text.			
14. Listening tests should avoid inference questions that rely on test-taker intelligence.			
15. Errors in grammar or spelling may impact scores in listening evaluations.			
16. Listening cloze tasks promote targeted or selective listening.			
17. Tasks involving phonemic contrasts, such as minimal pairs, are considered integrative assessments.			

18. Evaluating note-taking responses is generally straightforward.			
19. Discrete-point listening tests typically assess surface-level or literal understanding.			
20. Selective listening requires learners to identify specific pieces of information.			

ASSESSING WRITING

21. Providing two prompts and asking students to choose one promotes scoring consistency.			
22. Analytic scoring reveals learners' strengths and weaknesses.			
23. Scoring scales and scores remain consistent across different levels.			
24. In the final stages of writing assessment, mechanical issues like punctuation and spelling should be considered in grading.			
25. Students are typically required to respond to at least two writing tasks in the exam.			
26. Assessing writing skills by giving learners a reading text to use as a basis for their written response is a valid method.			
27. Asking learners to present and justify their viewpoints in written form is an effective way to measure writing proficiency.			
28. Using visual prompts to support writing can be problematic.			
29. Holistic scoring is used to assess overall writing ability at the end of a term.			
30. When it comes to assessing written work, using analytic scoring tends to provide more consistent results than holistic methods.			

ASSESSING SPEAKING

31. Role plays are employed when the emphasis is on evaluating discourse skills.			
32. Assessing speaking skills may require more than one task.			
33. Interlocutors showing interest through signals can be problematic.			
34. Ending a task when a learner can't reach the criterion level is appropriate.			
35. Simultaneously using holistic and analytic scales is problematic.			
36. Oral reading is an appropriate method for evaluating speaking proficiency.			
37. Checklists can be used to assess in-class oral presentations.			
38. Interactive tasks with more than two learners can pose difficulties.			
39. Assessing speaking skills can involve learners participating in discussions.			
40. Speaking exams assess both production and comprehension together.			

SECTION 5. Professional Developmental Topics

From the list of topics related to language testing and assessment below, please indicate the extent to which you consider each one essential for inclusion in a teacher training program you would be willing to attend.

Topics	Not at all essential	Little essential	Somewhat essential	Essential	Very essential
1. Evolution and background of language assessment					
2. Selection and determination of test content					
3. Appropriate techniques for evaluating language proficiency					
4. Designing and constructing language evaluations					
5. Formulating detailed test guidelines					
6. Evaluating test efficiency through item performance analysis					
7. Categories and functions of test elements and formats					
8. Organizing and executing language evaluations					
9. Analyzing examinee results and achievement levels					
10. Evaluating comprehension abilities in reading and listening					
11. Measuring expressive abilities in speaking and writing					
12. Testing combined language skills in integrated tasks					
13. Assessing specific language aspects: grammar, vocabulary, and pronunciation					

14. Evaluating learners with diverse educational needs					
15. Using portfolio-based methods for student assessment					
16. Implementing learner-driven self-evaluations					
17. Facilitating peer-led evaluation processes					
18. Statistical methods applied in test analysis					
19. Foundational concepts in language evaluation					
20. Applying detailed scoring rubrics for speaking and writing					
21. Using overall impression scales for productive skill rating					
22. Supporting students in preparation for formal exams (e.g., FCE)					
23. Moral and professional responsibilities in test design					
24. Communicating student outcomes and giving feedback					
25. Familiarity with CEFR proficiency descriptors					
26. Using assessment outcomes to enhance teaching strategies					

APPENDIX B. Semi Structured Interviews (SSI)

Pre-Interview Questions:

1. How would you describe your understanding of the concept of quiz study in the context of language assessment?
2. Have you had any prior experience with collaborative activities focused on developing language assessments?
3. What are your expectations for engaging in the quiz study model as a professional development activity for language assessment?
4. Are there specific goals you hope to achieve through participating in the quiz study process?
5. From your perspective, what potential benefits do you anticipate for your language assessment literacy from participating in a collaborative quiz study?
6. How do you think quiz study can contribute to your proficiency in designing effective language assessments?
7. How comfortable are you with collaborating with colleagues on the design and evaluation of language assessments?
8. What skills or qualities do you think are important for effective collaboration within a quiz study group focused on language assessment?
9. Briefly describe your experience in designing language assessments for your EFL classes.
10. What challenges, if any, have you faced in the past when designing language assessments?

Post-Interview Questions:

1. Reflect on your overall experience with the quiz study model in the context of language assessment. What aspects of the process did you find most valuable?
2. How has participating in quiz study influenced your perceptions of collaborative professional development in the field of language assessment?
3. In what ways, if any, did the quiz study model impact your approach to designing language assessments?
4. Can you provide specific examples of changes or improvements you made to your assessments based on insights gained from the quiz study?
5. Evaluate the effectiveness of communication and collaboration within your quiz study group, particularly concerning the design and evaluation of language assessments.
6. How did the collaborative aspects of quiz study contribute to the development and refinement of the language assessments?
7. How did the feedback received from your colleagues during the post-implementation meeting influence your understanding of the language assessments you designed?
8. Reflect on the role of reflective practises in quiz study. How did reflection contribute to your professional growth in language assessment?
9. Do you see elements of the quiz study experience that you can apply to future language assessment planning and design?
10. How might the quiz study model continue to impact your language assessment practises beyond this study?

APPENDIX C. Reflective Journal on the implementation of Lesson Study Model as a tool to Professionally Develop Test Writers' Language Assessment Literacy

Reflective Journal on the implementation of Lesson Study Model as a tool to Professionally Develop Test Writers' Language Assessment Literacy

How to Reflect Effectively

There are two kinds of reflection: reflection-in-action and reflection-on-action.

According to Schön, reflection in action is the cognitive process of perceiving a situation and evolving an action in response to it while the action is unfolding. In contrast, reflection on actions refers to the cognitive process of mentally reconstructing the behaviours implicated in an action while considering the conditions surrounding the actual execution of the action.

Please try to take the following into account while writing your Reflection:

Before the Experience

- Consider outcomes that could have taken place.
- What challenges do you sense might come up?
- What could you do to prepare for these experiences?

During the Experience

- Observe what is happening in the moment when you make a choice.
- Is it all going according to plan? Are you managing the challenges?
- Is there anything you ought to do, say or think to make this successful?

After the Experience

1. Right after the event or at some later time when you have gained emotional distance, what are your thoughts?
2. Would you consider doing anything differently if you were to be involved in a similar event in the future?
3. What lessons do you take from this experience?

Please write your reflection on the process answering the following two questions:

1. How does this process contribute to your language assessment literacy level as a test writer?
2. How does its implementation assist you to improve professionally?