Second Language Learning and Teaching

Sahbi Hidri
Christine Coombe *Editors*

Evaluation in Foreign Language Education in the Middle East and North Africa

Springer

Second Language Learning and Teaching

Series editor

Mirosław Pawlak, Kalisz, Poland

About the Series

The series brings together volumes dealing with different aspects of learning and teaching second and foreign languages. The titles included are both monographs and edited collections focusing on a variety of topics ranging from the processes underlying second language acquisition, through various aspects of language learning in instructed and non-instructed settings, to different facets of the teaching process, including syllabus choice, materials design, classroom practices and evaluation. The publications reflect state-of-the-art developments in those areas, they adopt a wide range of theoretical perspectives and follow diverse research paradigms. The intended audience are all those who are interested in naturalistic and classroom second language acquisition, including researchers, methodologists, curriculum and materials designers, teachers and undergraduate and graduate students undertaking empirical investigations of how second languages are learnt and taught.

More information about this series at http://www.springer.com/series/10129

Sahbi Hidri · Christine Coombe
Editors

Evaluation in Foreign Language Education in the Middle East and North Africa

Springer

Editors
Sahbi Hidri
English Language Institute
University of Jeddah
Jeddah
Saudi Arabia

and

Faculty of Human and Social Sciences
of Tunis
Tunis
Tunisia

Christine Coombe
Higher Colleges of Technology
Dubai
United Arab Emirates

ISSN 2193-7648 ISSN 2193-7656 (electronic)
Second Language Learning and Teaching
ISBN 978-3-319-43233-5 ISBN 978-3-319-43234-2 (eBook)
DOI 10.1007/978-3-319-43234-2

Library of Congress Control Number: 2016947024

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Preface

Evaluation is everywhere. It has been attributed a great deal of attention because of its germane role for different stakeholders, such as educators, policy-makers, students, teachers, parents, administrators, curriculum developers, book designers, and evaluation practitioners in general. Evaluation has repercussions for the individual, societal, economic, cultural, and political levels. It also has an ethical side, and it is tailored to the needs of these different people to remain abreast of the effectiveness and efficiency of programs. It is only in this regard that evaluation has to be implemented carefully by the right people.

Evaluation plays a very important role in different fields, such as language programs, quality assurance, teaching, and testing. For instance, in any educational reform (Leung & Rea-Dickins, 2007), evaluation is needed to improve teaching, testing, accreditation, and curriculum reform based on students' self-assessment (Ladyshewsky & Taplin, 2015). Evaluation can also be implemented for sustainable development that could be linked to quality assurance. Kiley and Rea-Dickins (2005, p. 6) define evaluation as a "form of enquiry, ranging from research to systematic approaches to decision-making." Evaluation is implemented from different perspectives: Students' learning and evaluation (Golding & Adam, 2014; Nygaar & Belluigi, 2011), teachers being evaluated by students, faculty staff evaluation, conceptions and practices of assessment (Hidri, 2015), formative versus summative evaluation, dynamic assessment, language program evaluation, doctoral dissertations or theses (Kyvik & Thune, 2015), teacher professional development evaluation, classroom observation to improve the teaching practices (Wei, 2015), evaluation of text genres, test evaluation (placement, diagnostic, progress, achievement, CBT), evaluation of teaching methods and methodologies and assessment literacy, etc. All these evaluation perspectives are contrived for decision-making purposes, such as maintaining "evaluation practices" (West, 2015). Tests and test validations are used for evaluation purposes in that they are evaluated for their test usefulness qualities, such as validity, reliability, authenticity, practicality, interactiveness, and impact (Bachman & Palmer, 1996, 2010).

To cope with the changing needs of the different stakeholders, many types of evaluations have witnessed tremendous changes: Formative vs. summative evaluation in improving teaching practices (Wei, 2015), classroom-based assessment, large-scale assessment, course evaluation (Bailey & Brown, 1996; Brown & Bailey, 2008), standardized examinations, dynamic assessment in mediating the test-takers to overcome their testing difficulties (Hidri, 2014; Lantolf & Poehner, 2011), frameworks of reference, teacher evaluation, and quality assurance and its sustainable development (Shiel, Filho, do Paço, & Brandli, 2015). Such evaluation facets require well-defined standards deemed necessary and relevant to demarcate the different benchmarks against which any evaluation judgment could be made.

Many researchers (e.g., Campbell & Mark, 2015; Hart, Diercks-O'Brien, & Powell, 2009; Jin, 2010; Kiley & Rea-Dickins, 2005; Swanwick, 2007) have investigated evaluation because of its cultural and prominent role in preparing and evaluating language teachers, focusing on the quality of learning and teaching, addressing language programs, the curriculum and changing and improving the teaching and learning qualities and processes. To enhance the effectiveness of evaluation facets, a wide range of data collection methods could be utilized, such as questionnaires, interviews, observation checklists, text analyses, and examinations all of which are geared toward enhancing the effectiveness of such facets.

In the Middle East and North Africa (MENA) settings, many stakeholders (e.g., educators, policy-makers, students, teachers, parents, and administrators) have cautioned against the overlooking of evaluation. The MENA region has gone through different and significant educational changes with pinned hopes on attaining better education quality and, therefore, meeting international standards. While there is a plethora of studies on evaluation, there is a paucity of research on the status of evaluation in the MENA context, especially evaluation of English language programs, examinations, text genre analyses, assessment accountability (Hart, Diercks-O'Brien, & Powell, 2009), learning, teaching and quality assurance. Unfortunately, such types of evaluation have not gained momentum for practical and most often political reasons. A Google search on evaluation in the MENA context yields fewer, if not, poorer results that could not amount to official evaluation enterprises of educational programs. There is a dramatic scarcity of research on evaluation in the MENA region.

This book presents myriads of evaluation themes the first of which is *Teacher and Faculty Staff Evaluation*. In Chapter "Teacher Evaluation: What Counts as an Effective Teacher?," Mazandarani and Troudi talked about the multidimensional aspects of evaluation in shaping the teacher profile that is perceived as central to the attention of policy-makers and administrators alike. This exploratory study looked into teachers' perceptions of how an "effective second/foreign language teacher" should be. Results showed five major categories of what it means to be an effective teacher and perhaps what is relevant about this chapter is its model of effective teaching that embraces teachers' personal, cognitive, metacognitive, pedagogical, and professional skills deemed necessary for teacher development. In the same vein, Alamouldi and Troudi, in Chapter "EFL Teacher Evaluation: A Theoretical Perspective," tackled teacher evaluation and its relevant importance in signposting the necessity for teachers to keep abreast of the different trends that promote teacher development. This chapter suggested different ways in which teachers could be evaluated. Al-Fattal, Chapter "Faculty Performance Evaluation and Appraisal: A Case from Syria," foregrounded performance evaluation of academic staff members in Syria and its relevance to the educational context. The study triangulated different sources of data collection tools and methods, such as interviews, documentary analysis, and observation. One of the results relevant to this study was the idea of "judgmental" evaluation. The author called for a recommendation tip: Standardizing evaluation procedures that would help faculty members to be aware of the relevance and importance of performance evaluation.

The second important part of this book is concerned with *Assessment Practices*. In Chapter "Ethicality in EFL Classroom Assessment: Bridging the Gap between Theory and Practice," Torky and Haider, in an empirical study in Egypt, stressed the notion of ethicality in EFL classroom-based assessment context and its relation to test fairness. Based on a fifty-item questionnaire on teachers' perceptions of ethicality, findings of this study demonstrated that such concepts were controversial, thus calling for the use of multiple sources of measuring students' ability to set up some matching between the assessment methods, curriculum objectives and classroom activities. For the authors, one of the practical ways to reach assessment fairness is teacher training. In Chapter "Problematizing Teachers' Exclusion from Designing Exit Tests," another equal ethical issue related to assessment fairness was investigated by Dammak who questioned the exclusion of teachers from designing examinations, such as exit tests in the UAE context. This issue was undertaken from the perspective of two main stakeholders: teachers and policy-makers. Qualitative results bespoke that teachers' self-awareness of assessment was relevant in evaluating the course objectives. This assessment literacy stood in sharp contrast with the assumptions that teachers lacked testing competence, thus denying the motif to exclude them from test design. Albaiz, in Chapter "The Voice of Classroom Achievement towards Native and Non-native Educators in English Language Teaching: An Evaluative Study," evaluated the students' class performance as undertaken by native and non-native speakers (NNS) in a Saudi context. Results indicated that NNS teachers were likely to face different problems related to the misconceptions of some subject-matter teaching key concepts. To remedy this, the author suggested the implementation of appropriate teaching strategies to help learners to develop their language ability in their learning environment.

Part III of the book addresses *Text Genre Analysis Evaluation* by presenting two controversial cases from Tunisia. In Chapter "Evaluation of Generic Structure of Research Letters Body Section: Create a Research Letter Body Section Model," Melliti focused on the Evaluation of the Organizational Structure of Research Letters sections. Based on a sentence-by-sentence content analysis, the author found that a "Create A Research Letter Body Model (CARL)" contained 58 sentences of which 49 are obligatory. Implications of this study could be related to curriculum design and implementation as well as the teaching of writing to ESP researchers, students, and curriculum developers. In the same vein, text genre analysis evaluation, in Chapter "Genre Analysis and Cultural Variations: A Cognitive Evaluation of Anglo-American Undergraduate Personal Statements," in a contrastive analysis of 60 British (n = 30) and American (n = 30) personal statements written by undergraduates in different disciplines, such as business, physics, and biology, Hajji maintained that both Anglophone undergraduates yielded rhetorical and linguistic similarities and differences in the analyzed corpora, which was due to the sociocultural context where these undergraduates were operating. This study is also relevant in probing the genre aspects of personal statements.

In Part IV, *Assessment of Productive Skills*, Ben Maad, in Chapter "Learner Differences: A Trojan Horse Factor in Task-Based Oral Production Assessment?," tackled the role of learner differences in an oral production assessment mode from a task-based assessment approach. The author listed three assessment criteria in any oral performance: Fluency, accuracy, and complexity necessary for a successful speaking production. In Chapter "Assessing ESL Students' Paraphrasing and Note-Taking," Soheim, in assessing paraphrasing and note-taking, spotlighted the necessity to go through different steps to design a writing test for undergraduate ESL learners in Egypt, such as initial planning, test specs, and note-taking skills. All such phases were meant to make students aware of the necessity to avoid plagiarism, while concentrating on the obligation of academic writing through the implementation of paraphrasing and note-taking for Arab learners in an ESL program. In Chapter "Criteria for Assessing EFL Writing at Majma'ah University," Yahya studied the necessity of implementing quality assurance standards in the Saudi universities considered as relevant for teaching and assessment purposes. The author wondered whether the students' writings were standardized using defined marking rubrics from three perspectives: Instructional experience of the faculty, academic levels of writing courses, and type and nature of writing. Findings showed that there were no statistically significant differences in the implementation of the marking rubrics.

In Part V of the book, *Textbook and ICT Evaluation*, Hermessi, Chapter "An Evaluation of the Place of Culture in English Education in Tunisia," evaluated the cultural instances found in official educational documents, eight textbooks, and seven teacher guides produced by Tunisians who were involved in designing curricula and materials design in English. Results pointed that curriculum developers and textbook writers did not have any preconceived ideas on excluding culture from the English program although approaching such cultural instances was done in a non-systematic way. In Chapter "Evaluation of ICT Use in Language Education: Why Evaluate, Where to Look, and with What Means?," Derbel reviewed the status of ELT from a computer-assisted language learning (CALL) angle. The study highlighted the intricacies of learning and teaching in an ICT environment, and it concluded by mentioning future research venues for evaluating ICT in the MENA educational context.

Part VI deals with *Evaluation of ELT Certificates and Programs* in Sudan. In Chapter "Evaluating the Certificate of Teaching English as a Foreign Language (CTEFL): A Way to Quality," Nur and ElSaid Mohamed probed the evaluation of the *Certificate of Teaching English as a Foreign Language* (CTEFL) administered by the graduate unit in the English Language Institute in Khartoum, Sudan. Data from a questionnaire survey revealed that the service beneficiaries (students) were satisfied with the program, since it contained interesting modules that met their needs. Students also praised the quality of their instructors. In the same context, Alhassan and Holi Ali, in Chapter "An Evaluation of the Challenges of Sudanese Linguistics and English Language-Related Studies' Ph.D. Candidates: An Exploratory Qualitative Study," in an exploratory qualitative study, evaluated ELT studies of the Sudanese Ph.D. candidates. The study put focus on different facets to improve the quality of Ph.D. supervision in this very context. Results suggested that the quality of this supervision was fraught with myriads of problems and challenges among which were scarcity of resources and organization. What was missing from this chapter was whether the nature of the supervisors' expertise had a significant impact on the quality of dissertation.

Part VII of the book, evaluation of *Quality Assurance and ESP Needs Analysis*, Staub, Chapter "Quality Assurance and Foreign Language Programme Evaluation," addressed the relevance of quality assurance and foreign language program evaluation in Turkey. Many stakeholders, such as universities, faculties, and instructional programs, were striving to achieve institutional quality that was valued by external and internal stakeholders. In addition, the author stressed the fact that both quality assurance and evaluation were increasingly becoming "critical activities for EFL programs wishing to demonstrate their worth." In Chapter "Evaluation in Tunisia: The Case of Engineering Students," Jamly investigated the status of evaluating needs analysis of engineering students from an ESP dimension. Based on data collection on the perceptions of the TOEIC test in an engineering program in Tunisia administered to students and teachers a results denoted that current employees, who were former students of this engineering program, did not think of the TOEIC as an important test to evaluate their language ability in English, since, according to their replies, the test contents did not meet their needs. The author recommended that perceptions of evaluation embrace the needs of all stakeholders from an outcome-based measuring perspective "to build the bridge between learning objectives and learner evaluation in ESP."

In Part VIII of the book, *Assessment Literacy and Dynamic Assessment*, Bouziane, in Chapter "Why Should the Assessment of Literacy in Morocco Be Revisited?," investigated the reading and writing assessment literacy in Morocco, which could be a replica of the testing situation in North African, such as Tunisia, Libya, and Algeria. The author stressed the negative washback effect of assessment on students "both during and after school." Unfair types of assessment, as the author claimed, rested upon two major aspects: The use of a very narrow range of skills in testing reading, which could be amounted to the notion of construct fuzziness, and teachers' rating inconsistencies in writing. To remedy these shortcomings, the author called for a more comprehensive definition of the reading and writing constructs to include other types of test items that could be objectively scored. In addition, for this similar situation to be improved, the author highlighted the fact that there should be a reconsideration of ELT research and teacher training in how to produce useful and fair tests in similar-related contexts. In Chapter "Specs Validation of a Dynamic Reading Comprehension Test for

EAP Learners in an EFL Context," Hidri evaluated the theoretical and practical aspects of designing dynamic assessment specs of a reading comprehension examination for learners of English in an Omani context. Using both qualitative and quantitative approaches, results of the study showed that students' performance improved in the presence of mediation and support. The study concluded with highlighting a list of specs that test designers, in similar contexts, might consider to help their learners overcome their testing difficulties through appropriate use of dynamic assessment.

In the MENA context, evaluation challenges are numerous. For instance, one of the challenges rests on the stakeholders' ability to implement specific evaluation standards that would serve the interests of all parties. Evaluation has to be contextualized and relevant and it has to be regarded from this perspective. For instance, how can evaluation be employed to improve the curriculum and other related programs? The current situation in the MENA context should consider different parameters to enhance the relevance of evaluation to different stakeholders, such as teacher training, teaching content, teaching methodologies, students' needs, curriculum design, and writing relevant test specs that meet the curriculum objectives and materials design. There should be a careful consideration of the people who should be well versed in evaluation to implement, collect, analyze, and report on data. Evaluation has to be contextualized by highlighting its purpose, content, usage, and method (Nygaar & Belluigi, 2011) and its results should be treated with caution to avoid any misuse or harm.

The MENA context boasts itself for being "unified" at the level of language, culture, and religion. However, there exists a big challenge on whether the different stakeholders of the MENA context, especially the Arab countries, are capable of developing a common Arab framework of reference to improve the educational standards and practices. Research in this field should focus on student evaluation, teacher evaluation, and summative versus formative evaluation. In addition, further research is needed to evaluate the quality of the graduate programs, given the eventuality of the increasing number of graduates sitting for their master's and Ph.D. programs everywhere in the MENA context. There should be an open debate among these countries to maintain some educational sustainability and improvement. Many challenges for evaluation appear straightforward, such as the sociocultural context, the right people to implement evaluation, and the different ways to align evaluation to well-defined parameters that reflect the very sociocultural context of the MENA region. Even though this book bears great significance to evaluation in general, still addressing other facets of evaluation in the MENA context remains unexplored.

Sahbi Hidri
English Language Institute, University of Jeddah
Jeddah, Saudi Arabia; and
Faculty of Human and Social
Sciences of Tunis, Tunis, Tunisia

Christine Coombe
Higher Colleges of Technology
Dubai, United Arab Emirates

References

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Bailey, K. M., & Brown, J. D. (1996). Language testing courses: What are they? In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 236–256). Clevedon, UK: Multilingual Matters.

Brown, J. D., & Bailey, K. M. (2008). Language testing courses: What are they in 2007? *Language Testing, 25*(3), 349–384.

Campbell, B., & Mark, M. M. (2015). How analogue research can advance descriptive evaluation theory: Understanding (and improving) stakeholder dialogue. *American Journal of Evaluation*, *36*(2), 204-220. doi:10.1177/1098214014532166

Golding, C., & Adam, L. (2014). Evaluate to improve: Useful approaches to student evaluation. *Assessment and Evaluation in Higher Education*, *41*(1), 1–14. doi:10.1080/02602938.2014.976810

Hart, D., Diercks-O'Brien., & Powell, A. (2009). Exploring stakeholder engagement in impact evaluation planning in educational development work. *Evaluation, 15*(3), 285–306. doi:10.1177/1356389009105882

Hidri, S. (2014). Developing and evaluating a dynamic assessment of listening comprehension in an EFL context. *Language Testing in Asia*, *4*(4). doi:10.1186/2229-0443-4-4

Hidri. S. (2015). Conceptions of assessment: Investigating what assessment means to secondary and university teachers. *Arab Journal of Applied Linguistics*, *1*(1), 19–43.

Jin, Y. (2010). The place of language testing and assessment in the professional preparation of foreign language teachers in China. *Language Testing*, *27*(4), 555–584. doi:10.1177/026553220935143

Kiley, R., & Rea-Dickins, P. (2005). *Program evaluation in language education*. Palgrave: Macmillan.

Kyvik, S., & Thune, T. (2015). Assessing the quality of Ph.D. dissertations. A survey of external committee members. *Assessment and Evaluation in Higher Education*, *40*(5), 768–782. doi:10.1080/02602938.2014.956283

Ladyshewsky, R., & Taplin, R. (2015). Evaluation of curriculum and student learning needs using 360-degree assessment. *Assessment and Evaluation in Higher Education*, *40*(5), 698–711. doi:10.1080/02602938.2014.950189

Lantolf, J. L., & Poehner, M. E. (2011). Dynamic assessment in the classroom: Vygotskian praxis for second language development. *Language Teaching Research, 15*(1), 1–23. doi:10.1177/1362168810383328

Leung, C., & Rea-Dickins, P. (2007). Teacher assessment as policy instrument: Contradictions and capacities. *Language Assessment Quarterly*, *4*(1), 6–36. doi:10.1080/15434300701348318

Nygaard, C., & Belluigi, D. Z. (2011). A proposed methodology for contextualized evaluation in higher education. *Assessment and Evaluation in Higher Education*, *36*(6), 657–671. doi:10.1080/02602931003650037

Shiel, C., Filho, W. L., do Paço., & Brandli, L. (2015). Assessing and evaluating sustainable development in higher education. *Assessment and Evaluation in Higher Educatio*n, *40*(6), 783–784. doi:10.1080/02602938.2015.1073028

Swanwick, T. (2007). Introducing large-scale educational reform in a complex environment: The role of piloting and evaluation in modernizing medical careers. *Evaluation*, *13*(3), 358–370. doi:10.1177/1356389007078624

Wei, W. (2015). Using summative and formative assessments to evaluate EFL teachers' teaching performance. *Assessment and Evaluation in Higher Education*, *40*(4), 611–623. doi:10.1080/02602938.2014.939609

West, S. E. (2015). Evaluation, or just data collection? An exploration of the evaluation practice of selected UK environmental educators. *The Journal of Environmental Education*, *46*(1), 41–55. doi:10.1080/00958964.2014.973351

Acknowledgments

I would like to thank all the authors who volunteered to take part in this project. This book would not have come to life without their dedication and commitment. A special thank you would go to my co-editor, Christine Coombe, who contributed a lot to the final shape of this book. Thank you all for making it possible.

Sahbi Hidri

Contents

Part I Teacher and Faculty Evaluation

Part II Assessment Practices

Part III Text Genre Analysis Evaluation

Part IV Assessment of Productive Skills

Part V Textbook and ICT Evaluation

Part VI Evaluation of ELT Certificates and Programs

Part VII Quality Assurance, ESP Needs Analysis

Part VIII Assessment Literacy and Dynamic Assessment

Editors and Contributors

About the Editors

Sahbi Hidri is an assistant professor of applied linguistics at the Faculty of Human and Social Sciences of Tunis, Tunisia. Recently, he joined the English Language Institute, University of Jeddah, KSA. He worked in Oman for three years (College of Applied Sciences and the Humanities Research Centre at Sultan Qaboos University). Dr. Hidri is the founder of Tunisia TESOL and the *Arab Journal of Applied Linguistics*. Currently, he is chair the TESOL Arabia Research Special Interest Group. His research interests include language assessment, testing and evaluation, assessment literacy, test-taking strategies, statistics, measurement, specs validation of the language skills, SLA, and dynamic assessment. His work has appeared in Nile TESOL and TESOL Arabia publications, *Arab Journal of Applied Linguistics, Language Testing in Asia* where he serves as an editor along with other journals, such as *Assessment for Effective Intervention*. Dr. Hidri has also authored two entries (Item analysis and Discrete vs. integrative testing) in *TESOL Encyclopedia of English Language Teaching* (to appear 2016) along with other forthcoming publications. Email: sahbihidri@gmail.com

Christine Coombe has a Ph.D. in foreign/second language education from The Ohio State University. She is currently on the English faculty of Dubai Men's College. She is the former testing and measurements supervisor at UAE University and assessment coordinator of Zayed University. Christine is co-editor of *Assessment practices*; co-author, *a practical guide to assessing English language learners*; co-editor, *evaluating teacher effectiveness in EF/SL contexts*; co-editor, *language teacher research in the Middle East leadership in English language teaching and learning; applications of task-based learning in TESOL*; *the Cambridge guide to second language assessment*, and *reigniting, retooling and retiring in English language teaching*. Her forthcoming books are on research methods in EF/SL and life skills education. Dr. Coombe has lived and worked in the Arabian Gulf for the past 23 years where she has served as the president and conference chair of TESOL Arabia and as the founder and co-chair of the TESOL

Arabia Testing Special Interest Group. Christine is also the founder and chair of the TESOL Arabia Leadership and Management SIG. Dr. Coombe has won many awards including 2002 Spaan Fellowship for Research in Second/Foreign Language Assessment; 2002–2003 TOEFL Outstanding Young Scholar Award; TOEFL Board Grant for 2003–2004, 2005–2006, 2007–2008, and 2009–2010. Most recently, she served on the TESOL Board of Directors as Convention Chair for Tampa 2006 and was the recipient of the Chancellor's Teacher of the Year for 2003–2004. She served as TESOL President (2011–2012) and was a member of the TESOL Board of Directors (2010–2013). Dr. Coombe received the British Council's International Assessment Award for 2013. Email: ccoombe@hct.ac.ae

Contributors

Nawal Abdul Sayed Haider has completed a master's in teaching english as a foreign language from the University of Illinois at Urbana-Champaign (UIUC) with a focus on testing and classroom assessment. She has taught EFL and ESP courses in Kuwait and the USA. She has also participated in many projects during her career mainly pertaining to the assessment committee and the coarse development committee at PAAET. She has presented papers on teaching and assessment at conferences in the USA. Her latest research explores the assessment methods being used at the Ministry of Education in Kuwait. Email: nawala33@hotmail.com

Khadija Alamoudi is an academic and a lecturer at the English Language Institute of King Abdulaziz University. She is a Ph.D. researcher at the University of Exeter. Her interests lie in the area of EFL teacher evaluation in higher education. She has published some articles in English linguistics and she acts as a peer reviewer for some academic online journals on language and applied linguistics. Email: khaalamoudi@kau.edu.sa

Tahany Albaiz is an associate professor in instruction and curriculum of EFL. Obsessed with the leadership and development of English language programs, Dr. Albaiz has had the privilege of attending and presenting at many conferences and events at many international destinations, such as Russia, Hawaii, France, and the Netherlands. One of her goals is to help others get great opportunities to learn and develop so they feel like partners in the workplace. Dr. Tahany is a full-time faculty member and the vice-dean of the English Language Institute at the University of Jeddah. Her research interests include instruction and assessment of EFL and the inclusion of thinking routines in EFL classrooms. Email: talbeiz@uj.edu.sa

Anas Al-Fattal received his Ph.D. and master's degrees from the University of Leeds in the UK. The focus of his study and research has been educational management, marketing, applied psychology, and consumer behavior (student choice). He is the author of a book called Marketing Universities. He lectures in marketing, management, and research. Dr. Al-Fattal currently lectures at the College of

Banking and Financial Studies in Oman. He has awarded several administrative positions at post-secondary educational institutions. He worked for an ESRC project on visual research methods “Building Visual Capacity.” Dr. Al-Fattal has been a member of Higher Education Reform Experts (HERE) team for ERASMUS+ for several years. Email: anasfat@hotmail.com

Awad Alhassan is an assistant professor of applied linguistics and TESOL at the Department of English, Faculty of Arts, University of Khartoum. He has an M.A. and Ph.D. in applied linguistics from the University of Essex, UK. Dr. Alhassan’s teaching and research interests include TESOL, EAP, academic writing, academic literacy and critical EAP, corpus linguistics, the use of corpora in ELT and translation, the use of EMI in higher education, and ethnographic and qualitative research methodology in applied linguistics and ELT. Email: awad_alhassan@hotmail.com

Holi Ibrahim Holi Ali is a lecturer at Rustaq College of Applied Sciences in the Sultanate of Oman. He has an M.A. in applied linguistics, CELTA, and he is currently pursuing his Ph.D. in applied linguistics in the University of Huddersfiled, UK. He has presented widely in regional and international conferences and published extensively in peer-reviewed journals. He is primarily interested in ESP/EAP, EMI, ELT, academic writing, writing for publication, assessment, critical discourse analysis, assessment, translation, and plagiarism detection software and technology. He has a general interest in critical pedagogy, academic literacy, sociolinguistics, and linguistic landscaping. Email: Holi.ibrahim.rus@cas.edu.om

Mohamed Ridha Ben Maad is currently teaching at the Institut Supérieur des Cadres de l’Enfance, University of Carthage, Tunisia. Dr. Ben Maad’s main research interests are in applied linguistics, psycholinguistics, and alternative SLA theory. He actively contributes to research projects related to early childhood education. Email: ridhamaad@hotmail.com

Abdelmajid Bouziane holds a doctorate in education. He is a professor at Hassan II University of Casablanca, Morocco. He has participated in many national and international conferences, workshops, and projects. He has published on different issues related to ELT and has reviewed books, CDROMs, and Web sites. His main interests include the following: (Quantitative) classroom-oriented research, ICT in education (especially in the teaching of languages), literacy (in ESL/EFL), governance and quality in higher education, teacher training, and NGOs. He is the president of the Moroccan Inter-University Network of English (MINE). He is also the editor of ELTeCS AME List (English Language Teaching Contacts Scheme for Africa and the Middle East). Email: abdelmajid.bouziane@gmail.com

Abderrazak Dammak is an “All But Dissertation” (ABD) doctoral candidate in TESOL at the University of Exeter’s Graduate School of Education. He is a multilingual scholar with a wide range of experience in the field of applied linguistics and TESOL and a senior lecturer and researcher as well as program team leader in the English Department (Academic & ESP Sections) of ADNOC Technical Institute, Abu Dhabi, United Arab Emirates. In addition, Abderrazak has over 20 years of experience in teaching, ESL curriculum design, educational development, and academic leadership. He is a member of various academic and professional associations including TESOL Arabia and on the International Editorial Board of the *Journal of Somali Studies* (JOSS). He has extensively presented in institutional, regional, and international academic conferences. His current research projects are related to teacher empowerment and development, ESL/EFL teaching and learning in the Arabian context, design, and application of remedial courses for slow learners, as well as research methods in TESOL. Email: damarazak66@gmail.com

Faiza Derbel is an assistant professor of English and linguistics at the Faculty of Letters, Arts and Humanities, Manouba (FLAHM), Tunis. Dr. Derbel is course leader of the Professional Master in English Applied to Business and Communication at FLAHM and the Master of English didactics at the High Institute of Education and Continuous Development (ISEFC). She is also coordinator of the Agrégation Exam Preparation Program at Manouba. Her research interests include ELT (teaching of skills, teacher cognition and learner variables, curriculum and instruction in ELT), Business and Professional English, and computer-assisted language learning (CALL). Dr. Derbel is the former president of Tunisia TESOL. Email: fderbel26@gmail.com

El-Sadig Yahya Ezza is an associate professor of English at Majma'ah University in Saudi Arabia. Dr. Ezza teaches undergraduate courses and conducts action research in different aspects of EFL. His research interests include academic writing, EFL pronunciation, and lexicography. Email: e.ezza@mu.edu.sa

Ghada Hajji is currently an English teacher at the Faculty of Arts and Humanities and the Higher Institute of Computer Sciences at Sfax. She was graduated from the University of Social and Human Sciences of Tunis. She got her B.A. in English Literature, Linguistics and Civilization in 2011. Then in 2014, she was awarded her master's in applied linguistics and genre analysis. She has participated in several conferences on modern linguistics at the faculties of Tunis and Sfax. Email: hajjighada@hotmail.com

Tarek Hermessi teaches psycholinguistics, TEFL, and research methodology at postgraduate level. His research interests include motivation and L2 learning, culture and L2 education, globalization and L2 education, and acculturation and L2 education. He is currently the head of the English Department at Institut Supérieur des Langues de Tunis, Tunisia. Email: hermestic@yahoo.com

Rym Jamly is a part-time English teacher at the Higher Institute of Legal and Political Studies of Kairouan, Tunisia. She received a B.A. degree in English from the Higher Institute of Languages of Gabes and a master's degree in English from the Faculty of Human and Social Sciences of Tunis. As a master's student, she participated in a number of national and international conferences and study days on various topics such as needs analysis in English for specific purposes, learner assessment, curriculum development, and course evaluation. Throughout her three years of teaching experience, she has taught oral expression and pronunciation to first-year English students and legal as well as business English to law students. Her research interests include ESP, language testing, discourse analysis, academic writing, and professional communication. Email: rym.jamly@gmail.com

Omid Mazandarani is an assistant professor in TESOL at the Department of English Language Teaching, Islamic Azad University, Aliabad Katoul Branch, Aliabad Katoul, Iran, where he teaches undergraduate and postgraduate courses in TESOL and supervises master's students on various topics in TESOL/applied linguistics. Dr. Mazandarani holds a Ph.D. in TESOL from the University of Exeter, England, and is an Associate Fellow of the Higher Education Academy (AFHEA). His research interests include teacher education, development, and evaluation, research methodologies, educational technology, and critical issues in TESOL. Email: omazandarani@gmail.com

Mimoun Melliti is a lecturer of English at University of Kairouan, Tunisia. He wrote two books entitled "*Globality in global textbooks: Principles and applicability*" and "*The perceived value of English: The case of Tunisian university students*." His main research interests are genre analysis, evaluation of academic discourse, evaluation of ELT and ESP materials, and discourse analysis. He is the founder and president of *Tunisian Association of Young Researchers* and editor-in-chief of its journal *TAYR Quarterly*. Email: mimoun_melliti@yahoo.com

Abuelgasim Sabah Elsaid Mohammed is an assistant professor at the University of Khartoum. Currently, Dr. Elsaid Mohamed works as the head of the Postgraduate Studies Unit at the English Language Institute. He also works as a freelance consultant in curriculum design and he has participated in curriculum design projects run by the UNDP, World Bank, and tertiary-level institutions. Dr. Elsaid Mohamed's interests are materials design, evaluation, EAP, and ESP. Email: abuelgasims@gmail.com

Hala Salih Mohammed Nur is an associate professor at the University of Khartoum. In 2011, Dr. Nur was appointed as the founding director of the English Language Institute, the first institute of its kind in Sudan. Since the establishment of the ELI, Dr. Nur has become involved in curriculum development and program assessment. She also works as a national consultant and a teacher trainer. Email: halasalih64@gmail.com

Yasmine Soheim is an English instructor, currently working at the Rhetoric and Composition Department at the American University in Cairo. She has been teaching for six years in different language institutions. She has a passion for research in the field of assessment and pragmatics. She earned her master's in TESOL from the American University in Cairo in 2014, where she has worked as a research fellow to her assessment professor for more than a year. Her teaching experience and her enthusiasm for research have also been thrived by being a teaching fellow for almost two years at the Intensive English Program at the same university. Email: soheim@aucegypt.edu

Donald F. Staub is an assistant professor in the Department of Psychology at Işık University in Istanbul, Turkey. He also serves as the coordinator of the Quality Assurance Unit within the School of Foreign Languages. Prior to moving to Turkey in 2011, Dr. Staub spent five years as the director of a student retention project that included his leadership in successfully developing and implementing an institution-wide outcomes assessment plan. Dr. Staub has a doctorate in educational leadership from Eastern Michigan University (EMU), and an M.A. (TESOL) and B.A. (Literature) from Michigan State University. Dr. Staub is a reviewer for The Commission on English Language Accreditation. His primary research interests lie in EFL quality assurance, student retention, and the delivery of English medium instruction. Email: staubdonald2@yahoo.com

Shaimaa A. Torky is an associate professor of TEFL at the National Center for Educational Research and Development in Egypt. Dr. Torky has taught EFL and ESP courses at many accredited universities in Egypt and Kuwait and has also worked as a teacher trainer for many years. With a special interest in testing and language assessment, Dr. Torky has published papers that address modern and innovative methods of teaching and testing language skills and has presented seminal papers on ELT and assessment at conferences in Egypt, Kuwait, and the USA. As part of her career, she carried out some training programs to train EFL teachers on designing test item specs, developing test items, and utilizing holistic and analytic methods to assess writing skills. Dr. Torky's latest research explores the implications of differentiated instruction for teaching reading comprehension and teaching writing using Web 2.0 tools. Email: shaimaatorky@gmail.com

Salah Troudi is an associate professor at the Graduate School of Education of the University of Exeter where he is the director of the Doctorate in TESOL in Dubai and the supervisory coordinator of the Ph.D. in TESOL. Dr. Troudi's teaching and research interests include language teacher education, critical issues in language education, language policy, curriculum development and evaluation, and classroom-based research. Dr. Troudi has published articles in several TESOL and language education journals and edited a number of books. Email: s.troudi@exeter.ac.uk

Part I
Teacher and Faculty Evaluation

Teacher Evaluation: What Counts as an Effective Teacher?

Omid Mazandarani and Salah Troudi

Abstract The issues of teacher evaluation and teacher professional development have always been at the centre of policymakers and administrators' attention. As a multidimensional phenomenon, teacher evaluation research has closely been intertwined with other concepts and notions such as teacher effectiveness that tends to serve as an important element in teacher evaluation systems. However, whereas there is a wealth of research on effective teaching in mainstream education, research studies in EFL/ESL contexts and particularly in higher education are rather sparse. Drawing upon the findings of a recent exploratory study, this chapter reports on Iranian EFL lecturers' perceptions of the qualities and characteristics of an effective second/foreign language (L2) teacher. The statistical and thematic analyses of the data led to the emergence of five major categories each of which included a number of subcategories. Building on the existing frameworks in the literature, a newly modified model of effective teaching based on teachers' personal (behavioural), cognitive, and metacognitive qualities as well as pedagogical, and professional skills is proposed. The analysis of the subcategories provided some new insights into areas the main stakeholders need to address such as policymakers, administrators, and teachers.

Keywords Teacher development · Teacher education · Teacher effectiveness · Teacher evaluation · Mixed methods research

O. Mazandarani (✉)
Department of English Language Teaching, Aliabad Katoul Branch,
Islamic Azad University, Aliabad Katoul, Iran
e-mail: omazandarani@gmail.com

S. Troudi
Graduate School of Education, University of Exeter, Exeter, UK
e-mail: s.troudi@exeter.ac.uk

S. Hidri and C. Coombe (eds.), *Evaluation in Foreign Language Education in the Middle East and North Africa*, Second Language Learning and Teaching,
DOI 10.1007/978-3-319-43234-2_1

1 Introduction

Teacher effectiveness research (TER), which was erratically researched in the first half of the 20th century (e.g., Ryans, 1949), started to gain researchers' attention in the 1970s (e.g., Doyle, 1977; Good, 1979; McKeachie, Lin, & Mann, 1971). The issues of teacher effectiveness and its pertinent appraisal mechanism have culminated in different definitions of effective teaching in the 1990s and 2000s (e.g., Darling-Hammond, 2000; Stronge, 2007; Stronge, Tucker, & Hindman, 2004; Stronge, Ward, Tucker, & Hindman, 2007) for which different models and frameworks have thus far been proposed by researchers for evaluating teacher effectiveness (e.g., Campbell, Kyriakides, Muijs, & Robinson, 2004a; Cheng & Tsui, 1998, 1999; McBer, 2000). More recently, research in this area has fervently been pursued by several scholars (e.g., Darling-Hammond, Newton, & Wei, 2013; Garrett & Steinberg, 2014; Koedel & Betts, 2010; Leigh, 2010; Stronge, Ward, & Grant, 2011). However, as it will be argued in this chapter, TER in non-Western contexts has yet to be established, given the fact that research into teacher effectiveness in the Middle-eastern EFL context is rather sparse. Drawing upon the existing research on effective teaching, the study reported in this chapter aims to explore Iranian EFL lecturers' perceptions of the characteristics of an effective language teacher of which the importance is threefold. First, the bulk of research on TER, as it will be discussed shortly, appears to be rooted in Western contexts, thereby casting doubts on the extent to which current understanding of effective teaching in non-Western contexts, especially the Middle-eastern context, is similar to that of Western ones. Second, in a similar vein, much of the current awareness of teacher effectiveness and the qualities of an effective teacher is drawn from research in primary and secondary education bringing up some concerns as to whether the existing findings can be applied to the higher education context which tends to carry with it different values, considerations, sensitivities, etc. Third, whereas there is a wealth of research on effective teaching in mainstream education, there is a dearth of research in EFL/ESL contexts. There is more to such concerns than meets the eye, in that EFL lecturers' challenges could be rather different from those of their counterparts in mainstream education. To address such concerns, this chapter reports some findings of a recent study (Mazandarani, 2014) which delved into EFL lecturers' perceptions of the qualities and characteristics of an effective EFL lecturer as one of its objectives. With this end in view, the main purpose of this study, amongst others, is to propose a modified model for evaluating effective L2 teaching.

1.1 Effective Teaching: An Enigmatic Concept

Whereas teacher effectiveness and effective teaching were erratically investigated in the first half of the 20th century, research on the qualities of an effective teacher became more systematic in the second half of the century and henceforth extended by the outset of the third millennium. It can be argued that such endeavours were more or less in response to policymakers and governments' accelerating quest for the issue of 'quality'. The emergence of some international organizations dealing with the concept of excellence of education such as The Organization for Economic Cooperation and Development (OECD) clearly substantiate the increasing demand for high quality education.

As a multifaceted phenomenon, research on effective and quality teaching has always been challenged by some rudimentary but nonetheless important concerns and questions. Arthur, Bennett, Edens, and Bell (2003, p. 235) maintain that evaluating the effectiveness of a training programme needs to be followed by the question "effective in terms of what?" Similarly, it appears that discriminating effective teachers from less effective ones is highly contingent upon the identification of the characteristics or qualities of an effective teacher for which setting a benchmark seems to be the very first step. However, as it will be argued shortly, the main challenge mostly lies with the notions of standards or criteria per se, given the fact that different stakeholders may have different understandings of such criteria. The literature also gives evidence to discrepancies over the nature of the characteristics of effective (EFL) teachers among researchers and practitioners. Teachers' little awareness of the standards against which they are evaluated by administrators and the lack of transparency are other aspects of teacher appraisal that tend to add to the teacher evaluation predicament. This is extremely important on the grounds that teacher appraisal systems have been found to exert influence on teachers' effectiveness. As Kelly, Ang, Chong, and Hu (2008, p. 39) maintain, appraisal schemes can create an impact on teachers' behaviours, and hence their performance which can potentially exercise undesirable effects on students' achievement. There is ample evidence that policymakers' stances towards teachers' duties have also changed, i.e., from the traditional approaches focusing on 'within-the-classroom' activities (Cheng & Tsui, 1998) to new approaches in the newly emerged era which expect teachers to 'guide' students rather than merely transfer knowledge to them (Korthagen, 2004).

More importantly, the question which has thus far been left unclear rests with the dynamics of effective teaching. This raises some concerns as to whether or not effective teachers are the product of an effective education system within which they gradually become effective. In other words, little research has been done to explicate the mechanism through which (ineffective/less effective) novice teachers evolve. Moreover, how effective teaching is bound up with what Stronge (2007, p. 5) calls "prerequisites" for effective teaching such as "verbal ability" is another dimension of effective teaching which needs to be paid more attention. There is more to such concerns than meets the eye in that effective teaching seems to be the outcome of intricate interactions between teachers' individualistic attributes, e.g., talents, potentials and environmental (workplace) variables, e.g., successful initial teacher education programmes (ITEP) and teacher professional development programmes (TPDP).

2 Theoretical and Conceptual Framework

As it was mentioned in the previous section, research on teacher effectiveness was rather hit-and-miss until the 1950s after which it became more systematic. Nevertheless, since the history of research on teacher evaluation was not clear before 1970 (Shinkfield & Stufflebeam, 1996, p. 37), it can be argued that much of current awareness of teacher evaluation seems to have theoretically and conceptually evolved since the 1970s; thereafter, the contentious notions of effective teaching and teacher effectiveness have always been a point of controversy among researchers and practitioners. As stated elsewhere, many of these discrepancies have their roots in the very concept of 'effectiveness'. Evaluation of teacher quality tends to be contingent upon some 'conceptual foundations' which can elucidate what is meant by teacher quality (Ingvarson & Rowe, 2008).

As Zhu and Zeichner (2013, p. v) put forth, teachers in the new century of information blast have increasingly been recognized as the most seminal element in education systems in more and more countries throughout the world. The related research is also replete with evidence for teachers' centrality to students' learning outcomes (e.g., Ladd, 2012, p. 216) and their responsibility for fostering education (Giroux, 2014, p. 495). Teachers are the most important agent in making changes and innovation in education (Bakkenes, Vermunt, & Wubbels, 2010). As research shows, teacher quality is a key factor in students' achievements (Rockoff, 2004, p. 251). And to this end, teacher quality is the foremost asset of which teachers need to avail themselves.

A review of the bulk of the research substantiates that much of the current understanding of effective teaching and its pertinent evaluation systems was established in the 1990s and the 2000s during which TER witnessed a plethora of studies most of which focused on mainstream primary/secondary education (e.g., Campbell et al., 2004a; Campbell, Kyriakides, Muijs, & Robinson, 2003, 2004b; Cheng & Tsui, 1998, 1999; Darling-Hammond et al., 2013; Ellett & Teddlie, 2003; Kyriakides, Campbell, & Christofidou, 2002; McBer, 2000; Muijs, 2006; Seidel & Shavelson, 2007; Stronge et al., 2011). Accordingly, several frameworks and models have been so far proposed during the past two decades for evaluating effective teaching (e.g., Campbell et al., 2003; Cheng & Tsui, 1998, 1999; McBer, 2000).

Since most of the above-mentioned published works, amongst others, have mostly been germane to mainstream education, there has always been a question as to whether such findings can be applied to L2 education contexts in which teachers are required to teach in a language other than the students' mother tongue. Addressing a similar concern, Crandall (2000, p. 34) maintains that language teacher education is a microcosm of teacher education and tends to be informed by similar trends in theory and practice in general teacher education. Thus, it seems quite wise to concur with the view that the current understanding of effective language teaching is largely informed by conceptual frameworks of effective teaching in mainstream (general) education. Nevertheless, several L2 researchers have endeavoured to address the dynamics of teacher effectiveness and its pertinent evaluation approaches in EFL/ESL contexts (e.g., Bailey, 2006; Brosh, 1996; Coombe, Al-Hamly, Davidson, & Troudi, 2007; Farrell, 1999; Farrell & Jacobs, 2010; Freeman, 1989; Freeman & Johnson, 1998; Richards & Nunan, 1990).

From among the existing frameworks and models proposed for teacher effectiveness and effective teaching in the literature, the Hay McBer model of teacher effectiveness (DfEE, 2000), Cheng and Tsui's (1999) multi-models of teacher effectiveness, Korthagen's (2004) Onion model, and Campbell et al.,'s (2004a) differentiated model seem to be of seminal importance in shaping TER. Perhaps one of the most influential attempts to inquire into teacher effectiveness during the past 15 years by which this study has been informed is the Hay McBer's (2000, 2002) research report entitled "research into teacher effectiveness; a model of teacher effectiveness". According to McBer (2000, p. 6), students' progress is highly influenced by three main elements all of which are within the teachers' control, viz. "teaching skills, professional characteristics, and classroom climate". The findings of McBer's study indicated that a combination of the aforementioned categories predicts over 30 % of the variance in students' achievements (2000, p. 9). From McBer's perspective, teaching skills or what he calls "micro-behaviours" can be grouped under Ofstead inspection headings (2000, p. 10). As he continues, professional characteristics or teachers' "deep-seated patterns of behaviour", are an amalgamation of some 16 characteristics classified into five clusters (p. 19). Finally, referring to how students feel in a particular teacher's classroom, classroom climate may have a number of dimensions such as clarity, fairness, etc. (see McBer, 2000). Although the Hay McBer's model successfully places teacher performance on three levels, it fails to give a clear picture of how the levels of teacher performance are connected with students' cognitive progress (Campbell et al., 2004a, p. 13).

2.1 Lack of TER in Tertiary EFL Contexts

The accelerating movement of mass education around the world in recent years has brought with it a dire demand for effective teaching. Students' unbounded enthusiasm for evaluating their teachers/professors along with the emergence of online databases such as "ratemyprofessor.com" (Clayson, 2013) clearly signifies the extent to which effective teaching has been in the popularity stakes. In line with other dimensions of education, effective teaching has evolved throughout the history of TER. Such evolvement has been a kind of transition from the so-called traditional conception of effective teaching to a modern understanding of what makes an effective teacher. Whereas, on one extreme of this continuum, some researchers (e.g., Angelo, 1990, p. 75; Angelo & Cross, 1993, p. 3) maintain that "teaching without learning is just talking", others maintain that teachers need to adopt responsibilities which are beyond the aforementioned traditional conceptions of teaching, that is 'transferring knowledge'.

Another aspect of teacher evaluation which seems to be a problem in EFL contexts and the context of this study is no exception, is the relatively little appreciation of the importance of teacher education research as compared to other areas in second/foreign language research. It is unfortunate that only 9 % of the featured articles in TESOL Quarterly from 1990 to 1997 were focused on teacher education (Freeman & Johnson, 1998, p. 397). A quick review of the related literature simply echoes Freeman and Johnson's argument, given the fact that much of the research done seems to be mostly revolving around students and learning strategies. Perhaps, part of this lack of awareness emanates from the false adage that "*Those who can, do; those who can*'*t, teach*" which has nowadays been belied by ample evidence that teachers are a key element in education being always in the forefront.

Despite its relatively rich history and thorough attention it has received, research on effective teaching remains controversial due to the lack of a consistent benchmark against which teacher effectiveness and effective teaching can be defined and hence appraised. This inconsistency can be also traced in the literature. As Campbell et al. (2004a, p. 2) contend, teacher effectiveness, school effectiveness, and educational effectiveness are used interchangeably in the literature. Any mismatches among different stakeholders involved in teacher evaluation could potentially exert undesirable effects on the quality of education and may hinder the achievement of educational objectives. Moreover, a brief review of the literature indicates that 'transparency' of the evaluation scheme/policy and the 'clarity' of the standards against which teachers tend to be evaluated have usually been a point of contention in most education systems especially the context of this study. This is important as teachers' inadequate awareness of their appraisal can lead to idle speculation about the 'fairness' of the entirety of teacher evaluation system in a context.

One problem associated with teacher effectiveness and effective teaching remains with the contextualized understanding of teacher evaluation, as stated in the Introduction. As the literature shows, the seminal works underpinning the theoretical framework draw heavily upon studies done in mainstream (general education) rather than those conducted in EFL/ESL contexts. Finally, it is worth mentioning that much of the current understanding of effective teaching is based on the findings obtained from primary and secondary education which are less likely to be axiomatic in universities and higher education institutions. Teaching in higher education as a complex phenomenon (Postareff, Lindblom-Ylänne, & Nevgi, 2008) differs from other levels of education in that as Teichler and Kehm (1995, p. 119) put forth, higher education institutions tend to serve as training grounds for scholars expected to be the teachers of future students. Higher education is different as it prepares students for their future professional practice (Gregori-Giralt & Menéndez-Varela, 2015, p. 1).

Having considered the above-mentioned lacunae in research on teacher evaluation with respect to the elements proposed in this research, i.e., the '*Middle-eastern EFL higher education*' context, this research has been done to probe into effective teaching and inquire into the challenges ahead of effective teaching through teachers' lenses.

2.2 Understanding Effective Teaching in Iran

The significance of this study lies in an increased understanding of the very nature of effective teaching in the Iranian higher education context on the grounds that little is known about teacher effectiveness and effective teaching in non-Western L2 contexts, especially that of the Middle East. As Campbell et al. (2004b, p. 451) contend, much of the research pertaining to educational effectiveness is conducted in Western contexts, i.e., U.S., the Netherlands, and the UK in which students' achievement is considered as a significant benchmark for effectiveness.

Given the impact of teachers' conceptions of teaching on their stances towards teaching (Postareff et al., 2008, p. 30), this study provides teachers with an opportunity to have their say without the fear of any repercussions. Informed by McBer's (2000) model of teacher effectiveness and taking into account Arthur et al.,'s (2003) question of "effective in terms of what", this study can hopefully deepen the insights into the challenging nature of L2 teaching and the benchmark (standards) against which effective teaching can be evaluated. Moreover, this research project will propose a modified framework for effective teaching and teacher effectiveness (e.g., McBer, 2000) which is expected to add to the existing knowledge base of effective teaching in the context of the study.

Given the fact that research on effective L2 teaching is rather sparse in the Middle-eastern context, this study is expected to partially fill the gap and raise the awareness of policymakers, administrators, and teachers of different skills, characteristics, duties, and responsibilities which teachers need to be cognizant of in order to teach in a more effective and efficient way. It is also hoped that this study can bring to the fore the peculiarities and idiosyncratic features of teaching in a language other than the students' mother tongue. This study was an attempt to answer the following main research questions:

a. What are the qualities/features of an effective EFL teacher?
b. What skills does an EFL teacher need to master in order to be more effective?

3 The Study

3.1 Context and Participants

The context of this study was universities and higher education institutions in Iran. In order to have a more profound understanding of the nature of EFL lecturers' perceptions of effective teaching, the universities and higher education centres selected for this study included the ones with English Language Teaching (ELT) departments in which TEFL courses were taught. The participants included a purposive sample of EFL lecturers who were affiliated to ELT departments. For quantitative and qualitative phases of the study, 43 and 14 lecturers were selected respectively to participate in the study. They were all Iranians and native speakers of Persian (Farsi). Although this study is not interested in generalizing its findings which is in line with its paradigmatic standpoint, every endeavour was made to choose lecturers with different backgrounds, e.g. gender, academic qualification and experience. This was an important criterion for the selection of the participants as lecturers who are affiliated to ELT departments at universities in Iran are from different educational (academic) backgrounds, TEFL/TESL, Applied Linguistics, Translation, Literature, and Linguistics.

3.2 Research Instruments

In order to explore the research questions set forth in this study, an adapted exploratory sequential design (Creswell, 2012, p. 543) was adopted to collect quantitative and qualitative data. The data were collected sequentially using two instruments. Adopting a mixed methods approach towards the phenomenon, this study utilized semi-structured interviews and a researcher-developed Likert scale questionnaire, which were piloted beforehand. In line with the adopted design, data collection was implemented in two phases with a particular 'emphasis on qualitative data' (Creswell, 2012). The design allows data collection from more than one perspective, numerically and qualitatively in a sequence that is appropriate to the context and conditions available to conduct this study. In the first stage, lecturers were required to respond to the questionnaire which included close and open-ended sections. In the second stage, some of the lecturers were invited to participate in follow-up interviews. All ethical dimensions and procedures were observed and participants were assured of anonymity and confidentiality.

The analyses of quantitative and qualitative datasets were conducted using SPSS and Nvivo, respectively. Both instruments were developed thematically whereby identifying the similarities and/or differences between the findings extracted from two sets of data could be done with ease. Statistical (descriptive) analysis of the quantitative data and thematic analysis of the qualitative data were performed at different stages. Upon the collection of quantitative data, all information was entered into the SPSS for statistical analysis. Collecting participants' background information such as age, gender, etc. provided an opportunity to identify and explain the relationships between their personal attributes and their responses to the statements of the questionnaire.

4 Major Findings

The thematic analysis of the interviews led to the emergence of five major categories as shown in Fig. 1 as follows.

Each of the above-mentioned categories included a number of sub-categories which represented some concrete aspects of teaching suggested by the participants as prerequisites for quality and effective teaching. In the following sections, we will report the most important subcategories. It is worth clarifying that whereas some ideas had support only from interviews, others emerged through the statistical analysis of the questionnaires. Nevertheless, there did exist a number of subcategories, which had support from both datasets. Pseudonyms were used to refer to the research participants.

4.1 Personal Attributes (Traits)

The research into effective teaching is replete with a number of qualities perceived as personal attributes of effective teachers (e.g., Meijer, Korthagen, & Vasalos, 2009; Stronge, 2007; Stronge et al., 2011). However, the literature indicates that teachers' personal qualities are somehow ignored in research on effective teaching (Patrick & Smart, 1998). In total, eight qualities were formulated for this category as shown in Table 1, as follows.

It is worth highlighting that the majority of the participants (95.4 %) concurred with a statement introducing teacher personal qualities as an influential factor in

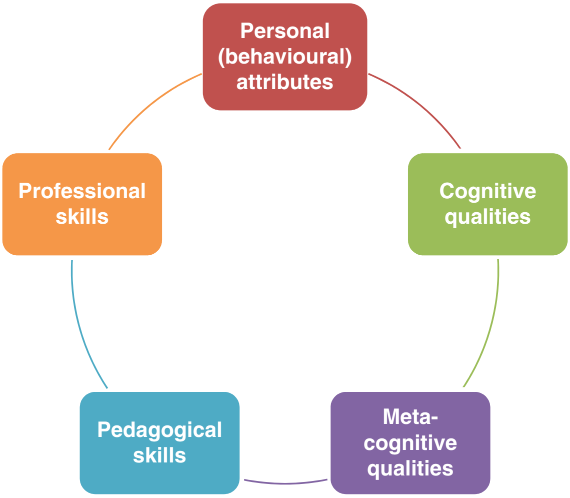


Fig. 1 Major categories of the qualities of an effective teacher

Table 1 Personal attributes

Sub-category (quality)	Source interview/open-ended questionnaire	Close-ended questionnaire
1. Friendly and approachable	✓	✓ (93.1 %)
2. Dedicated and informative	✓	✓ (81.4 %)
3. Respectful	✓	✓ (97.7 %)
4. Inspirational and self-motivated	✓	
5. Confident	✓	
6. Patient	✓	✓ (95.4 %)
7. Open and adaptable	✓	✓ (86.1 %)

their teaching effectiveness. As Table 1 shows, qualities such as friendliness, patience, and respectfulness seem to be the mainstay of teachers' desirable personal characteristics. As shown in Table 1, most of the subcategories received support from both datasets, i.e., questionnaire and interviews/open-ended questionnaire. For instance, as Parham, one of the interviewees, maintained:

> He [an effective teacher] should be approachable; students can access him very easily, be friendly with the students, constructs a friendly atmosphere, a stress-free atmosphere in the class in order to motivate them, encourage them to study, be honest to students, and treat them as his brothers and sisters, that's it.

Niloofar who is an experienced lecturer placed emphasis on students' motivation and encouragement and stated:

> They must always believe that motivation in the greatest way and if they put barrier on this motivation the results may not be so satisfactory and I want to say that some are too idealists, some want to make it quite a miracle and some are quite bored and exhausted with years of struggle, something in between something in average.

Similarly, referring to the role of 'confidence' in effective teaching, Maria who is also an experienced lecturer contended that:

> Taking relevant measures to boost my confidence as a person and of course as a teacher teaching subjects that I am interested in.

'Patience' as one of the subcategories that received support from both interviews and the questionnaire. In line with the strong support it received from the statistical analysis of the questionnaire (95.4 %), patience was introduced in the interviews as an important characteristic of an effective teacher. For instance, Mersedeh posited:

> S/He [an effective teacher] must be very patient and stimulate and encourage the students to have interaction with each other.

Finally, referring to teachers' growth, Shahab maintained that an effective teacher should learn from others and stated:

> Engendering within himself/herself a strong desire to learn from the learners.

Nevertheless, it is important to remember that each of such qualities is highly influenced by the norms and conventions of the country within which teaching transpires and to that end tends to be culturally oriented. For example, whereas addressing a student in his or her first name might be quite normal and acceptable in one country, this could be relatively unacceptable or even offensive in another. Therefore, it seems that we need to have an operationalized definition of such features with reference to the context of the study.

4.2 Cognitive Qualities

The obtained findings gave evidence to the importance of the teacher's knowledge base in effective teaching. Teacher cognitive abilities and more specifically teachers' knowledge have been extensively investigated in the literature (e.g., Freeman, 2002; Freeman & Johnson, 1998; Shulman, 1987, 1999), as many consider teachers' knowledge as the mainstay of teaching practice. Although teachers' generic knowledge can embrace any aspect of teaching even those related to pedagogy and professionalism, in this study four major dimensions of teachers' knowledge were identified and fully supported by the participants as different aspects of the knowledge base required for effective language teaching. Notwithstanding the fact that language teachers share different dimensions of knowledge base for effective teaching, they need to have some extra knowledge such as English language proficiency (Richards, 1998, p. 7) as a certain threshold is needed for non-native speaker language teachers in order to teach effectively. The category of cognitive quality refers to the linguistic and general knowledge competencies, including ICT, that teachers need as part of their general knowledge.

Table 2 summarizes the cognitive qualities as follows.

As indicated in Table 2, all the subcategories had support from both sets of data. Features such as teachers' good command of English, knowledge of use of technology in the classroom and also knowledge of curriculum and syllabus

Table 2 Cognitive qualities

Sub-category (quality)	Interview/open-ended questionnaire	Source close-ended questionnaire
8. Language proficiency	✓	✓ (83.8 %)[a]
9. Subject matter knowledge	✓	✓ (67.4 %)
10. Knowledge of curriculum and syllabus	✓	✓ (88.4 %)
11. ITC literacy	✓	✓ (90.7 %)

[a] 83.8 % of the respondents were opposed to the idea that language proficiency does not contribute to teacher effectiveness

development were of high priority from participant lecturers' points of view. As to the role of teachers' language proficiency, for instance, Ali stated:

> S/He [an effective teacher] should have good command of English (especially good command of spoken English).

Sepher, one of the experienced participant lecturers, referred to the importance of teachers' knowledge of curriculum and syllabus design and stated:

> Try to design and prepare teaching materials according to any given teaching context, considering students' cognitive, social, and emotional characteristics.

Similarly, teachers' subject knowledge was raised in the interviews as shown in the following Excerpt:

> (…) and he should be a knowledgeable person in his own field ok (…) should be a knowledgeable person in that field to for example to be able to answer the students' questions.

Although the statistical analysis of the questionnaires firmly introduced teachers' knowledge of computer and technology as one of the characteristics of effective teachers, the analysis of the interviews did not corroborate the idea, given the fact that only one interviewee referred to this very important aspect of teachers' cognitive abilities.

4.3 Meta-Cognitive Qualities

Having its roots in John Dewey's works followed by Donald Schön (Craig, 2010), 'reflection' as a metacognitive strategy which is argued to be 'thinking about thinking' (Jonassen, Mayes, & McAleese, 1993), and critical thinking emerged though the analysis of the interviews which are shown in Table 3, as follows.

Kian considered 'reflection' as a precious tool for an effective teacher as a life-long learner as follows:

> A life-long learner, willing to learn from experience and share it with colleagues, continual self-assessment, monitoring what truly works and what does not, evaluating the quality of lessons being delivered, reflecting on and improving the teaching practice on a regular basis, with a wide range of pedagogical skills.

Bahar, one of the participants, also stated:

> An effective EFL teacher should be a good critic to [of] her/his performance.

Table 3 Metacognitive qualities

Sub-category (quality)	Interview/open-ended questionnaire	Source close-ended questionnaire
12. Reflective practice and critical thinking	✓	

4.4 Pedagogical Skills

The analysis of both sets of data culminated in a number of qualities and skills categorized under the overarching notion of pedagogical skills. The minutiae of different pedagogical skills recognized by the participants as to be influential in their effectiveness are shown in Table 4. However, it is worth mentioning that 88.4 % of the respondents advocated the centrality of pedagogical skills in its entirety to teacher effectiveness. The importance of pedagogical skills especially for Iranian lecturers is undeniable, for EFL lecturers recruited by Iranian universities include those who major in TEFL-specific or non-TEFL disciplines. And herein lies a problem in that lecturers with non-TEFL backgrounds such as translation studies, English language literature, and linguistics tend to show little awareness of pedagogical skills as compared to their colleagues with TEFL-specific backgrounds. It is unfortunate that even lecturers with academic backgrounds in TEFL and its pertinent fields such as TESOL, Applied Linguistics, etc. may be vulnerable to what Watzke (2007, p. 66) calls "a process of wash out" during their transition from pre-service education to real teaching.

As Table 4 suggests, most of the formulated categories were raised in the interviews and open-ended questionnaires. However, qualities such as teachers' ability to engage students in classroom activities as well as their assessment seem to be of high importance for effective teaching. Pedagogical skills are indeed teachers' tools which help them fulfil the educational objectives determined by universities. Although experienced teachers might be more or less aware of such strategies and *tricks*, novice teachers really need to be acquainted with such skills which, as stated elsewhere, are rarely provided to them in their pre-service education. As Johnson (2009, p. 18) argues teachers' awareness of "pedagogical recourses" can help them teach more appropriately.

With regard to teachers' use of alternative instructions, Sepehr and Sarah maintained that:

Table 4 Pedagogical skills

Sub-category (quality)	Interview/open-ended questionnaire	Source close-ended questionnaire
13. Instructional planning and alternative instructions	✓	
14. TEFL-driven understanding of teaching		✓ (76.7 %)
15. Simplicity/tailoring material to students' needs	✓	
16. Needs analysis	✓	
17. Student engagement		✓ (97.7 %)
18. Attentive to affective filter	✓	✓ (88.4 %)
19. Transferring knowledge	✓	–
20. Assessment	✓	✓ (93 %)

> (…) [Effective teachers] make use of a wide range of instructional strategies.
>
> (…) [Effective teachers] have a lesson plan, be familiar with new sources.

The lecturers call for a TEFL-driven understanding of teaching (76.7 %) simply substantiates the above-raised features in that issues such as lesson planning, alternative instructions and assessment, etc. tend to be important aspects of TEFL.

Teacher as needs analyst was an important characteristic proposed by interviewees for teacher effectiveness. For instance, Rima maintained that:

> (…) S/He should try to meet students' needs and recognize that students learn at different rates.

In a similar vein, Ali touched upon the importance of teachers' ability to predict students' future needs:

> (…) an effective teacher or a good teacher should predict what their needs are in that special courses and try to link or make a link between what they are studying and what they are expected to do in the future. So a practical type of teaching.

Interestingly, the thematic analysis of the interviews and the open-ended questionnaires revealed that the so-called traditional conception of effective teaching, i.e., 'transferring knowledge', is still a prevalent belief among language teachers. As Parham posited:

> An effective teacher is the one who is able to transfer his knowledge to his students. I know some people who are very knowledgeable but don't have the skill of transferring their knowledge to their students so one aspect of teaching effectiveness for me is ability to transfer your knowledge.

As to the role of teachers' knowledge and use of assessment strategies, Kian and Rima maintained that:

> It appears that assessment should not only be "of learning", but also "for learning" for it to be effective.
>
> According to (…) ZPD, a teacher should scaffold students' learning effectively and promote effective interaction in the class.

4.5 Professional Skills

Four major categories of the qualities of an effective teacher have so far been presented. However, as the literature suggests, for language teachers to possess personal, cognitive, metacognitive, and pedagogical skills do not suffice for effective teaching. Professional characteristics appear to be more sophisticated as compared to pedagogical skills in that they seem to be determined by the intrinsic nature of teachers as human beings, and to that end might be less teachable as compared to pedagogical skills. As shown in Table 5, the analysis of the data led to the emergence of nine professional characteristics as follows.

Table 5 Professional skills

Sub-category (quality)	Interview/open-ended questionnaire	Source close-ended questionnaire
21. Sense of responsibility and accountability	✓	✓ (81.4 %)
22. Creativity and innovation	✓	✓ (100 %)
23. Authority and management skills	✓	✓ (55.9 %)
24. Building rapport and trust	✓	
25. Fulfilling ILOs (Intended Learning Outcomes)	✓	
26. Improving students' performance	✓	✓ (51.2 %)
27. Networking	✓	
28. Overcoming problems	✓	✓ (97.7 %)
29. Fair evaluation	✓	

As indicated in Table 5, the respondents extolled the virtues of professional qualities such as innovation with 100 % support which clearly shows how such skills are currently needed, especially in the context of this study and in this era of mass education, economic uncertainties and high competition for employment. It is also clear that students' performance is no longer a robust indicator of teachers' effectiveness as it used to be in the past decades. These issues will be further discussed in the upcoming sections.

From Thelma's perspective, an effective teacher is a creative one who provides students with what they want:

> A person who is creative, we need creativity, and well as it goes down and it boils down to a person who actually wants the learners to learn something, that gives the thing that the learners want, not the thing that he or she thinks is true.

As to teachers' authority and management skills, the idea did not receive strong support from the respondents. However, some interviewees maintained that such a feature is an important asset to effective teaching. For instance, Mersedeh commented on this as follows:

> (…) a teacher must control and handle the class. I mean, based on the characteristics and behaviours of the students, there are many different factors.

Connected with the above-mentioned quality, an effective teacher needs to be able to face and overcome challenges and problems that might happen in the classroom especially the unpredicted ones for which he or she should be prepared to take impromptu actions. Ali touched on this and stated:

> My answer according to what I have learned from my professors is that you should be an artist at the first stance. I mean you have to know how to overcome the problems, how to treat the students, how, for example (…) use your even very limited facilities around you to make that class as active as possible. This is an art; I believe in that.

As shown in Table 5, some participants of this study contended that the notion of teacher effectiveness is imbued with teachers' 'fulfilment' of the course objectives or intended learning outcomes (ILOs). As Maria posited:

> I believe teacher effectiveness on the whole means producing the intended result which is reaching the aims established by the curriculum.

As stated earlier, an effective teacher is deemed to be possessed of metacognitive abilities such as the 'power of reflection'. In a similar vein, as a professional quality effective teachers seem to be the ones who tend to get into the habit of 'networking'. Sohrab referred to this as follows:

> The second one is the teachers' network (…), getting I mean one another's experience of teaching. This may be a good I mean habit if you want to, for example, teach Pragmatics okay you ask other colleagues to know about their way of teaching, the I mean materials they are introducing to the students and many other factors that can be solved by teachers networking or let's say teaches network.

The teachers' ability to build trust and conduct fair evaluations of students was introduced by some of the participants as crucial professional features of effective teachers. For instance, Maria believed that:

> (…) and at the same time building the kind of rapport with the students that can make learning and teaching pleasant and successful.

Similarly, Shahab voiced his concerns and stated:

> The minimum level of expectation is that teachers must consider principles of morality as a human being in the class special a unique case to apply power is given to them and they must always remind themselves not to misuse or abuse that special position.

It is worth mentioning that professional qualities such as *fairness* and *trust* tend to be important criteria for effective teaching in almost any education system regardless of the context in which teaching transpires.

5 Discussion

The characteristics of an effective EFL teacher were the centrepiece of this study. However, as argued earlier, such characteristics need to be identified and interpreted within a larger context of a teacher evaluation system and hence need to be informed by policymakers and administrators' approaches and understanding of 'what counts as an effective teacher'.

One central point raised in this study was the dynamic nature of teaching. Although EFL teachers, similar to their counterparts in mainstream education, do share some basic and rudimentary characteristics for teaching more effectively, research in L2 teacher education, offers compelling evidence that EFL teachers tend to face more challenges as compared to teachers in mainstream education. It has been argued that, in L2 contexts EFL teachers especially the novice ones are likely to be in a predicament. This is why there appears some dissenting voices among teachers as to the so-called 'one-size-fits-all' evaluation which can potentially lead them to take issue with administrators' approaches to teacher evaluation.

This study reported on qualities and skills for an effective language teacher some of which appear to have been given scarce attention in the literature. As it was explained, each of the reported categories, i.e., personal, cognitive, metacognitive, pedagogical, and professional qualities embrace other features which can be discussed from two perspectives. First, it alludes to teachers' transition from a 'single-role' to a 'multi-role' agent in education; from the traditional conception of an effective teacher, e.g., transferring knowledge, to a more modern understanding of the qualities of an effective teacher holding more empowered roles in education.

5.1 Personal Traits

Personal characteristics were the first category reported in the findings in this chapter which included a number of distinct features such as friendliness, motivation, respectfulness, openness, etc. most of which corroborated those of the literature. However, there remains one central question which seems to take little heed in research on TER, i.e., the extent to which such traits can be engendered in teachers especially the novice ones. Whereas some personal traits are innate features of a teacher as a human being, there appears a concern as to whether intrinsic features such as friendliness can be learned or improved by teachers over time? However, as the literature suggests, teachers' personality and (inter)personal skills are important elements in effective teaching (e.g., Jones, Jenkin, & Lord, 2006; McBer, 2000). Hence, the obtained findings more or less corroborate those in the literature. For instance, teachers' enthusiasm for teaching and learning has been found to be crucial for effective teaching. (e.g., Minor, Onwuegbuzie, Witcher, & James, 2002; Stronge, 2007, p. 27). Therefore, it seems very unlikely to expect a teacher to be effective without the ability to demonstrate suitable interpersonal skills. As reported in the findings, qualities such as friendliness and respectfulness were found to command sheer support on the part of participant lecturers which simply testifies to their centrality to effective teaching.

In addition to the above-raised issues, there remains the complexity of drawing an operationalized definition of every personal characteristic proposed in this study and the literature. A review of the literature gives evidence to a dearth of specific benchmarks for personal attributes of effective teachers. Given the personal, individual and qualitative nature of these attributes, teachers are not likely to receive guidance or directive instruction on how much friendly, respectful, dedicated, open, etc. one should be to be considered as an effective teacher.

5.2 Teachers' Knowledge

Teachers' knowledge as a precious asset to effective teachers has been extensively researched especially in mainstream education (e.g., Christison & Murray, 2014; Murray & Christison, 2010, 2011; Shulman, 1987, 1999), even though it has been more or less researched in L2 education (e.g., Freeman & Johnson, 1998). The findings of this study as to the centrality of teacher knowledge to successful teaching echo those in the literature. As Stronge et al. (2004, p. 11) contend, knowledgeable teachers do better in making connections between the "real world" and the topics embedded in the curriculum. Not only is teachers' knowledge an essential factor in teacher evaluation but also an important criterion in effective teacher recruitment. As Jones et al. (2006, p. 119) argue, candidates with "good curriculum and subject knowledge" are more likely to be recruited.

Perhaps one of the most influential works on the teacher knowledge base is that of Shulman (1987, 1999) which addressed different categories of teacher knowledge base. As reported in the findings, teachers' knowledge explored in this study included teachers' language proficiency, subject matter knowledge, knowledge of curriculum and syllabus design, and finally ICT literacy. The reported findings in this study are in line with the literature. For instance, Campbell et al. (2004a, p. 40) maintain that teacher 'subject knowledge' impacts teacher effectiveness, even though impacts on teacher effectiveness and student achievement are rather mixed. Teachers' knowledge of curriculum is also in line with Connelly and Clandinin's (1988) conception of "teachers as curriculum planners". From among the four categories reported in the findings, one category seems to be yet a point of challenge in the context of this study, i.e., English language proficiency which will be discussed in this section.

English language proficiency was found to be an important dimension of teacher knowledge. The analysis of the interviews showed that lecturers were so keen on language proficiency maintaining that an effective teacher should have a good command of English. Such a stance is in line with Park and Lee's (2006, p. 236) study in which English proficiency was perceived by teachers as a characteristic of an effective English teacher and ranked the highest among three categories. As part of language proficiency, fluency is an extremely important function especially in the context of this study, for hardly can teachers attain the fluency needed for effective teaching. Such lack of knowledge might have its roots back in teachers' academic and pre-service education during which language proficiency might not have been given much attention. In other words, language teachers, by convention, are expected to show good command of English which is supposed to be obtained in their BA and MA programmes. Surprisingly, this is not the case for some EFL teachers who might feel less confident with their language proficiency. And herein we raise the question as to why language teachers do not gain solid knowledge of the English language and how can we expect a language teacher to be effective in L2 classrooms while having problems with English. Surely this would exert negative influences on one's confidence and hence classroom performance.

5.3 Pedagogical Skills

There has always been a question as to whether the findings of research into teachers in mainstream education context can be applied to the language education context. As Crandall (2000, p. 34) argues general education theory and practice have been influential on the direction of language teacher education which can be considered as a microcosm of teacher education. As a consequence, it is quite reasonable to conclude that the majority of pedagogical and professional skills perceived by EFL teachers corroborate those perceived by mainstream (general) teachers, given the fact that the theoretical and practical trends in language teacher education are much influenced by mainstream education.

As stated earlier, 'pedagogical skills' as one aspect of effective teaching received strong support from the participants of this study. One reason for such a demand might emanate from lecturers' lack of pedagogical knowledge. The need for the improvement of language teachers' pedagogical skills in universities in Iran is manifestly evident, for universities currently not only recruit lecturers with TEFL and pertinent backgrounds but also those with non-TEFL backgrounds such as Translation studies, English language literature, and Linguistics which bear little resemblance to TEFL. The latter groups of lecturers are expectedly not well informed about pedagogical skills during their academic education. These lecturers tend to have few opportunities to become familiar with principles and theories of language learning and teaching as compared to EFL lecturers even though they might have a native like and good command of English. As Llurda (2005, p. 144) contends, the mastery of language has nothing to do with pedagogical skills. Therefore, it appears that universities need to be more sensitive to teachers' pedagogical skills providing them with several opportunities both prior to and during their professional career. Solid pedagogical knowledge and skills will help teachers gain more confidence, an important factor which tends to be usually missing especially among novice teachers.

5.4 Professional Skills

Professional skills as "deep-seated patterns of behaviours" (McBer, 2000, p. 19) are a seminal dimension of teacher effectiveness. Whereas pedagogical skills can supposedly be taught during pre-service and in-service programmes (academic programme, initial teacher education programme, in-service teacher development programme), the question as to whether professional skills, similar to the pedagogical ones, can be taught and learned has been contentious. Moreover, a review of the literature shows that such skills have been given scant attention as compared to pedagogical (teaching) skills on which research has been done for a long time. In line with the literature, a review of the teacher development programmes simply substantiates the idea that policymakers and administrators have been more concerned with teachers' teaching skills rather than their professional lives. There is more to this than meets the eye, given the fact that not only effective teaching is bound with effective pedagogical skills but also the professional ones most of which have their roots in one's social life and hence may not be easily instructed. For instance, one emerged category was teachers' ability to overcome problems in the classroom.

Some of teachers' professional skills are beyond the very profession of teaching and can be applied to other jobs. For instance, the ability to overcome problems, as one professional skill, can be a desirable professional skill for a teacher, manager, engineer, etc. Therefore, administrators should pay special attention to this group of skills which need to be continually revisited. As Richards and Farrell (2005, p. 1) argue, teachers need "ongoing renewal of professional skills", given the fact that they are not necessarily provided with everything they need to know in their pre-service education. From among several professional skills reported earlier, some are of the utmost importance with reference to lecturers in the Iranian universities which will be discussed in this section.

Teachers' reflection was among the characteristics proposed by the participants as a quality of an effective teacher. Although this category was reported as a metacognitive category in the previous section, it is indeed a professional quality. It was very unfortunate that the notion did not receive much attention, despite its popularity in the literature. Indeed, research on teacher education, development, and evaluation is replete with ample evidence for the centrality of reflection on teachers' practices. However, it seems that reflection practice is not enshrined in the existing teacher education and teacher development programmes. Perhaps, this is why many participant lecturers did not pay much attention to such an important aspect of effective teaching.

It was found in this study that innovation as a professional characteristic received strong support with 100 % of the participants concurring with the idea. There could be a number of scenarios underlying such a strong demand on the part of lecturers. It appears that in line with the global movement in mass education and hence the consequent increase in the number of applicants for academic jobs in Iran, universities are adding new essential and desirable capabilities surplus to the requirements. Lecturers have entered a new intensified and tightened competition for recruitments for which they need to be accountable for new roles and duties, and be possessed of capabilities such as innovation, creativity, etc.

6 Theoretical and Pedagogical Implications

This study has brought to the fore some important theoretical and pedagogical implications. Nevertheless, due to the limitations and delimitations of the study especially the methodological ones, such implications should be approached with caution. Important amongst others is that the obtained findings are highly contingent upon teachers' views. Although teachers are said to be the most important stakeholders involved in the teaching phenomenon, it is a good idea to bring in the voices of other stakeholders such as administrators, students, etc. in which case a broader understanding of the notion of effective teacher could have been gained.

As to the theoretical and practical/pedagogical implications, this study has been able to contribute to the existing body of knowledge both in theory and practice. Overall, the drawn implications can be applied to three main parties involved in teacher evaluation, i.e., policymakers, administrators, and teachers. The question of 'what makes an effective teacher?' is a key component of teacher appraisal which is central to any given teacher evaluation system designed, developed, and imposed by policymakers mostly at the ministerial level in many countries including Iran. Thus the findings of this study can deepen policymakers' insights into the qualities and skills teachers need to be equipped with in order to teach effectively. And to this very end, such qualities need to be incorporated into national teacher appraisal schemes as standards for teacher effectiveness. This also meets the previously mentioned concern of a lack of transparency over the notion 'effective in terms of what?'

Teachers' dire call for sound pedagogical and professional skills simply indicated their longing for continual professional support on the part of administrators. As the findings suggested, it appears that teachers are not provided with what they need in their future career during their academic education. As Richards and Farrell (2005, p. 1) remind us, teachers need "an ongoing renewal of professional skills", on the grounds that teachers are rarely provided with everything they need to know in the pre-service stage. As a process in which teachers' cognitive and emotional assets are actively required, teachers' professional training may have different structures such as workshop, course (Avalos, 2011, p. 10), seminars, etc. As such, the responsibility of providing teachers with such developmental opportunities, i.e., initial teacher education programmes and teacher development programmes for pre-service and in-service teachers respectively, remain with administrators and teacher educators.

Research shows that since the outset of the new century, teachers have started to adopt more and more roles, responsibilities, and duties, e.g., teachers as researchers, as counsellors, as practitioners, etc. for which teachers need to constantly reflect on their own practices and improve their cognitive and metacognitive capabilities as well as their personal, pedagogical, and professional skills. Such reflective behaviour is indeed a formidable task of which teachers are deemed to be more cognizant. The findings suggested that an effective language teacher should exhibit all the qualities addressed in this study which are more or less emphasized in the literature. Nevertheless, the findings demonstrated that some qualities such as reflection which in Loughran's (2002, p. 33) words provides "ways of questioning taken-for-granted assumptions" are precious assets for language teachers to be more effective in their practices.

7 Conclusions and Recommendations for Further Research

This study addressed the rudimentary but nonetheless challenging question of 'what counts as an effective EFL Teacher', and with this end in view probed into the qualities and skills needed for effective language teaching. Following ample ideas and categories garnered in the findings, five major categories were proposed as indicators of effective teaching. These included personal attributes, cognitive, and metacognitive qualities, pedagogical, and professional skills. It is worth noting that each and every one of the aforementioned categories embraced a number of *specific* features which carry specific meanings and associated actions. Although some of the categories reiterate the existing ideas in the literature, the analysis of the subcategories brought into the fore new insights into the issue of the evaluation of teacher effectiveness in relation to the context of this study. This is important in that much of the awareness of what constitutes effective teaching needs to be interpreted with regard to the context within which teaching takes place.

As stated earlier, the findings in this study are heavily informed by teachers' views and perceptions. Although it is argued that teachers are the main and the most important stakeholders in teacher evaluation, it is wise to seek out the views of other stakeholders such as students and administrators, whether individually or collectively, for other parties might have their own contemplation and contentions upon the qualities of an effective EFL teacher. More interestingly, one can explore any similarities or contradictions among different stakeholders over the concept of 'effective teaching'. Such awareness is critically important especially in cases of discrepancies among different stakeholders, for any conflicts of interest may exert negative impacts on teachers' performance and hence that of the students.

References

Angelo, T. A. (1990). Classroom assessment: Improving learning quality where it matters most. *New Directions for Teaching and Learning, 1990*(42), 71–82. doi:10.1002/tl.37219904208

Angelo, T. A., & Cross, K. P. (1993). *Classroom assessment techniques: A handbook for college teachers*. San Francisco, CA: Jossey-Bass Publishers.

Arthur, W, Jr., Bennett, W, Jr., Edens, P. S., & Bell, S. T. (2003). Effectiveness of training in organizations: A meta-analysis of design and evaluation features. *Journal of Applied Psychology, 88*(2), 234–245. doi:10.1037/0021-9010.88.2.234

Avalos, B. (2011). Teacher professional development in teaching and teacher education over ten years. *Teaching and Teacher Education, 27*(1), 10–20. doi:10.1016/j.tate.2010.08.007

Bailey, K. M. (2006). *Language teacher supervision: A case-based approach*. Cambridge: Cambridge University Press.

Bakkenes, I., Vermunt, J. D., & Wubbels, T. (2010). Teacher learning in the context of educational innovation: Learning activities and learning outcomes of experienced teachers. *Learning and Instruction, 20*(6), 533–548. doi:10.1016/j.learninstruc.2009.09.001

Brosh, H. (1996). Perceived characteristics of the effective language teacher. *Foreign Language Annals, 29*(2), 125–136. doi:10.1111/j.1944-9720.1996.tb02322.x

Campbell, J., Kyriakides, L., Muijs, R. D., & Robinson, W. (2003). Differential teacher effectiveness: Towards a model for research and teacher appraisal. *Oxford Review of Education, 29*(3), 347–362. doi:10.1080/03054980307440

Campbell, J., Kyriakides, L., Muijs, D., & Robinson, W. (2004a). *Assessing teacher effectiveness: Developing a differentiated model*. London: Routledge.

Campbell, J., Kyriakides, L., Muijs, D., & Robinson, W. (2004b). Effective teaching and values: Some implications for research and teacher appraisal. *Oxford Review of Education, 30*(4), 451–465. doi:10.1080/0305498042000303955

Cheng, Y. C., & Tsui, K. T. (1998). Research on total teacher effectiveness: Conception strategies. *International Journal of Educational Management, 12*(1), 39–47. doi:10.1108/09513549810195893

Cheng, Y. C., & Tsui, K. T. (1999). Multimodels of teacher effectiveness: Implications for research. *The Journal of Educational Research, 92*(3), 141–150. doi:10.1080/00220679909597589

Christison, M. A., & Murray, D. E. (2014). *What English language teachers need to know (Vol. III): Designing curriculum*. New York, NY: Taylor & Francis.

Clayson, D. E. (2013). What does ratemyprofessors.com actually rate? *Assessment and Evaluation in Higher Education, 39*(6), 678–698. doi:10.1080/02602938.2013.861384

Connelly, F. M., & Clandinin, D. J. (1988). *Teachers as curriculum planners: Narratives of experience*. New York, NY: Teachers College Press.

Coombe, C., Al-Hamly, M., Davidson, P., & Troudi, S. (Eds.). (2007). *Evaluating teacher effectiveness in ESL/EFL contexts*. Ann Arbor, MI: University of Michigan Press.

Craig, C. (2010). Reflective practice in the professions: Teaching. In N. Lyons (Ed.), *Handbook of reflection and reflective inquiry* (pp. 189–214). New York, NY: Springer.

Crandall, J. (2000). Language teacher education. *Annual Review of Applied Linguistics, 20*, 34–55. doi:10.1017/S0267190500200032

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Boston, MA: Pearson Education.

Darling-Hammond, L. (2000). How teacher education matters. *Journal of Teacher Education, 51* (3), 166–173. doi:10.1177/0022487100051003002

Darling-Hammond, L., Newton, S., & Wei, R. (2013). Developing and assessing beginning teacher effectiveness: The potential of performance assessments. *Educational Assessment, Evaluation and Accountability, 25*(3), 179–204. doi:10.1007/s11092-013-9163-0

Department for Education and Employment (DfEE). (2000). *Research into teacher effectiveness, a model of teacher effectiveness. Report by Hay McBer to the Department for Education and Employment* (p. 67). London: DfEE.

Doyle, W. (1977). Paradigms for research on teacher effectiveness. *Review of Research in Education, 5*(1), 163–198. doi:10.3102/0091732x005001163

Ellett, C., & Teddlie, C. (2003). Teacher evaluation, teacher effectiveness and school effectiveness: Perspectives from the USA. *Journal of Personnel Evaluation in Education, 17*(1), 101–128. doi:10.1023/A:1025083214622

Farrell, T. S. C. (1999). Reflective practice in an EFL teacher development group. *System, 27*(2), 157–172. doi:10.1016/S0346-251X(99)00014-7

Farrell, T. S. C., & Jacobs, G. (2010). *Essentials for successful English language teaching*. London: Continuum.

Freeman, D. (1989). Teacher training, development, and decision making: A model of teaching and related strategies for language teacher education. *TESOL Quarterly, 23*(1), 27–45. doi:10.2307/3587506

Freeman, D. (2002). The hidden side of the work: Teacher knowledge and learning to teach. A perspective from north American educational research on teacher education in English language teaching. *Language Teaching, 35*(01), 1–13. doi:10.1017/S0261444801001720

Freeman, D., & Johnson, K. E. (1998). Reconceptualizing the knowledge-base of language teacher education. *TESOL Quarterly, 32*(3), 397–417. doi:10.2307/3588114

Garrett, R., & Steinberg, M. P. (2014). Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students. *Educational Evaluation and Policy Analysis,*. doi:10.3102/0162373714537551

Giroux, H. A. (2014). When schools become dead zones of the imagination: A critical pedagogy manifesto. *Policy Futures in Education, 12*(4), 491–499. doi:10.2304/pfie.2014.12.4.491

Good, T. L. (1979). Teacher effectiveness in the elementary school. *Journal of Teacher Education, 30*(2), 52–64. doi:10.1177/002248717903000220

Gregori-Giralt, E., & Menéndez-Varela, J. (2015). Validity of the learning portfolio: Analysis of a portfolio proposal for the University. *Instructional Science, 43*(1), 1–17. doi:10.1007/s11251-014-9327-4

Ingvarson, L., & Rowe, K. (2008). Conceptualising and evaluating teacher quality: Substantive and methodological issues. *Australian Journal of Education, 52*(1), 5–35. doi:10.1177/000494410805200102

Johnson, K. E. (2009). *Second language teacher education: A sociocultural perspective*. New York, NY: Taylor & Francis.

Jonassen, D., Mayes, T., & McAleese, R. (1993). A manifesto for a constructivist approach to uses of technology in higher education. In T. Duffy, J. Lowyck, D. Jonassen, & T. Welsh (Eds.), *Designing environments for constructive learning* (Vol. 105, pp. 231–247). New York, NY: Springer-Verlag.

Jones, J., Jenkin, M., & Lord, S. (2006). *Developing effective teacher performance*. London: Sage Publications.

Kelly, K. O., Ang, S. Y. A., Chong, W. L., & Hu, W. S. (2008). Teacher appraisal and its outcomes in Singapore primary schools. *Journal of Educational Administration, 46*(1), 39–54. doi:10.1108/09578230810849808

Koedel, C., & Betts, J. R. (2010). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the rothstein critique. *Education Finance and Policy, 6* (1), 18–42. doi:10.1162/EDFP_a_00027

Korthagen, F. A. J. (2004). In search of the essence of a good teacher: Towards a more holistic approach in teacher education. *Teaching and Teacher Education, 20*(1), 77–97. doi:10.1016/j.tate.2003.10.002

Kyriakides, L., Campbell, R. J., & Christofidou, E. (2002). Generating criteria for measuring teacher effectiveness through a self-evaluation approach: A complementary way of measuring teacher effectiveness. *School Effectiveness and School Improvement, 13*(3), 291–325. doi:10.1076/sesi.13.3.291.3426

Ladd, H. F. (2012). Education and poverty: Confronting the evidence. *Journal of Policy Analysis and Management, 31*(2), 203–227. doi:10.1002/pam.21615

Leigh, A. (2010). Estimating teacher effectiveness from two-year changes in students' test scores. *Economics of Education Review, 29*(3), 480–488. doi:10.1016/j.econedurev.2009.10.010

Llurda, E. (2005). Non-native TESOL students as seen by practicum supervisors. In E. Llurda (Ed.), *Non-native language teachers* (pp. 131–154). New York, NY: Springer.

Loughran, J. J. (2002). Effective reflective practice: In search of meaning in learning about teaching. *Journal of Teacher Education, 53*(1), 33–43. doi:10.1177/0022487102053001004

Mazandarani, O. (2014). *EFL Lecturers' perceptions of teacher effectiveness and teacher evaluation in Iranian universities.* Doctoral thesis. University of Exeter, Exeter.

McBer, H. (2000). Research into teacher effectiveness: A model of teacher effectiveness. Research Report No. 216. Nottingham: Department for Education and Employment, UK.

McBer, H. (2002). Teacher effectivness. Hay McBer report. In S. Hutchinson, B. Moon, & A. S. Mayes (Eds.), *Teaching, learning and the curriculum in secondary schools: A reader* (pp. 49–63). London: Routledge.

McKeachie, W. J., Lin, Y. G., & Mann, W. (1971). Student ratings of teacher effectiveness: Validity studies. *American Educational Research Journal, 8*(3), 435–445.

Meijer, P. C., Korthagen, F. A. J., & Vasalos, A. (2009). Supporting presence in teacher education: The connection between the personal and professional aspects of teaching. *Teaching and Teacher Education, 25*(2), 297–308. doi:10.1016/j.tate.2008.09.013

Minor, L. C., Onwuegbuzie, A. J., Witcher, A. E., & James, T. L. (2002). Preservice teachers' educational beliefs and their perceptions of characteristics of effective teachers. *The Journal of Educational Research, 96*(2), 116–127. doi:10.1080/00220670209598798

Muijs, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational Research and Evaluation, 12*(1), 53–74. doi:10.1080/13803610500392236

Murray, D. E., & Christison, M. A. (2010). *What English language teachers need to know (Vol. I): Understanding learning*. New York, NY: Routledge.

Murray, D. E., & Christison, M. A. (2011). *What English language teachers need to know (Vol. II): Facilitating learning*. New York, NY: Routledge.

Park, G.-P., & Lee, H.-W. (2006). The characteristics of effective english teachers as perceived by high school teachers and students in Korea. *Asia Pacific Education Review, 7*(2), 236–248. doi:10.1007/BF03031547

Patrick, J., & Smart, R. M. (1998). An empirical evaluation of teacher effectiveness: The emergence of three critical factors. *Assessment and Evaluation in Higher Education, 23*(2), 165–178. doi:10.1080/0260293980230205

Postareff, L., Lindblom-Ylänne, S., & Nevgi, A. (2008). A follow-up study of the effect of pedagogical training on teaching in higher education. *Higher Education, 56*(1), 29–43. doi:10.1007/s10734-007-9087-z

Richards, J. C. (1998). *Beyond training: Perspectives on language teacher education*. Cambridge: Cambridge University Press.

Richards, J. C., & Farrell, T. S. C. (2005). *Professional development for language teachers: Strategies for teacher learning*. Cambridge: Cambridge University Press.

Richards, J. C., & Nunan, D. (1990). *Second language teacher education*. Cambridge: Cambridge University Press.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review, 94*(2), 247–252. doi:10.2307/3592891

Ryans, D. G. (1949). The criteria of teaching effectiveness. *The Journal of Educational Research, 42*(9), 690–699. doi:10.1080/00220671.1949.10881737

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454–499. doi:10.3102/0034654307310317

Shinkfield, A. J., & Stufflebeam, D. L. (1996). *Teacher evaluation: Guide to effective practice*. Boston, MA: Kluwer Academic Publishers.

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1–23.

Shulman, L. (1999). Knowledge and teaching: Foundations of the new reform. In J. Leach & B. Moon (Eds.), *Learners and pedagogy* (pp. 61–77). London: Sage.

Stronge, J. H. (2007). *Qualities of effective teachers* (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

Stronge, J. H., Tucker, P. D., & Hindman, J. L. (2004). *Handbook for qualities of effective teachers*. Alexandria, VA: Association for Supervision & Curriculum Development.

Stronge, J. H., Ward, T. J., & Grant, L. W. (2011). What makes good teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education, 62*(4), 339–355. doi:10.1177/0022487111404241

Stronge, J. H., Ward, T. J., Tucker, P. D., & Hindman, J. L. (2007). What is the relationship between teacher quality and student achievement? An exploratory study. *Journal of Personnel Evaluation in Education, 20*(3–4), 165–184. doi:10.1007/s11092-008-9053-z

Teichler, U., & Kehm, B. M. (1995). Towards a new understanding of the relationships between higher education and employment. *European Journal of Education, 30*(2), 115–132. doi:10.2307/1503524

The Organisation for Economic Cooperation and Development (OECD). About. Retrieved March 3, 2015, from http://www.oecd.org/about/

Watzke, J. L. (2007). Foreign language pedagogical knowledge: Toward a developmental theory of beginning teacher practices. *The Modern Language Journal, 91*(1), 63–82. doi:10.1111/j.1540-4781.2007.00510.x

Zhu, X., & Zeichner, K. (2013). Preface. In X. Zhu & K. Zeichner (Eds.), *Preparing teachers for the 21st century* (pp. v–vii). London: Springer.

EFL Teacher Evaluation: A Theoretical Perspective

Khadija Alamoudi and Salah Troudi

Abstract In this chapter, we theoretically tackle the issue of evaluating English language teachers in educational institutions. In some parts of the world and out of all structural elements in institutes of education, teachers seem to receive the least amount of support and/or opportunities to assess their proficiency or to evaluate their teaching skills. Where teachers' performance evaluation does receive sufficient attention, cases sometimes can be expected where educators might be unaware of the latest up-to-date information concerning the methods that are utilized or the purposes behind the evaluation systems in their context and any policies related to it. The objectives of this chapter are to review and contribute to the current debates on the purposes and methods of evaluating language teachers' performance.

Keywords Teacher evaluation · Evaluation purposes · Evaluation methods · Language teacher evaluation

1 Introduction

Teachers are known to have a very significant influence on their students' achievement and to raise the interest of the pupils in the subject they are teaching. It is for that reason that high quality teaching is the goal of all language teachers. This might lead us to the importance of the evaluation of teachers and specifically EFL teachers in our case. Danielson (2001) assures that educators have realized that a well-designed system of evaluation is needed in order to improve their educational practices and to ensure a standard quality of teaching. Therefore, this chapter will be

K. Alamoudi (✉)
English Language Institute, King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: khaalamoudi@kau.edu.sa

K. Alamoudi · S. Troudi
Graduate School of Education, University of Exeter, Exeter, UK
e-mail: s.troudi@exeter.ac.uk

S. Hidri and C. Coombe (eds.), *Evaluation in Foreign Language Education in the Middle East and North Africa*, Second Language Learning and Teaching,
DOI 10.1007/978-3-319-43234-2_2

useful to understand the current situation of EFL teacher evaluation and to offer insights to improve the existing practice of the system of evaluation.

2 The Concept of Evaluation in Education

In the field of education, many attempts have been made to clarify the concept of evaluation and to distinguish between evaluation and other closely related concepts such as measurement or assessment. In the following section, we will provide some of the established definitions for the term "evaluation" and how it has emerged in the field of education in order to distinguish it from the other related concepts.

2.1 Evaluation: Operational Definitions

As an educational concept, evaluation has received much attention in the literature and many definitions have been provided in order to help people conceptualize this significant notion. Into the context of evaluation, Ralph W. Tyler, a leading figure in educational evaluation, associates evaluation with the concept of objectives. According to Tyler (1950), evaluation is "the process of determining to what extent educational objectives are actually being realized" (p. 69). His objectives model had a lasting impact on evaluation conceptions. However, the model was criticized for the inability to present a method to assess educational objectives themselves. Cronbach (1963, as cited in Verma & Malick, 1999), on the other hand, links evaluation to decision-making instead of objectives and defines it as "the collection and use of information to make decisions about an educational program" (p. 47). His work involves evaluation in three different layers of educational decisions: Administrative regulation, course improvement, and decisions about individuals.

Although Cronbach's definition seems effective in guiding decision making, his model was criticized for equating evaluation to only one of its various roles. Another definition is provided by Rossi, Lipsey and Ferma (2004) who identify the concept of evaluation as the "use of social research methods to systematically investigate the effectiveness of social intervention programs to improve social conditions" (p. 16). Having a systematic method of evaluation is vastly considered in their definition; however, it seems that the developmental-based approach would benefit the most out of their model of evaluation. The idea of the systematic tactic has been taken further by Patton (2008) who defines evaluation as the "systematic collection of information about the activities, characteristics, and results of programs to make judgments about the program, improve or further develop program effectiveness, inform decisions about future programming, and/or increase understanding" (p. 38). Patton has provided not only a systematic method in his definition but the definition also embodies an inclusive description for various purposes. While all the previously mentioned definitions differ in their details and the ways they conceptualized the term “evaluation”, the decision to choose one of the definitions may depend on some other important factors, such as the evaluation context, research questions, and the issues to be addressed.

This chapter is informed by the last definition offered by Patton (2008) for a number of reasons. First, the definition is comprehensive in the sense that it includes a variety of purposes. Second, Patton considers evaluation as a systematic way to collect information about different aspects.

2.2 *Emergence of Evaluation in Education*

In the field of education, it seems that there is a consensus that the history of evaluation began before the turn of the 20th Century (Glasman, 1986; Guba & Lincoln, 1981; Norris, 1990). Glasman argues that the history of educational evaluation can be divided into three distinct phases: The first continued until the 1930s, the second lasted until the 1960s and the third is still going on. It seems that expansion rather than substitution of the old ideas is the main characteristic of the development of educational evaluation throughout those three periods. Evaluation was seen first as measurement in education and the focus was initially on the level of intelligence measurement for learners and their ability to learn a specific subject (Glasman, 1986). Glasman claims that educational evaluation before the 1930s was used widely in the life and physical sciences. On the other hand, Guba and Lincoln (1981) argue that during the last decade of the ninetieth century, Joseph Rice who is known as the father of educational research devised some achievement tests supporting his debate about the insufficient use of school time. His published test in 1904 has become the base for almost all tests that measure intelligence since then. However, the publication of Fredrick Taylor’s *The principles of scientific management*, can be considered as the core effect of the ideas about standardization and systematization on industry which offers a systematic methodology for educational administration (Norris, 1990). Despite the fact that Ralph Tyler’s contribution in the field of educational evaluation in the 1930s keeps evaluation synonymized with measurement, he is regarded by many as the father of educational evaluation and the invention of the term “evaluation” was attributed to him (Norris, 1990). This idea was opposed by Guba and Lincoln (1981) who argue that Tyler’s method of evaluation has a distinctive advantage over the measurement-directed methods that were popular at that time. The reasoning in Tyler’s approach is systematic in nature. This can be true given that Tyler’s focus was on refining of programs and curricula in particular by means of examining educational objectives that can be considered as an essential impetus for evaluation.

2.3 The Changing Landscape of Teacher Evaluation

Medley, Coker and Soar (1984) briefly depict the teacher evaluation change of the twentieth century. They divide it into three main phases: (1) Questing for Great Teachers; (2) Determining the Quality of Teachers by Students' Learning; and (3) Observing Teaching Performance. In 1896, the issue of Great Teachers was evoked with a study conducted by Kratz who asked 2411 students in Iowa to define the features of the best teachers (Medley, 1979). Kratz was thinking of establishing a benchmark that all teachers can be judged against. In his study, "helpfulness" was labelled as the most significant characteristic of a great teacher and "personal appearance" was reported as the next important feature. This can be accepted if one just considers the students apart from other methods when evaluating teachers. That idea was not accepted by Barr (1948) who claimed that supervisors' assessment of teachers was the actual choice metric. However, some researchers started to examine student achievement and use students' learning to infer about teacher quality assuming that supervisors' opinions of teachers do not reveal anything about students' learning. For instance, Domas and Tiedeman's (1950) review of more than 1000 studies of teacher characteristics indicated that for evaluators, there is no clear direction. The notion of using students' achievement to evaluate their teachers was, however, rejected by Getzels and Jackson (1963) who argue that many of the tests were inappropriate to address the effectiveness of teachers. Medley, Coker and Soar (1984) support this opinion claiming that students' achievement may vary and achievement tests can be poor measures of the success of the students themselves. This is true especially because students' achievement can be linked to a wide range of distinct considerations.

The era of Observing Teaching Performance focused on detecting effective teachers' behaviours that cause student learning. Brophy and Good (1986) argue that learners who receive quality instruction by their teachers achieve more than those who work independently or receive poor instruction. Clark and Peterson (1986) do not only concur with this view but also go further claiming that good teachers tend to adapt their instructions to their students' needs. However, Powell and Beard (1984) argue that subjective judgment can be found when comparing one domain in teacher performance to another. Their bibliography of teacher evaluation research between 1965 and 1980 remains a valuable reference. From the time when it was first commenced until recently, teacher evaluation based on teacher performance has gone through different changes and many concerns have been detected "including evaluation inflation, highly subjective instruments, and a lack of objective measures" (Nagel, 2012, p. 33). Noticeably, the previous overview reflects that despite the fact that there are many methods to assess the quality of teachers; each one has its own limitations. Notwithstanding the restraining factors, the fact may remain that better student learning can be a result of effective instruction (Darling-Hammond, 2000).

3 Why Conduct Evaluation of Teachers?

According to McGreal (1983), evaluation is expected to serve two fundamental needs: Accountability (summative evaluation) and improvement (formative evaluation). The push for both accountability and improvement has resulted in supervision relying on integrated models of formative and summative evaluation (Gullat & Ballard, 1998, p. 16). However, both purposes of teacher evaluation cannot be satisfied by only one system (Towe, 2012). If one system is claimed to satisfy both purposes, one of them is expected to have more weight than the other. Danielson and McGreal (2000) argue that formative evaluation is conducted with the importance placed on teacher improvement, growth, and development. In line with this, Bailey (2007) argues that formative evaluation is conducted mainly to offer feedback or for the purpose of improvement. It might be claimed, then, that formative evaluation can be used to feed professional development decisions. Peterson (2000) supports this and claims that formative assessment data may be used as feedback to shape performances, build new practices or alter existing practices. Summative evaluation, on the other hand, is the summary of evaluation that serves decision-making. Its focus is on ranking, rating, and making judgments about the adequacy of teachers' performance (Danielson & McGreal, 2000). Bailey (2007) argues that the results of summative evaluation help to determine if the funding is going to be continued. Summative evaluation to her is "a final assessment, a make-or-break decision at the end of a project or funding period" (p. 184). However, teachers are not often directly involved in this kind of evaluation.

According to Daresh (2001), a diagnostic evaluation can be considered as a third purpose for teacher evaluation. According to him, this type of evaluation is used to "determine the beginning status or condition (…) prior to the application or intervention or treatment" (p. 281). As such, Bailey (2007) argues that before any attempt to change and in order to provide data about the current status, diagnostic evaluation can be carried out. She also claims that it seems sensible to start with a diagnostic evaluation, followed by systematic formative evaluation, and then a summative evaluation can be conducted after an extended period of formative evaluation. As a sequence, this seems to be logically adequate, however, all three types can be given a different amount of attention and significance depending on the context, objectives, and the rationale of the evaluation system adopted in the educational institution.

4 How to Evaluate Teachers?

In the wide range of literature on teacher evaluation, there have been various methods to evaluate teachers, such as student ratings, peer observation, self-evaluation, and teaching portfolio. In the following section, we will present some of them. They will be presented randomly so that the order does not indicate priority or significance of one of them over the others. Yet, it depends on the educational institution's needs and characteristics to adopt one or more of them to satisfy the purpose or purposes of EFL teacher evaluation in that particular institution.

4.1 Student Ratings

Student ratings are commonly used to evaluate the performance of teachers. Seldin (2006) argues that it is expected that everyone thinks the ratings of students are all that we need to evaluate the effectiveness of teaching. It may be widely known that students as the product of educational systems have a very close and extended interaction with their teachers; hence, their judgment can be valuable and genuine. Despite the fact that students are seen as a significant source to evaluate the performance of a teacher in a wide range of educational institutes, their ratings as a tool have their own limitations.

Most of the students might not be well prepared nor have enough experience that enables them to evaluate their teachers. Accordingly, they might concentrate on the teacher's personality and give it more attention than academic and teaching skills. Arreola (2007) argues that students in compulsory maths and science courses tend to rate teachers harshly. In line with this argument, students might tend to evaluate EFL teachers harshly when English language is compulsory. In such a case, student ratings can be more beneficial for professional development programs. Accordingly, inclusive evaluation systems will need to consider research findings before counting on student ratings solely. In their study on Japanese university students rating of teaching, Burden and Troudi (2010) support this view and argue that other evaluative methods, such as self-evaluation could be introduced in order to encourage more professional development input.

4.2 Peer Observation

Peer observation can be a useful tool to reflect on the performance of teachers inside their classrooms. It can be more precise, objective, professional and effective than student ratings to develop the instructional practice at educational institutions. Teachers may make use of checklists and forms for peer review that are provided in Braskamp and Ory (1994), Chism (1999), and Weimer, Parrett, and Kerns (2002). Seldin (2006), however, suggests three phases for peer observation; pre-visit consultation where visitor reviews the syllabus and other relevant materials, the visit itself where the visitor observes the performance of the teacher, and the follow-up visit where both of them discuss ideas and observations.

Arguably, serious weaknesses might be highly related to peer observation as a method to evaluate teachers' performance. For instance, how can one make sure that the piece of teaching that is being observed is representative of everyday practice? Another drawback could be the presence of the observer him/herself that can affect or even alter the class environment and disturb learning. In an attempt to help address previous disadvantages, Arreola (2007) argues that scheduling multiple visits, training peer observer teams, preparing the students, preparing the instructors, and scheduling a post-observation conference might be useful.

4.3 Self-evaluation

Though not widely used as a method for teacher evaluation in many educational systems, self-evaluation can be a good method to evaluate teachers. Teachers themselves perceive the lack of self-evaluation as a weakness in any teacher evaluation system (Towe, 2012). Self-reflection can be significant, since teachers are able to analyse their own instructional practices, which will help towards their professional growth. A major criticism of this evaluative method might be that teachers tend to give themselves higher ratings than they deserve. Besides, this method cannot be used in decisions like promotion (Centra, 1980). Brandt (2010) partly supports this argument and claims that “self-evaluation is a formative, not a summative, activity” (p. 208). This might be true, yet teachers can be more aware than anyone else about their own contributions and hence might be better able than others to annually report their own progress.

4.4 Teaching Portfolio

Teaching portfolios can be used as a means to collect materials and to provide evidence and documents showing the teaching effectiveness of the teacher. Portfolios may also reflect the individuality of teaching. Seldin (2006) argues that “developing a teaching portfolio allows the faculty member to connect theory with practice” (p. 114), which provides the teacher with “a natural outcome of improvement” (p. 114). The key problem with this approach is that it depends on how teachers present their work in the portfolio; accordingly, a very high trust level is required between teachers and principals (Arreola, 2007). In fact, teachers should be trusted especially in reporting and documenting their own work for appraisal purposes in order to have better and more effective teaching. It might be argued that portfolios have more advantages over observation since they represent larger accounts of teaching, yet they might be seen as difficult to deal with from an evaluators’ point of view (Alwan, 2010). Accordingly, there needs to be clear criteria and standards to construct portfolios effectively.

To conclude this section on the methods of teacher evaluation, a comprehensive inclusive teacher evaluation system in any educational system needs to consider all the above-mentioned methods along with others (if needed depending on the purpose of the evaluation) in order to have an adequate evaluative tool. Multiple sources of teacher evaluation techniques can be very useful for principals and administrators of educational institutes to evaluate, improve, and enhance the effectiveness of their teachers. Consequently, a well-designed teacher evaluation system is expected to identify the features of effective teaching and to allocate their effective teaching criterion and accordingly develop the outcome of the whole educational institute. When taking EFL teacher evaluation into account, special concerns may arise for both evaluators and teachers who are being evaluated. The following section will highlight some of the major issues related to EFL teacher evaluation and special attention will be directed to the context of higher education.

5 EFL Teacher Evaluation

In their conceptual articles, Brown and Crumpler (2013) claim that there is no agreement as to what makes any assessment method effective. This problematic issue affects the teachers of foreign languages in particular. They argue that foreign language teacher evaluation has more challenges, especially for the evaluators who do not have sufficient knowledge about second language acquisition and the case becomes worse when those evaluators do not speak the target language that is used inside the classroom. Despite the fact that in this case it is challenging to judge the content knowledge of the teacher and the degree of students' understanding, principals very frequently observe teachers' performance in foreign language classrooms using checklists that contain the content knowledge of the foreign language teacher as one of the criteria for teacher performance assessment. For this particular reason, Brown and Crumpler (2013) call for a change in foreign language teacher evaluation.

Brown and Crumpler (2013) developed a model that positions assessment of peers at top priority of foreign languages instructors' evaluation to shift evaluation towards more learning and progression. Their assessment portfolio model, in Fig. 1, offers an inclusive and wide-ranging instructor's performance assessment that is informed by "multiple sources of evidence, which leads to a more complete and authentic evaluation" (p. 145). They also argue that due to their busy schedules, administrators cannot supervise and evaluate foreign language teachers properly. In fact, their model can be seen as adequate in contexts where self-assessment, as a method for teacher evaluation, is marginalized since this model overlooks the self-evaluation of a foreign language teacher where the teacher him/herself diagnoses his/her teaching in an attempt to improve the quality of his/her own performance.

In an attempt to investigate the main criteria of in-service English language teachers' evaluation, Akbari and Yazdanmehr (2011) conducted an exploratory study in five private language institutes in Iran. Interviews with the supervisors along with analysis of application forms, observation sheets and other relevant documents illuminated the procedures and criteria of teacher assessment in the

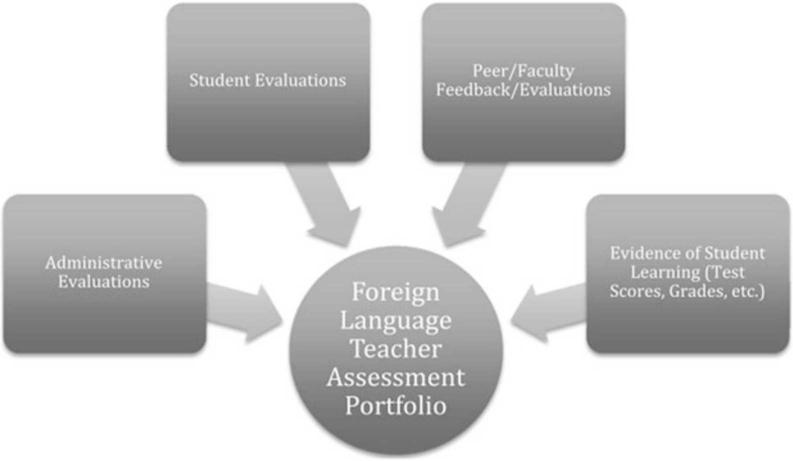

Student Evaluations
Peer/Faculty Feedback/Evaluations
Administrative Evaluations
Foreign Language Teacher Assessment Portfolio
Evidence of Student Learning (Test Scores, Grades, etc.)

Fig. 1 Brown and Crumpler's (2013) model

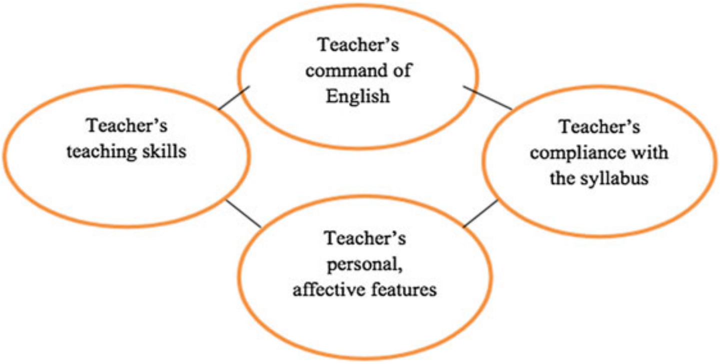

Teacher's command of English
Teacher's teaching skills
Teacher's compliance with the syllabus
Teacher's personal, affective features

Fig. 2 Akbari and Yazdanmehr's (2011) model

target setting. Their procedures in assessing in-service English language teachers' performance are categorized into four groups of teacher's command of English, teaching skills, compliance with the syllabus and personal/affective features. The model they developed exclusively for English language teachers is presented in Fig. 2.

In Akbari and Yazdanmehr's (2011) model, the teacher's command of English involves: Accuracy of speech, structure, pronunciation, and performance in discourse along with fluency in speech. Personal/affective features include: Punctuality, rapport with learners, tolerance in error treatment, enthusiasm and dynamism in involving learners. Teacher's compliance with the syllabus comprises: Expected content to be covered, educational goals to be achieved, and the way to present the material to be followed. Teaching skills involve: Communication skills, classroom management techniques and task management. Their model might be seen as distinctive and uniquely designed for EFL teachers, however, it does not consider other social and administrative skills besides community service and research activities that might be essential parts of EFL faculty members' activities that should not be overlooked.

By surveying 457 post-secondary foreign language teachers, Bell (2005) examines teacher perceptions on the teaching attitudes and behaviours contributing to effective foreign language teaching. Her study demonstrated a strong positive agreement on all five standards for foreign language teaching. Other categories that teachers agree with the majority of items include: Qualifications of teachers, general theories related to the communicative approach to foreign language teaching, the significance of small group activities, and negotiation of meaning and strategies in foreign language classes. In fact, the study is more concerned with the teachers' behaviour and attitude towards aspects highly related to language acquisition rather than contributing to the effectiveness of foreign language teachers and teaching.

Brown (2006) investigates students and teachers' perceptions of effective teaching in foreign language classrooms that, he argues, are distinctive from other subjects. The findings are the result of analysing a questionnaire distributed amongst 49 university teachers and 1400 of their students. From the teachers' perspective, engaging students in information gap activities, assessing group tasks, being as knowledgeable about culture as language, and having students respond to physical commands are the main characteristics of effective foreign language teachers. Concerning students' opinions, correcting oral errors indirectly, being as knowledgeable about culture as language, having students respond to physical commands, addressing errors with immediate explanation, presenting grammar with real-world context, speaking with native-like control of language, using real-life materials in teaching language culture, and engaging students in information gap activities are the most prominent features of effective foreign language teaching. Arguably, Brown's study can be seen as much concerned with instructional practices and disregarded the other areas that can be used to evaluate language teachers.

Al-Hammad (2011) conducted a study aiming at examining the teaching performance level of 18 English language teachers from the intermediate-level schools in the city of Hail, Saudi Arabia, according to the teaching quality standards. By employing an analytical and descriptive approach, Al-Hammad utilized a controlled observation method on teaching standards and found that the use of teaching aids, and class management skills were highly achieved. In that study and within her sample, students' assessment and lesson delivery were satisfactorily accomplished. Lesson planning was, however, the lowest quality standard achieved by her participants. Al-Hammad's study reinforces the importance of conducting in-service training sessions on teaching quality standards for English language teachers mainly in the three dimensions of: Planning, implementation and assessment. Despite the fact that the sample was solely English language teachers, the dimension and the criteria were not subject-content oriented and could be applicable to teachers of any other subjects. In the Gulf context Al Mahrooqi, Denman, Al-Siyabi, and Al-Maamari (2015) compared Omani school students and teachers' perceptions of the characteristics of good EFL teachers. One hundred and seventy-one Omani students and 233 English teachers took part in the study which showed general agreement between students and teachers about the importance of all characteristic categories, with special importance to English language proficiency and equality in treating students.

6 Summary and Implications

In this chapter, we have shown how EFL teacher evaluation might be different from other types of teacher evaluation. Purposes and methods may look the same for all subject teachers; however, when conducted for language teachers, evaluators need to consider the uniqueness and the nature of the subject matter. Different concerns for EFL teacher evaluation have been discussed separately and need to be taken into account before determining and constructing academic systems to evaluate language teachers.

References

Akbari, R., & Yazdanmehr, E. (2011). EFL teachers' recruitment and dynamic assessment in private language institutes of Iran. *Journal of English Language Teaching and Learning, 8*, 29–51.

Al-Hammad, Y. (2011). *Teaching performance assessment for the intermediate school English teachers in the city of Hail as per the Teaching Quality Standards.* Master's Thesis, Imam Muhammad Bin Saud Islamic University, Saudi Arabia.

Al-Mahrooqi, R., Denman, C., Al-Siyabi, J., & Al-Maamari, F. (2015). Characteristics of a good EFL teacher: Omani EFL teacher and student perspectives. *SAGE Open*, pp. 1–15.

Alwan, F. (2010). Teacher voice in teacher evaluation: Teaching portfolios in the United Arab Emirates. In C. Coombe, M. Al-Hamly, P. Davidson, & S. Troudi (Eds.), *Evaluating teacher effectiveness in ESL/EFL contexts* (pp. 232–244). Michigan: The University of Michigan Press.

Arreola, R. (2007). *Developing a comprehensive faculty evaluation system: A guide to designing, building, and operating large-scale faculty evaluation systems*. San Francisco: Jossey-Bass.

Bailey, K. (2007). *Language teacher supervision: A case-based approach*. Cambridge: Cambridge University Press.

Barr, A. (1948). The measurement and prediction of teaching efficiency: A summary of investigations. *Journal of Experimental Education, 16*(4), 203–283.

Bell, T. (2005). Behaviours and attitudes of effective foreign language teacher: Results of a questionnaire study. *Foreign Language Annals, 38*(2), 259–270.

Brandt, C. (2010). Giving reflection a voice: A strategy for self-evaluation and assessment in TESOL teacher preparation. In C. Coombe, M. Al-Hamly, P. Davidson, & S. Troudi (Eds.), *Evaluating teacher effectiveness in ESL/EFL contexts* (pp. 199–212). Michigan: The University of Michigan Press.

Braskamp, L., & Ory, J. (1994). *Assessing faculty work: Enhancing individual and institutional performance*. San Francisco: Jossey-Bass.

Brophy, J., & Good, T. (1986). Teacher behaviour and student achievement. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328–375). New York: Macmillan.

Brown, A. (2006). *Students' and teachers' perceptions of effective teaching in the foreign language classroom: A comparison of ideals and ratings.* Doctoral Dissertation, Graduate College, University of Arizona.

Brown, I., & Crumpler, T. (2013). Assessment of foreign language teachers: A model for shifting evaluation toward growth and learning. *The High School Journal, 96*(2), 138–151.

Burden, P., & Troudi, S. (2010). An evaluation of student ratings of teaching in a Japanese university context. In C. Coombe, M. Al-Hamly, P. Davidson, & S. Troudi (Eds.), *Evaluating teacher effectiveness in ESL/EFL contexts* (pp. 152–166). Michigan: The University of Michigan Press.

Centra, J. (1980). The how and why of evaluating teaching. *Engineering Education, 71*, 205–210.

Chism, N. (1999). *Peer review of teaching: A sourcebook*. Bolton: Anker.

Clark, C., & Peterson, P. (1986). Teachers' thought process. In M. Wittrock (Ed.), *Handbook of research on teaching* (pp. 255–296). New York: Macmillan.

Danielson, C. (2001). *New trends in teacher evaluation: Educational leadership*. Alexandria, VA: Association of Supervision and Curriculum Development.

Danielson, C., & McGreal, T. (2000). *Teacher evaluation to enhance professional practice*. Alexandria, VA: Association of Supervision and Curriculum Development.

Daresh, J. (2001). *Supervision as proactive leadership*. Prospect Heights, IL: Waveland Press.

Darling-Hammond, L. (2000). Teacher quality and student assessment: Are view of state policy evidence. *Educational Policy Analysis Archives, 5*(1), 1–44. doi:http://dx.doi.org/10.14507/epaa.v8n1.2000

Domas, S., & Tiedeman, D. (1950). Teacher competence: An annotated bibliography. *Journal of Experimental Education, 19*(99), 101–218.

Getzels, J., & Jackson, P. (1963). The teacher's personality. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally.

Glasman, N. (1986). *Evaluation-based leadership: School administration in contemporary perspective*. Albany: State University of New York.

Guba, E., & Lincoln, Y. (1981). *Effective evaluation: Improving the usefulness of evaluation results through responsive and naturalistic approaches*. San Francisco: Jossey-Bass Inc.

Gullatt, D., & Ballard, L. (1998). Choosing the right process for teacher evaluation. *American Secondary Education, 26*(3), 13–17.

McGreal, T. (1983). *Successful teacher evaluations*. Alexandria, VA: Association of Supervision and Curriculum Development.

Medley, D. (1979). The effectiveness of teachers. In P. Peterson & H. Walberg (Eds.), *Research on teaching concepts, findings and implications* (pp. 11–27). California: McCutchan Publishing Corporation.

Medley, D., Coker, H., & Soar, R. (1984). *Measurement-based evaluation of teacher performance: An empirical approach*. New York: Longman.

Nagel, C. (2012). *Teachers' perceptions regarding portfolio-based components of teacher evaluation*. Dissertation. Illinois State University.

Norris, N. (1990). *Understanding educational evaluation*. London: Kogan Page Ltd.

Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). London: Sage Publication Inc.

Peterson, K. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices*. Thousand Oaks, CA: Corwin.

Powell, M., & Beard, J. (1984). *Teacher effectiveness: An annotated bibliography and guide to research*. New York: Garland.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7th ed.). London: Sage Publication Inc.

Seldin, P. (2006). *Evaluating faculty performance: A practical guide to assessing teaching, research, and service*. San Francisco: Jossey-Bass.

Towe, P. (2012). *An investigation of the role of a teacher evaluation system and its influence on teacher practice and professional growth in four urban high schools*. Dissertation. Seton Hall University.

Tyler, R. (1950). *Basic principles of curriculum and instruction*. Chicago: University of Chicago Press.

Verma, G., & Malick, K. (1999). *Researching education: Perspectives and techniques*. London: Falmer Press.

Weimer, M., Parrett, J., & Kerns, M.-M. (2002). *How am I teaching? Forms and activities for acquiring instructional input*. Madison: Atwood.

Faculty Performance Evaluation and Appraisal: A Case from Syria

Anas Al-Fattal

Abstract The purpose of this chapter is to broaden our understanding about the issue of performance evaluation and appraisal in the particular context of higher education. Several theoretical issues are discussed, presented and reflected on in a practical case from Syria. The study employs a case study methodology for one of the private universities in Syria. Interviews, documentary analysis and field observation are the data collection techniques utilized. The interviews are semi-structured triangulating findings from three groups of respondents. The findings highlight that several evaluation methods are used. Judgmental evaluation methods are more commonly used in the context of private higher education in Syria. The findings help provide useful insights about the practices related to performance evaluation in that case. Further research could investigate the issue in a wider population including more case studies. People responsible for human resources and performance need to pay attention to standardizing evaluation procedures. They should also empower academic members of staff to be more aware of performance evaluation and its purposes and processes. There is a lack of research on performance evaluation in higher education contexts in Syria which makes this study of key importance to the literature.

Keywords Performance evaluation and appraisal · Human resources management · Syria · Private higher education · Evaluation methods

1 Introduction

The issue of performance evaluations is rapidly gaining significance for many businesses and organizations. Such importance is also witnessed in educational institutions especially those implementing strategies and techniques from the business field (Al-Fattal & Ayoubi, 2012). There has been an urgent need for

A. Al-Fattal (✉)
College of Banking and Financial Studies, Muscat, Oman
e-mail: anasfat@hotmail.com

S. Hidri and C. Coombe (eds.), *Evaluation in Foreign Language Education in the Middle East and North Africa*, Second Language Learning and Teaching,
DOI 10.1007/978-3-319-43234-2_3

several higher education institutions to adopt and embrace such strategies and techniques in order to correspond to the pressure placed on them by officials and their quality concerns or university rating systems. For example, several quality management systems, such as the well-known ISO (International Standardization Organization), require an institution to evaluate performance. When talking about performance evaluation, there are different levels of evaluation, e.g., organizational, personnel or staff performance. This chapter focuses on evaluating staff performance and more particularly academic members of staff.

There is a current noteworthy trend for educational institutions in the Arab world to improve and develop their teaching staff's performance. Such a trend is powered by several factors, e.g., a desire to improve position or rank on international indices such as the Shanghai Index. Another factor is improving position in local and regional markets in order to attract more or better students (Al-Fattal & Ayoubi, 2013). This, perhaps, is more evident in the private or for-profit institutions. Besides, the fact that educational institutions are service industries (Kotler & Fox, 1995) which place a critical role for its personnel to reflect on the overall success of these organizations. This makes most institutions aware of the importance of investing in their human resources. It has been noticed that most educational institutions in the Arab world import ready-made evaluation and appraisal systems from the West. There is an urgent need to test the applicability of such systems on the Arab culture. Finally, there is a paucity of similar peer-reviewed published research in the Arab World. This chapter is aimed at bridging this particular gap.

The chapter starts by defining performance evaluation and appraisal and linking it to the educational context. The importance of performance evaluation in the educational context is also presented. Then the complications surrounding performance evaluation in education are discussed. This is followed by some discussions on the evaluation process and data collection tools and methods. The final part of the chapter presents an empirical case study on performance evaluation from a private university in Syria.

2 Performance Evaluation and Appraisal

Appraising and evaluating performance is believed to have started early in history yet through informal or unsystematic methods. Khanka (2003) highlights that during World War I the concept started to develop where military people were assessed for merit rating purposes. It is believed that the concept started establishing formal grounds in the business domain in Japan with the quality revolution during the fifties and sixties (Bernardin, 2010). Since then the concept has gone through substantial changes and developments and has become a significant strategic option for organizational improvement and success.

The literature on human resource management provides a number of definitions for performance evaluation and appraisal. There have many attempts to define the term. It is defined as the process of measuring what employees contribute to the organization (Stewart & Brown, 2009). It is the action of determining an employee's work and outcomes in relation to the job in a particular setting. This definition, though comprehensible, lacks depth and purpose. Khanka (2003), on the other hand, defines the term as a systematic and objective assessment of an individual's performance in order to assess the changing needs, potential for promotion, or salary review. It is a way of judging the relative worth or ability of an employee in performing their tasks.

Performance evaluation and appraisal normally involves assessing how an individual employee is doing their job against a set of criteria or standards, e.g., personal competencies, behavioural characteristics or achievements. It also involves providing feedback and creating a development plan. It generates information that may be used for several organizational purposes: (1) Administrative purposes such as rewards, promotions, transfers and terminations, (2) developmental purposes such as training and development, coaching and career planning and possibly (3) research purposes such as validation selection procedures and evaluating the effectiveness of training (Stone, 2011). The evaluation and appraisal process is summarized by Gomez-Mejia, Balkin, and Cardy (2010) as having three categories: Identification, measurement and management of human performance in an organization. In this regard, performance evaluation and appraisal is not limited only to identifying and measuring performance as highlighted earlier by Stewart and Brown (2009); it also reflects on and benefits managerial practices and decisions.

3 Importance of Performance Evaluation and Appraisal

There are several reasons to evaluate employees' performance and the level of competency with which they are performing their jobs. The importance of performance evaluation and appraisal is shown in its uses and applications in an organization. Lunenburg and Ornstein (2004, p. 596), for example, highlight some uses that are related to institutional *effectiveness and efficiency*. Institutions need to check not only their own effectiveness and efficiency but also their employees'. The latest global financial recession and crises has reflected dramatically on many educational institutions. This is most evident at financially independent educational institutions, e.g., British universities or Middle Eastern private universities. With the financial problems and shortage of income at hand, institutions are being encouraged to perform more effectively and efficiently. For example, and as a result, some private institutions adopted a pay-on-merit system (Podgursky & Springer, 2006) where teachers are offered more courses and higher salaries based on the income they generate for their institutions.

Other imperatives for performance management and evaluation relate to supervising employees and *improving their performance*. Evaluation data, for example, might highlight areas of weakness, whether individual or institutional. Particular attention and professional development could be focused on these weaknesses. One strategy to ensure better performance suggests that the process of evaluation should start from the early stages where job performance is usually discussed with an employee during a one-to-one meeting. This could provide a useful opportunity for verbal communication between an institution and its employees. Through these meetings, goals are set for both the employees and the institution. Goals regarding employees are those related to professional development where they usually aspire to develop their skills and competences. Regarding institutional goals, performance evaluation and management establishes objectives for contributing to the departmental and institutional mission. It is also important that an employee's professional development goals and the institutional and departmental ones are geared together for the betterment of the organization and possibly the community. Further and more detailed discussion on the process of conducting evaluation and appraisal is provided later in this chapter.

Other uses for performance evaluation and appraisal relate to *attitudes* about work. Gomez-Mejia et al. (2010) believe that in order to improve employees' job-related attitudes they need first to know where they are in terms of individual performance; only then they can know where they want to go in developing their performance. Lunenburg and Ornstein (2004, p. 596) add that such benchmarking and measurement for performance is usually used to *motivate employees*. It is believed that by doing so as it can encourage competition among staff; it stretches goals to foster innovation (Coleman & Glover, 2010). A commonly used practice in this regard is the 'teacher of the year' award. Some institutions might offer financial incentives for their teacher of the year, yet it is usually the status and emotional uplift that have the higher motivational impact. The 'teacher of the year' initiative has received major criticism especially if wrongly used. The complications and criticism against teacher performance evaluation and appraisal is further discussed in the following section.

Other uses and reasons to evaluate performance relate to the human resources management practices. In other words, data provided from the evaluation aid the human resource or personnel department in *making decisions* (Bernardin, 2010). For example, directors might need information to help them decide who to promote. Other more strategic decisions might include those of staffing or training.

4 Performance Evaluation Complications

There is a considerable amount of criticism of and even *opposition* to the application of performance evaluation and appraisal at educational institutions and more particularly to the teacher's job (Al-Fattal, 2011; Gallagher, 2004; Milanowski, 2004; Ramsden, 1991). A number of ideas support this position. Opposition probably stems from the *nature of the job* and the kind of accountability, liability and authority teachers might enjoy (Avalos & Assael, 2006). This, perhaps, is more evident in Middle Eastern contexts and cultures where teachers and, more specifically, academics enjoy higher hierarchal social positions and their jobs are 'not to be questioned'. For example, several educational institutions in this context are unable to perform such evaluations due to the severe opposition from their academic members of staff. In this instance, opposition occurs from the bottom of the organization (Stone, 2011). However, opposing the evaluation sometimes stems from the top. Perhaps, as members of the top management at educational institutions are educators in the first place, they might carry some prejudice against teacher performance evaluation.

Another complication regarding teacher performance evaluation is related to *employee reaction*. Performance evaluation is a multi-purpose process and these purposes often conflict, probably resulting in the prevention of the evaluation process from achieving its goals and benefits to the institution (Boswell & Boudreau, 2000). One major purpose of the evaluation is to improve performance; nonetheless, evaluation could work against this creating a negative reaction for employees and counterproductive performance (Stewart & Brown, 2009). Among possible undesired reactions, an evaluated employee might react in a relatively defensive or aggressive manner. Such an attitude might be created as an employee might feel s/he is a victim of the higher authorities and being exposed to criticism or even abuse. Middlewood and Cardno (2001) comment that the evaluation process is a sensitive one and if it is not carried out properly it possibly reflects negatively on the institution. Teacher reaction about performance evaluation could also have some *cross-cultural complications*. Academics from different cultures might have different perceptions and attitudes towards performance evaluation. For example, in some cultures, it is not acceptable for a younger person to evaluate an older or more experienced one. It should also be mentioned that evaluation practices are more accepted and used in the West. This could be a result of the long experience in such practices there.

Performance evaluation in other businesses seems less *problematic* than it is in education. It is much more straightforward and easier to evaluate other professionals than it is with teachers. For example, a sales person might be evaluated based on the number of items sold or the revenues s/he generates for the company; a factory worker is evaluated based on the quantity and quality of items s/he produces. In the business of education such measurement is, unfortunately, not possible; the sophisticated nature of the education business presents further complications to the evaluation process. Defining good teaching and a good teacher is not straightforward and has always been a matter of debate (Middlewood & Cardno, 2001). For example, some educators consider teaching as an art and others as a profession. One of the possible complications for this issue results in creating the evaluation forms and measures discussed in a later section. The *complexity* of defining good teaching and the good teacher also relates to the earlier complication regarding culture where different cultures might have different perceptions and attitudes to the teaching job. Different cultures have different expectations of a teacher. Some cultures, for example, perceive good teaching to be the ability to make students achieve higher scores in examinations and other forms of formal assessment; other cultures perceive good teaching to be the ability to develop creative and independent skills for the learners; and other cultures perceive good teaching to be the ability to control students and convey discipline in class

(Al-Fattal, 2010). Between these, different and sometimes conflicting attitudes and perceptions teacher performance evaluation seems truly problematic. Thrupp (1999) criticizes the teacher evaluation process asking the simple question of who is to be blamed when a student fails. Answering this question differs from one culture to another as some might blame the student; other cultures might blame the teacher, school or the system.

One final complication in performance evaluation relates to bias and data error (Khanka, 2003). For example, data collected to evaluate teacher X could be inaccurate and unrepresentative about his/her performance. One bias error is data contamination which occurs when items that should not be measured are included in a teacher's performance evaluation. For instance, some teachers are evaluated or judged by their performance and behaviour during off-work times; for some educational institutions, teachers hired should represent a highly prestigious social role model or figure. In the Middle Eastern context, the case might go even more extreme where a teacher might be asked to leave his/her job because he/she had been seen doing second 'inappropriate' jobs, e.g., taxi-driver, bartender or musician. Another type of error is deficiency. This error occurs when items that should be included in a teacher's performance evaluation are not included. Sometimes evaluators apply these two types of bias error intentionally and others occur accidentally, due to the sophisticated nature of the teaching job mentioned earlier. Perhaps, all these complications mentioned in this section make performance evaluation in the top of the undesirable duties for managers (Stewart & Brown, 2009, p. 292).

5 The Evaluation Process

With all the complications mentioned above it is understood that the performance evaluation of academics is a sensitive issue and the process should be designed and carried out very carefully. It is attention grabbing how the literature sets out different steps for the design process. Bernardin (2010) believes that this variation is a result of the fact that different institutions might perform the process in different manners. However, reflecting on what is mentioned in the literature, it is understood that the six steps are commonly shared between the models. The steps are (1) establishing performance standards, (2) communicating performance evaluation to employees, (3) measuring actual performance, (4) comparing actual performance with standards, (5) discussing the appraisal with the employees and (6) initiating corrective actions (Khanka, 2003).

In the first step, *establishing performance standards*, the supervisor (evaluator) decides what is to be measured, e.g., teaching skills and academic achievements. These standards are usually fed by the job description documents stating the performance and behaviour required of an employee. The standards should be clear and measurable. On the contrary, many institutions make the mistake of producing vague and sometimes extremely brief job description documents. The second step is *communicating the evaluation standards* to employees. This step is important in order to make the employees aware of what is expected of them. For example, at some universities teachers are expected to be available in their offices during non-teaching hours. Office availability, or attendance discipline, might be taken as a performance standard in this case. The following step is *measuring actual performance* where data about the teachers are collected. Different methods of data collection could be used, e.g., classroom observations, student surveys, reports. This stage is a critical and sensitive one as it is vulnerable to error, e.g., evaluator bias or data error. It is extremely important that the evaluator be objective and measure actual performance based on facts and findings. In step four and as data about actual performance are collected they are to be *compared with the evaluation standards*. It is important to check whether or not the employee is meeting the standards and performing what is required of him/her. Any deviation from the standards set out in step one, whether in the negative or the positive direction, should get the evaluator's attention. Results of the comparison are used in the fifth step which is *discussing the appraisal with the teacher*. At this stage, the evaluator communicates the feedback or findings of the appraisal with the employee. Ahmad (2002) believes this stage to be the most critical in the whole process. It is challenging for the evaluators to present an accurate report and make the employee accept it in a constructive manner (Khanka, 2003). This, in fact, is more challenging in the field of education, and more particularly in higher education, as some academics might feel too proud and might respond defensively to any negative opinions. Communicating the appraisal usually provides an opportunity for the employee to realize areas of strength and weakness. The final step in the process is to *initiate corrective actions*. These actions could have different directions. For example, some of these actions might impact on the employee in a direct way, leading for example to internal promotion, financial incentives, or employee termination. Other actions might reflect on the institution's strategies and practices, e.g., developing employee contracting policies, or service development.

6 Types and Methods of Collecting Data

Several types and methods of collecting data to appraise staff performance are used. Each of them have different strengths and weaknesses. Organizations with different preferences might use different methods or approaches. Lunenburg and Ornstein (2004) group these into three categories: The judgmental approach, the absolute standards approach, and the result-oriented approach.

The *judgmental approach* is the oldest approach, whereby the appraiser follows comparison strategies in which s/he compares the employee's performance with other employees. Several traits or behavioural aspects are compared. Within this approach, four main methods are used. (1) The *graphic scale* uses several traits that are assigned values and added up and totalled to indicate and rate an employee. This is usually used with the help of assigned tables or graphs. (2) *Ranking* is another method in which the appraiser ranks employees from best to worst in particular traits and qualities. This method is usually used for promotions or assigning particular tasks (Redman & Wilkinson, 2009). For example, the most organized member of faculty might be assigned to organize a conference for his/her university. (3) *Paired comparison* is also another method in the judgmental approach that is somehow similar to ranking. An appraiser here, however, compares two employees only at the same time to indicate who is superior in a particular area or trait. This method might be helpful in making decisions when selecting a particular member. (4) The final method in this approach is *forced distribution*. In this method, an employee's performance is rated against normal statistical distribution of all other employees' performance. The problem with the judgmental approach is that it allows scope for the appraiser's (or supervisor's) own judgment and subjectivity since it depends heavily on the appraiser's assessment or even opinion. However, this approach could be useful in a particular context, e.g., assessing job performance in an area that is difficult to measure (Stewart & Brown, 2009). This means that this approach could be useful in assessing a teaching job.

The *absolute standards approach* is another significant one in evaluating performance. The main difference in this approach is that an employee's performance is compared not to other employees, but they are compared to and rated against certain established standards. This means that methods and tools in this approach depend on job analysis that will describe actual behaviour necessary for effective performance. There are three main methods within this approach. (1) The *Checklist* is a commonly used method. It requires less effort on the part of the appraiser as it offers a ready list of criteria. However, establishing the list of questions or items might consume more time and effort. Evaluating an employee with the checklist method offers a numerical rating that is used for personal decisions such as salary or promotions (Lunenburg & Ornstein, 2004). (2) Another method within the absolute standards approach is the *essay*. This method is very simple; yet time consuming, as the appraiser produces a narrative and descriptive report (essay) about an employee's performance. This report could highlight strengths, weaknesses, potentials and even particular incidents. (3) The *Critical incidents* method in which the appraiser focuses on a key critical behaviour that affects the job performance in a noteworthy manner. Such critical incidents could be effective (positive) or ineffective (negative). These critical incidents, usually done annually, are reported and recorded for later analysis. They might highlight training needs or distinguished performance.

The last category is the *goal-oriented approach*. This approach focuses on evaluating the results achieved rather than the employee's behaviours or qualities. The *goal setting* method is the most common in this approach in which a meeting is held to discuss and set goals. In some cases, goals are set individually (each member is assigned his/her own goals) or collectively (similar goals are assigned for all members of organization). One example of the goal setting method is when teachers are required to achieve high levels of student retention. Another example is that some universities set goals for their academics in publications where they are supposed to publish a particular number (quantity) of research papers. Other universities might require publications in high quality journals with high influence and impact factor ratings (quality). Meeting these goals is a major indicator of a member's performance.

7 The Case Study

This section presents empirical research by means of a case study research strategy. The case study investigates the issue of performance evaluation at one private university in Syria. Different tools are used to collect and triangulate data. These are interviews, documentary analysis and observations. In addition, data are collected from different groups: Administrative members of staff, academic members of staff and students.

Al-Alam University (pseudonym) is a private shareholding university. It has six faculties and has plans to initiate two more in the next two years. It is among the pioneering private universities in Syria since it was licensed in the mid-2000s. It began offering courses four years later. During these four years, there was much work and preparation on all levels and this reflected positively on the university performance; this is most evident in the rapid increase in student population. The university student population in the first years was just a few hundred; however, in 2011 it grew to thousands. The administrative structure, which is large and complicated, has attempted to cater for the rapid growth. There are about two hundred administrative members of staff distributed over several councils, boards and directorates to facilitate the university's work. The administrative structure comprises a board of trustees, a university council, a board of directors, faculty councils, and ten directorates: Admission and enrolment, finance, human resources, information resources, information technologies, maintenance and service, professional training, public relations, quality and accreditation and student affairs.

Regarding academic members of staff, the University has more than 250 teaching members whose teaching ranks range from teacher assistant, teacher, lecturer, assistant professor, associate professor, and professor. The number of academic staff is a key factor as the university claims to offer the best teacher/student ratio compared to other private or even public universities in the country. The University has followed a strategy of recruiting high calibre members of faculty, as this has been helpful in marketing the university and attracting more students. There have been aims at recruiting distinguished, important and well-known people, e.g., former ministers or even public figures. The university competes with other institutions to attract the best academics, bearing in mind the scarcity of qualified personnel for such jobs in Syria. There is a common belief that the high quality of their academic staff is the 'benchmark' that distinguishes this university from competitors. The method and style of teaching used is also significant as most of the academic staff are educated in the West; they speak English fluently and use modern teaching methods. The university claims no influence or nepotism in selecting, recruiting or assigning members of faculty. They are assigned merely on the basis of their merits and established reputation. Meanwhile, the public sector and several other private universities in Syria suffer the severe negative effects of influence and nepotism. Before a new academic member of staff is assigned, he/she is evaluated through certain processes that include extensive interviews, a study of previous work profiles, and in most occasions a presentation demonstrating the capabilities possessed.

7.1 The Evaluation Process

The procedure of evaluating academic performance is delegated to the colleges' deanship, particularly the deans' assistants. Supposedly, the process starts establishing performance standards. The university's performance standards, however, are basically established as 'taken for granted ideas and common knowledge' in the heads of directors. For example, teachers should not be late or miss scheduled classes. As a result, these standards are vaguely and briefly mentioned in some documents, e.g., employment contracts and accreditation documents. It would help the process to have standards better documented. Moving to the next step, communicating standards with academic members of staff is done initially through the employment contracts. Another method of communicating standards is done occasionally through periodical meetings, e.g., departmental meetings. It is understood that the standards are generally implied by these two channels. Sharing standards in a more explicit manner (e.g., one-on-one meetings) might help develop performance and the evaluation process. In some instances, some executives thought sharing standards with academics is unnecessary as these standards are universal and taken for granted. Measuring actual performance is the most critical stage in the process and the university uses several methods to accomplish this. Further discussion on this stage and its data collection methods is presented in the following section. In the fourth stage, comparing actual performance to standards, data collected are analysed to compare with the university standards. However, as the standards are not well documented, this allows more room for subjectivity as the supervisor's opinion about what should be done might differ from one person to the next, or in some instances from one faculty member to another. The final stage is about taking corrective action. These might have some direct effects. For example, an academic member might be asked to change a teaching method. In other instances, he/she might be reminded of the university's regulations about particular 'do's and dont's'. Perhaps, the most important issue that relates to this stage is about renewing employment contracts bearing in mind that in the private sector in Syria such contracts are renewed annually.

7.2 Evaluation Methods and Tools

The Human Resources Department has developed a solid performance evaluation data collection system which includes input from four sources following the judgmental and the absolute approaches. The methods used are critical incident reports, class observations, student course surveys, and informal talks with students. (1) In the *critical incident report*, a supervisor creates a sheet (Word Document) for each member of faculty. This sheet has a table of two columns; one is assigned for positive incidents (titled positives initiatives), and the other column is assigned for negative incidents (titled negative behaviour). A quick review of these sheets shows the number of negative incidents outweighs the positive ones. This could indicate that this tool is more assigned for disciplinary purposes. (2) *Class observation* is another very important tool. This is particularly important for newly assigned members as supervisors need to establish initial ideas about performing in class, teaching methods, and methods of dealing with students. It is an unstructured style of observation where the supervisor does not have a schedule or checklist. Once a class is observed the supervisor produces a document/essay of about 500 words. The document structure and analysis style follows the well-celebrated SWOT analysis (strengths, weaknesses, opportunities, and threats) model. In this regard, it highlights strengths and possible opportunities in an encouraging and supportive manner with the purpose of boosting and empowering such traits and areas. The document also draws on weaknesses and threats with the aim of solving such weaknesses and eliminating such threats. Once a class is observed, the teacher is asked for a meeting with the supervisor, usually on the same day, to discuss the results. Supervisors work on delivering and discussing the results in a positive, supportive and productive manner with academics. In some situations, sensitivity or even tension could be felt, e.g., if the evaluator is younger than the person being evaluated or at a lower academic/scholar level.

The third method is the (3) *student course survey*. Student evaluations of teaching are conducted towards the end of the semesters for each subject taught, to evaluate and assure the quality of in-service faculty. All students express their views of the different faculty members in each subject. There have been several criticisms and concerns about this evaluation tool in the literature (Centra, 1993; Kember et al., 2002; Liaw & Goh, 2003). Among the main concerns are those of bias and relationship between survey scores and learning achievements. The survey is one page following the 5-point Likert scale ranging from “strongly disagree” to “strongly agree”. It covers twelve items covering areas of receiving support, empowering autonomous learning, knowledge delivery, allowing discussions, having disciplined class, punctuality, office hour availability, assigning and marking homework, speaking English in class, class enjoyment and the overall satisfaction of the course. There is also an open-ended question for students to make comments. On conducting the survey, a member of administrative staff, usually the deans’ secretaries, visits the class. S/he asks the teacher to leave the classroom and wait outside. The person administers the survey starting by telling the students about the questionnaire and about its purposes in developing the university. The whole process takes less than ten minutes. The survey administrator then asks the teacher back to class to continue his/her session. Data from the questionnaires are uploaded into an SPSS file for analysis. Several descriptive and inferential statistical analysis and tests are conducted, e.g., frequencies, means, SDs, ANOVA tests and factor analysis. Following the results, members of faculty are ranked in several themes and areas following the variables in the questionnaire. Unlike the earlier tool, results from this method are not shared with the faculty members. On the other hands, all faculty members expressed a strong desire to know these results. The university, however, does not share the results as the university administration believes that sharing the results might create sensitive and even negative feelings (e.g., rivalry or aggressive and counterproductive competition) among the faculty members.

The last method is (4) *informal talks with students*. In this method, as the university intends to establish ‘good and positive’ relationships with students, supervisors (e.g., deans, deans’ assistants, or head of departments) have casual conversations with students. These conversations could happen anywhere, e.g., student cafeteria, corridors, or university transportations. A supervisor starts by asking a student generally about their studies and the university life. S/he asks about teachers and the ones they prefer and why. Supervisors try not to make the students feel that they are seeking any information to evaluate any member of academic staff as this might reflect negatively on the data. Serious and disciplined students are usually approached in this method, as they are believed to offer more mature and realistic views. The last two methods of data collection, student course surveys and informal talks with students, place heavier weight on students’ opinions in evaluating academics. This is possibly because Al-Alam University is a private fee-paying university. It is clear that private educational institutions offer students customer sovereignty (Al-Fattal, 2011). This could be justified as student tuition fees are, probably, the only source of income for the university. It is also important to mention that offering such influence for students to evaluate teachers and academics is very uncommon in Syrian culture. This, perhaps, is a result of the short experience of private education in the country. Members of the faculty have different opinions about this. For example, some academics felt insulted to be evaluated by students. Those people might be the ones who have a longer history in public higher education where such privileges have never been offered to students.

8 Conclusion

This chapter has discussed the issue of performance evaluation and appraisal, focusing on a case from the private education sector in Syria. It has shed light on the drive for improving performance on organizational and personnel levels in the higher education context. Performance evaluation for members of faculty has been the particular area of inquiry for the chapter. Performance evaluation and appraisal in this regard is the systematic and objective assessment of an academic’s performance; it is a method of judging the relative worth or ability of academics in performing their tasks. The importance of performance evaluation and appraisal is highlighted in reflecting on personnel and organizational performances through improving effectiveness and efficiency. It also reflects on staff attitudes and motivation. This chapter has also presented some complications regarding academic staff members' performance evaluation and appraisal. The major complication relates to the nature of the teaching job and the complexity of defining professional roles. Other complications relate to staff reaction, cross-cultural disparities, and errors. This has been followed by some discussions on the evaluation process and its steps of establishing performance standards, communicating performance evaluation to employees, measuring actual performance, comparing actual performance with standards, discussing the appraisal with the employees and initiating corrective actions. This chapter has also documented some methods of data collection approaches and methods.

The second part of the chapter has presented a case study of a private university in Syria. Performance evaluation of academic members of staff has been investigated through this case study. This section has mainly discussed two issues: The process of evaluation and data collection methods. Matching the findings from the case study to the literature and theory, Al-Alam University has achieved an advanced level in performance evaluation. This also comes as a result of a quick comparison of evaluation practices in other higher education institutions in Syria. No doubt, Al-Alam University's performance evaluation could be developed and improved. The areas requiring improvement are in particular the first two stages in the evaluation process, namely establishing performance standards and communicating performance evaluation to employees. The university is recommended to establish more standardized procedures for these stages. For example, the university council could agree on issuing particular expectations, standards or codes for good practice in this regard. These need to be well communicated to all members of faculty, especially newly assigned ones. One possible practice several institutions could conduct is the orientation programme where newly assigned members are extensively informed about performance expectations, standards and codes through more than one communication channel.

References

Ahmad, A. (2002). Supervisory and non-supervisory employees' attitudes and perceptions towards performance appraisal in the Malaysians public sector. *Journal of King Abdulaziz University: Economy and Administration, 16*(2), 3–17. doi:10.4197/Eco.16-2.1

Al-Fattal, A. (2010). Understanding student choice of university and marketing strategies in Syrian private higher education. Retrieved on December 31, 2012 from etheses.whiterose.ac.uk/…/PhD_Thesis_Anas_Al-Fattal_SID200229252_Education.pdf

Al-Fattal, A. (2011). *Marketing universities: Understanding student choice of university and marketing strategies at private higher education in Syria*. Saarbrucken: Lambert Academic.

Al-Fattal, A., & Ayoubi, R. (2012). Understanding consumer buyer behaviour in the EFL market: A case study of a leading provider in Syria. *Education, Business and Society: Contemporary Middle Eastern Issues, 5*(4), 237–253. doi:10.1108/17537981211284425

Al-Fattal, A., & Ayoubi, R. (2013). Student needs and motives when attending a university: Exploring the Syrian case. *Journal of Marketing for Higher Education, 23*(2), 204–225. doi:10.1080/08841241.2013.866610

Avalos, B., & Assael, J. (2006). Moving from resistance to agreement: The case of the Chilean teacher performance evaluation. *International Journal of Educational Research, 45*(4–5), 254–266. doi:10.1016/j.ijer.2007.02.004

Bernardin, J. (2010). *Human resource management: An experimental approach* (International 5 ed.). Boston: McGraw-Hill.

Boswell, W., & Boudreau, J. (2000). Employee satisfaction with performance appraisals and appraisers: The role of perceived appraisal use. *Human Resources Development Quarterly, 11* (3), 283–299. doi:10.1002/1532-1096(200023)11:3<283:AID-HRDQ6>3.0.CO;2-3

Centra, J. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness (The Jossey-Bass higher and adult education series)*. San Francisco: Jossey-Bass.

Coleman, M., & Glover, D. (2010). *Educational leadership and management: Developing insights and skills*. Maidenhead: McGraw-Hill Education.

Gallagher, H. A. (2004). Vaughn elementary's innovative teacher evaluation system: Are teacher evaluation scores related to growth in student achievement? *Peabody Journal of Education, 79* (4), 79–107. doi:10.1207/s15327930pje7904_5

Gomez-Mejia, L., Balkin, D., & Cardy, R. (2010). *Managing human resources* (Global 6 ed.). New Jersey: Pearson.

Kember, D., Leung, D., & Kwan, K. (2002). Does the use of student feedback questionnaires improve the overall quality of teaching? *Assessment and Evaluation in Higher Education, 27* (5), 411–425. doi:10.1080/0260293022000009294

Khanka, S. (2003). *Human resources management: Text and cases*. New Delhi: S Chand & Company.

Kotler, P., & Fox, K. (1995). *Strategic marketing for educational institutions* (2nd ed.). Englewood Cliffs, N.J.: Prentice-Hall.

Liaw, S., & Goh, K. (2003). Evidence and control of biases in student evaluations of teaching. *International Journal of Educational Management, 17*(1), 37–43. doi:10.1108/09513540310456383

Lunenburg, F., & Ornstein, A. (2004). *Educational administration: Concepts and practices* (4th ed.). Belmont: Thomson Wadsworth.

Middlewood, D., & Cardno, C. (2001). The significance of teacher performance and its appraisal. In D. Middlewood & C. Cardno (Eds.), *Managing teacher appraisal and performance: A comparative approach* (pp. 1–16). London: Routledge Falmer.

Milanowski, A. (2004). The Relation between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education, 79*(4), 33–53.

Podgursky, M. & Springer, M. (2006). *Teacher performance pay: A review*. A Working Paper for National Centre for Performance Incentives, Department of Educations; Institute of Education Sciences, USA. www.performanceincentives.org

Ramsden, P. (1991). A performance indicator of teaching quality in higher education: The course experience questionnaire. *Studies in Higher Education, 16*(2), 129–150. doi:10.1080/03075079112331382944

Redman, T., & Wilkinson, A. (2009). *Contemporary human resources management: Texts and cases* (3rd ed.). Essex: Prentice Hall & Financial Times.

Stewart, G., & Brown, K. (2009). *Human resource management: Linking strategy to practice*. New Jersey: John Wiley & Sons.

Stone, R. (2011). *Human resource management* (7th ed.). Milton: John Wiley & Sons.

Thrupp, M. (1999). *Schools making difference: Let's be realistic*. London: Open University Press.

Part II
Assessment Practices

Ethicality in EFL Classroom Assessment: Bridging the Gap between Theory and Practice

Shaimaa A. Torky and Nawal Abdul Sayed Haider

Abstract In recent years, the renewed focus on classroom or assessment for learning and the role it plays in gauging and supporting the learning process has entailed adopting an alternative approach to fairness that takes into consideration the open ended, dynamic nature and informality of most of these practices. The current study tackles EFL university teachers' perception of the ethicality of various classroom assessment practices to uncover the hidden code of ethics they ideally refer to, and to determine its conformity to codes endorsed by previous research. A sample of 28 teachers from the English department at the Public Authority of Applied Education (PAAE) at Kuwait University, College of Science, was selected. A survey in the form of a fifty-item questionnaire comprising five dimensions was utilized to assess teachers' assessment practices and their perceptions of assessment ethicality. Results imply that many areas were considered controversial for most teachers. One of these areas is using multiple forms for assessing students. Another issue is consistency between the assessment methods used and the curriculum objective and classroom activities. Equity issues also seem to be blurred for most teachers. Results of the current study testify to the value of training that is particularly focused on fair assessment and ethicality dilemma.

Keywords Classroom assessment · Assessment ethicality · Assessment for learning · Ethical conflicts · Teachers' perceptions of assessment

S.A. Torky (✉)
The National Centre for Educational Research, Cairo, Egypt
e-mail: shaimaatorky@gmail.com

N.A. Sayed Haider
Public Authority for Applied Education and Training, Adailiyah, Kuwait
e-mail: nawala33@hotmail.com

S. Hidri and C. Coombe (eds.), *Evaluation in Foreign Language Education in the Middle East and North Africa*, Second Language Learning and Teaching,
DOI 10.1007/978-3-319-43234-2_4

1 Introduction

The concept of assessment has undergone a recognizable shift throughout educational history. Originally, assessment was conceptualized as testing learners to provide evidence of accountability for stakeholders. In fact, this view is thought to ignore the fact that assessment should be geared first to assist students and provide them with the necessary feedback regarding their progress and efficiency of their adopted strategies (Buzzelli & Johnston, 2002). Most scholars consider this approach as "assessment of learning" (A*o*L).

Yet, with the advent of concepts like lifelong learning, social construction of knowledge and learner's autonomy, the term assessment has evolved to endorse an alternative conceptualization that views assessment as the most important means of providing help to learners and fostering their self-awareness. Akin to this paradigm shift, unconventional means of assessment that drastically change power relationships in the classroom have emerged; the call for using self-assessment, peer assessment and participatory assessment is a clear reflection of a more democratic or participatory trend (Tierney, 2010). This quite recent notion is termed formative assessment or according to Dann (2002) and Tierney (2010) assessment for learning (A*f*L).

However, although teachers may spend as much as one third of their time in assessment-related activities (Stiggins, 2002), pre-service and in-service training in Kuwait do not require EFL teachers to take a course, or demonstrate competency in the area of assessment, suggesting that teachers often lack formal training in assessment. Even if some training is provided in this regard, it is almost tackling the procedures of designing and applying assessment tools. Hardly any guidance is provided that touches on ethical dilemmas EFL teachers might encounter as they attempt to strike a balance between the two competing demands of tuning into students' needs on the one hand, and meeting the demands for accountability, on the other hand. This lack of preparation in assessment related ethical issues is problematic because ethical reasoning in assessment can barely develop by everyday experience (Green, Johnson, Kim, & Pope, 2007). Hence, left to their own devices, teachers' decisions or classroom practices pertinent to preparing, designing tests and grading students' on various assessment tools are mostly based on intuition.

2 Literature Review

Classroom assessment is small-scale assessment prepared and implemented by teachers in classrooms. Classroom assessment includes traditional assessment and alternative assessment. Traditional assessment refers to predetermined testing measures such as selected-response tests (e.g., multiple-choice questions, true/false questions, matching questions), brief constructed-response tests (e.g., short-answer questions), and essay questions. Alternative assessment refers to authentic assessment tasks/forms such as oral questioning, teacher observation, performance tasks, and student self-assessment (Airasian, 2005; McMillan, 2007).

Traditionally, a fair test was considered one, which is free from bias, partiality, discrimination and favouritism (Tierney, 2010). Adopting this point of view, the Educational Testing Service (ETS) (2002) has added a new section addressing the issues of eliminating bias and ensuring equity in the testing processes. Notably, the fairness concept adopted by these standards is highly technical, and so it was thought that evidence of fairness could be obtained via statistical procedures such as validity and reliability (Camili, 2006; Volante, 2006). From another perspective, a broader procedural conceptualization for fairness that goes beyond statistics has started to emerge. This includes defining a clear purpose for the test, developing test specs, reviewing test content and, finally, conducting a field test of the examination (Plake & Jones, 2002). Noticeably, ethical issues are not directly tackled according to this approach.

Attempting to address the ethical issues of assessment more deeply, Messick (1995) stressed the interrelatedness of ethicality and validity. According to his viewpoint, test developers need to minimize construct-irrelevant test variance, which can be the result of the test response being based on factors irrelevant to the objectives being assessed and thus might distort results. Fair testing has also become a pursuit for standardized testing practices (JCSEE, 2003, p. 3). Fairness is, also, a critical consideration for good testing practice in ILTA's (International Language Testing Association) Draft Code of Practice (2000). Noteworthy, however, this discussion of fair testing is very limited and disregards ethical assumptions underlying day-to-day classroom assessment.

Thus, in recent years, the renewed focus on assessment for learning has entailed adopting an alternative approach to fairness that reemphasizes the value-laden aspects that have long been neglected in education (Blanchard, 2008). Providing an operational concept of fairness that takes into consideration the unique aspects characterizing assessment for learning, Airasain (2005), Camili (2006), McMillian (2007), Shepard (2005, 2007) and Zhang and Burry-Stock (2003) argued that fairness includes setting clear learning expectations, helping students learn how to do the assessment task, ensuring equity and avoiding bias, using varied approaches for eliciting learning, accommodating special needs and providing detailed and balanced feedback for learners.

This approach to fairness, as seen from the aforementioned criteria, places the learner and his interests at the heart of the assessment procedure. Therefore, great emphasis is placed on preparing the learner for the assessment. Noticeably the assessment is no longer viewed as an end in itself, rather it is seen as a tool to drive the learning process forward. Nonetheless, Tierney (2010, p. 63) views these standards as lacking sufficient description or empirical evidence that attest to their validity. Also missing, according to his viewpoint, is the emphasis on the protection of privacy, communication of results and the use of multiple evaluators.

From another perspective, the concept of fairness can be quite related to the ethical dilemma teachers face in their relationships with the individuals they interact with in their professional life. Reviewing previous research, some codes could be gleaned, such as the no harm principle, avoiding score pollution, equity, transparency and consistency. In fact, all these codes are intertwined so that it renders it difficult to draw clear broader lines to separate them.

The concept of no harm, unlike the abstract concept of fairness which may fall short of revealing teachers' practices, stimulates their awareness of their mal-practices regarding classroom assessment. Examples of harm can be unexpected items on or offensive items in a language test. Avoiding score pollution is an application of the principle of "do no harm" to assessment. It is defined as any practice that improves test performance without concurrently increasing actual mastery of the content tested (Payne, 2003). When teachers take into consideration during grading students' work are factors, such as their effort, behaviours or punctuality, the scores students obtain may overstate or understate their actual skills. Inequity, as opposed to equity, occurs when the teacher offers different opportunities and makes different decisions in the same environment. For instance, teachers sometimes unjustifiably increase test time for some students or change students' answers (Gipps & Murphy, 1994; Hidri, 2015). Transparency and consistency can also affect test fairness. Transparency implies involving students in the process of determining the evaluation criteria and methods of assessment (Popham, 2000). Consistency basically necessitates compatibility between assessment tools and the purpose for which they were designed (Pope, Green, Johnson, & Mitchell, 2009).

2.1 Teachers' Perceptions of Assessment

As Chang (2006) argues, few studies attempted to probe into their underlying perspective and ethical convictions that inform their various decisions. Nevertheless, examining previous studies, two research trends could be discerned: The first aimed at investigating teachers' perspectives of their assessment practices in general; the second, however, focused on examining teachers' ethical convictions underlying their practices.

Addressing the first trend, Pelly and Allison (2000) explored primary school teachers' perspectives on the assessment of the use of the English language in Singapore. The findings revealed that teachers were markedly divided and uncertain in their views of the efficacy of current tests. Zhang and Burry-Stock (2003) investigated teachers' assessment practices across teaching levels and content areas. Results showed that, regardless of teaching experience, teachers with measurement training reported a higher level of self-perceived assessment skills than those without measurement training. In the same way, Chan (2008) investigated elementary teachers' beliefs and practices of multiple assessments. Results showed that most teachers considered using multiple assessments a positive experience. Yip and Cheung (2005) pinpointed that teachers expressed their concerns about the consistency of assessment whether among different teachers or consistency in each individual teacher's assessment practices.

Two studies used the Teachers' Conceptions of Assessment (TCoA) inventory to investigate teachers' assessment conceptions. The first is Hidri's study (2015) and it revealed that teachers harbour wrong and conflicting assessment conceptions. The second is Gebril and Brown's study (2013), which suggested that greater changes to the examination system are warranted if teacher beliefs are expected to be more positive about the priority of formative, improvement-oriented uses of assessment.

To address the second trend, some researchers attempted to examine teachers' ethical beliefs and the ethical dilemmas they grapple with during everyday classroom practices. For instance, Szpyrka (2001) explored the relationships between equitable assessment practices and actual classroom assessment practices. Results showed discrepancies between teachers; whereas some teachers believe that tasks should be modified to enable each learner to be successful, others think that all students should abide by the same standards. Lu (2003) investigated the beliefs and practices of assessment by two university English instructors. The results showed that there was a high consistency and a very slight inconsistency between the instructors' beliefs and their assessment practices.

In the same way, employing a web-based survey, Green et al. (2007) conducted a study that exposed teachers to a set of classroom scenarios to examine their implicit ethical perspectives. Findings suggest that assessment is currently an educational realm without professional consensus. Likewise, Tierney (2010) and Simon, Chitpin, and Yahya (2010), found, throughout studies that aim at reaching a better understanding of classroom assessment fairness, that teachers' fairness relies on their ability to understand students and to reflect on both the interaction and decisions made in the classroom. Moreover, the studies revealed that group work, test failure, fairness, multiple assessment opportunities, and academic enablers were key areas of concern.

Pope et al., (2009) conducted a study to document ethical conflicts faced by teachers regarding the assessment of students. Critical incidents generated by teachers revealed a majority of reported conflicts related to score pollution, and conflicts frequently arose between teachers' perceptions of institutional demands and the needs of students. Using an introspective critical approach, Simon, Chitpin and Yahya (2010), examined pre-service teachers' perceptions of classroom assessment. The researchers found that group work, test failure, accommodation, fairness, multiple assessment opportunities, and academic enablers were key areas of concern for most teachers.

The only study that approached classroom assessment from the students' perspective was the study of Bursuck, Munk, and Olson (1999) who attempted to determine the students' perceptions about fairness of grading their final report. Students indicated that they could accept differentiation in teacher's responses to an assessment task, yet they cannot accept adaptation in the assessment task to accommodate various students' needs.

2.2 Writing Assessment

Being highly prone to teachers' personal judgment or to students' self -judgment, the accuracy and reliability, and hence fairness, of writing assessments are looked at with a great deal of skepticism. Fairness in writing assessment was tackled from two perspectives. On the one hand, teachers' perspectives on the fairness of the criteria they adapt to grade students' performance are tackled. On the other hand, students' self-assessment as a means of realizing equality is also probed.

Addressing the first perspective, Zoeckler (2005) intended to understand the moral aspect of grading writing by examining English language teachers' assessment artifacts and by interviewing and taking field notes. Similarly, Graham (2005) conducted a two-year study on pre-service teachers to track the development in their assessment theories and practices. Teachers agreed that evaluating writing is a challenging process and their comments regarding fairness centered on providing constructive feedback for weak students to provide them with the best possible support.

Similarly, Dann (2002) conducted a case study that aimed at examining the fairness of grading students' writing projects. A participatory form of classroom assessment was adopted where self, peer and teachers' assessments were embraced. Though most students expressed their overall satisfaction with the scores they obtained, they expressed a sense of having the scores imposed on them that shows that participatory classroom assessment can be controversial.

Therefore, it can be concluded from the previous literature review that, except for the study of Green et al. (2007) and Pope et al. (2009), few studies have investigated EFL teachers' perceptions of assessment ethicality or the ethical dilemmas that can bear upon their classroom practices. Noteworthy, also, research examining teachers' perceptions of assessment in general and ethical considerations pertinent to assessment in particular has either resorted to using indirect methods, such as interviews, scenario techniques and incident techniques or direct methods, e.g., direct classroom observation, artefacts or examining students' perceptions. Most of these studies have widely reflected the paradoxical stance most teachers experience when they assess students. Issues such as equity, fairness or score pollution and distinction between what constitutes fair assessment versus unfair assessment are still blurred for most teachers. Furthermore, the concept of assessment for learning with its implications has not yet been well assimilated by teachers, and hence their notion of equity is rather confined to the traditional concept of assessment of learning. Thus, the purpose of the current study is to lay more emphasis on the concept of ethicality and how teachers perceive this concept as far as their assessment practices are concerned.

3 Purpose of the Study

The purpose of the current study is twofold. On the one hand, it is meant to examine EFL university teachers' perceptions of the ethicality of various classroom assessment practices to uncover the hidden code of ethics they ideally adhere to, and to determine its conformity to codes endorsed by previous research. On the other hand, the study aims at examining EFL teachers' current practices regarding various ethical issues in the realm of classroom assessment to identify the discrepancy between those practices and teachers' ethical beliefs. The study aimed at answering the following questions:

(a) What are EFL teachers' perceptions regarding the ethicality of the identified classroom assessment practices?
(b) What are the actual assessment practices carried out by those teachers in light of ethicality considerations?
(c) To what extent are teachers' perceptions and classroom assessment practices consistent with ethicality norms?

4 Method

The current study used both qualitative and quantitative methods to investigate teachers' perceptions of ethicality of their classroom assessment practices. The perspectives of the participants are of the utmost importance as the researchers sought to understand and describe the participants' experiences as seen by the participants themselves. For these reasons, therefore, the application of the qualitative paradigm was considered critical to the study. This took the form of analysing teachers' answers on each questionnaire item by item to discern their way of thinking and locate compatibility, or lack of it, between their endorsed ethical codes and classroom practices. Quantitative methods were utilized, however, to analyse obtained data and get generalizations about teachers' perspectives.

4.1 Participants

Purposive sampling was utilized to locate teachers who were willing to converse about their experiences with classroom assessment practices. A sample of 28 teachers, 16 females and 12 males, from the English department at the Public Authority of Applied Education (PAAE) at Kuwait University, College of Science, was selected. Respondents had taught for an average of fifteen years. Current grade level taught was university levels. Some respondents (30 %) had a bachelor's degree; (25 %) held a Master's degree and (45 %) held a Ph.D. About 82 % of the teachers had had at least one measurement course. Teachers were involved in a multitude of assessment activities including administering regular exams, grading writing, evaluating students' oral performance and applying final university mandated exams. Thus, it was thought that the sample of this study would have adequate experience or background knowledge of classroom assessment, and so they would be able to judge the ethicality or otherwise of various practices. First, Tables 1, 2 and 3 present summary information on respondents by teaching experience, assessment experience and obtained training.

It is clear from Table 1 that the teachers' teaching experience ranged from less than 7 to more than 22 years, with most teachers having less than 7 years of experience. As Table 2 shows, teachers' experience in assessment ranged from 1 to 20 years, with most teachers having from 16–20 years of experience. Table 3 indicates that about half of the teachers (46.7 %) had not received any training courses in assessment, (21.7 %) received pre-service training and (10.7 %) received in service training. Yet, (21.4 %) reported that they received training throughout other means such as conferences, individual reading and workshops.

Table 1 Teaching experience

EFL experience (in years)	Number	Percent	Cumulative percent
<7	13	46.7	46.7
15–21	7	25	68.4
22–	8	28.6	100
Total	27	100	

Table 2 Assessment experience

Experience (in years)	Number	Percent	Cumulative percent
1–10	10	35.7	35.7
11–15	7	25	57.1
16–20	11	39.2	100
Total	28	100	

Table 3 Professional assessment related training

Assessment training	Number	Percent	Cumulative percent
No training courses	13	46.7	46.7
Pre-service training	6	21.4	67.9
In service courses	3	10.7	78.6
Other means of training	6	21.4	100
Total	28	100	

4.2 Instruments and Procedure

A survey in the form of a questionnaire, consisting of 50 items, comprising five dimensions, was utilized to assess teachers' assessment practices and their perceptions regarding assessment ethicality. The instrument was developed within the theoretical framework delineated by the literature on classroom assessment and fair testing.

Teachers were asked to mark their responses to the same 50 items on two different rating scales: The practice scale and the ethicality scale. The practice scale was designed to measure teachers' assessment practices on a 3- point scale (*1 = never practiced, 2 = sometimes practiced* and *3 = usually practiced*). The ethicality scale was designed to measure teachers' ethicality perceptions on a three-point scale (*1 = unethical, 2 = somewhat ethical* and *3 = ethical).* Negatively-keyed items were "reverse-scored" before computing students' total scores. These included all items subsumed under the dimension of score pollution as well other items with the sign (-) as shown in the Appendix. Two data sets were produced, one on assessment practices and the other on perceived ethicality assessment skills. The items of the questionnaire are presented in the Appendix.

The questionnaire comprised five dimensions reflecting ethicality in assessment. Though an overlap in the underlying dimensions may exist, each dimension contains a certain degree of uniqueness. The first dimension is transparency and confidentiality; it subsumed 5 items addressing the teachers' openness regarding assessment objectives, techniques and correction methods. It also subsumed two items tackling the examinees' right for privacy. The second dimension comprised 8 items- and it implied making sure that the assessment adheres to its purpose and to what was taught in the classroom. The third dimension is avoiding score pollution, and it comprised 12 items measuring teachers' awareness of the fact that the score a student receives should tightly reflect what he/she mastered. The fourth dimension is A*f*L-comprising 8 items. This dimension covered aspects, such as using peer evaluation, using multiple assessment methods and avoiding looking at testing as the sole high-stakes assessment device. The last dimension addressed in this questionnaire is equity-including 17 items-, which addressed bias avoidance, providing equal chances to students and catering for various students' needs.

The scenario technique was adapted to present teachers with a set of classroom assessment situations reflecting various stances of classroom assessment, including various sorts of formal and alternative classroom assessment. The assessment situations tackled were derived from everyday classroom experience and were categorized under a set of main areas: Preparation for assessment, developing assessment tools, administering assessment, grading and feedback and reporting or communicating grades.

The first draft of the survey consisted of 62 scenarios and 6 questions about demographic information. To establish validity of the questionnaire, selected professors from the field of assessment and experienced EFL teachers in Kuwait University were asked to review the survey questions. Some items were deleted and others were modified according to the jury viewpoint. Subsequently, a pilot survey was conducted with 14 participants. Participants were given oral instructions on how to answer questions. The results of the pilot survey were reviewed and six items that appeared confusing were modified or replaced. Reliability analysis yielded a Cronbach's of 0.75, for the survey items. Reliability of all dimensions ranged from 0.55 to 0.72. This proves the consistency of the survey and of its various dimensions. The final 50-item survey was administered in the summer of 2013. The instrument along with a cover letter was distributed to the teachers by both researchers, while some were sent via email.

5 Results

5.1 Teachers' Perceptions and Practices

Determining teachers' practices of ethical assessment was based on the respondents' scores on the practice scale of a questionnaire containing 50 items. Similarly, teachers' perceptions of ethicality were based on their scores on the ethicality scale of the same questionnaire. The scores were a sum of these 50 items.

Frequencies and percentages were used to summarize teachers' rating of each situation in terms of the frequency of practice as well as in terms of ethicality. From these percentages, implications could be drawn about ethical issues that were controversial. Moreover, teachers' malpractices or misconceptions regarding these issues were also pinpointed to identify areas warranting more focus. First, to analyse teachers' scores on the survey, descriptive statistics were obtained. Furthermore, to test whether EFL teachers' perceptions of multiple assessments were related to their practices, Pearson product-moment correlation coefficient was computed as shown in Table 4.

Table 4 shows that on the practice scale, teachers' scores ranged from a low of 81 to a high of 112. On the perception of ethicality dimension, teachers' scores ranged from a low of 100 to a high of 122. Teachers' overall mean score on the practice scale was (96.5) and the SD was (5.3). On the other hand, the overall mean score on the ethicality scale was (110) and the SD was (4.7). Given that the total score was 150, the mean score shows that teachers' practices can hardly be considered fair or ethical, even if their perception of ethicality shows that they were

Table 4 Teachers' practices and perception of assessment ethicality

	N	Total	Minimum	Maximum	Range	Mean	SD	Pearson correlation	Sig. (2-tailed)
Total ethicality	28	150	100	122	22	110	4.787	0.219	0.433
Total practice	28	150	81	112	31	96.5	5.364		

relatively aware of what constitutes ethical versus unethical assessment practices. The Pearson correlation was computed between EFL teachers' practices and perception practices, yielding a value of 0.219. The result showed that the relationship between beliefs and practices was not significant, $p = 0.433$.

Considering the scale sub-dimensions, further insights could be drawn. First, a Pearson product-moment correlation coefficient was computed to assess the relationship between teachers' practices and perspectives regarding transparency and confidentiality.

Teachers' mean scores on the practice aspect of the transparency and confidentiality dimension was (12.7) and the SD was (1.2). Similarly, teachers mean score on the ethicality aspect of the same dimension was (10.8) and the SD was (2.1). Pearson product-moment correlation coefficient between the teachers' practices and perceptions of ethicality was not significant at 0.05, $r = 0.008$, $p = 0.996$ (Table 5).

Regarding teachers' responses to the sub-items subsumed under the first dimension, descriptive statistics were obtained. In particular, in terms of preparation for assessment, most of the teachers (68 %) assigned high ethical value to unveiling their grading schemes to the students; this was also reflected in the practice of (68 %) of the study sample. As for assessment development, (28.6 %) of the teachers seemed unconvinced with the importance of sharing with students the rubric according to which a written task will be corrected, whereas (46.4 %) were unsure of whether to consider this practice as ethical or unethical. This uncertainty was also reflected in teachers' practice, i.e., only (39.3 %) indicated that they would entirely avoid hiding information about the writing rubric.

As far as the communication of results is concerned, only (32.2 %) of the teachers recognized the unethicality of limiting feedback to students' strengths; yet the majority (46.7 %) were unsure how to categorize such a practice. When it comes to practice, most of the teachers (53.7 %) reported that they would limit their feedback to students' strengths, whereas (39.3 %) would totally avoid that. The percentages of teachers who admitted to the unethicality of disclosing students' scores to their partners or to other parties were (35.7 %) to (42.9 %) for both cases respectively; however, in everyday practice, the majority of teachers agreed that they would never announce students' scores in front of their partners (67.9 %) or disclose a student's academic information to their peers (89.3 %).

As far as consistency is considered, descriptive statistics were obtained and a Pearson product-moment correlation coefficient was computed to assess the relationship between teacher's practices and perspectives.

Table 5 Teachers' practices and perceptions of transparency and confidentiality

	N	Minimum	Maximum	Range	Mean	SD	Pearson correlation	Sig. (2-tailed)
Practice	28	11	15	4	12.7	1.2	0.008	0.996
Ethicality	28	7	15	8	10.8	2.1		

Note Correlation is significant at the 0.01 levels

Table 6 Teachers' practices and perceptions of assessment consistency

	N	Minimum	Maximum	Mean	SD	Pearson correlation	Sig. (2-tailed)
Practice	28	15.00	22.00	19.11	1.99927	0.527	0.018
Ethicality	28	15.00	22.00	18.066	2.12020		

Table 6 shows that teachers' mean score on the practice aspect of the dimension of consistency was (19.1) and the SD was (1.99). Similarly, teachers' mean score on the ethicality aspect of the same dimension was (18.06) and the SD was (2.12). Since total score on this dimension is 24, it can be concluded that teachers adopted quite fair practices in terms of the conformity of the assessment utilized to both the purpose of assessment and the teaching methods adopted. Pearson product-moment correlation coefficient between teachers' practices and perceptions of ethicality was significant at 0.05, ($r = 0.53$), ($p = 0.018$). Since the coefficient of determination (r^2) = 0.28, the correlation between both constructs is considered low to moderate.

Analysing sub-items subsumed under this dimension, it appeared that, in terms of preparation for assessment, although most of the teachers (68 %) assigned high ethical value to practices pertinent to consistency, such as training students on test taking skills and administering a parallel form of the test, only (26.7 %) reported that they would usually administer a parallel form of the test.

With regard to developing assessment tools, (57 %) believed that any test has to be designed with reference to the curriculum objectives; this conviction was also reflected in the practice of (68 %) of the respondents. Similarly, (53.3 %) avoided incorporating methods that students have not encountered before, and (68 %) avoided using surprise items in their assessment. However, it seemed that teachers were quite unsure of the ethicality or otherwise of these practices; basically, only (46.7 %) thought that both practices are unethical. Similarly, the majority of teachers (80 %) reported that they usually try to incorporate assessment activities similar to those practiced in the classroom, which conformed to the beliefs of (80 %) of the study sample. When assessing oral proficiency, only (47.7 %) of the teachers would refrain from solely relying on classroom observation. These practices seem to conform to the teachers' perception of ethicality, i.e., only (26.7 %) perceived this practice as unethical; the rest were either unsure (46.7 %), or certain that classroom observation was sufficient to judge students' oral competence (13.3 %).

As far as grading is concerned, surprisingly, a high percentage of teachers (73.3 %) reported that they would sometimes grade reading comprehension based only on two multiple-choice tests. This practice is compatible with the ethical perspectives of all respondents since no teacher could perceive the unethicality of using a method that underrepresents students' competence.

Similarly, descriptive statistics were obtained and a Pearson product-moment correlation coefficient was computed to assess the relationship between teacher's practices and perspectives regarding the issue of avoiding score pollution.

Table 7 Teachers' practices and perceptions of score pollution

	Minimum	Maximum	Mean	SD	Pearson correlation	Sig. (2-tailed)
Practice	20.00	28.46	24.08	2.366		
Ethicality	17.00	25.00	22.38	2.471	0.172	0.054

As Table 7 shows, the mean score the teachers obtained on the practice aspect of "avoiding score pollution" dimension was (24.08) and the SD was (2.36). This mean is considered low relative to the total score which is 36. In the same vein, teachers' mean score on the ethicality aspect regarding score pollution was (25), which is low as well. It also shows that the teachers lack awareness of what constitutes ethical versus unethical assessment practices regarding score pollution. Pearson product-moment correlation coefficients between teachers' practices and perceptions of ethicality was not significant at 0.05, $r = 0.172$, $p = 0.541$.

Regarding teachers' responses to the sub-items subsumed under this dimension, frequencies and percentages of teachers' responses showed that, in terms of preparation for assessment, most of the teachers seemed uncertain as to the ethicality of various practices relevant to avoiding score pollution. In particular, some teachers (33.3 %) could not perceive the unethicality of practices aiming at pre-exposing students to parts of an upcoming test, while (47.6 %) were unsure or felt divided. Yet, in practice more teachers (46.7 %) reported that they would totally avoid training students on specific activities included in the actual assessment. Moreover, (46.7 %) of the teachers did not find it ethically problematic to draw students' attention to certain materials to prepare for an exam, while (33.3 %) were unsure about such a practice. Similarly, practice-wise, most teachers decided that they would draw students' attention to important material either on a regular basis (60 %) or sporadically (33.3 %). Notably, most teachers (93 %) agreed that in actual situations, they would not deduct more points for a wrong answer than for leaving the answer blank, though they were somewhat sceptical as to the ethicality of such practice.

As regards assessment development, providing clues to help students figure out the answer was considered unethical by (46.7 %) of the teachers which was also reflected in the practice of (40 %) of the respondents. Most teachers (66.7 %) found it totally unethical to pinpoint the correct answers through using a higher voice pitch, yet in practice only (40 %) reported that they would entirely refrain from alluding to the correct answer, whereas (53.3 %) decided that they would sometimes allude to correct answers. Notably, although (53.3 %) reported that they would refrain from vocally placing more emphasis on certain parts of test instruction to allude to the right answer, teachers seemed to have a blurred vision of what constitutes ethicality in that regard; only (33.3 %) referred to that practice as unethical, the rest were either unsure (33.3 %) or judged the practice as ethical (33.3 %).

As for grading and providing feedback, teachers seemed to hold a great deal of ethical misconception regarding the fairness of the grade students are awarded. All teachers (100 %) considered it ethical to deduct scores for late work and to count students' effort or participation in the final grade. Teachers also seemed to have a blurred vision regarding whether to fail students for missing an exam (80 %). Others (80 %) believed it is ethical to grade students for exhibiting mastery even if class work was not completed. These beliefs were somehow reflected in the practices of (73 %) of the teachers who decided to count how late the homework was handed in, how much effort the student exerted (100 %), and in the case of group work, they would count other students' effort (86.6 %). In the same way (66.7 %) of the teachers reported that they would assign students a good mark for showing content mastery regardless of course assignment completion. Likewise, (46.7 %) of the teachers reported that ethicality-wise, a student who missed an exam should deserve the same score as that of a failing student. This was reflected in the practice of (60 %) of the respondents.

5.2 Assessment of Learning

To analyse students' scores, descriptive statistics were obtained and a Pearson product-moment correlation coefficient was computed to assess the relationship between teachers' practices and perspective regarding the issue of consistency. As Table 8 shows, the mean score the teachers obtained regarding their practice on the dimension of A*f*L was (15.9) and the SD was (2.1); this mean is considered low relative to the total score which is 24. The scores ranged from a low of 13 to a high of 19. In the same way, teachers' mean score on the ethicality aspect of A*f*L was (16.8), which is also quite low. The SD was (2.09). The scores ranged from a low of 14 to a high of 20. This gives indication to the fact that teachers normally do not adopt practices that reflect the use of assessment to enhance the learning process, which might allude to a lack of understanding of what, constitutes ethical versus unethical assessment practices. Pearson product-moment correlation coefficients between teachers' practices and perceptions of ethicality was not significant at 0.05, $r = 0.506$, $p = 0.054$.

Regarding teachers' responses to the sub-items subsumed under the fourth dimension (A*f*L), in terms of preparation, most of the teachers (86.7 %) seemed certain of the necessity of comprehensively covering the content before considering assessing students' mastery. This was also clearly reflected in the practice of the majority of the respondents (86.7 %) who agreed that they would not assess students until they had made sure they have covered the intended material. In the same way,

Table 8 Teachers' practice and perception of assessment for learning (AFL)

	N	Range	Minimum	Maximum	Mean	SD	Pearson correlation	Sig. (2-tailed)
Practice	28	6.00	13.00	19.00	15.9	2.144	0.506	0.054
Ethicality	28	6.00	14.00	20.00	16.8	2.09		

(73.3 %) believed in the ethicality of using multiple means for assessment; this was also reflected in the practice of (60 %) of the teachers.

When it comes to grading and providing feedback, some uncertainty as to whether to consider peer evaluation of oral performance an ethical practice could be observed, only (6.7 %) thought it is fair or ethical to include peer assessment in the students' final grade, whereas (73.3 %) were unsure or quite divided in their opinions; correspondingly only (13.3 %) agreed to regularly include that kind of rating in assessment. However, when it comes to writing assessment, (46.7 %) teachers seemed to be more tolerant to accept peer rating as an ethical practice. Nevertheless, a discrepancy can be discerned when examining teachers' practice. Only (13.3 %) of the teachers reported that they usually resort to peer rating to correct either oral reports or writing performance; others employed peer assessment sporadically—(23 %) and (33 %) for both oral and written performance respectively. This shows that teachers are still reluctant about involving students in the assessment process; that also explains why many teachers reported that they usually (46.7 %) or occasionally (40 %) weigh tests heavily compared to other means of assessment. This was backed by a strong conviction that testing should be given precedence; (86.6 %) of the teachers thought it is totally or somewhat ethical to weight tests heavily.

As regards to communicating students' results, considerably although student-teacher conferencing was considered ethical by (73.3 %), only (40 %) confirmed that they would regularly perform conferencing sessions, while (33.3 %) reported that they would use conferencing occasionally. Similarly, slowing down the teaching pace according to students' results was considered ethical by (66.7 %) of the teachers. This was reflected in the practices of (33.3 %) of the respondents, who pinpointed that they would act responsively to students' results, and also in the practice of (53.3 %) who indicated that they would occasionally adopt that responsive action. Categorizing students and labelling them as high, low, at risk was thought to be unethical by only (33.3 %) of the study sample, and was as well reflected in the responses of (40 %) of the teachers. The rest decided to label students according to their level either occasionally (46.7 %) or regularly (13.3 %).

To analyse students' scores on the dimension of equity, descriptive statistics were obtained and a Pearson product-moment correlation coefficient was computed to assess the relationship between teacher's practices and perspectives regarding equity. As indicated in Table 9 shows, teachers' mean score on the ethicality aspect of the "equity" dimension was (39.3) and the SD was (3.5); this mean is considered moderate relative to the total score which is 51. In the same way, teachers' mean score on the practice aspect of the "equity" dimension was (37.1), and the SD was (3.9). The scores ranged from a low of (27) to a high of (44). This gives indication to the fact that teachers somehow adopted assessment practices to ensure fairness and equity among students. Pearson product-moment correlation coefficients between teachers' practices and perceptions of ethicality was not significant at 0.05, $r = 0.262$, $p = 0.346$.

As for teachers' response to the sub-items subsumed under fifth dimension-equity, percentages and frequencies show that, in terms of assessment

Table 9 Teachers' practice and perception of equity

	N	Range	Minimum	Maximum	Mean	SD	Pearson correlation	Sig. (2-tailed)
Ethicality	28	10.00	37.00	47.00	39.3	3.514	−0.262	0.346
Practice	28	17.00	27.00	44.00	37.1	3.99		

development, most of the teachers (60 %) seemed certain that any test should cater for students' interests, yet the majority (60 %) occasionally reported that they would follow that practice. Providing help to weak students was considered unethical by (40 %) of the teachers, yet (46.7 %) were sceptical as to the ethicality of such practices. Nonetheless, as far as practice is concerned, (60 %) of the teachers reported that they would avoid giving extra clues to weak students. Taking into account the needs of students with special needs was considered ethical by (60.7 %), yet only (40 %) of the teachers indicated that they would attune, via assessment methods, to students' special needs.

As for administering assessment tools, (46.7 %) considered it ethically unproblematic to draw a student's attention to an item he has missed by mistake. Correspondingly, only (32.3 %) would regularly resort to that practice or would do so occasionally (32.3 %). This viewpoint regarding missed items was not endorsed when it comes to items students answered wrongly; most of the respondents (71.4 %) agreed that it was unethical to draw students' attention to incorrectly answered questions, yet in actual classroom situations only (32.3 %) would totally avoid correcting students' incorrect answers. In the same way, reminding students of what was learned during a test was considered unethical by (73.3 %). The practices of both translating difficult words during a test and giving slow students extra time were considered unfair by only (20 %) of the respondents. With regard to practice, (40 %) of the teachers reported that they would entirely avoid translating words, while only (32.3 %) would avoid giving slow students extra time.

In terms of providing feedback, teachers seemed to view bias towards weak students as an ethical practice, i.e., only few teachers admitted to the unethicality of giving extra marks to a weak class (33.3 %), however, a higher number of teachers (53.3 %) agreed on the unethicality of bias against an advanced class. As far as practice is concerned, only (40 %) would entirely refrain from being less strict with a weak class, whereas (53.3 %) of the teachers would not deprive high-level classes from getting chances for getting extra marks. Unexpectedly, teachers' stance towards the ethicality of relying on teachers' discretion in assigning grades was not quite definite; only (33.3 %) agreed that relying on the teacher's own impression is unethical. On the contrary, when queried about their practice, most teachers (53.3 %) decided that they would avoid unsubstantiated conclusions even if they sounded self-evident.

Addressing students' individual circumstances formulated a part of teachers' perceived ethical dilemma. For instance, although only (40 %) of the respondents thought it was unethical to be biased toward a student due to his unprivileged economic condition, about (60 %) would completely refrain from giving extra marks due to economic hardships. Similarly, (40 %) of the teachers did not ethically accept bumping up students' marks to make up for temporary circumstances. On the other hand, (40 %) were not sure of how to perceive such cases. As far as practices are concerned, only (46.7 %) made the decision to avoid giving an unprivileged student the unfair advantage of getting extra marks. Even though hiding students' identity during grading students' work to exclude any chances of bias or favouritism-was considered ethical by the majority of the teachers (66.7 %), only (26.7 %) admitted that they would regularly follow that practice in real classroom conditions.

5.3 Relationship between Beliefs/Practices and Other Factors

EFL teachers in this study possessed quite distinct ESL teaching and assessment experience and received different types of training. Teachers were divided according to their (a) years of teaching experience (1–9, 10–19 and 20–30); (b) years of assessment experience (1–10, 11–15 and 16–20) and according to the (c) type of assessment training they were exposed to (no training, pre-service, in-service and other training strategies). Differences between mean scores on both ethicality and practice were examined using 3 × 1 univariate ANOVA (F) tests to look at each dependent variable (ethicality and practices) to see if the three independent variables have a significant impact on them as displayed in Table 10.

Table 10 shows that except for the main effect of teachers' experience on practice, no main effects were found for the three independent variables on both

Table 10 Univariate analysis of variance: main and interactional effects

Source	Dependent variable	Sum of squares	d.f.	Mean squares	F	P	Partial eta squared
Teaching exp1.	Ethicality	27.678	2	13.83	0.450	0.646	0.110
	Practice	309.875	2	154.9	8.502	0.003*	0.92
Assessment exp.	Ethicality	127.804	2	63.90	2.079	0.160	0.360
	Practice	29.687	2	14.84	0.815	0.461	0.163
Training	Ethicality	102.026	3	34.00	1.106	0.377	0.240
	Practice	127.074	3	42.35	2.324	0.116	0.472
Experience * exp_assess	Ethicality	135.458	2	67.72	2.203	0.145	0.379
	Practice	20.159	2	10.07	0.553	0.586	0.125
Experience * training	Ethicality	3.000	1	3.00	0.098	0.759	0.060
	Practice	21.333	1	21.33	1.17	0.296	0.173
Exp_assess * training	Ethicality	9.375	1	9.375	0.305	0.589	0.081
	Practice	0.453	1	0.453	0.025	0.877	0.053
Error		273.342	15	18.22			

Note *Interaction between two or more variables. *Exp* experience; *assess* assessment

Table 11 Mean difference between years of teaching and practice in Tukey post hoc test

	1–9 years	10–19 years	20–30 years
1–9 years		7.1*	9.3*
10–19 years	7.1*		2.3
20–30 years	9.3*	2.3	

Note *The locations of significant group differences in both Tukey HSD post hoc multiple comparisons

teachers' ethicality and practice. In other words, no statistically significant differences were found between teachers of distinct assessment experiences or differences between teachers exposed to various types of training in terms of ethicality, $p = 0.160$ and $p = 0.377$ in both cases respectively. Similarly, participants of different assessment and training experiences did not exhibit tangible differences in terms of their assessment practice, $p = 0.461$ and $p = 0.116$ for training and assessment experience respectively. Noticeably, no interaction at 0.05 between the study independent variables was found.

Therefore, as indicated in Table 10, it appears that teachers' previous experience has significant univariate main effect on teachers practice, F (2.15) = 8.5, $p = 0.003$, partial eta squared = 0.092. This means that participants in the three groups with different years of EFL teaching experience varied significantly in their mean scores on practices of ethical assessment. To examine the location of group differences, the statistical procedures of post hoc multiple comparisons were applied, as this study did not propose hypotheses about specific group differences. The Tukey HSD test was used.

In the post hoc multiple comparisons test of this study, since group sizes were unequal, harmonic mean sample size was used. In terms of the relationship between EFL teachers' years of English teaching and ethical assessment practice, the results showed that teachers from 1 to 9 years of experience and those with 10–19 differed significantly at $p < 0.05$; teachers with 10–19 years of experience performed better (M = 111.4) than those with 1–9 years of experience (M = 99) with respect to practice. Certainly also, teachers with 20–30 years of teaching experiences performed better (M = 113) than those with 1–9 years of experience (M = 99). Yet, notably, there were no statistically significant differences between teachers with 10–19 and those with 20–30 years of experience in terms of assessment practice, M = 223, (Table 11).

6 Discussion

The current study aimed at identifying the consistency between the teachers' ethical beliefs and their classroom assessment practices. Generally speaking, although teachers seemed somehow aware of what constitutes ethical versus unethical assessment practices, a discrepancy between their ethical perceptions and the course of actions they chose to adopt could be detected. In other words, EFL teachers' notions of ethical assessment did not significantly bear upon their assessment practices. Principally, the teachers have to face the main dilemma of striking a balance between providing maximum support to individual learners and being honest to ensure fairness and support long term learning. In addition, it can be induced that intuition and discretion were given precedence when judging assessment fairness. Therefore, when asked to provide justifications for their answers, teachers were unable to apply ethicality standards and they appeared to be more governed by official considerations. The teachers reported also that external factors, such as time and curriculum constraints and mandated assessment policies, might affect assessment fairness or their adherence to ethical beliefs. Notwithstanding these remarks, in some cases teachers seemed to resort to ethical behaviour even though they could not perceive the underlying ethical motive of their actions. It seemed that teachers never used reflection to think of their assessment-oriented practices.

In particular, issues of confidentiality and transparency were well dealt with by most teachers and somehow borderlines were drawn between what should be public and what should be private. Nevertheless, some discrepancies between teachers' convictions and actions could be discerned. Most teachers agreed to the ethicality of stating how a task will be graded; yet they did not have the same attitude about sharing rubrics with students which might be ascribed to the teachers' belief that grading is an exclusive teacher's responsibility. Moreover, the teacher's image as a guardian might have caused many teachers to think that only points of strength should be discussed with a student; that is why many teachers did not realize the unethicality of hiding any information from the student. Noticeably, although teachers' answers reflected their unawareness of the unethicality of disclosing confidential information about students' academic achievement, in practice most of them showed a tendency to keep students' information somewhat confidential. This indicates that teachers were in many cases driven by their rational intuition or "practical wisdom" in Tierney's (2010) terms.

As far as *consistency* or conformity between assessment and curriculum objectives is concerned, a moderate correlation could be discerned between what teachers believed and how they tended to act. Generally, the respondents acknowledged the importance of assessing students on material they knew students had mastered which was reflected in their practice. However, many areas of discrepancy existed between what teachers believed and how they tended to act. For instance, even though teachers believed that students should be exposed to activities that enable them to anticipate the format of the assessment procedures, some teachers refrained from adopting such practices. Interestingly, most teachers had blurred vision regarding the consistency between the course objectives and methods of assessment adopted which was reflected also in their practices. Sometimes both teachers' perceptions and practices reflect their lack of awareness of what constitutes ethical assessment. For instance, when it comes to grading students, teachers seemed not well cognizant that the scores the students receive should tightly reflect their mastery of the skills stated by the objectives.

The present study also gave some indication that ethical dilemmas centring on *score pollution* made up the majority of incidents. These findings are consistent with the findings of Green et al. (2007). Issues that did not yield a great deal of conflict were those related to training students on certain exam questions and the unjustifiable deduction of scores to minimize guessing. Furthermore, teachers tended to ethically justify practices that reflect incorporation of students' non-academic performance such as effort, participation, improvements, laziness (…) etc. This indicates that teachers normally do not always ensure that the assessment tools and grading procedures employed actually reflect students' targeted competences. Thus, perhaps with clearer guidelines about what constitutes score pollution and why score pollution is unethical, these ethical dilemmas would not be so prevalent.

As for *AfL*, teachers were inclined to use multiple methods of assessment, yet a disproportionate heavy weight was allotted to testing as the best method for assessing students. Peer evaluation and self-evaluation were looked at with a lot of suspicion and integrating them in the classroom assessment plan was considered unethical by nearly most of the teachers. Many teachers thought that it is important to give students tasks that suit them and that not all students should be tested in the same way. Discrepancy between teachers' convictions and actions were obvious in that the high ethical value they accorded to practices such as slowing the teaching pace to adapt to students' needs and conferencing with students is not transferred to their everyday practice. In sum, teachers' practices in this respect contracticted the concept of A*f*L.

Regarding *equity*, teachers' seemed to hold a clearer vision regarding ethicality, even if they were hesitant to apply what they perceived in their daily practices. In other words, the ethicality or otherwise of some practices, such as addressing students' interests, providing more than one format for a test, avoiding clues, seemed to be well settled and agreed upon by most teachers. However, in spite of teachers' apparent ethical approach, their practices fell short of reflecting their way of thinking. This might be due to fact that teachers are constrained by many factors that direct their practices. For instance, teachers perceived that students' special needs should be addressed; yet practically they found it difficult to address these needs. One interpretation is that teachers might have felt that any adaption to the assessment process should be the responsibility of other stakeholders, rather than the teacher himself. In some cases, teachers' ethical practices were driven by their rational intuition, even if they were inconsistent with their ethical convictions. For example, most teachers avoided providing clues to weak students, even though they could not ethically justify their sound practices.

The results of the current study give also some indication that teachers' perceptions of ethicality was not affected by their teaching or assessment experience. This can be attributed to the fact that most training programs focus on the practicalities of the assessment process and pay no heed to ethical issues underlying teachers' actions. Nonetheless, it was proved that teaching experience has a significant impact on teachers' ethical practices regardless of the training received. This contradicts what was suggested by previous research that ethical reasoning in assessment does not develop on the job (Green et al., 2007). Yet, it seems that subsequent to fifteen years of experience, teachers tended to get accustomed to certain practices and that no remarkable change in their ethical decisions can be discerned.

7 Recommendations and Limitations

Results of the current study imply that many areas were considered controversial for most teachers. One of these areas was using multiple forms for assessing students. Another issue was consistency between the assessment methods used and the curriculum objectives and classroom activities. Equity issues also seem to be blurred for most teachers. Most teachers tended to adopt an over protective stance towards students regardless of whether or not this stood in sharp contrast to their own beliefs of what constitutes ethical assessment. Ensuring equity by avoiding bias toward certain groups such as students with limited ability or disability is also not well substantiated for teachers. In other words, although some rules were morally self-evident for the teachers, discrepancies between what teachers believed to be fair and what they got used to doing in class was obvious.

The generalizability of the specific results of this study may be limited by its use of a self-report survey and the limited number of the participating sample. Future studies may use multiple methods of data collection including classroom observation, analysis of teacher-made tests, and teacher interviews to validate teacher self-reports. In the future, also, the survey should be sent to a more representative sample selected from a variety of geographic regions across the country. The current data suggest that more time needs to be spent in confronting the ethical dilemmas of assessment and methods of approaching and resolving these dilemmas.

Results of the current study imply that general measurement training by itself cannot compensate for novices' lack of experience in terms of fair assessment. Nevertheless, the findings testify to the value of training that is particularly focused on fair assessment and ethicality dilemmas.

In light of previous results, it is recommended that teachers should be directed to put into consideration score pollution issues by providing clearer guidelines about what constitutes score pollution and why score pollution is unethical. Furthermore, explicit instruction in ethical concepts, such as equity, consistency, transparency and confidentiality, ought to be part of teacher pre-service training program as well in-service programs with ample chances for putting these ethical codes into practice by directly relating to the daily work in which teachers engage. Thus, the current study suggests pre-service and in-service training should address the issue of how to strike a balance between knowing a lot about students and avoiding biases. Teachers' awareness of the discrepancy between their roles as assistants to students and their roles as agents who takes part in establishing an accountable educational system should as well be raised.

Accordingly, continued research is needed to define more clearly the ethical issues teachers face as regards assessment. In addition, self-reflection practices should be encouraged among teachers; this can be accomplished by requiring teachers to report their regular ethical dilemmas pertinent to classroom assessment using reflection logs, diaries or group discussion.

Appendix

Teacher's perception and practice survey

	How often you do that			Ethicality		
	U	S	N	E	SE	U
A. Preparation for assessment						
1. I state how I will grade a task when I assigns it (transparency+)						
2. I spend a class period to train students on test-taking skills (e.g., not spending too much time on one question, eliminating impossible answers, guessing) (consistency+)						
3. To prepare students to an upcoming test, I administer a parallel form of the test. The parallel form is another version of the test; however, the items are not the same as those on the final form of the achievement test (consistency+)						
4. Based on my review of a university final test framework, I create learning activities with specific questions that are included in the annual achievement test (score pollution)						
5. I don't assess students until I make sure that I comprehensively covered the material and that students possess all the skills needed (assessment for learning)						
6. I tell students what materials are important to learn in preparing for a classroom assessment (score pollution)						
7. To minimize guessing, I announce that I will deduct more points for a wrong answer than for leaving the answer blank (pollution)						
B. Assessment development						
8. I use a previously designed test without referring to the objectives of the syllabus (consistency−)						
9. I assess my students' knowledge by using many types of assessments: multiple-choice tests, essays, projects, portfolios (equity+)						
10. When I develop a rubric for correcting students' written composition I hide the information about it as highly confidential (transparency−)						

(continued)

(continued)

	How often you do that			Ethicality		
	U	S	N	E	SE	U
11. In a reading test, I include texts that are relevant to the interests of various students' sub-groups (equity+)						
12. In a vocabulary test, I use assessment methods that students have never encountered before, for examples, drawing, giving examples, fill in a table (consistency−)						
13. I make sure that the activities included in a test were quite similar to activities presented in class (consistency+)						
14. I assess oral proficiency only through observing students during classroom discussion (consistency−)						
15. For the final exam, I use a few surprise items about topics that were not on the study guide (consistency−)						
16. If I have a blind student in my class, I design a recorded version of the test (equity+)						
17. I use more than one format of the same test to prevent cheating (equity+)						
18. In a vocabulary test, I put some clues in each item to help students find the answers easily (score pollution)						
19. For MCQ, I try to make the correct answer longer than others to help weak students answer better (equity−)						
C. Administering assessment						
20. If I notice that a student has skipped a question. I stop at his/her desk and show the him/her where to record the answer he is working on (equity−)						
21. While administering a test, when I notice that a student has missed a problem that he obviously knows, I stand by the student's desk, taps my finger by the incorrect problem, shake my head, and walk on to the next desk (equity−)						
22. While applying a test, I remind any student who stumbles on a question of what we learned by giving him or her a hint (equity−)						
23. Upon students' request, I would translate a difficult word that hinder students' understanding of a reading comprehension text (equity−)						
24. On a final exam, I would read all the test instruction orally with some emphasis on the key parts to help all students answer the questions easily (score pollution)						
25. For a slow student, I allot extra test time even if the test time has passed (equity)						
26. In a listening comprehension test, I read the text loudly to all students trying to highlight key parts by using a higher voice tone (score pollution)						

(continued)

(continued)

	How often you do that			Ethicality		
	U	S	N	E	SE	U
27. I would allow a student with a learning disability, i.e., a blind student, to use a tape-recorder when the student answers the essay questions on a grammar test (equity+)						
D. Grading and feedback						
28. I lower grades for late work by one score or more for each day (score pollution)						
29. I consider student effort when determining grades class (score pollution)						
30. In case of teaching two or more classes, I try be less strict in grading a class whose students I believe are weaker or slower (equity−)						
31. In an advanced reading class, I would assess reading based on students' final semester grade on two multiple choice tests (consistency−)						
32. For a group project, I base each student's grade on the group's product and a heavily weighted individual component (score pollution)						
33. To encourage lively discussion in English, I count class participation as part of the final grade (multiple assessment)						
34. I weigh tests heavily in determining students' final grades compared with other methods, i.e., homework, discussion, projects, presentation (multiple assessment−)						
35. I would give a student an F for the course because he/she missed the final exam (score pollution)						
36. I use student's peer assessment as a part of a final grade on an oral report (multiple assessment+)						
37. I lower class grades for disruptive behavior (score pollution)						
38. I would give extra scores to a student to make up for his/her underprivileged economic conditions (equity−)						
39. If I know a student had a bad week because of problems at home, I would bump his/her participation grade up a few points to compensate for his bad score on a quiz (equity−)						
40. If I believe that that students' work is rarely perfect in one of classes, I would make the decision of assigning very few grades of "A" to my class (equity−)						
41. I would change one student's course grade from a B+ to an A because tests and papers showed he/she had mastered the course objectives even though he had not completed some of his homework assignments (score pollution)						

(continued)

(continued)

	How often you do that			Ethicality		
	U	S	N	E	SE	U
42. I would offer extra credit opportunities to all the classes I teach except the advanced class (equity−)						
43. I hide the identity of the students (by concealing the name) whose essay test I'm grading so I won't identify them (equity+)						
44. I use peer evaluation with reference to certain rubric to help correct writing essays quickly (assessment for learning)						
E. Communication of results and feedback						
45. To enhance self-esteem, I address only students' strengths when correcting students' writing (transparency−)						
46. I spend time conferencing with each student to explain points of strength and weakness in their writing performance (assessment for learning)						
47. Based on the students' results, I would slow down my teaching pace to adapt to students' needs (assessment for learning)						
48. To motivate students to perform better, I would announce that I'm passing out scored tests to students in order of points earned, from the top score to the bottom score (confidentiality−)						
49. To calm the fears of distraught parents, I would compare their child's achievement scores with the results of the student's cousin who is also in the class (confidential−)						
50. I categorize students by labelling them to low level, high level, at risk (equity−)						

Usually (U); sometimes (S); never (N); ethical (E); somewhat Ethical (SE); unethical (U)

References

Airasian, P. (2005). *Assessment in the classroom: A concise approach* (2nd ed.). Boston, MA: McGraw-Hill Company.

Blanchard, J. (2008). Learning awareness: Constructing formative assessment in the classroom, in the school and across schools. *Curriculum Journal*, *19*(3), 137–150. doi:10.1080/09585170802357454

Brookhart, S. M. (2004b). Classroom assessment: Tensions and intersections in theory and practice. *Teachers College Record*, *106*(3), 429–458. doi:10.1111/j.1467-9620.2004.00346.x

Bursuck, W. D., Munk, D., & Olson, M. (1999). The fairness of report card grading adaptations: What do students with and without learning disabilities think? *Remedial and Special Education, 20*(2), 84–92. doi:10.1177/074193259902000205

Buzzelli, C., & Johnston, B. (2002). *The moral dimensions of teaching: Language, power, and culture in classroom interaction.* London: Routledge Falmer.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 221–256). Westport: American Council on Education & Praeger.

Chan, Y. (2008). Elementary school EFL teachers' beliefs and practices of multiple assessments. *Reflections on English Language Teaching*, *7*(1), 37–62. Retrieved from: http://www.nus.edu.sg/celc/research/relt.php

Chang, Ch-W. (2006). Teachers' beliefs towards oral language assessment in Taiwan collegiate EFL classrooms. A paper presented at 2006 International Conference on English Instruction and Assessment. Retrieved from: http://fllcccu.ccu.edu.tw/conference/2005conference_2/download/C03.pdf

Dann, R. (2002). *Promoting assessment as learning: Improving the learning process.* London: Routledge Farmer.

Educational Testing Service. (2002). *ETS standards for quality and fairness*. Princeton: Educational Testing Service.

Gebril, A., & Brown, G. T. L. (2013). The effect of high-stakes examination system on teacher beliefs: Egyptian teachers' conceptions of assessment. *Assessment in Education: Principles, Policy and Practice, 21*(1), 16–33. doi:10.1080/0969594X.2013.831030

Gipps, C., & Murphy, P. (1994). *A fair test? Assessment, achievement and equity*. Buckingham: Open University Press.

Graham, P. (2005). Classroom-based assessment: Changing knowledge and practice through pre-service teacher education. *Teaching and Teacher Education*, *21*(6), 607–621. doi:10.1016/j.Tate.2005.05.001

Green, S., Johnson, R., Kim, D., & Pope, N. (2007). Ethics in classroom assessment practices: Issues and attitudes. *Teaching and Teacher Education*, *23,* 999–1011. doi:10.1016/j.tate.2006.04.042

Hidri, S. (2015). Conceptions of assessment: Investigating what assessment means to secondary and university teachers. *Arab Journal of Applied Linguistics, 1*(1), 19–34.

International Language Testing Association. (2000). *Code of ethics for ILTA*. Retrieved on March 30, 2009 from http://www.iltaonline.com/code.pdf

JCSEE. (2003). *The student evaluation standards*. Arlen Gullickson, Chair. Thousand Oaks, CA: Corwin.

Lu, Ai-ying. (2003). *Teachers' beliefs and classroom assessments: A case study of two university instructors of English* (Unpublished master's thesis). National Taiwan Normal University, Taiwan, R.O.C.

McMillan, J. (2007). *Formative classroom assessment: Theory into practice*. New York: Teachers College Press.

Messick, S. (1995). Validity of Psychological Assessment: Validation of Inferences from Persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. Retrieved from: http://dx.doi.org/10.1037//0003-066X.50.9.741

Payne, D. A. (2003). *Applied educational assessment* (2nd ed.). Belmont, CA: Wadsworth.

Pelly, C., P., & Allison, D. (2000). Investigating the views of teachers on assessment of English language learning in the Singapore education system. *Hong Kong Journal of Applied Linguistics, 5*(1), 81–106.

Plake, B., & Jones, P. (2002). *Ensuring fair testing practices. The responsibilities of test sponsors, test developers, test administrators, and test takers in ensuring fair testing practices.* A paper presented at the February 2002 meeting of the Association of Test Publishers, Carlsbad, CA. Retrieved from: http://www.testpublishers.org/assets/documents/Ensuring%20Fair%20Testing%20volume%204%20issue%201%20070202.pdf

Pope, N., Green, S., Johnson, R., & Mitchell, M. (2009). Examining teacher ethical dilemmas in classroom assessment. *Teaching and Teacher Education 25,* 778–782. doi:10.1016/j.tate.2008.11.013

Popham, W. J. (2000). *Modern educational measurement: Practical guidelines for educational leaders*. Needham, MA: Allyn & Bacon.

Shepard, L. (2005). Linking formative assessment to scaffolding. *Educational Leadership*, *63*(3), 66–70.

Shepard, L. (2007). Formative assessment: Caveat emptor. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 279–303). Mahwah, NJ: Lawrence Erlbaum Associates.

Simon, M., Chitpin, S., & Yahya, R. (2010). Pre-service teachers' thinking about student assessment issues. *International Journal of Education* *2*(2), 1–20. doi:10.5296/ije.v2i2.490

Stiggins, R. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan*, *83*(10), 758–765. doi:10.1177/003172170208301010

Szpyrka, D. (2001). Exploration of instruction, assessment, and equity in the middle school science classroom. *Dissertation Abstracts International*, *62*(10), Section: A, 3287–3548.

Tierney, R. (2010). *Insights into fairness in classroom assessment: Experienced English teachers share their practical wisdom* (PhD dissertation). University of Ottawa, Canada (UMI No 69109).

Volante, L. (2006). Reducing bias in classroom assessment and evaluation. *Orbit*, *36*(2), 34–36. Retrieved from: https://ezpa.library.ualberta.ca/ezpAuthen.cgi?url=search.proquest.com/docview/213742716?accountid=144744

Yip, D. & Cheung, D. (2005). Teachers' concerns on school-based assessment of practical work. *Journal of Biological Education*, *39*(4), 156–162. doi:10.1080/00219266.2005.9655989

Zhang, Z., & Burry-Stock, J. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. *Applied Measurement in Education*, *16*(4), 323–342. doi:10.1207/S15324818AME1604_4

Zoeckler, L. (2005). Moral dimensions of grading in high school English. *ProQuest digital dissertations.* UMI No. AAT3183500.

Problematizing Teachers' Exclusion from Designing Exit Tests

Abderrazak Dammak

Abstract This paper investigates teachers' exclusion from designing exit tests and the justifications of different stakeholders. Teachers and decision-makers can justify exclusion from different perspectives. This small-scale critical exploratory study, which was conducted in a vocational institute in the United Arab Emirates, aims at problematizing the issue of depriving teachers from designing exit tests. It also intends to raise teachers' awareness about this issue. According to proponents of the critical theory, questioning perpetuated situations and raising others' awareness about similar experiences can lead to a change in the dominating culture of many workplaces. In this study, the researcher used questionnaires and semi-structured interviews as tools of data collection to problematize this issue and compare the various justifications of the two main stakeholders: Teachers and decision makers. Results of this critical exploratory study showed that most teachers are not allowed to participate in designing exit tests. Results also revealed that most of the excluded teachers are assessment literate, aware of the objectives and principles of testing, which may refute the alleged assumptions about teachers' incompetence. Moreover, results of the study showed that the impact of the study was immediate as most of the excluded teachers expressed their intention to discuss this issue with decision makers.

Keywords Critical language testing · Teachers' exclusion · Assessment literacy · Fair evaluation

1 Introduction

Throughout the last few decades, teacher participation in decision-making has emerged as a controversial issue. Critical theorists started questioning perpetuating situations in an attempt to raise teachers' awareness and change dominating cultures

A. Dammak (✉)
Exeter University, Exeter, UK
e-mail: damarazak66@gmail.com

S. Hidri and C. Coombe (eds.), *Evaluation in Foreign Language Education in the Middle East and North Africa*, Second Language Learning and Teaching,
DOI 10.1007/978-3-319-43234-2_5

in educational workplaces. Critical language testing, in particular, questioned the roles of different participants in the testing process. This critical exploratory study aims at problematizing the issue of depriving teachers from designing exit tests. It also intends to raise teachers' awareness about this issue. My initiative to discuss this issue and therefore raise my colleagues' awareness stems from my personal experience of exclusion. As an English language teacher developing the curriculum and designing daily quizzes and weekly tests, being deprived of being involved in designing exit tests is an issue that should be raised and discussed. To my knowledge, except for one study (Troudi, Coombe, & Al-Hamly, 2009), no published research on teachers' exclusion from designing exit tests has been conducted in this region. I thought that questioning perpetuated situations and raising others' awareness about similar experiences can lead to a change in the dominating culture of my workplace. To discuss this topic, the first part of this paper will deal with the theoretical background. It will be followed by sections on methodology, results, and discussion respectively. The conclusion will focus on the limitations of the study and recommendations for future research.

2 Theoretical Background

2.1 Critical Applied Linguistics and the Problematizing Stance

Questioning teachers' exclusion from assessing students cannot be understood in depth without situating teachers' exclusion under the umbrella of evaluation in general and Critical Language Testing (henceforth CLT) in particular.

According to Pennycook (2010, p. 16.3), "critical applied linguistics is more than just a critical dimension added on to applied linguistics: It involves a constant scepticism, a restive problematization of the givens of applied linguistics, and presents a way of doing applied linguistics that seek to connect it to questions of gender, class, sexuality, race, ethnicity, culture, identity, politics, ideology, and discourse." Influenced by the Marxist theory of class struggle, critical applied linguistics tackles the previous concerns to unveil dominant hegemonies, ways of perpetuating existing relations and interests, and seeks to change power relations. Pennycook (2001) raised a number of concerns to be able to reach a better understanding of critical applied linguistics. According to him, applied linguistics, praxis, being critical, micro and macro relationships, critical social inquiry, critical theory, problematizing givens, self-reflexivity, preferred features, and heterosis are the concerns of critical applied linguistics. He looks at applied linguistics in all its "contents as a constant reciprocal relation between theory and practice" (ibid, p. 3). He argues that critical applied linguistics is a "way of thinking and doing" and contends that one of the challenges for critical applied linguistics is to find ways of mapping the micro and macro relationships. It is fundamental to understand the relations between the concepts of society, ideology, education, and classroom conversation discourses and second language acquisition. Therefore, critical applied linguistics should not only highlight the relationship between language contents and social contexts but should also view these relationships as problematic. Critical applied linguistics should go beyond exploring correlational relationships between language and society and raise questions of power, access, desire, and resistance. This is to assume that critical work with a sceptical eye can help to raise such questions as it engages with problems of inequality, injustice, and human rights. The ability to raise questions and be critical cannot be achieved without questioning the givens and asking critically about their assumptions. Pennycook (ibid, p. 7) holds that "a critical component of critical work is always turning a sceptical eye toward assumptions, ideas that have become naturalized, notions that are no longer questioned." In addition to concerns, Pennycook gave an overview of domains comprising critical applied linguistics. They are critical discourse analysis and critical literacy, critical approaches to translation, language teaching, language planning and language rights, language literacy and workplace settings. Language testing is another domain of critical applied linguistics, especially considering Shohamy's view that language testing follows a political agenda. She argues that language tests were used as "triggers and vehicles through which bureaucratic agendas could be achieved" (Shohamy, 1997, p. 346).

2.2 *Critical Language Testing*

Shohamy (2001) discusses the main features of Critical Language Testing. She claims that CLT assumes that the act of language testing is not neutral as it is the product and agent of political, educational, social, and ideological agendas that determine the life and future of the different test stakeholders. She adds that CLT views test takers as political subjects and encourages them to criticize and critique the value inherent in tests as they are embedded in educational, cultural, and political contexts. Moreover, CLT asks about the agendas behind implementing tests. It also asks questions about "what knowledge tests are based on (…) is it something that can be negotiated, challenged, and appropriated?" (ibid, p. 132). Adding to that, CLT challenges the psychometric traditions of language testing and advocates interpretive ones "whereby different meanings and interpretations are considered for test scores" (ibid, p. 132). CLT challenges the reliance on tests as the only and sole instrument of assessment and suggests the use of other assessment tools and procedures to gauge and interpret the knowledge of learners. Furthermore, CLT admits that the knowledge of testers is not comprehensive and that there is a need to rely on other sources to obtain a more solid, valid, and accurate description and interpretation of knowledge. Furthermore, CLT examines the stakeholders of tests and asks about the parties involved in designing and producing tests. It calls for a more democratic process where different stakeholders including policy makers, test writers, students, parents, and teachers are involved in designing tests.

In addition to features, she highlights the powerful uses of tests. She argues that decision makers attribute importance to tests as they “allow those in authority to control and manipulate knowledge” (Shohamy, 2001, p. 38). She emphasizes that, rarely challenged, tests serve the needs of those in power to perpetuate their dominance to enforce policies. In her critical observation, she expresses concerns about “the power of tests and their uses in society” (ibid, p. 5) and draws our attention to the fact that the voices of test takers are silent, that tests have detrimental effects on test takers, and that tests are used as disciplinary tools. She contends that using tests as disciplinary tools is “an extension of the manipulation of tests by those in authority-policy makers, principals and teachers-into effective instruments for policy making.” This issue of using tests as disciplinary tools is also discussed by Foucault (1979, p. 184) who states that “at the heart of procedure of disciplines, it manifests the subjection of those who are perceived as objects and the objectification of those who are subjected.” Parallel to this, McNamara discusses the issue of tests being used as weapons of policy reform and immigration policy, claiming that “language tests have a long history of use as instruments of social and cultural exclusion” (McNamara, 2000, p. 68).

McNamara (ibid, p. 76) states that the principles and practices of testing that have “become established as common sense or common knowledge are actually ideologically loaded to favour those in power.” He adds that critical language testing “is best understood as an intellectual project to expose the role of tests in this exercise of power.” Similarly, Shohamy (2001, p. 131) connects between the use of tests and power and justifies placing the domain of testing within the broad area of critical pedagogy by stating that critical testing “implies the need to develop critical strategies to examine the uses and consequences of tests, to monitor their power, minimize their detrimental force, reveal their misuses and empower the test takers.” She claims that critical testing attempts to criticize the field of testing, monitor, and limit the powerful uses of tests. This criticism includes regarding the act of testing as biased and not neutral, as it shapes the lives of teachers and learners. Critical testing examines the stakeholders of tests, their agendas, testing methods, and the ideology delivered through the test. Shohamy also maintains that one of the issues that critical testing problematizes revolves around the persons included in designing tests, an issue that has been scarcely tackled in the literature but which I will try to illuminate in the subsequent section.

2.3 Teachers’ Exclusion

Problematizing teachers’ exclusion from designing exit tests has scarcely been discussed. The feelings of mistrust, marginalization, and humiliation caused by exclusion have been highlighted by Shohamy (2001) and Rea-Dickins (1997). Shohamy connected exclusion to the democratization of educational systems. For her, it all revolves around power, trust and trustworthiness. She argues that the “selection of the testing body can also provide a good indication of the extent to which the educational system trusts the teachers and is willing to grant them professional authority" (2001, p. 30). She presents the experience of introducing new reading comprehension tests and the way teachers were "humiliated by the system which viewed them as potential cheaters and untrustworthy" (ibid, p. 57) by forcing them out of their classrooms and denying them access to any information about the test. Shohamy highlights on the criticality of the effect of such exclusion on the image of teachers and wonders "about the message conveyed to students when their teachers are not trusted by the system" (ibid, p. 57). Shohamy (2005, p. 106) showed more interest in teachers' exclusion and their subservient role when they "are viewed as bureaucrats; (…) [and] are being used by those in authority to carry out testing policies and thus become servants of the system."

Similarly, Rea-Dickins (1997, p. 304), who defines stakeholders as "those who make decisions and those who are affected by those decisions", relates teachers' inclusion or exclusion to the issues of power and democratization and highlights the harms of exclusion. She argues that "in terms of obvious power, some stakeholders are more important than others: The more important ones make the decisions and take action while the less important are those affected by those decisions" (ibid, p. 305). Instead of exclusion, Rea-Dickins claims that consulting and involving the different stakeholders "in the process of test development and test use reflects a growing desire among language testers to make their own tests more ethical" (ibid, p. 304). To further discuss the problem, she asked the following question: "How much control do teachers have of the assessment procedures and the tests they administer?" (ibid. p. 307). Instead of exclusion because of incompetence, Rea-Dickins proposes appropriate preparation and empowerment of teachers as potential solutions. She elucidates that teachers' participation, among other factors, can promote greater fairness in the testing process and advocates "democratization of assessment processes through greater stakeholder involvement" (ibid, p. 3). She argues that teachers can become competent at designing tests if they are heard and given opportunities to develop their understanding of the assessment process.

Hearing teachers' voices was what Troudi et al. (2009) tried to highlight in their study, which is important for two main reasons. First, it is a recent research study that elicited teachers' voices and reduced their feelings of marginalization. Second, it was conducted in the Gulf region, where this actual study is taking place. In this study conducted in the UAE and Kuwait, Troudi et al. (2009) investigated issues of assessment design and implementation in these two Gulf countries. The researchers tried to explore teachers' assessment philosophies and their roles in student assessment. Results of the study showed that the teachers' role in assessment is minor because of "the top-down managerial approaches to education and a concern for validity and quality assurance in large programmes" (ibid, p. 546). The researchers argue that exclusion was the recurrent theme and that many of such instances were noticed. Results also showed that teachers' opinions were not solicited and that they were excluded from designing assessment tools because they were "perceived not to have expertise in this area" (ibid, p. 550). What is interesting in the study is the ability of researchers to present reasons of both parties to justify teachers' exclusion from designing assessment tools. Teachers, for example, expressed how they felt distrusted and disrespected. Those who are in power, on the other hand, argue that assessment should be centralized for reasons of practicality, efficiency, and reliability. Moreover, they expressed a fear that teachers may be inclined to help their students because of these latter's involvement in teacher evaluation.

Except for the previous study, I was not able to find any study focusing on the issue of teachers' exclusion from designing exit tests in the Gulf region. The paucity of related research may grant importance to the actual study and may contribute to fill this gap about teachers' exclusion from assessment decisions. It may illuminate reasons of exclusion and shed light on the justifications of the different stakeholders. Succinctly, this study aims at answering the following research question: "Can problematizing the issue of excluding teachers from designing exit tests help to raise teachers' awareness?" In order to be able to tackle this issue, the way should be paved by answering the following sub questions: "how do classroom teachers justify their exclusion from designing exit tests?" and "how do decision makers justify teachers' exclusion?"

3 Method

3.1 Context of the Study

This study was conducted in a technical institute in the UAE. It is a vocational institute that trains students to become future technicians in the oil and gas industry. Students' age in the foundation program ranges from seventeen to twenty-two. Most of the students are Emirati with a very small number of Omani students.

Students usually spend three terms (levels 1, 2, and 3) in the foundation program and pass the exit level 3 tests to be able to join the technical program. Each term lasts a study period of six months. During the foundation course, English is the medium of instruction. All subjects, Math, Science, Lab, and the four English language skills, are taught in English. Students receive approximately six hours of instruction daily for five days a week. They are tested biweekly in all subjects. In the reading course, for example, level three teachers, design these biweekly tests collaboratively. The biweekly tests, which are administered at the end of every unit, include reading comprehension passages and vocabulary exercises and carry an assessment weight of 70. The remaining 30 are awarded to participation (10), vocabulary journals (10), and short comprehension quizzes (10). Similarly, level three teachers also design these short quizzes in a collaborative way. Teachers correct the daily and weekly tests and give students the opportunity to see and discuss their mistakes. By the end of the term, students should reach the cumulative average of 70 in each subject to be eligible to sit for the final exam. At the end of each term, students must pass an exit test to be promoted to the next level. These exit tests include maths, science, reading, writing, and communication skills.

During the final exams, students are tested one subject per day. After exams, students consult teachers to verify some answers or enquire about certain questions.

3.2 Participants

3.2.1 First Participants: Teachers

The choice of the participants was purposeful (Creswell, 2008). I targeted the teachers of the foundation program to which I am affiliated. I started questioning the existing testing policy and examined the situation with sceptical eyes. I decided to design a questionnaire about the issue of designing exit tests in order to raise teachers' awareness of issues of inclusion and exclusion in designing these tests. The teachers of the foundation program that answered the questionnaire were from different countries: U.S.A, Canada, England, Sudan, Tunisia, Egypt, Jordan, Morocco, Algeria, Kenya, and Iraq. Most of them hold a Masters' degree and have at least five years of teaching experience. They teach different subjects: Maths, Science, Reading, Writing, Communication, and Lab courses. The identity of the participants in the questionnaire was not revealed as they were promised confidentiality and anonymity.

3.2.2 Second Participant: The Head of the English Department

The second targeted participant consisted of the head of the English department. He is a Masters' degree holder with more than thirty years of experience in the field of education. His role as head consists of supervising, developing curriculum, hiring new recruits, and deciding on assessment policy. Through the interview, I aimed at questioning teachers' exclusion from designing exit tests and highlighted his justifications of teachers' exclusion from assessment practices.

3.3 Methodology

This critical exploratory study is compatible with the critical research paradigm as it aims at questioning an existing situation and raising teachers' awareness about the issue of designing exit tests. This raised awareness, if achieved, may help teachers to get involved and change future practices. Ontologically, reality in the critical research paradigm is described within a political, cultural, historical, and economic context. Mertens (2008, p. 74–75) states that the "transformative-emancipatory ontology assumption holds that there are diversities of viewpoints with regard to many social realities but that these viewpoints need to be placed within political, cultural, historical, and economic value system to understand the basis for the differences." Epistemologically, in the critical theory researchers emphasize the importance of the interactive relation between the researcher and the participants and the impact of social and historical factors that influence them. Mertens (Ibid, p. 99) also holds that the "interaction between the researchers and the participants is essential and requires a level of trust and understanding to accurately represent viewpoints of all groups fairly." Because of the transformative emancipatory assumption and the importance of interaction between researchers and participants, critical methodology is directed to raise the awareness of participants and interrogate accepted injustice and discrimination. Critical theorists are "concerned with action rather than discovery" (Edge & Richards, 1998, p. 341). From this point of view, critical researchers have an agenda of change, to improve the lives and situations of the oppressed. Likewise, raising teachers' awareness, that may lead them to question previously accepted assumptions about designing exit tests and making their voices heard, was the objective of conducting this research.

3.3.1 Data Collection Tools

Mertens (2008) argues that critical researchers may use qualitative, quantitative or mixed methods but should be aware of the contextual, historical and political factors related to the topic under study. She states that within the assumptions associated with the transformative paradigm, several of these approaches can be combined in the mixed methods design that means the use of qualitative and quantitative methods. Accordingly, critical researchers use the data collection methods that best work and serve their critical enquiry to enable them to critically study situations from cultural, economic, political, and historical perspectives. With this in mind, I used questionnaires and a semi-structured interview to investigate the issue of designing tests and raising teachers' awareness about this issue. The use of these two research tools enabled me to triangulate data, which is defined by Cohen, Manion, and Morrison (2003, p. 112) "as the use of two or more methods of data collection in the study of some aspect of human behaviour."

3.3.2 Questionnaires

Brown (2001, p. 6) defines questionnaires as "any written instruments that present respondents with a series of questions or statements to which they are to react either by writing out their answers or selecting from among existing answers." Dornyei (2003, p. 14) states that the "typical questionnaire is a highly structured data collection instrument, with most items either asking about very specific pieces of information (…) or giving various response options for the respondent to choose from, for example by ticking a box. This makes questionnaire data particularly suited for quantitative, statistical analysis."

With the above-mentioned advantages of questionnaires in mind, I chose them as my first instrument to collect data (Appendix 2). The choice was purposeful as they enabled me to gather as much data as possible in a very short period (Gillhman, 2000). They also allowed me to gather quantitative data that can be easily classified and analysed. To obtain qualitative data, I added an open-ended question at the end of the questionnaire to allow participants to express their points of views and write about issues that the questionnaire may have overlooked. Dornyei (2003, p. 15) defends this option by stating "that some partially open-ended questions can play an important role in questionnaires."

The title was clear and the instructions at the beginning were short, informative, and well-pitched. Dornyei (2003) suggests that these may determine respondents' feelings toward the questionnaires. In the instruction section, I informed respondents about the study and reasons for conducting the questionnaire.

Before designing the questionnaire, I conducted a group discussion during the professional development days in my institute as a part of my preparation. The aim was to brainstorm, elicit ideas, and come out with a short list regarding the issue of designing exit tests. This discussion along with the available literature helped me to identify the critical concepts and provided me with information on the relevant points and issues that I needed to address in the questionnaire.

After modifying the questionnaire, I piloted it with five colleagues from another institution where the testing practices are similar and where teachers are not involved in designing exit tests. The results of the pilot study were encouraging as respondents answered without complaining about ambiguity. Encouraged by the results of the pilot project, I distributed twenty-six questionnaires to my colleagues and gave them ample time to hand them back to me. Finally, only three of the respondents failed to answer and apologized for declining to answer the questionnaires. The content of the questionnaire elicited the following information: Demographic information about the participants, teachers' roles in evaluation, teachers' assessment literacy, teachers' involvement in designing exit tests, and reasons of teachers' exclusion from designing exit tests.

3.3.3 Semi-structured Interview

In order to triangulate data and create equilibrium between quantitative and qualitative data, I used a semi-structured interview as a second tool of data collection. Punch (2009, p. 144) states that the "interview is the most prominent data collection tool in qualitative research." Drawing from that, I decided to use a semi-structured interview to gather data from the decision maker, the head of department (Appendix 1). The interview was conducted after analysing the results of the teachers' questionnaires as some of the questions were based on the questionnaires' results. It consisted of eight prompts that question the issue of testing and teachers' exclusion from designing exit tests. The content of the questions consisted of demographic information about the interviewee, the extent of teachers' involvement in evaluation, reasons of excluding teachers from designing exit tests, and the interviewee's comments on the reasons of exclusion presented by teachers. I conducted the interview ten days after coding and analysing data in the questionnaire. Interviewing the head of the department was easy and smooth. He already had prior knowledge about the issue as he helped me to obtain the consent of the administration to conduct the study. I could say that this prior knowledge, the timing, and the venue (his office), were appropriate conditions to conduct the interview. After transcribing the interview, I gave the data to two of my colleagues who volunteered to check the transcription. First, I asked my colleagues to listen to the interview, check the transcription and highlight any possible discrepancy in the content. Their feedback confirmed my initial transcription. Later, the final version was given back to the interviewee to check the content. I asked him to read the transcription and make sure that I did not add any information that he had not mentioned in his interview or left out any. He was satisfied with the conformity between what he said and what I transcribed.

3.3.4 Data Analysis and Validation

Lather (1986) contends that the qualities of rigor and care can be achieved by adopting measures of conventional ethnography. She advocates using triangulation, systematized reflexivity, member checks and catalytic validity which "refers to the degree to which the research process re-orients, focuses, and energizes participants (…) [and] knowing reality in order to better transform it" (ibid, p. 67). In this study, I analysed the qualitative data inductively and adopted member check techniques to build credibility. I provided the interviewee with the transcription of the interview and asked him to check the content. Moreover, I used auditing to achieve dependability and confirmability. I gave the interview and its transcription to two independent data coders to check the content. As for catalytic validity, I think that it was achieved as 100 of the participants, who were not previously involved in designing exit tests, expressed their intention to discuss the issue of designing exit tests with their supervisors. Moreover, I used Miles and Huberman's (1994) techniques to organize data that consists of data reduction, data display, and conclusion drawing and verification. In order to reduce and organize data, I utilized data reduction, which is the process of selecting, focusing, and transforming the gathered information. I coded and classified data into themes and used data display, which includes the use of charts and graphs, to organize information. Moreover, conclusion drawing and verification refers to my efforts to give meaning and interpret data. In addition, I compared the data from the questionnaires and interview for evidence of convergence and divergence.

4 Results

The analysis of questionnaires' data yielded the following results.

4.1 Teachers' Qualifications and Roles in Assessment

The twenty-three teachers who answered the questionnaires were mostly M.A or Ph.D. holders (78.3) with a teaching experience of more than twenty years (60.9). They were teachers of English, or Math and Science working in the foundation program. Fifty-two percent were either PET or KET examiners. In their answers to questions about designing daily, weekly, and exit tests, these teachers provided the following data: 34.8 design daily quizzes and 43.5 correct them. As for weekly tests, 65.2 respondents state that they design and 73.9 responded that they correct these tests. Finally, only 26.1 of the respondents designed exit tests whereas 73.9 of teachers were deprived from designing exit tests.

4.2 Teachers' Assessment Literacy

The second section of the questionnaire was meant to elicit teachers' awareness of assessment issues, their perception of testing, and the different variables that should or should not be considered in designing tests.

Only 26.1 of the respondents strongly agree that weekly tests can be formative while 60.9 just agree with the statement. Over forty-three percent of respondents strongly agree that weekly tests can have an impact on teaching materials. In a similar way, the same percentage agrees that weekly tests can have an impact on teaching materials. Moreover, 56.5 of teachers strongly agree that weekly tests can help them modify teaching materials while 47.8 strongly agree that weekly tests can have impact on teaching practices.

The second set of questions was meant to elicit teachers' awareness and conceptions about exit tests. Only 21.7 strongly agree that exit tests can be formative whereas 47.8 simply agree with the statement. Comparatively, only 34.8 agree that exit tests can have an impact on teaching materials, teaching practices, and may help teachers modify teaching materials. As for the reliance on external examiners to design exit tests, 64.7 strongly agree that external designers should be aware of all testing issues of validity, reliability, and tests specs; yet 41.2 strongly disagree that test designers can prepare exit tests without consulting classroom teachers, while 88.2 strongly agree that these examiners should consult classroom teachers before designing tests.

In their responses to statements about testing purposes, 87 of the respondents strongly agree that teachers should be aware of testing purposes. Moreover, respondents were aware of test validity issues as 82.6 strongly agree that teachers should test what they teach, 69.6 strongly agree that teachers should tackle course objectives while designing tests, and 78.3 strongly agree that items should measure the intended point to be tested.

Concerning reliability, respondents showed the same amount of awareness as 69.6 strongly agree that teachers should design reliable tests that enable students to perform regularly if the test is given at different times. A closer high percentage was reached when 65.2 of respondents strongly agreed that teachers should provide learners with multiple opportunities to show what they know and can do. Awareness of reliability issues was less evident in the last statement where only 52.2 strongly agreed that teachers should include more test items to yield more reliable scores.

In their response to a statement about the level of proficiency, 60.9 of respondents agree that teachers should consider the varying levels of learners' proficiency while designing tests; on the other hand, 56.5 disagree with the idea that teachers should design tests according to the level of low performers against 47.8 who disagree that teachers should design tests according to the level of high performers.

As far as test specs are concerned, most of the respondents strongly agree with most of the statements. For example, 69.6 strongly agree that teachers should be aware of the duration of tests and 73.9 strongly agree that teachers should be aware of the importance of wording in test design. Moreover, 65.2 of respondents strongly agree that words in questions should be familiar to students and that teachers should consider ways of presenting tests. Furthermore, 56.2 of respondents strongly agree that teachers should consider content and cultural differences while designing tests as well as the way students are expected to answer. Similarly, 47.8 of respondents strongly agree that students should be familiar with types of questions and that teachers should expose learners to exam question types. Finally, only 34.8 strongly agree that test questions should be short whereas 52.2 strongly agree that the language for directions should be simple.

4.3 *Teachers' Justification of Exclusion*

This section was preceded by a direct question about teachers' involvement in designing exit tests (question 37). Out of twenty-three responses, seventeen teachers representing 73.9 stated that they were excluded. Only these excluded teachers were asked to complete the subsequent part of the questionnaire. In this section of the questionnaire, I provided them with possible reasons for being excluded from designing exit tests. Sixty-four percent strongly disagree that they were not involved because they are not qualified in testing compared to 58.8 who strongly disagree that their exclusion is related to their lack of knowledge about designing exit tests. Results show that 47.1 strongly deny their need for special training to design exit tests. Teachers' disagreement with potential proposed reasons for exclusion continued as 35.3 remained neutral and 47.06 disagreed with the statement that their exclusion is because of their heavy teaching load. Moreover, 47.06 disagree that they are excluded because designing tests is time consuming. A high percentage of neutral answers was returned on the question about trust and test leakage. For example, 35.3 remained neutral about the reason that they cannot be trusted while the same percentage disagreed with the statement. On test leakage, 41.18 remained neutral in their reply. They were undecided and only 29.42 agreed that teachers are not involved for fear of test leakage. Finally, only 35.3 agreed that teachers are not involved in designing exit tests because the institute is using a standardized test in the final exam.

With regard to their involvement in designing exit tests, 94.2 of the respondent teachers think that they should be involved since they are aware of the variables discussed in section 2 (reliability, validity, test specs). However, 64.7 strongly agree and 35.3 agree that classroom teachers should be involved in designing exit tests.

Although I provided respondents with many possible reasons for being excluded from designing exit tests, I gave myself the opportunity to gather more qualitative data and the respondents the possibility of expressing themselves by asking them to mention any other possible reasons for exclusion. Teachers’ responses to this question, though sometimes a repetition of the reasons I proposed, were informative and provided some reasons that the questionnaire did not address. Table 1 clarifies the different reasons that teachers presented.

respondents the possibility of expressing themselves by asking them to mention any other possible reasons for exclusion. Teachers’ responses to this question, though sometimes a repetition of the reasons I proposed, were informative and provided some reasons that the questionnaire did not address. Table 1 clarifies the different reasons that teachers presented.

Apart from lack of trust, security, time constraints, and lack of competence, which were mentioned in the questionnaire, new issues were raised. Four teachers drew my attention to the issue of the policy of the institute. One respondent states that “test institution policy fear of results and their impact in other words, tests may be made easier than the course.” Another respondent writes that it is “the wish of stakeholders (director) to assign the test designing job to a decision maker (coordinator, supervisor) who might not be fully aware of the actual teaching/learning process and not aware of the test designing tools.” Another respondent raised the issue of the impact of designing exit tests on teaching practices. He asserts that “if classroom teachers design exit tests, this might unconsciously direct and influence the choice of the information they focus on in the class.” A respondent attributed teachers’ exclusion from designing exit tests to the “over reliance on habit.” He adds that “someone has always assigned exit tests at rote-learning times and it used to work for the institution. Things did not evolve. Even with the change of curricula that same person will always be in charge of test designing.”

The absence of the culture of communication was another reason presented by one respondent. He notes that the rupture between teachers and “those who are in full control of decision making” is based on the assumption that the teachers’ role is “to teach and someone else designs tests, which is wrongly thought to be beyond teacher’s abilities and knowledge.” Teachers were blamed by one of the respondents who thought they were responsible for their exclusion. This respondent writes that teachers claim that “they know about testing, without really developing their skills beyond the basics they studied in university.” According to this teacher, this development would help them “to find fault with the inaccuracies of any test and communicate them to their direct decision maker (…) absence of action on the part

Table 1 Teachers' justification for exclusion from designing exit tests

Reasons of exclusion	Number of teachers and answers				
	1	2	3	4	5
Policy and power				√	
Lack of trust					√
Security		√			
Accreditation	√				
Habit	√				
Absence of culture of communication	√				
Teachers' acceptance of passive roles	√				
Time constraints	√				
Lack of competence		√			
Designing tests means directing teaching	√				

of teachers rendered their role passive and kept old habits of test designing monopoly and teachers' exclusion lingering this long."

In their response to the last question (question number 52) which was intended to gauge the impact of the questionnaire on teachers, all teachers who were not involved in designing exit tests expressed their willingness to discuss this issue with their supervisors as a future action.

4.4 Results of the Interview

Results of the interview yielded answers to the second sub question and presented decision makers' justification of teachers' exclusion. Overall, the results of the interview confirmed the data collected from teachers. As a decision maker, the interviewee denied total exclusion of teachers as he emphasizes that they are involved but monitored. In response to my question about teachers' involvement in designing weekly tests, he responds that they "design tests in coordination with unit coordinators and the final product is endorsed by the unit coordinator to make sure that these assessment tools are in line with our performance indicators in our framework." Contrary to weekly tests, teachers' exclusion from designing exit tests is total as the interviewee confirmed what respondents in the previous section had declared. He justifies that exit tests "are designed by the assessment and testing specialists in coordination with the Head of academic studies to make sure that students had the linguistic aptitude to be promoted to the next level." He adds that the "components and layout of exit tests are shared with all instructors so that the assessment procedures will be valid." He concludes by stating that excluding teachers is a "choice to have minimal teachers' involvement in order to have valid tests that focus on aptitude rather than achievement." He justifies exclusion by the necessity to avoid teaching to the test.

Apart from avoiding teaching to the test, the interviewee did not deny incompetence as another reason for exclusion. When I referred to the results of the questionnaire that showed teachers' awareness of most assessment issues, he changed his position and said that "involving competent teachers in the assessment procedure will bring benefit (…) but the majority of teachers are not competent." He justifies their incompetence by their lack of training to design tests. He declares that they "may know about basic concepts but they should be trained." In his response to my question about reconsidering his decision about teachers' exclusion, he insists that it cannot be reconsidered because of the "nature of the program, the context." He refers again to validity by claiming that the adoption of this policy stems from considering these exit tests as a type of standardized, external measurement tools to gauge that "what we are doing is valid."

The impact of raising the issue on the head of the department seems to be minimal. He expressed his willingness to involve some "competent" teachers in designing exit tests in the future but rejects reconsidering the whole policy because of "the nature of the program, the context." In any case, I think that his approval to involve "competent" teachers is a step toward a minor change in the policy. At least the interview was able to draw his attention to one of the critical issues that are causing controversy in his department.

5 Discussion

5.1 Status Quo: Inclusion and Exclusion

Results of the questionnaires and the interview reflected the status quo of teachers' exclusion from designing exit tests. Exclusion is a fact and above all a choice imposed by the powerful stakeholders on teachers. Though they are trusted to teach, design, administer, and correct daily and weekly tests, teachers are not trusted to design exit tests. Teachers' exclusion rate (73.9) is in sharp contrast with their inclusion in designing weekly tests. These findings seem to be in harmony with several previous studies, which highlighted this issue of teachers' exclusion from designing exit tests. This exclusion has persisted for decades and in different settings (Rea Dickins, 1997; Shohamy, 2001, 2005; Troudi et al., 2009). Despite countless appeals to involve teachers in designing tests to guarantee test fairness (McNamara, 2012), instances of exclusion still persist.

In addition to exclusion, the issue of power was consciously present in the interviewee's statements as he refers to his dominance as the head of the academic section on the whole process (Shohamy, 2005; Spolsky, 1997). The testing policy is built on a hierarchy that is monitored by him, the decision maker. He states that teachers are involved in weekly tests but their contribution is monitored by their unit coordinators, whose contribution is itself monitored by the head of the section.

5.2 Teachers' Assessment Literacy

The second section of the questionnaire was meant to discuss teachers' competence and awareness of the different issues of assessment as it has always been one of the main reasons presented by policy makers to exclude teachers from designing exit tests (Rea-Dickins, 1997; Shohamy, 2001; Troudi et al., 2009). Contrary to the interviewee who states that "the majority of teachers are not competent", teachers' responses to the statements about the washback effects on teaching materials and practices reflect their acceptable awareness and competence, a factor that decision makers deny to teachers. A total percentage of 86 agree that weekly tests should be formative and have an impact on teaching materials and practices, a majority of 64.7 agree that external designers should be aware of validity, reliability, and test specs, and 88.2 agree that these examiners should consult classroom teachers before designing exit tests.

Teachers' responses to statements about reliability and validity, levels of proficiency, and test specs reflected their awareness about these issues. Hence, lack of competence does not seem to be a good reason for exclusion. For example, teachers' responses to statements about the targeted levels of proficiency reflected their awareness about criterion-referenced and norm-referenced testing as 60.9 of respondents agree that testing should be criterion-referenced and not norm-referenced compared to 47.8 who disagree with designing tests according to the level of high performers. Moreover, teachers showed awareness of testing specs such as test duration, importance of wording, ways of presenting tests, length of questions, and content and cultural differences. Except for test content, this awareness is in harmony with the interviewee's position who states that teachers' inclusion in final tests should be restricted to their awareness about the layout and components of tests.

The findings of the second section of the questionnaire were important as I used them in the interview with the head of the section to deny incompetence and lack of awareness as reasons for exclusion. I anticipated the possibility for these reasons to surface during the interview and I wanted to have counter evidence. In general, teachers' competence and awareness about most of the testing issues seem to be high and therefore not a sufficient and acceptable reason for exclusion. Moreover, in case of lack of competence, policy makers should act as post method educators and recognize teachers' voices and visions instead of excluding them (Kumaravadivelu, 2001).

In a similar way, Rea-Dickins (1997, p. 312) suggests greater teacher involvement and states that "if teachers are given opportunities, starting through dialogue and working with the materials to develop a greater understanding of assessment processes, then, they will become better skilled at constructing tests." Parallel to this, Shohamy (2005) argues that the lack of awareness cannot be a reason for exclusion and suggests developing teachers' skills. She suggests abandoning the culture of viewing teachers as bureaucrats carrying out orders but rather as professionals who take part in testing policies so as to "develop critical strategies to examine the uses and consequences of tests, control their power, minimize their detrimental forces, reveal their misuses, and empower test takers" (ibid, p. 108) in a bid to empower teachers and foster a more democratic and inclusive approach to testing.

The interviewee's and respondents' justifications for exclusion may mean two major realities. First, exclusion is a fact, an issue, and a policy. Second, the reasons for exclusion are justified differently. From their perspective, teachers denied incompetence, lack of knowledge, the need for training, and time consumption as being adequate reasons for excluding them from designing exit tests. This denial is in sharp contrast with the interviewee's position who views teachers as incompetent and in need of training. This same justification of incompetence was presented by decision makers in the study of Troudi et al. (2009). Moreover, both parties denied the lack of trust as a reason for exclusion, but from different perspectives. Teachers' high percentage of neutrality concerning the issue of trust and test leakage reflects undecided positions. They were neither able to agree nor disagree with the statement. They seem to be torn between their position as teachers who would like to be involved and the alleged accusations of mistrust, dishonesty, and being sources of test leakage. The interviewee was more explicit when he stated that teachers' inclusion in designing exit tests might lead to teaching to the test, which in principle should be avoided. His position was acknowledged by Shohamy (2001) who reported the impact of tests on educational behaviour of teachers who changed teaching emphasis and whose instruction became test-like.

Similarly, one of the teachers agreed with the interviewee by acknowledging that "if classroom teachers design exit tests, this might unconsciously direct and influence the choice of the information they focus on in the classroom." The use of the word "unconsciously" is very important as it dispels any alleged accusation of conscious direction of teaching toward testing and clears the teachers of any deliberate dishonesty.

What is noticeable is the interviewee's use of validity to justify teachers' exclusion. He considers exit tests as a type of standardized measurement tool that may gauge the validity of what is being done, thus his belief that test design should be done by the testing and assessment specialists only. The justification presented by the interviewee is in harmony with the one presented by decision makers in the study of Troudi et al. (2009). Decision makers in both studies agree that quality assurance and validity are reasons to justify teachers' exclusion. Such a view ignores the fact that excluding teachers from designing exit tests deprive their students from fair evaluation. Test specialists cannot be aware of the dynamics of classroom, proficiency level of students, and the importance of wording in questions.

The respondents' and the interviewee's justifications for exclusion seem to overshadow a deeper conflict of power, in which assessment tools are the arms that every party would like to control. The interviewee, for example, affirms that exit tests are in the hands of testing specialists and that tests are designed hierarchically (Foucault, 1979; McNamara, 2000; Rea-Dickins, 1997; Shohamy, 2001). It is the head who monitors the work of teachers and unit coordinators. However, some teachers seem to be aware of this issue of power in their responses. One of them refers to this as "habit". Another one defines it as the "wish of stakeholders to assign the test designing job to a decision maker." Another respondent explains that teachers' role is to teach and that "someone else [should] design tests." As for lack of competence, two teachers agreed with the interviewee and justified the exclusion in terms of lack of competence. One of them even criticizes teachers for their refusal to develop beyond the basic skills they studied at university. He claims that this passivity and the absence of "action on the part of teachers rendered their role passive and kept old habits of test designing monopoly and teachers' exclusion lingering this long."

As far as the impact of the study is concerned, Cohen et al. (2003) state that catalytic validity embraces the critical theory paradigm. It informs that research will lead to action. It should reveal injustice, dominance and help participants to understand and change situations. The impact on teachers seems to be obvious as 100 of the teachers who were excluded from designing exit tests expressed their intention to raise the issue and discuss it with their supervisors. The questionnaire helped to raise the respondents' consciousness and consequently realize the injustice of being excluded from designing exit tests. By so doing, these teachers expressed their desire to cease being soldiers and servants of the system (Shohamy, 2005) to become post method teachers who have a say in policy making (Kumaravadivelu, 2001). The impact on the head of the department was not as obvious as that on teachers. The policy maker in this study, though he totally refused to change the testing policy because of the "nature of the program, the context", nevertheless expressed his acceptance to involve "competent" teachers in the future testing process. Results of this study showed that the given issue of testing was problematized (Pennycook, 2010) and that a process of conscientisation (Freire, 1970) started occurring in my work place. I expect a more overt bargaining of power to take place sooner rather than later, an issue I may explore in future research.

6 Limitations and Recommendations

This small critical exploratory study tried to problematize the issue of depriving classroom teachers from designing exit tests and raising their awareness about this issue. Located within the critical paradigm, this study was guided by a critical agenda. Depriving teachers from designing exit tests was a political more than a pedagogical choice. The results of the questionnaire denied all the alleged accusations of teachers' incompetence, dishonesty, and untrustworthiness. The head of the department, representing power in this study, acknowledges that the decision to exclude teachers was dictated by the context. Teachers' marginalization and exclusion is an exclusion of the dominated group that seem to lack the tools to defend its rights. Moreover, the results showed that excluded teachers' awareness was raised after answering the questionnaires' questions. Excluded teachers' intention to discuss the issue of exit tests with their supervisors reflects the success of the study to raise their awareness about one of the injustices in their workplace. This study also shows that teachers need concretization and raising awareness campaigns to be empowered. Being involved in designing exit tests may be interpreted as recognition of their competence and trustworthiness. It is also a step to involve other stakeholders that may lead to fair evaluation.

Limiting the study to my work place diminishes the chances of forming a wider picture of the situation in other educational institutions. Future research should include larger samples from different schools and universities to obtain a wider image. Future research should also seek to highlight instances of exclusion from assessment practices and the necessity to fight for equal opportunities of the different stakeholders to reach fair evaluation.

Appendix 1

Interview prompts

Good morning

1. Can you please introduce yourself?
2. Do classroom teachers design weekly tests?
3. Are classroom teachers involved in designing exit tests?
4. Why are classroom teachers excluded from designing exit tests?
5. What are your comments about the following results from teachers' questionnaires?
6. Will you reconsider the decision of exclusion?
7. Are there any issues you would like to add or talk about?

Thank you.

Appendix 2

Questionnaire

Dear colleagues

I kindly request you to help me by answering the following questions concerning the issue of who should design exit tests. This questionnaire is conducted for the purpose of research as part of my doctoral studies at the University of Exeter. This is not a test so there are no "right" or "wrong" answers and you don't even have to write your name on it. The outcome of this questionnaire will be used for research purposes. I am interested in your personal opinion. Please give your answers sincerely as only this will guarantee the success of the investigation. I will collect the questionnaires next week. In case you need any help, you can contact me at: damarazak@yahoo.com; Tel: 0551611205

Thank you very much in participation

Section 1: Demographic information

Please, put (X) where you think appropriate

1. What is your highest qualification?
   Diploma (   ) B.A ( ) M.A (   ) PhD (   ) other: (...)

2. How many years have you been teaching?
   1 to 5 years ( ) 6 to 10 years ( ) 11 to 15 years (   )

   16 to 20 years ( ) more than 20 years (   ) No answer:

Now, answer the following questions:

3. What subject do you teach?..
4. What level do you teach?...

5.

Are you an examiner?	KET	PET	IELTS	NO

		YES	NO
1	Do you design daily quizzes?		
2	Do you correct daily quizzes?		
3	Do you design weekly tests?		
4	Do you correct weekly tests?		

Section 2: Testing and tests' variables

The purpose of this section is to elicit your perception of testing and the different variables involved/not involved in designing tests. The following are a number of statements with which some people agree and others disagree. I would like you to indicate your opinion after each statement by putting an 'X' in the box that best indicates the extent to which you agree or disagree with the statement.

Strongly Disagree (SD)
Disagree (D)
Neutral (N)
Agree (A)
Strongly Agree (SA)

		SD	D	N	A	SA
	Weekly tests					
1	Weekly tests can be formative					
2	Weekly tests can have an impact on teaching materials					
3	Weekly tests can help you modify your teaching materials					
4	Weekly tests can have an impact on teaching practices					
	Exit tests					
5	Exit tests can be formative					
6	Exit tests can have an impact on teaching materials					
7	Exit tests can help you modify your teaching materials					
8	Exit tests can have an impact on teaching practices					
	Reliability and validity					
9	Teachers should be aware of test purposes (formative or summative)					
10	Teachers should test what they teach					
11	Teachers should tackle the learning standards while designing tests					
12	Items should measure the intended point to be tested					
13	Teachers should design reliable tests that enable students to perform regularly if the test is given at different times					
14	Teachers should include more test items to supply more reliable scores					
15	In designing tests, teachers should provide learners with multiple opportunities to show what they know and can do					
	Levels of proficiency					
16	While designing tests, teachers should consider the varying levels of proficiency					
17	Teachers should design tests according to the level of low performers					
18	Teachers should design tests according to the level of high performers					
	Tests' specs					
19	Teachers should be aware of the duration of tests.					
20	Teachers should be aware of the importance of wording					
21	Words in questions should be familiar					
22	Test questions should be short					
23	Language for directions should be simple					
24	Students should be familiar with types of questions					
25	Teachers should expose learners to exam question types					
26	Teachers should consider content and cultural differences					
27	Teachers should consider how tests will be presented (booklets, test papers, lab based)					
28	Teachers should consider how students are expected to answer: Answer sheets, writing on test papers, using computers					

Section 3: Exit tests

Do you design exit tests?	YES ( )	NO ( )

If your answer is “**NO**”, respond to the following statements:

		SD	D	N	A	SA
1	You don’t design exit tests because you are not qualified					
2	You cannot design exit tests because you did not study techniques of designing tests					
3	You don’t design exit tests because you need special training					
4	You don’t design exit tests because you have a heavy teaching load					
5	You cannot design exit tests because it is time consuming					
6	You don’t design exit tests because you cannot be trusted					
7	Teachers are not involved in designing exit tests for fear of test leakage					
8	You don’t design exit tests because the institute is using a standardized test in the final exams					
9	You should design exit tests because you are aware of most of the variables discussed in section 1					
10	Classroom teachers should be involved in designing exit tests					
11	External test designers should be aware of all the variables discussed in section 1					
12	External test designers can design reliable and valid tests without teaching					
13	External test designers should consult classroom teachers before designing exit tests					

Section 4: Open ended question

Please answer the following question

Can you mention any other possible reasons for not involving classroom teachers in designing exit tests?

………………………………………...

Section 5: Future action

Please answer with YES or NO

		Yes	No
	Will you discuss this issue with your supervisor/colleagues?		

Thank you very much for devoting time to answer this questionnaire. I will provide you with a brief summary of the findings if you are interested.

References

Brown, J. D. (2001). *Using surveys in language programs*. Cambridge, UK: Cambridge University Press.

Cohen, L., Manion, L., & Morrison, K. (2003). *Research methods in education* (5th ed.). London: Routledge.

Creswell, J. W. (2008). *Educational research* (3rd ed.). New Jersey: Pearson.

Dornyei, Z. (2003). *Questionnaires in second language research: Construction, administration, and processing*. New Jersey: Lawrence Erlbaum Associates.

Edge, J., & Richards, K. (1998). May I see your warrant please? Justifying outcomes in qualitative research. *Applied Linguistics, 19*(3), 334–356.

Foucault, M. (1979). *Discipline and punish*. New York: Vintage book.

Freire, P. (1970). *Pedagogy of the oppressed*. New York: Continuum.

Gillham, B. (2000). *Developing a questionnaire*. London: Continuum.

Kumaravadivelu, B. (2001). Toward a post method pedagogy. *TESOL Quarterly, 35*(4), 537–560.

Lather, P. (1986). Issues of validity in openly ideological research: Between a rock and a soft place. *Interchange, 17*(4), 63–84.

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

McNamara, T. (2012). Language assessment as shibboleths: A Poststructuralist perspective. *Applied Linguistics, 33*(5), 564–581.

Mertens, D. M. (2008). Mixed methods and the politics of human research: The transformative-emancipatory perspective. In V. P. Clark & J. W. Creswell (Eds.), *The mixed methods reader* (pp. 68–104). California: Sage.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis*. California: Sage.

Pennycook, A. (2001). *Critical applied linguistics: A critical introduction*. London: LEA.

Pennycook, A. (2010). Critical and alternative directions in applied linguistics. *Australian Review of Applied Linguistics, 33*(2), 16.1–16.16.

Punch, K. F. (2009). *Introduction to research methods in education*. London: Sage.

Rea-Dickins, P. (1997). So, why do we need relationship with stakeholders in language testing? A view from the UK. *Language Testing, 14*(3), 304–314.

Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? Are they fair? *Language Testing, 14*(3), 340–349.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Harlow: Pearson Education.

Shohamy, E. (2005). The power of tests over teachers: The power of teachers over tests. In D. J. Tedick (Ed.), *Second language teacher education: International perspectives* (pp. 101–111). New Jersey: Lawrence Erlbaum Associates.

Spolsky, B. (1997). The ethics of gatekeeping tests: What have we learned in a hundred years? *Language Testing, 14*(3), 242–247.

Troudi, S., Coombe, C., & Al-Hamly, M. (2009). EFL teachers' views of English language assessment in higher education in the United Arab Emirates and Kuwait. *TESOL Quarterly, 43* (3), 546–555.

The Voice of Classroom Achievement towards Native and Non-native Educators in English Language Teaching: An Evaluative Study

Tahany Albaiz

Abstract Promoting active information acquisition among students is a challenging task for any teacher. However, it becomes even more complicated to empower learners to attain success when having a poor understanding of their culture or language. Although non-native speakers (NNS) of English can deliver rather efficient results as teachers, the challenges, which they have to overcome, are much more numerous; therefore, the performance rates of the students who are taught by NNS teachers, is likely to be considerably lower than those of the learners taught by NS instructors. The reasons for the specified assumptions concern not the level of subject mastery displayed by the NS and NNS teachers, but the performance rates displayed by their students. A study involving a qualitative analysis of the performance tendencies among the students of NS and NNS teachers displays that the latter are prone to the problems concerning misunderstanding or misinterpreting specific concepts regarding the subject than those having NS teachers as instructors. The study also shows that the specified tendencies can be addressed and the performance of students instructed by NNS teachers can be improved significantly once adequate strategies concerning the use of proper teaching tools are incorporated into the lesson design. Particularly, the use of visual aids, as well as other means of getting the message across in a manner as clear and efficient as possible, needs to be considered.

Keywords Assessment · Performance · NS · NNS · Evaluation · Achievement · Teaching · EFL

T. Albaiz (✉)
English Language Institute, University of Jeddah, Jeddah, Saudi Arabia
e-mail: talbeiz@uj.edu.sa

S. Hidri and C. Coombe (eds.), *Evaluation in Foreign Language Education in the Middle East and North Africa*, Second Language Learning and Teaching,
DOI 10.1007/978-3-319-43234-2_6

1 Introduction

NNS versus NS has been always an issue in the teaching of foreign languages for many years and across many EFL theories. Starting from the very beginning of the concept, facilitating high-quality education in the environment shaped by globalization is a tricky task. The quality of staff performance defines the success of learners to a considerable degree; therefore, with the development of a multicultural environment, in which non-native speaking teachers could instruct students and provide them with the necessary information, questions regarding the possibility for students to understand the source material have been raised (Wong & Barrea-Marlys, 2012). Particularly, it is doubted that non-native speakers (NNS) can attain the voice of classroom (VoC) in the course of the teaching process. Although NNS may deliver less impressive results than NS as far as the achievement of proper VoC rates is concerned, the incorporation of the teaching approaches focused on the development of independence and self-directed learning among students will help NNS create the environment, in which the students will be able to evolve and become independent learners.

The process of teaching has clearly become more complex in the environment of global education. Particularly, the significance of meeting the needs of students as individuals and adapting towards their learning styles, thus, improving the quality of education significantly, can be viewed as a key change. However, apart from the above-mentioned alterations, significant changes seem to have occurred to educators and the very theory of teaching; particularly, the introduction of non-native speaking teachers (NNS) into the target background can be interpreted as one of the essential modifications of the learning environment. The specified change was inhibited partially by the changes in demographics, or, to be more exact, by the drastic increase in the number of learners (Klemencic & Fried, 2012). Therefore, the existence of NNS in education is a part and parcel of the modern reality. However, due to the language issues that may emerge in the course of the learning process, the efficacy of NNS can be questioned. Moreover, because of the obvious need to introduce an increasingly large number of NNS into the designated environment, the design of an approach that will help improve the overall quality of their teaching strategies will be required.

2 Theoretical Background

2.1 Where Control Theory Meets Constructivism

In order to address the issues raised in the research, one will have to consider the changes in the students' performance from two key perspectives, i.e., the Control Theory, which creates premises for understanding how the process of information transfer may occur between the students and the NNS teacher, and the Constructivist approach, which focuses on the process of knowledge building. The adoption of two separate theories as the tools for addressing the situation under analysis can be justified by the fact that the variables in question, i.e., the VoC, which is measured by evaluating the students' performance rates (Stairs, Donnell, & Dunn, 2011), are placed in two different environments, i.e., the ones, in which an NS and an NNS perform the roles of teachers. In other words, two different experimental settings are created, which calls for the adoption of two appropriate theories.

2.2 Control Theory as the Means to Assess the Impact of NNS Teachers

First and most obvious, the above-mentioned issue regarding the communication process between NNS teachers and their students' needs to be addressed as the most ambiguous one. The given issue can be approached from the tenets of Control Theory, as it states clearly that the process of knowledge acquisition can be managed by learners along with the teacher. Indeed, seeing that there is an obvious language issue in the scenario in question, one may assume that the learners have just as much, if not more, control over the course of the lesson and, therefore, the transfer of information. According to a recent study, the Control Theory can be defined as the "theory of motivation that ties learning to what a person wants most at the given time (Papa, 2011, p. 96)." Therefore, the specified approach allows the teacher to focus on the individual needs of each student and, thus, provides them with information that they actually need.

According to the existing definition, Control Theory suggests that the process of learning is enhanced not by the external stimuli but by the current needs and requirements of the student (Schwarzer, 2014). In other words, Control Theory suggests that learners should be just as active in the process of defining the course of the lesson and the information that they will need later on to evolve in the designated area as their instructors should (Carey, 2012).

2.3 Constructivism as the Tool for Approaching the NS Teachers' Success

The Constructivism Theory, in its turn, can be used as the means of evaluating the efficacy of the results delivered by the NS teachers (Fostnot, 2013). The assessment of the teachers' efficacy in getting the key messages across to the students and helping them acquire the necessary knowledge will be assessed based on the students' ability to understand the messages in question and construct the corresponding knowledge on their own, therefore, building awareness regarding their learning process (Akyol, 2012).

The Constructivism approach, being the strategy that helps "using our words to further our own understandings as well as those of others (Bentley, 2013, p. 44)," allows enhancing the process of acquiring the necessary information and understanding it so that the learners could solve complex tasks based on the theory in question.

A combination of the two theories described earlier will allow for deeper insight into the changes, which the NS and NNS teachers have contributed to. Both theories will shed some light on the importance of the clarity of the message conveyed by the instructor (Benton, 2014), as well as the necessity for the teacher to provide the visualization of the required concepts. In other words, the theories listed earlier will help state whether the learners are capable of constructing the necessary image and, therefore, a concept based on their ability to navigate and even guide the process of knowledge acquisition (Wang, 2011).

3 Research Problem

As it has been stressed, the efficacy of NNS as educators can be questioned because of the obvious concerns regarding the communication process. Indeed, a closer look at the specifics of information transfer from an NNS to NS will reveal that the pieces of data transmitted in the process may undergo significant changes when passed from a NNS teacher to a learner; as a result, the latter's concept of the subject matter may not coincide with the desirable one. As a result, the students' performance may drop due to the poor quality of the communication process between the teacher and the learners. More importantly, the issues that the students may experience when communicating with an NNS teacher may snowball to the point where the former may lose an opportunity to become a proficient language user. Recent studies (Braine, 2010; Mahboob, 2010) say that the aforementioned concerns are not far-fetched at all. According to a 2012 study by Wong and Barrea-Marlys, they explain that NNS teachers are very likely to face the issues related to language and understanding on a regular basis in the target environment without proper skills:

> If teacher candidates are asked to think about methodology and their own L2 acquisition, one can conclude from the findings in this study that their assumptions about what is effective teaching will be based on prior experience. Once teacher candidates reflect on CLT versus more traditional methodologies with which they are familiar, a positive or negative conclusion toward CLT is established. (p. 71)

Therefore, the problem exists and needs to be addressed. More to the point, the issue in question may be viewed from a different angle. Because of the concerns about the progress of students, people tend to overlook NNS as a valuable resource

(Kumaravadivelu, 2013). Therefore, a negative paradigm leading to the evolution and further blossoming of prejudice in the educational setting is created. Unfortunately, the specified instances are not quite rare among students and instructors alike (Jang, 2015); therefore, detailed tests will have to be carried out to prove that the efficacy of NNS teachers depends on a variety of factors and, more to the point, can and will be improved once the corresponding strategies are implemented. The study, thus, aims at proving that the aforementioned VoC can be achieved in both the learning environment created by a NS and the one created by a NNS, as well as the fact that the process of VoC achievement can be enhanced with the help of strategies aimed at improving the efficacy of communication in the setting, where NNS instruct learners.

4 Goals and Significance

Defining the key goals of the study, one must mention that the research focuses on the analysis of the differences in the performance of the students, who were instructed by NS, and those, who were instructed by NNS teachers. Particularly, the research addresses the achievement of the voice of classroom (VoC) in the designated settings and the comparison of the rates thereof among the NS and NNS correspondingly.

An analysis of the VoC rates in the settings headed by NS and NNS is the key objective of the research. To identify the problems, which both NS and NNS face in the specified environment, one will have to consider the fact that NNS teachers have the largest percentage of B and, unfortunately, D and F students, whereas the NS instructors seem to have very high rates of A and B-students. The latter type of teachers, in fact, has the smallest number of D and F students. The identification of the issues that inhibit the development of VoC in the NS and NNS setting can be viewed as another strategy of the study. The development of the strategies that will help NS and especially NNS achieve a significant increase in the VoC rates is the final objective of the research. Particularly, the approaches aimed at enhancing the efficacy of communication among the students and the teacher will be considered.

The significance of the study can be deemed as rather high since it addresses one of the basic problems, which teachers and students alike have to face in the environment of global education, as well as outlines the possible solutions for it. Although claiming that the study in question will have a ground-breaking effect on the theory of teaching would be wrong, the research still provides rather deep insights into the performance of NS and NNS teachers, therefore, building the foundation for making further assumptions regarding the strategies that NS and NNS teachers need to adopt to achieve higher VoC rates in the learning environment.

5 Results and Discussion

In the course of the study, forty environments have been identified and studied; nineteen of them were created by native speaking teachers, whereas the rest were designed by non-native speakers. The statistical data regarding the students' performance was gathered; particularly, the number of learners achieving the scores of "A," "B," "C," "D," and "F" was estimated. Afterward, a statistical analysis based on the calculation of the elements such as the mean and the SD of the students' performance, as well as the regression analysis, have been conducted. The information retrieved in the course of the research serves not only as the proof of the increased efficacy of NS compared to NNS in the classroom setting but also informs the teachers on the further avenues to be taken as far as the teaching of the Arabic language is concerned. Specifically, the use of the opportunities for self-directed learning (Colin & Hammond, 2013) and metacognition (Vandergrift & Goh, 2012), which the learning environment created by NNS provides, deserves to be mentioned as a crucial outcome of the study.

5.1 NNS Teachers and Their Students' Progress

According to the results of the analysis, there is a significant lack of VoC in the learning environment, where NNS teachers are present. Although the performance of the students, instructed by the teachers in question, cannot be deemed as entirely negative, the correlation between the VoC rates among the students instructed by NNS and NS points to the need to adopt the tools that will allow raising the VoC achievement by improving the performance of learners and enhancing the efficacy of communication between students and their instructors. Particularly, the fact that most learners retrieved D and C marks during lessons deserves to be brought up. A closer look at the analysis results revealed that the rates of C, D and F students were much higher among the NNS teachers than the NS ones (Figs. 1, 2, 3, 4 and 5).

It is quite remarkable that the SD rates of VoC in each group vary to a considerable extent; at some point, the SD reaches 15.7, which can be considered a

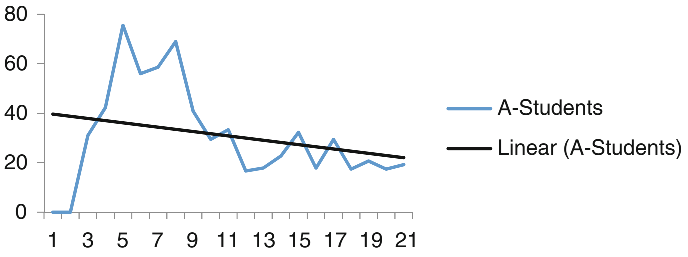


Fig. 1 A Students' percentage and tendencies among NNS teachers

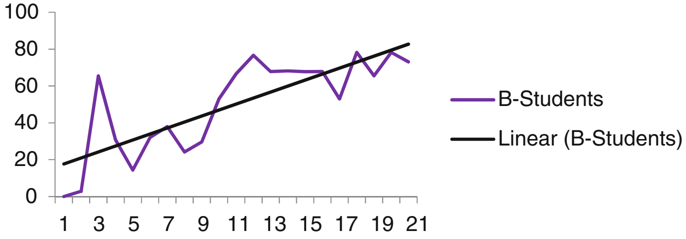


Fig. 2 B students' percentage and tendencies among NNS teachers

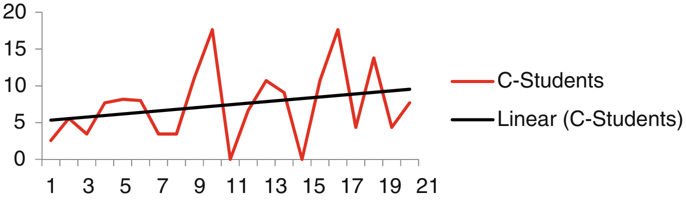


Fig. 3 C students' percentage and tendencies among NNS teacher

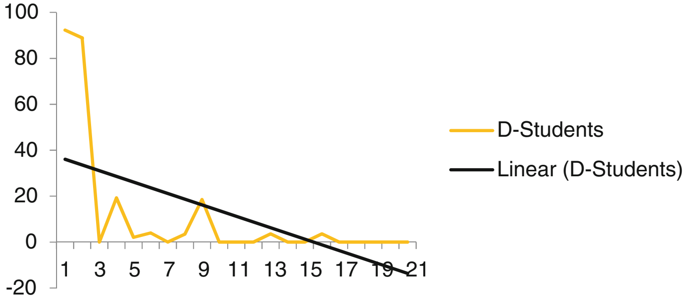


Fig. 4 D students' percentage and tendencies among NNS teachers

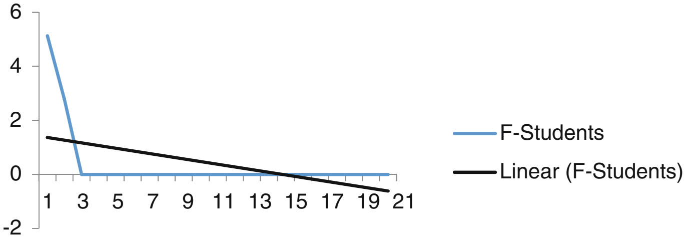


Fig. 5 F students' percentage and tendencies among NNS teachers

rather high difference given the fact that the experimental groups are rather small. According to the analysis carried out, there are obvious tendencies for an increase in the number of students performing positively; particularly, the average number of B- and C-students seems to be increasing gradually, as the trend lines in the corresponding graphs display. The number of A-students, however, is clearly declining, which means that NNS teachers should consider the communication tools that will increase the rates of VoC in the target environment. The research has also shown that the NNS teachers deliver less satisfying results than the NS ones; particularly, the number of D-students peaks in the specified environment, which means that significant changes have to be introduced.

It should be noted, though, that some of the data retrieved in the course of the analysis pointed to the fact that the NNS teachers may succeed at some aspects of teaching to a greater degree than the NS instructors do. Particularly, NNS teachers tended to promote independence among learners, thus, creating the environment, in which the students are supposed to digest the information provided to them in the course of the lesson in comfortable ways. In other words, NNS teachers did not foist a specific mode of thinking onto the students; instead, they left the lecture material free for interpretation, therefore, triggering an increase in creative thinking among learners. It should be noted, though, that the specified phenomenon typically occurs among the students that already have a substantive background regarding the subject matter and, therefore, are capable of forming an opinion on their own (Fig. 6).

The chart provided indicates clearly that there were significant differences in the VoC rates among the students belonging to a single group. While one group may have a rather large number of learners with high performance rates, another group may display the students' inability to grasp the problem suggested by the teacher by adopting the theory that was learned prior to completing the assignment. NNS teachers, therefore, may also achieve VoC in the environment of the classroom despite the language issues that they are most likely to have in the setting of a Saudi educational environment.

The study has also shown in a rather evident manner that the proficiency in the English language, though being admittedly important for an English language instructor, did not define the success of conveying the material to the students entirely. As the study showed, a number of NNS teachers reached the required VoC rates despite the language issues, which they must have been having with the students in the process.

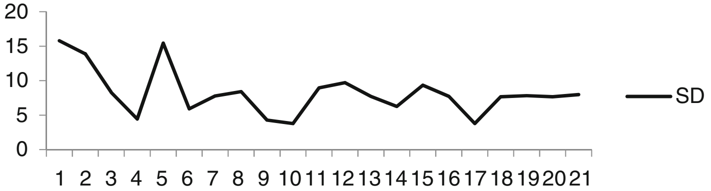


Fig. 6 SD of Percentage of A, B, C, D and F students in NS groups

5.2 NS Teachers and Their Impact on Learners

The results of the study showed clearly that the rates of VoC were clearly much higher in the groups led by NS. The specified results can be explained by the fact that NS teachers are most likely to find the ways to express themselves with the help of both verbal and nonverbal elements of communication in the manner that the students are most likely to understand. The above-mentioned phenomenon aligned with the principles of the Constructivism Theory. Seeing that the latter presupposes that the learners should be provided with an opportunity to build their understanding on the basis provided by the instructor and, therefore, be led by the teacher, so that they could expand their knowledge base, it will be reasonable to assume that the learners had many more chances for receiving decent scaffolding instructions and assistance from the teacher, who could convey the message on a variety of levels (Figs. 7 and 8).

It is quite remarkable that the number of B-students among NS teachers is slowly declining. Therefore, it can be assumed that the VoC rates are growing in the specified environment. Much like the previous result, it could be explained by the fact that the Constructivism-related processes, which promote creative and logical thinking among learners, allowed them to process the information provided by the teacher in the manner that linked their background knowledge and vision of the world to the data offered by the instructor.

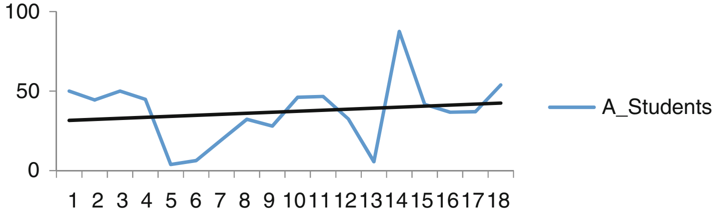


Fig. 7 A students' percentage and tendencies among NS teachers

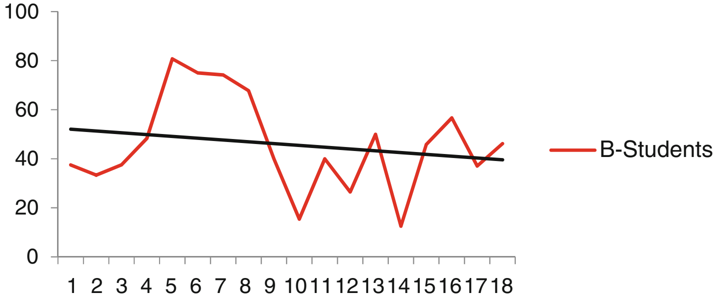


Fig. 8 B students' percentage and tendencies among NS teachers

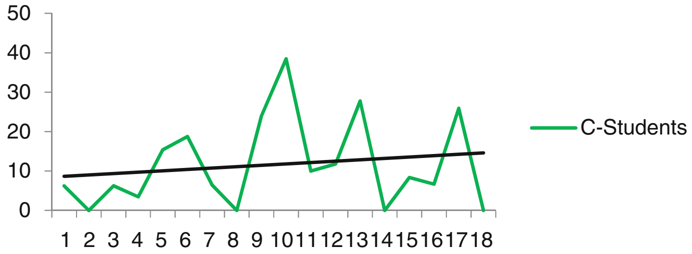


Fig. 9 C students' percentage and tendencies among NS teachers

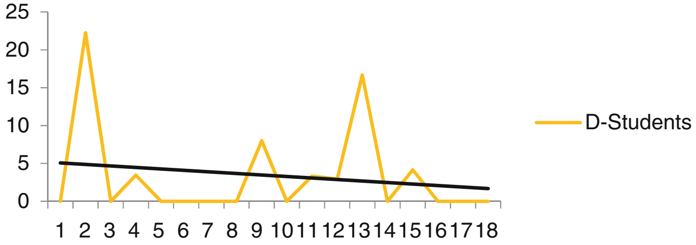


Fig. 10 D students' percentage and tendencies among NS teachers

The Constructivist Theory of learning holds that people learn by constructing their own understanding and knowledge of the world through experience and reflecting upon that experience. We are active creators of our knowledge, reconciling our previous ideas as we encounter new experiences and information (Harasim, 2012, p. 47) (Fig. 9).

Figures 10 and 11 make it quite clear that the number of D- and F-students is slowly declining. Thus, the VoC rates among the students that were instructed by NS teachers could be deemed as rather high due to increasingly high performance rates among the learners in question. According to the information retrieved in the course of the analysis, 37 % of the students received A marks on average. The same could be said about B-students; as the results displayed in a rather graphic manner, at least 46 % of learners achieved the specified mark. The number of C, D and F students, on the contrary, was very low, the means being 12, 3, and 22 % correspondingly. Moreover, the study revealed that the SD rates among the A-students in the designated groups were comparatively low (4.1), which means that the number of the students, who perform well, is basically similar in all eighteen groups supervised.

Unfortunately, the SD rates among the learners of different performance rates in the target groups were also rather low, as Fig. 12 shows. The specified characteristics of the experimental groups, however, could be attributed to the fact that the number of students in each of them was comparatively low (Rasinger, 2013); therefore, the slightest changes registered as very high in the given setting.

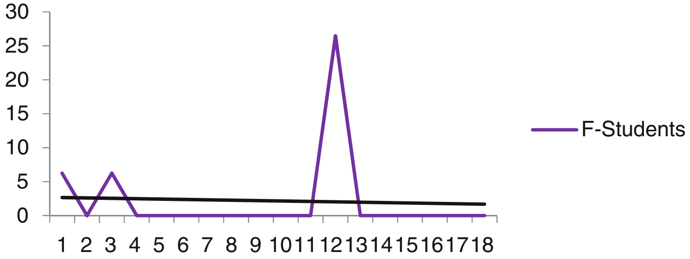


Fig. 11 F Students' percentage and tendencies among NS teachers

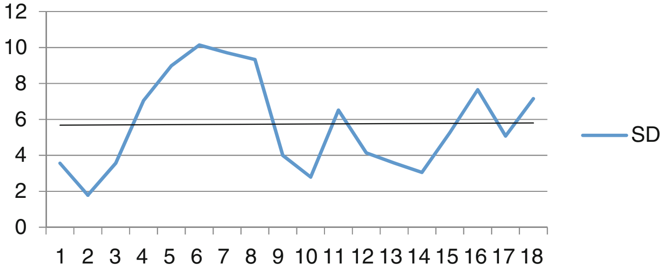


Fig. 12 SD of percentage of A, B, C, D and F students in NS Groups

Hence, it can be assumed that VoC is achieved by NS teachers in most cases in the above-mentioned setting. Although the VoC rates achieved by NS might be viewed as questionable when considering the issues that some of the students might have due to the low performance rates as displayed in Figs. 10, 11, and 12, the number of A- and B-students was still far higher in the groups headed by NS than in those that were instructed by NNS. Indeed, according to Fig. 12, the differences in the VoC rates were considerably lower than the groups led by NS. More importantly, the general tendency, which could be traced in the above-mentioned charts, could be defined as quite stable. Although the lack of growth as displayed by the trend line in Fig. 12 may be considered somewhat discouraging and the absence of negative tendencies in the designated area should be interpreted as the perfect grounds for the promotion of new learning tools among the students and the process of testing new teaching approaches. The appendices attached have numerical comprehensive presentations of all findings.

6 Suggestions and Recommendations

It can be suggested, therefore, that the NNS teachers should focus on the development of stronger ties with students by engaging in communication processes during lessons. Additionally, the fact that the students of NNS teachers were exposed to the environment, in which they are also capable of controlling the process of information transfer, made it necessary to make sure that the target audience should be able to navigate the learning environment along with teachers, therefore, contributing to the development of lessons and steering the latter in the direction that they felt necessary to digest specific bits of information.

As far as the improvement of the score delivered by the NS teachers is concerned, it is advised that the instructors should consider the tools that would help learners enhance their metacognition skills; particularly, the students would need to understand exactly how they acquired communication skills and adopted the identified approaches to their English language practice and their needs. Seeing that communication is not an issue in the teaching process facilitated by NS, it is essential to identify the obstacles that blocked learners' way to gaining the required knowledge and skills by improving the process of communication between teachers and learners. Once the latter are capable of identifying and naming the problems that they have in the course of learning, instructors will be capable of identifying the patterns that would help the students learn the necessary information in a manner as efficient and expeditious as possible (Hartman, 2013).

As the study has shown, the relationship between the native language of the teachers and the success of their students is obvious. Although the endeavours of NNS teachers are worth appreciating, the number of students, who seem to have problems with understanding English as a subject and developing the skills required to solve the related problems are worth appreciating, there is no need to stress that the number of A-students is much higher in the groups led by the NS teachers.

Herein the need to reconsider the teaching strategies used by NS instructors to achieve VoC lies. Increasing the rates of VoC among the English native teachers is essential to the improvement of students' performance rates; as the analysis carried out above has indicated, the students' ability to interpret the information that is given to them depends largely on the instructor's ability to deliver the key facts and assumptions in the manner that learners will find accessible. It also depends on their experience with the language.

7 Conclusion

According to the research results, there is a direct correlation between the teacher's native language and the students' success rate. In other words, the connection between the VoC rates and the educator's ability to use the concepts and notions that the students actually understand exists. Moreover, this direct correlation defines the strategies that NS and NNS need to adopt in order to approach the learners efficiently and make sure that they are able to apply the information learned to solve a practical task.

Appendix 1: NS Teachers: VoC Rates

NS	A	B	C	D	F	A (%)	B (%)	C (%)	D (%)	F (%)	SD	Mean
	8	6	1	0	1	50	37.5	6.25	0	6.25	3.5637	3.2
	4	3	0	2	0	44.44444	33.33333	0	22.22222	0	1.7339	1.8
	8	6	1	0	1	50	37.5	6.25	0	6.25	3.5637	3.2
	13	14	1	1	0	44.82759	48.27536	3.448276	3.448276	0	7.0493	5.8
	1	21	4	0	0	3.846154	80.76923	15.38462	0	0	8.9333	5.2
	2	24	6	0	0	6.25	75	18.75	0	0	10.139	6.4
	6	23	2	0	0	19.35484	74.19355	6.451613	0	0	9.7057	6.2
	10	21	0	0	0	32.25805	67.74194	0	0	0	9.3381	6.2
	7	10	6	2	0	28	40	24	8	0	4	5
	6	2	5	0	0	46.15385	15.38462	38.46154	0	0	2.7923	2.6
	14	12	3	1	0	46.66667	40	10	3.333333	0	6.5192	6
	11	9	4	1	9	32.35294	26.47059	11.76471	2.941176	26.47059	4.1473	6.8
	1	9	5	3	0	5.555556	50	27.77778	16.66667	0	3.5777	3.6
	7	1	0	0	0	87.5	12.5	0	0	0	3.0496	1.6
	10	11	2	1	0	41.66667	45.33333	8.333333	4.166667	0	5.2631	4.8
	11	17	2	0	0	36.66667	56.66667	6.666667	0	0	7.6435	6
	10	10	7	0	0	37.03704	37.03704	25.92593	0	0	5.0794	5.4
	14	12	0	0	0	53.84615	46.15335	0	0	0	7.1554	5.2
SD	4.108416	7.094369	2.346601	0.916444	2.118237							
M	7.944444	11.72222	2.722222	0.611111	0.611111	37.0237	45.79778	11.63691	3.376575	2.165033		

References

Akyol, Z. (2012). *Educational communities of inquiry: Theoretical framework, research and practice*. Washington, DC: IGI Global.

Bentley, D. F. (2013). *Everyday artists: Inquiry and creativity in the early childhood classroom*. New York City, New York: Teachers College Press.

Benton, C. (2014). *Thinking about thinking: Metacognition for music learning*. New York City, New York: R & L Education.

Braine, G. (2010). *Non-native speaker English teachers: Research, pedagogy, and professional growth*. New York City, New York: Routledge.

Carey, T. A. (2012). *Control in the classroom: An adventure in learning and achievement*. Buffalo, New York: Living Control Systems Publishing.

Colin, M., & Hammond, R. (2013). *Self-directed learning: Critical practice*. New York City, New York: Routledge.

Fostnot, C. T. (2013). *Constructivism: Theory, perspectives, and practice* (2nd ed.). New York City, New York: Teachers College Press.

Harasim, L. (2012). *Learning theory and online technologies*. New York City, New York: Routledge.

Hartman, H. J. (2013). *Metacognition in learning and instruction: Theory, research and practice*. Berlin: Springer Science & Business Media.

Jang, L. J. (2015). Identity matters: An ethnography of two non-native English-Speaking teachers (NNESTS) struggling for legitimate professional participation. In *Advances and current trends in language teacher identity research* (pp. 116–131). New York City, New York: Routledge.

Klemencic, M., & Fried, J. (2012). Demographic challenges and the future of the higher education. *International Higher Education, 1*(1), 12–14.

Kumaravadivelu, B. (2013). Rethinking global perspectives and local initiatives in language teaching. In S. B. Said & L. J. Zhang (Eds.), *Language teachers and teaching: Global perspectives and local initiatives* (pp. 317–323). New York City, New York: Routledge.

Mahboob, A. (2010). *The NNEST lens: Non-native English speakers in TESOL*. Cambridge: Cambridge Scholars Publishing.

Papa, R. (2011). *Technology leadership for school improvement*. Thousand Oaks, California: SAGE.

Rasinger, S. M. (2013). *Quantitative research in linguistics: An introduction* (2nd ed.). A & C Black.

Schwarzer, R. (2014). *Self-efficacy: Thought control of action*. New York City, New York: Taylor & Francis.

Stairs, A. J., Donnell, K. A., & Dunn, A. H. (2011). *Urban teaching in America: Theory, research, and practice in K-12 classrooms*. Thousand Oaks, California: SAGE Publications.

Vandergrift, L., & Goh, C. C. M. (2012). *Teaching and learning second language listening: Metacognition in action*. New York City, New York: Routledge.

Wang, Y. (2011). *Education and educational technology*. Berlin: Springer Science & Business Media.

Wong, C. C. Y., & Barrea-Marlys, M. (2012). The role of grammar in communicative language teaching: An exploration of second language teachers' perceptions and classroom practices. *Electronic Journal of Foreign Language Teaching, 9*(1), 61–75.

Part III
Text Genre Analysis Evaluation

Evaluation of Generic Structure of Research Letters Body Section: Create a Research Letter Body Section Model

Mimoun Melliti

Abstract Research Letters (henceforth RLs) are short scientific papers reporting new and innovative research findings. Previous research has identified that they are shorter in terms of number of papers (Maci, 2008; Rutkowsky & Ehrenfest, 2012). However, studies have not focused sufficiently on the generic structure of this genre, which is the concern of this chapter. This paper aims at investigating the organizational structure of RLs Body sections and suggesting a model for their formation. The researcher resorted to content analysis to statistically evaluate the place of every sentential function and identify the different phases and the obligatory and optional kind of sentences required for this part of RLs. RL body sections were selected and analysed sentence by sentence. Each sentence was allocated a particular structural element or key. The occurrence and frequency of these functions in the Body sections of each RL were counted in order to identify the shared rhetorical patterns among the 37 randomly chosen RLs. The aim is to contribute to the effort of identifying the hidden structure of this new and under researched genre. The main result of this research paper is the identification of Create A Research Letter Body Model (CARL Body Model). It suggests that the Body of any publishable RL is to contain 58 sentences where 49 are obligatory and 9 are optional. Such a finding is important for it helps researchers in scientific disciplines in writing publishable RLs. Additionally, it supports ESP teachers in teaching writing to future researchers.

Keywords Genre analysis · Genre teaching · Research letters · Rhetorical functions

M. Melliti (✉)
University of Kairouan, Kairouan, Tunisia
e-mail: mimoun_melliti@yahoo.com

S. Hidri and C. Coombe (eds.), *Evaluation in Foreign Language Education in the Middle East and North Africa*, Second Language Learning and Teaching,
DOI 10.1007/978-3-319-43234-2_7

1 Introduction

This chapter will evaluate the organizational structure of RLs in an attempt to identify a model for their formation. It starts with revisiting previous studies related to genre analysis before exposing the methodology employed in this study. The article ends up with the results section where the model of formation of research letters body section is detailed.

Until the publication of Nwogu's article in (1997), studies evaluating features of medical discourse had tended to focus on the syntactic elements of texts (as cited in Helan, 2012). One here could mention the example of Pettinari (1981) who studied the functions of grammatical fluctuation in 14 surgical reports and Salager-Meyer (1985) who worked on the classificatory framework and rhetorical function of the professional medical English terminology.

However, needless to mention that some studies have attempted to evaluate the way information was organized in scientific research reports such as Adams-Smith (1984) who examined the subjective elements of the authors comments and Salager-Meyer (1994) who investigated the occurrence and distribution of category of hedges.

Several research studies have been conducted on the organization of genres such as grant proposals (Connor, 2000; Connor & Mauranen, 1999), research articles (Dubois, 1997; Holmes, 1997), direct mail letters (Upton, 2001), application letters (Henry & Roseberry, 2001), abstracts (Hyland, 2000), business faxes (Akar & Louhiala-Salminen, 1999), medical research papers (Li & Ge, 2009; Nwogu, 1997; Skelton, 1994), medical case reports (Helan, 2012), and various frameworks have been employed. Examples of these frameworks include genre analysis (Bhatia, 2004; Swales, 1990), the sociology of scientific knowledge (Bazerman, 1988; Berkenkotter & Huckin, 1995; Gilbert & Mulkay, 1984; Knorr-Cetina, 1981), and lately Systemic Functional Linguistics (Halliday & Martin, 1993; Martin & Veel, 1998; Martinez, 2001; Samraj, 2005).

The first study that analysed the Moves in the four sections of the articles has been done by Skelton (1994) who depicted the organization of research papers published in the *British Journal of General Practice.* His aim was to help researchers and educators write and teach writing medical research articles better (Skelton, 1994, p. 455). Nwogu (1997) analysed the Moves deployed in different sections of RAs from medical journals (*The Lancet*, *The British Medical Journal*, *The New England Journal of Medicine*, *The Journal of Clinical Investigation*, and *The Journal of the American Medical Association*) and identified eleven moves. Eight of these moves were identified as obligatory and three optional. Li and Ge (2009) worked on the frequency of occurrence of the 11 moves identified by Nwogu (1997) and demonstrated that the structural and linguistic features of medical RAs have changed. This chapter will evaluate the organizational structure of RLs in an attempt to identify a model for their formation.

2 Methodology

The aim of this study is to evaluate the way RLs are structured. The researcher resorted to Paltridge (1997) keys model in order to evaluate and extract the rhetorical patterns of 37 Research Letters taken from *Nature* journal. Paltridge (1997) keys model was implemented in addition to considering other possible keys to be extracted from the corpus. The reason behind choosing this model is its clarity and specificity in identifying the rhetorical steps compared to other models such as Swales' (1990) model. Additionally, this model was chosen for it allows flexibility of interaction between expected rhetorical conventions found in similar genres and functional elements directly extracted from the corpus.

RL body sections were selected and analysed sentence by sentence. Each sentence was allocated a particular structural element or key. The occurrence and frequency of these functions in the Body sections of each RL were counted in order to identify the shared rhetorical patterns among the 37 randomly chosen RL. The aim is to contribute to the effort of identifying the hidden structure of this new and under researched genre.

Analysing RLs in terms of the structural elements of each sentence is a form of content analysis. Krippendorff (2004) defines content analysis as "a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use (p. 18)." This definition shows that the content analysis is based on inferences made from interpretations of the content of texts in light of prescribed research questions. In the same vein, Carley (1990) asserts that content analysis "focuses on the frequency with which words or concepts occur in texts or across texts (p. 725)." Inspired by these definitions of content analysis, the present research paper studies the rhetorical organization of RLs' body sections at the sentence level.

In fact, the content analysis method is documented to have various advantages such as mixing qualitative and quantitative techniques (Carley, 1990). Therefore, the researcher settled for the content analysis method where sentences constructing RLs' Body sections were quantified and allocated a particular rhetorical function according to the keys of Paltridge (1997) and others extracted directly from the corpus. This means that the researcher applied the structural elements that Paltridge (1997) identified as components of environmental texts and elicited the new keys peculiar to the corpus under study.

The researcher explored the RLs' Body sections and allocated a structural element to each sentence. Each letter is five to seven pages long including the figures. The additional 'methods summary' and the 'supplementary information' sections were excluded from the analysis. They were not analysed because the methods summary is a repetition of the methodology described in the main content of the letter. Besides, the supplementary information section provides additional and detailed description of the way the results of the study were handled and treated. A priori keys found by Paltridge (1997) in the research articles he studied were adopted and new ones found in the present study corpus were identified. Based on

Swales (1990) moves and steps model, Paltridge (1997) developed a model to analyse research articles in environment studies.

In order to investigate the structural elements of the RLs body sections, the researcher had to either use a previously invented model for RLs analysis or to use a model used to analyse similar kinds of texts. Considering the absence of specific models focusing on the structural elements of RLs, the researcher chose the second option. This choice has been opted for viewing the closeness of the RA genre to the RL genre as the former seems to be a contracted form of the latter. It is so because the RL is considered a short RA that focuses primarily on the results and their implications with little consideration to the literature part (Gotti, 2007; Maci, 2008; Rutkowsky & Ehrenfest, 2012).

Searching for previous studies focusing on the structure of RLs body sections, the researcher found no single investigation of the structural elements of this genre except Maci (2008). In fact, the researcher explored several research engines, websites and databases in Tunisia and abroad that publish RAs, books, and theses such as Openthesis.org, Bookos.com, linguistlist.org, Google Scholar, Blackwell, Elsevier, Jstore, ProQuest, Emerald, and Ebscohost (…), etc.

For the reasons mentioned earlier the researcher chose the Paltridge (1997) model, which consists of a number of keys or terms symbolizing rhetorical patterns in the texts and aiming at indicating the frequency of their occurrence in the corpora:

BI	Background Information
JS	Justification for Study
IG	Indicating a Gap
PS	Purpose of Study
RS	Rationale for Study
QR	Question Raising
PR	Previous Research
CS	Context of Study
M	Materials
R	Results
C	Conclusions

In his study, Paltridge (1997) investigated the introduction part of the RA but considering the generality of his model. The researcher also kept the door open for new keys to be elicited from the corpus under investigation. The result of such an approach was the finding of new and hybrid keys in RLs organization. Such an approach in dealing with the structural elements of RLs seems to be coherent with calls to mix a priori models with text specific models in investigating generic forms.

In order to analyse the content of the Body sections (B) of the RLs, the researcher designed a table containing the title of the research letter investigated, the number of keys, the kinds of keys identified, the order of all keys, Paltridge (1997) keys (BI, JS, IG, PS, RS, QR, PR, CS, M, R, and C), and other new or mixed keys to be directly extracted from the letters investigated. The table exemplifies the way the RLs were analysed. In this example the researcher managed to identify two keys not mentioned previously by Paltridge (1997) in his keys model which are PR/BI and ME/R. This combination of previous research with background information and of methodology with results means that these keys were found mixed in one sentence. The allocation of different denominations to the structural elements in the analysed letters was carried out by the researcher based on the content of these sentences. Such a decision is based on what the researcher thinks the function of each sentence is and it is not a linguistic analysis. This approach has been adopted as

> [a]ttempts at employing linguistic criteria to the validation of psychological perspectives are not, however, a necessary condition for the maintenance of functional approaches to language description, and should not be seen as a threat to the central claims made by exponents of such approaches (Paltridge, 1994, p. 296).

Hence, the aim of this level of the study is to evaluate RLs body sections and identify the generic structure potential, being the description of the "total range of textual structures available within a genre (Hasan, 1984, p. 79)." The analysis aims at indicating as specified by Paltridge (1997) "what elements *must* occur; what elements *can* occur; where elements *must* occur; where elements *can* occur; and *how often* elements can occur (p. 66)."

The researcher classified the data obtained from Table 1 by creating a Microsoft Excel document showing the total number of mentions (TNM) of each key in the body sections of RLs analysed. The researcher calculated the Average Number of Mentions of each key in each RL Body sections (ANM/L), the Total Number of Mentions (TNM), the Number of all keys (NAK), and the Percentage of Mentions (PM).

The ANM/L has been calculated by counting all instances of mention of each key in each RL Body sections (i.e., TNM) and dividing the result by the number of analysed RLs (i.e., 37). The PM has been calculated using the following equation:

$$\mathrm{PM} = \mathrm{TNM} * 100/\mathrm{NAK}$$

Table 1 Moves found in medical research articles according to Nwogu (1997, p. 125)

Section	Discourse function
Introduction	Presenting background information
	Reviewing related research
	Presenting new research
Methods	Describing data collection procedure
	Describing experimental procedure
	Describing data-analysis procedure
Results	Indicating consistent observations
	Indicating non-consistent observations
Discussion	Highlighting overall research outcome
	Explaining specific research outcomes
	Stating research conclusions

The results have been, then, transformed into graphs to clarify them and prepare them for description and analysis. The aim was to expose the weight of each key in the RLs body sections.

3 Results and Discussion

The most important finding in this part of the study is the identification of the GSP of RLs Body sections. In fact, being a relatively new genre, RLs received little attention from genre specialists (Maci, 2008) as except for the Maci (2008) study no other investigation has focused on it. The researcher classified the keys found in the RLs' Body sections based on his knowledge and expertise analysing and processing this genre in addition to their logical place in the RLs' Body sections. This means that similar keys identified in the corpus of RLs investigated have been grouped together and their representation in the Body sections has been measured (Table 2).

This procedure has led to the emergence of a model for writing RLs Body sections that mixes the expert researcher moves and the actual moves existent in the RLs and written by the scientists. Such a strategy in generic conventions' classification maps with calls for marrying linguists' and experts' knowledge with real structures of genres as extracted from the texts subject of investigation (Swales, 1990). The researcher found three phases constructing the Body sections of the RL, which are the Introducing Phase (IP), the Contextualizing Phase (CP), and the Findings Phase (FP). Table 3 suggests the general structure of RLs Body.

Table 3 shows that RLs Body sections contain three phases. The importance of the three phases in the RLs Body sections varies according to their presence in the RLs. Figure 1 shows the ANM/L of phases in RL Body.

The graph shows dominance of FP in RLs Body section. Out of an average of 57 sentences composing each RL investigated, sentences dealing with the findings dominate 37. This definitely supports claims of centrality of FP related sentences in RLs as acclaimed, required, and anticipated by researchers and editors (Maci, 2008; Rutkowsky & Ehrenfest, 2012). This means that the FP in the Body sections of publishable RLs should dominate 65 % of the whole Body sections.

The ANM/L graph shows also that about 13 sentences of each RL deal with the CP. This means 22 % of the Body section of each RL. The relative importance of this rate too indicates that it is necessary to include in the structure of Body sections, sentences contextualizing the study. Both of FP and CP dominate 87 % of the whole structure of each RL. This leaves only 13 % of each RL to the IP. This finding further empowers the claims of importance of 'new findings' in RLs structure as a new and emerging genre (Maci, 2008) at the expense of reporting already found results in previous studies.

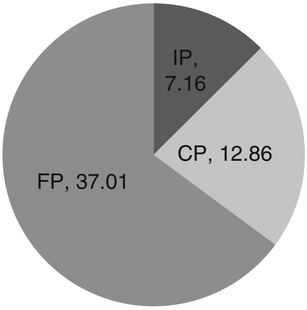
Table 2 Example of tables used to analyse the RLs body sections

Title of rsearch letter	Nb of keys	Kinds of keys identified	Order of all keys	BI	JS	IG	PS	RS	QR	PR	CS	M	R	C	Other
Control of substrate access to the active (…)	B:55	PR/BI IG PS R BI C ME/R R/C PR	(PR/BI PR/BI PR/BI IG PS R R BI C) (PR/BI BI BI R R R RRR IG ME/R RRRR R RRRR R RR C C R/C R RRR R RR PR BI R RRR C C C R PR R R R C CC R RRRR) (PS R R C C)	B:3	B:0	B:1	B:0	B 0:	B:0	B:2	B:0	B:0	B:38	B:8	PR/BI:B:1 ME/R: B:1

Table 3 The general structure of RLs body

IP	BI
	PR
	PR/BI
CP	JS
	IG
	PS
	RS
	QR
	CS
	M
	ME
	PR/IG
	BI/IG
FP	R
	C
	R/C
	PR/C
	ME/R
	FR

Fig. 1 ANM/L of RLs phases in the body



3.1 Phases of RLs Body Sections

For supplementary analysis and examination, the focus in the subsequent sub-sections will be laid on the phases of the Body sections in an attempt to suggest a final model for the generic structure i.e., GSP of the Body of RLs. The focus will be first on the most important phase statistically speaking.

The findings phase (FP) of Body sections is composed of 6 keys, which are R (40.34 % of each Body), C (17.05 % of each Body), R/C (3.78 % of each Body), PR/C (1.61 % of each Body), ME/R (1.15 % of each Body), and FR (0.36 % of each Body). The graph in Fig. 2 shows the keys constituting the FP of RLs Body sections and their PM.

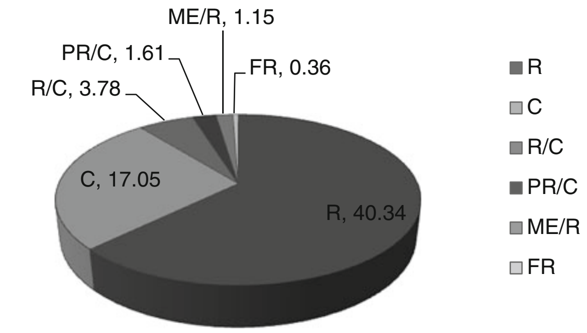


Fig. 2 PM of FP keys of RLs body sections

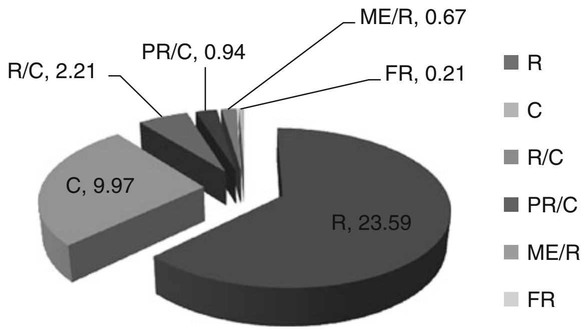


Fig. 3 ANM/L of FP keys of RLs body sections

The graph clearly exposes the dominance of Results, Conclusion, and Results melted with Conclusion in the structure of RLs body sections as they together constitute more than two third (61.17 %) of the whole body section. In terms of ANM/L, this dominance means that a publishable RL needs to encompass more than 23 sentences exposing the results in the body section, about 10 drawing conclusions, and more than 2 mixing results and conclusions as shown in Fig. 3.

The graph shows also that another sentence drawing a conclusion and mixing it with a previous study or a methodological point could exist in the FP. Thus, it could be deduced from the statistics obtained from the description, analysis, and processing of the RL investigated that the Body section of the RL is based on the FP.

The second phase in terms of importance after the FP is CP. CP of RLs Body sections is composed of 10 keys, which are JS (0 % of the whole Body), IG (1.01 % of the whole Body), PS (0.6 % of the whole introduction), RS (0 % of the whole Body), QR (0.5 % of the whole body), CS (0.03 % of the whole body), M (2.26 % of the whole body), ME (16.77 % of the whole Body), PR/IG (0.64 % of the whole Body), and BI/IG (0.02 % of the whole Body). Figure 4 clarifies these rates.

The graph shows that ME (16.77 % of the whole Body) and M (2.26 % of the whole Body) are the dominant keys in the CP of RLs Body. This means that they both represent 87 % of the keys forming the CP of the RL Body. Such dominance is also clear in the ANM/L of CP keys in the Body as shown in Fig. 5.

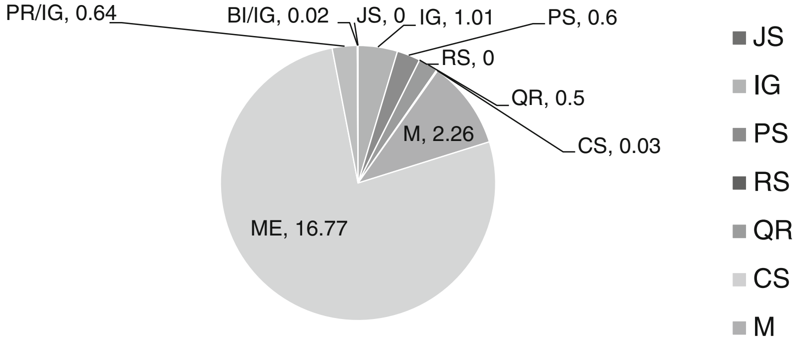


Fig. 4 PM of mention of CP keys of RLs body

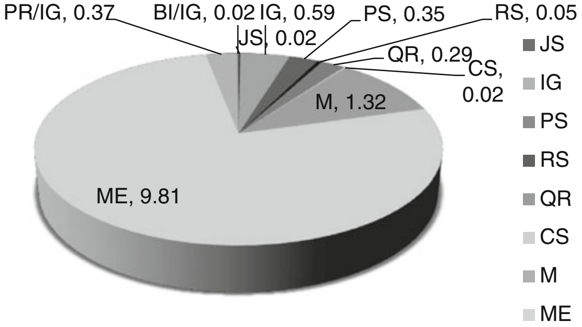


Fig. 5 ANM/L of CP keys in the body

Figure 5 shows that the body of RLs in its CP should contain about 10 sentences describing the methodology (ME) employed in the study conducted in addition to more than one sentence describing the materials (M). This phase could also possibly contain a sentence linking Identification of Gap (IG) to Purpose of Study (PS) or Previous Research melted with Identification of Gap (PR/IG).

IP of RLs Body sections is composed of three keys, which are BI (4.15 %), PR (7.39 %), and PR/BI (2.07 %) of the whole Body sections. Figure 6 clarifies them.

The graph shows that PR dominates the IP of the RLs body as it represents 7,39 % of the whole RL format which means 54 % of the IP keys. In terms of number of sentences (ANM/L), this means that there needs to be more than 4 sentences in the CP of RL body dealing with methodology (ME) as shown in Fig. 7.

The graph shows also that more than 2 sentences dealing with background information (BI) and more than 1 sentence linking Previous Research with Background Information (PR/BI) need to exist in the CP of RL body. It is remarkable in the PM of the keys in the IP, CP, and FP of the Body sections that there are highly represented ones and lowly represented ones. This scale will be

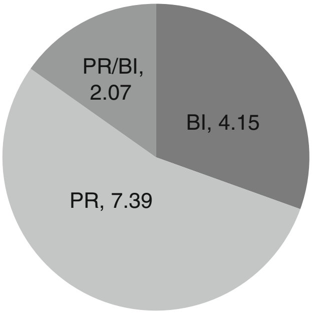


Fig. 6 IP of RLs body sections

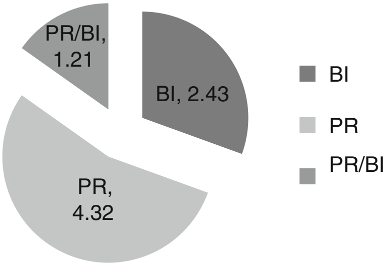


Fig. 7 ANM/L of CP keys in the body

used to categorize the keys present in the body sections into obligatory and optional ones.

3.2 Obligatory and Optional Keys of RLs Body Sections

The importance of each key in the structure of RLs differs in relation to the extent to which it is present in the RLs investigated. Considering this fact, the researcher managed to identify obligatory and optional keys in the structure of RLs.

Obligatory keys are those which their percentage of presence in the RLs body sections exceeds the average of mention of all the keys. The average is considered valid only when it exceeds 1 %. For example, the Average of the Total Percentage of Mention of All the Keys (henceforth ATPMAK) present in the IP of RLs body sections is 4.53 %. Thus, obligatory keys are those with a percentage of mention exceeding this rate. This means that PR (7.39) is the only obligatory key in the IP of RLs body sections.

Applying the same methodology, the researcher found that the ATPMAK present in the CP of the RLs body sections is 2.18 %. This means that only ME (16.77 %) and M (2.26) are obligatory keys in the CP of the RLs body sections. Continuing to apply the same principle, the researcher found that the ATPMAK present in the FP of the RLs body sections is 10.71 %. This means that R (40.34 %) and C (17.05 %) are the only obligatory keys in the FP of the RLs body sections.

The following diagram suggests a preliminary general generic structure of RLs body sections with obligatory keys highlighted with an asterisk.

Table 4 shows that the general structure of RLs Body sections contains obligatorily the following keys: Previous Research (PR), Materials (M), Methodology (ME), Results (R), and Conclusion (C). The existence of these elements in the structure of RLs Body sections could guarantee structurally speaking the publishability of RLs as these keys are of paramount importance in the structure of the generically investigated RLs. Optional keys in the RLs Body sections are those which their percentages of presence are less than the average of the total percentage of mention of all the keys (ATPMAK) in each phase.

Since the ATPMAK in the IP phase is 4.53 %, all keys with PM under this rate are considered optional. This means that BI (4.15 %) and PR/BI (2.07 %) are optional keys in the IP of the body sections. Concerning the CP, the ATPMAK in it is 2.18 %, which means that those under this rate are optional. This category concerns JS (0 %), IG (1.01 %), PS (0.6 %), RS (0 %), QR (0.5 %), CS (0.03 %), PR/IG (0.64 %), and BI/IG (0.02 %). As to the FP, the ATPMAK in it is 10.71 %. This means that the keys under this rate are optional, which are R/C (3.78 %), PR/C (1.61 %), ME/R (1.15 %), and FR (0.36 %).

It could be noticed that some keys are negligible in terms of mention that is why those under 1 % will not be considered as established optional keys. For this reason, the established optional keys are BI (4.15 %) and PR/BI (2.07 %) in the IP of the body sections. As to CP, the established optional key is only IG (1.01 %).

Table 4 Preliminary general generic structure of RLs body sections

Introducing phase (IP)	Background information (BI)
	Previous research (PR)*
	Previous research/background information (PR/BI)
Contextualizing phase (CP)	Justification of study (JS)
	Identification of gap (IG)
	Purpose of study (PS)
	Rationale for study (RS)
	Question raising (QR)
	Context of study (CS)
	Materials (M)*
	Methodology (ME)*
	Previous research/identification of gap (PR/IG)
	Background information/identification of gap (BI/IG)
Findings phase (FP)	Results (R)*
	Conclusion (C)*
	Results/conclusion (R/C)
	Previous research/conclusion (PR/C)
	Methodology/results (ME/R)
	Future research (FR)

NB Keys with asterisk (*) are obligatory

Finally, in the FP the established optional keys are R/C (3.78 %), PR/C (1.61 %), and ME/R (1.15 %). Hence, Table 5 clarifies the global generic structure of RLs Body section in terms of obligatory and established optional keys.

In order to better simplify the number of sentences that could be used in every phase the researcher created a chart that shows the phases suggested in Table 5 of RLs Body including their respective keys in relation to their TNM, the average of their mention, and the suggested number of sentences (SNS) for each key. Needless to mention that the ANM/L is calculated by dividing TNM by the number of all investigated RLs (i.e., 37). In order to suggest the SNS the researcher considered all numbers that are equal or above 0.5 as 1 and all numbers under 0.5 as 0. This aims at suggesting a logical number of sentences by the end as it is not possible to write 0.5 or 0.3 of a sentence. Table 6 shows that the identified obligatory and optional

Table 5 The global generic structure of RLs body section

Introducing phase (IP)	Background information (BI)**
	Previous research (PR)*
	Previous research/background information (PR/BI)**
Contextualizing phase (CP)	Identification of gap (IG)**
	Materials (M)*
	Methodology (ME)*
Findings phase (FP)	Results (R)*
	Conclusion (C)*
	Results/conclusion (R/C)**
	Previous research/conclusion (PR/C)**
	Methodology/results (ME/R)**

NB *Keys with one asterisk are obligatory
**Keys with two asterisks are optional

Table 6 TNM, ANM/L, and SNS of each key in RLs' body phases

	Keys	TNM	ANM/L	Suggested number of sentences (SNS)
IP	BI**	90	2.4	2
	PR*	160	4.3	4
	PR/BI**	45	1.2	1
CP	IG**	22	0.5	1
	M*	49	1.3	1
	ME*	363	9.8	10
FP	R*	873	23.5	24
	C*	369	9.9	10
	R/C**	82	3	3
	PR/C**	35	0.9	1
	ME/R	25	0.6	1
Total		2113	57.4	58

NB *Keys with one asterisk are obligatory
**Keys with two asterisks are optional

Table 7 Create a research letter body model (CARL body model)

Introducing phase (IP): 7 S	Background information (BI)**: 2 S
	Previous research (PR)*: 4 S
	Previous research/background information (PR/BI)**: 1 S
Contextualizing phase (CP): 12 S	Identification of gap (IG)**: 1 S
	Materials (M)*: 1 S
	Methodology (ME)*: 10 S
Findings phase (FP): 39 S	Results (R)*: 24 S
	Conclusion (C)*: 10 S
	Results/conclusion (R/C)**: 3 S
	Previous research/conclusion (PR/C)**: 1 S
	Methodology/results (ME/R)**: 1 S

NB *Keys with one asterisk are obligatory
**Keys with two asterisks are optional
S Refers to number of sentence(s)

keys in the RLs Body need to be represented the following way in order to be publishable (Table 7).

Thus, the present study suggests that scientists writing RLs should divide the Body into 3 phases, which are the Introducing Phase (IP), the Contextualizing Phase CP), and the Findings Phase (FP). The Introducing Phase (IP) could contain 2 optional Background Information (BI) sentences. Then, scientists must write 4 obligatory sentences dealing with Previous Research (PR). They could finish this stage with writing 1 optional sentence dealing with Previous Research melted with Background Information (PR/BI). Scientists then need to move to the following phase, which is the Contextualizing Phase (CP). They could start it by writing 1 optional sentence dealing with the Identification of Gap (IG). Subsequently, they must write 1 obligatory sentence describing the Materials (M). Then they must write 10 obligatory sentences describing the Methodology (ME).

Scientists need after that to move to the last phase, which is the FP. They must start this phase by writing 24 obligatory sentences dealing with Results (R) then 10 sentences dealing with Conclusions (C). Then, they could write 3 sentences melting Results with Conclusion (R/C). Afterward they could write 1 optional sentence mixing Previous Research with a Conclusion (PR/C). Finally, they could write 1 optional sentence mixing Methodology with Results (ME/R).

The evaluation of the structure of RLs and the identification of the CARL Body Model is important for research and pedagogy related reasons. This model could help researchers in scientific disciplines in writing publishable RLs. Having a model to follow when drafting the manuscript, researchers should find it easier to report their scientific findings and to publish them in specialized journals. Providing a model for beginners is invaluable as it equips them with a research-based and tested structure to follow in taking their first steps in the world of academic publication.

Moreover, evaluating RLs submitted for publication in scientific journals should be based on research findings such is the case of this work and not on editors' intuitions as far as the good RL is concerned. This certainly does not mean that the CARL model provides the best structural elements to construct a RL. Instead, it provides the actual structure of RLs that succeed (with originality being neutralized) in getting published in a leading scientific journal, which is *Nature*.

Additionally, the identification of the CARL Body Model is important for it supports ESP teachers in teaching writing to future researchers. It is an important step in setting the criteria for evaluating students' writings as far as RLs genre is concerned. In order to be publishable authors, to enrich knowledge in their disciplines, and obtain tenure, researchers need to be taught according to the established generic conventions. For this reason, the CARL Body Model could be of paramount importance to researchers and teachers. Such findings further strengthen the claims discussed in the literature review of this chapter that genre analysis is invaluable for ESP teachers and students.

Based on this study, a number of pedagogical recommendations emerge. They concern the students, the teachers, and educational authorities.

ELT teachers in science educational institutions are advised to sensitize students about the structure of different genres and especially the RL Body section. It is very important to provide students with basic generic background in order to empower them in a world of publish or perish. Practical application of this suggestion could be designing exercises helping students identify the generic structure of RLs Body section in order to find it easy to write their own RLs in the future. Just like they were taught how to write a good narrative, argumentative, or expository essay they need to learn how to write a good RL.

It is true that this strategy is related to the old school as modern strategies emphasize process related ones but the social constructivist approach in teaching writing shows that marrying the process and the product is better (Dudley-Evans & St John, 1998). For this reason, teachers could use the CARL Body Model to familiarize future scientists with the genre and its rhetorical structure. It has been found in this study that students need such tutoring in order to neutralize generic problems in the reviewing process when submitting their papers for publication. The CARL Body Model provides in detail the various kinds of sentences and their frequency of occurrence in the RL.

Another typical exercise to help future scientists write good RLs Body sections could be the transformation of some RAs into RLs using the CARL Body Section Model. This activity helps students write better RLs when they graduate and avoid rejection based on structural and generic reasons.

4 Conclusion

It is suggested for practicing scientists and researchers to obtain training in how to use the CARL Body Model to write better RLs generically speaking. Such training is highly important as it increases their chance in publishing in international journals. The claim is that the CARL Model has been developed based on the analysis of RLs successfully published by one of the leading scientific journals, which is *Nature*. Based on this study, a number of pedagogical recommendations emerge. ELT teachers in science educational institutions are advised to promote what this study calls Generic Structure Awareness, i.e., to sensitize students about the structure of different genres and especially the RL. It is very important to provide students with basic generic background in order to empower them in a world of publish or perish.

Practical application of this suggestion is designing exercises helping students identify the generic structure of RLs Body section in order to find it easy to write their own RLs in the future. Just like they were taught how to write a good narrative, argumentative, or expository essays they need to learn how to write a good RL. It is true that this strategy is related to the old school as modern strategies emphasize process-related ones but the social constructivist approach in teaching writing shows that marrying the process and the product is better.

References

Adams-Smith, D. E. (1984). Medical discourse: Aspects of author's comments. *The ESP Journal., 3*(1), 25–36.

Akar, D., & Louhiala-Salminen, L. (1999). Towards a new genre: A comparative study of business faxes. In F. Bargiela-Chiappini & C. Nickerson (Eds.), *Writing business: Genres, media and discourses* (pp. 207–226). Essex, UK: Pearson Education Ltd.

Bazerman, C. (1988). *Shaping written knowledge*. Madison: WI, University of Madison Press.

Berkenkotter, C., & Huckin, H. (1995). *Genre knowledge in disciplinary communication: Cognition, culture, power*. Hillsdale, NJ: Erlbaum.

Bhatia, V. K. (2004). *Worlds of written discourse: A genre-based view*. London: Continuum.

Carley, K. (1990). Coding choices for textual analysis: A comparison of content analysis and map analysis. Unpublished working paper.

Connor, U. (2000). Variation in rhetorical moves in grant proposals of US humanists and scientists. *Text, 20*(1), 1–28.

Connor, U., & Mauranen, A. (1999). Linguistic analysis of grant proposals: European Union research grants. *English for Specific Purposes, 18*(1), 47–62.

Dubois, B. L. (1997). *The biomedical discussion section in context*. Greenwich, Conn: Ablex Pub. Corp.

Dudley-Evans, T., & St John, M. (1998). *Developments in ESP: A multi-disciplinary approach*. Cambridge: Cambridge University Press.

Gilbert, G. N., & Mulkay, M. (1984). *Opening Pandora's Box: A sociological analysis of scientists' discourse*. Cambridge: Cambridge University Press.

Gotti, M. (2007). Identity and cross-cultural communication. In *Proceedings of the 72nd Annual Convention of the Association for Business Communication*. Washington, DC, October 10–12, 2007.

Halliday, M. A. K., & Martin, J. R. (1993). *Writing science: Literacy and discursive power*. Pittsburgh, PA: University of Pittsburgh Press.

Hasan, R. (1984). The nursery tale as a genre. *Nottingham Linguistics Circular, 13*(7), 1–102.

Helan, R. (2012). *Analysis of published medical case reports: Genre-based study*. Unpublished Ph. D. dissertation. Masaryk University, Czech Republic.

Henry, A., & Roseberry, R. L. (2001). A narrow-angled corpus analysis of moves and strategies of the genre: Letter of application. *English for Specific Purposes, 20*, 153–167.

Holmes, R. (1997). Genre analysis, and the social sciences: An investigation of the structure of research article discussion sections in three disciplines. *English for Specific Purposes, 16*(4), 321–337.

Hyland, K. (2000). *Disciplinary discourses: Social interactions in academic writing*. Harlow, England: Longman.

Knorr-Cetina, K. (1981). *The manufacture of knowledge: An essay on the constructivist and contextual nature of science*. Oxford: Pergamon.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Thousand Oaks, CA: Sage.

Li, L.-J., & Ge, G.-C. (2009). Genre analysis: Structural and linguistic evolution of the English-medium medical research article (1985–2004). *English for Specific Purposes, 28*(2), 93–104.

Maci, S. (2008). The research letter: An emerging medical genre. In G. Di Martino, V. Polese, & M. Solly (Eds.), *Identity and culture in English domain-specific discourse* (pp. 367–390). Napoli: Edizioni Scientifiche Italiane.

Martin, J. R., & Veel, R. (Eds.). (1998). *Reading science: Critical and functional perspectives on discourses of science*. London: Routledge.

Martinez, I. A. (2001). Impersonality in the research article as revealed by analysis of the transitivity structure. *English for Specific Purposes, 20*, 227–247.

Nwogu, K. N. (1997). The medical research paper: Structure and functions. *English for Specific Purposes, 16*(2), 119–138.

Paltridge, B. (1994). Genre analysis and the identification of textual boundaries. *Applied Linguistics, 15*(3), 288–299.

Paltridge, B. (1997). *Genres, frames and writing in research settings*. Amsterdam: John Benjamins.

Pettinari, C. (1981). The function of a grammatical alternation in fourteen surgical reports. *Applied Linguistics, 4*(1), 55–76.

Rutkowsky, J. L., & Dohan Ehrenfest, D. M. (2012). Research letters: A new editorial format for the rapid disclosure of innovative data and concepts, didactic demonstrations, and scientific discussions. *Journal of Oral Implantology, 38*(2), 101–103.

Salager-Meyer, F. (1985). Specialist medical English lexis: Classificatory framework and rhetorical functions. *EMP Newsletter, 2*(2), 5–18.

Salager-Meyer, F. (1994). Hedges and textual communicative function in medical English written discourse. *English for Specific Purposes, 13*(2), 149–170.

Samraj, B. (2005). An exploration of a genre set: Research article abstracts and introductions in two disciplines. *English for Specific Purposes, 24*, 141–156.

Skelton, J. R. (1994). Analysis of the structure of original research papers: An aid to writing original papers for publication. *British Journal of General Practice, 44*(387), 455–459.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Upton, T. (2001). Understanding direct mail letters as a genre. *International Journal of Corpus Linguistics, 7*(1), 65–85.

Genre Analysis and Cultural Variations: A Cognitive Evaluation of Anglo-American Undergraduate Personal Statements

Ghada Hajji

Abstract This chapter is a contrastive genre study that investigates the rhetorical structure of the British and American personal statements (PSs) written by undergraduate students. The corpus consists of 60 PSs (30 British and 30 American) selected from three different disciplines: Business, Physics and Psychology, and they were collected from four websites. The genre analysis of the collected data has been based on Ding's (2007) model as an analytical framework. This study seeks to test the applicability of this model on the corpus and examine the rhetorical and the linguistic resemblances and variations found between both cultures. Results of the genre analysis indicate that the analysed statements have revealed some rhetorical and linguistic similarities and differences between both corpora. The divergences and convergences between both corpora were attributed to certain socio-cultural and academic factors. The findings of the present study offered valuable insights regarding the genre features of PSs. Further, this research may fill in the gap in the rhetorical studies of the British and American academic genres.

Keywords Contrastive rhetoric · Genre analysis · New rhetoric · Personal statement · Cognitive evaluation

1 Introduction

The pupils' educational assessment can be achieved through the analysis of certain academic genres as they may reflect their rhetorical, linguistic and stylistic abilities. After much research on published academic texts, recent studies in English for Specific Purposes (ESP) have expanded to students' writings, focusing mainly on culminating genres of graduate students such as master's theses and doctoral dissertations, and to a lesser extent on undergraduate writings (Samraj & Monk, 2008).

G. Hajji (✉)
Faculty of Arts and Humanities, Tunis, Tunisia
e-mail: hajjighada@hotmail.com

S. Hidri and C. Coombe (eds.), *Evaluation in Foreign Language Education in the Middle East and North Africa*, Second Language Learning and Teaching,
DOI 10.1007/978-3-319-43234-2_8

Nonetheless, the genre of PSs has not been sufficiently addressed by previous genre research despite its significance in the evaluation process of the students' interests and writing skills. Indeed, analysing such an academic genre helps teachers to identify the kinds of problems students may be having in their writings (Hyland, 2007) and assess to what extent the applicants' interests correspond to the courses provided by the targeted program. In addition, this genre is poor in terms of comparative studies. In fact, no study has been carried out, as far as the literature consulted, to compare and contrast the generic and linguistic features of the British and American PSs. By reviewing the current literature, it is noticeable that there is a remarkable paucity in research on this genre's content and structures.

2 Theoretical Background

Contrastive research investigates differences and similarities in writing styles across cultures and disciplines. "It considers texts not only as static products but also as functional parts of dynamic cultural contexts (Connor, 2002, p. 493)." Kaplan (1966) was the first linguist to examine the rhetorical distinctions in discourse structure in various languages. Based on 600 English essays written by foreign students from various language backgrounds, Kaplan (1966) came to infer four discourse structures that contrast to the English predominantly linear style. He described the Semitic, Oriental, Romance and Russian languages discourse structure respectively by the following rhetorical pattern: Parallel, circular, digressive and parenthetical.

However, CR has come under sharp criticism in recent years and many linguists have proven the invalidity of his hypothesis. Since that time, various contrastive research studies have been conducted on different academic genres such as journal articles, abstracts, book reviews, dissertations conference, grant proposals etc. (…) which have been extensively investigated for their cross cultural and interdisciplinary rhetorical and linguistic features.

Despite their significant role in the school admission and evaluation process, graduate and undergraduate PSs have received little attention from researchers. Nevertheless, there were some studies, which have been made to examine the rhetorical structure and the linguistic realization of some statements in different disciplines and in different contexts. Brown's (2004) research was the pioneering study of the PS genre. In this project, Brown analysed the "T-units" in the school letters of application written for a clinical psychology program. After the rhetorical investigation of the selected PSs, Brown (2004) came out with different results. First, when comparing the successful and unsuccessful clinical PSs, Brown found out that the admitted PSs focused more on the moves describing their research experiences and interests than those that were refused. Second, successful applicants showed how their awareness and familiarization of their target discourse communities

Following Brown's (2004) findings, Ding (2007) conducted another study based on the application essays to medical and dental schools. She identified a rhetorical framework of five moves with three steps for each of the first two moves as the following: Move 1: Reasons for studying, Move 2: Credentials, Move 3: Relevant experiences, Move 4: Future goals and Move 5: Personality.

Samraj and Monk, (2008) conducted another research project based on the genre of PSs. The corpus of this study consisted of some samples of successful statements submitted to three master's programs: Linguistics, business administration and electrical engineering at a U.S university. Similarly, to Ding (2007), they established a rhetorical framework containing four essential moves with different steps. This study did not only utilize move analysis and interviews with the graduate programs chairs like the previous studies of Brown (2004) and Ding (2007), but it included a survey of printed books and websites of PS writing which revealed that information on writing statements for specific master's programs is not consistently available (Samraj & Monk, 2008).

Although the three reviewed studies have shown interesting findings of the genre of PSs, still there are certain deficiencies in some perspectives. First, all the mentioned research projects were based on limited data (from 30 to 35 PSs in each study) and a limited number of disciplines (from 1 to 3). This inadequacy of data and the small variation of disciplines may affect the reliability and generalizability of the findings of these studies. Even Samraj and Monk's (2008) study, though it appeared to choose three different disciplines, it neglected other disciplines worth investigation and analysis such as physics, biology, sociology, etc. (…). In addition, none of the mentioned studies dealt with the linguistic aspect of the genre of PSs. Another limitation of these research projects is that they relied on native speakers' PSs, as their research data. It would be rather interesting to conduct some research on non-native English speakers so that international students get some benefits from the obtained results in their process of application especially with the recent dramatic increase of international students applying for western universities.

In this respect this study aims to investigate the rhetorical structure of the British and American PSs written by undergraduate native speakers collected from three different disciplines: Business, Physics and Psychology. It adopts Ding's (2007) model as an analytical framework. In addition, it intends to deal with the major linguistic signals and strategies. Further, it attempts to answer the following research questions:

a. What are the rhetorical structures and the linguistic strategies realized in the British and American PSs?
b. Do the British and American PSs from the different programs conform to a similar rhetorical structure?
c. What is the role of the socio-cultural or educational backgrounds training in shaping the rhetorical pattern of the PSs?

3 Methodology

The corpus of the present study consisted of 60 personal statements written in English by undergraduate native English students (30 British and 30 American) selected from each of the following disciplines: Physics, Business and Psychology. All the personal statements of the corpora were collected from public websites. The British ones were downloaded from the Universities and Colleges Admissions Services website (UCAS), whereas the American ones were obtained from three different websites due to the absence of one official website like the British counterpart. The physics PSs were chosen from www.alumnus.caltech.edu, the psychology PSs from www.psych.uni.edu and the business ones were downloaded from www.eduers.com and www.essayforum.com.

The general criteria of selection of these letters of application were based on the objectives that this study was trying to achieve. All personal statements have to be written in English: 30 by British applicants and 30 by American ones. They were also supposed to be selected from the three mentioned disciplines: Business, physics and psychology

The method which was applied in the present research is “move analysis”, the purpose of which is to uncover the structure, style, content, and communicative purpose of the genre under investigation (Helal, 2013). As discussed by Ackland (2009), the identification of move strategies and boundaries in research articles was usually accomplished through two approaches. One is content-based called the “top down approach”, and the second is based on the linguistic signals called the “bottom up approach (Li, 2011)”. In this research, the overall organizations and boundaries of moves and sub-moves were identified essentially on the basis of content that is using the “top down approach.” Further, moves were identified relying, partly, on some frequent semantic meanings and linguistic features which were regularly present all through the analysed data set and, partly, on their communicative functions and rhetorical purposes.

After the identification of the main moves and strategies, the researcher analysed the use of the first personal pronoun “I”, together with its possessive form “my,” and their role in enhancing the applicants’ self-promotion. Thereafter, the whole corpus was processed using the software Ant Conc 3.2.4 2011 to conduct some statistical analyses. Some frequent words and linguistic features were identified such as the most frequent hedging and boosting devices.

The linguistic investigation of the collected PSs showed an intensive use of some linguistic elements that consisted mainly in the hedges and boosters. These elements help speakers or writers to express both interpersonal and ideational (or conceptual) information, allowing writers to communicate more precise degrees of accuracy in their truth assessments (Halliday, 1994). The selection of hedging and boosting devices in this study relied on those suggested by Hyland (2000). He noted that hedges and boosters could actually convey the major content of an utterance in carrying authorial judgments.

4 Results and Discussion

4.1 *Rhetorical Analysis*

The first move identified in Ding's (2007) model is "explaining reasons to pursue the purposed study." As it was indicated, the role of this move is to introduce the factors behind getting interested in the field. According to her, this move shares similar functions with Swales' (1990) move for article introductions namely, *establishing a territory*. In this move, the student explains the main reasons that motivated him/her to pursue the chosen course. It seems to be obligatory in both British and American PSs. It was present in all the statements and in almost all the disciplines. Indeed, it is noticeable that both groups of students showed a strong tendency to choose M 1 as a strong move in their PSs. In addition, it was present in all American statements (100 %) and in (93.3 %) of the British ones. This proves that both English and American applicants share a strong preference to open their PSs with stating the motivations behind their will to apply for the targeted study or discipline.

Further, even at the level of sentences number and percentage, there was no remarkable difference between both cultures. In the British PSs, the total number of sentences in Move 1 featured approximately 160 sentences in the three disciplines and it constituted around 24.2 %, on average, of the whole statement. In the same way, the statistics of the analysed American data showed no great difference from the British ones. Therefore, it seems clear that both British and American applicants share the same communicative intention in initiating their PSs directly with the same move. This can be explained by the same rhetorical paradigm that Anglo-American writers base their discourses or writings upon (Taylor & Tinguang, 1991).

Still in adherence to Ding's (2007) model, the second most prevalent move is "credentials." In this move, the writer is supposed to create his relevant self to promote the candidature by different strategies. It describes the candidate's academic, research professional and social qualifications and experiences. Further, according to Hsaio (2003), the purpose of this move is to select a relevant, positive and convincing self to persuade the admission committee to offer a place.

Being the core of the PS, this move seems to be of an obligatory nature in the British corpora. Indeed, it was present in almost 90 % of the British PSs. It featured a high frequency in all the disciplines. The importance of this move does not only reside in its high frequency which is marked in almost all the British PSs, but also in its length compared to other moves. It represented 42.3 %, on average, of the whole school application letter, which is certainly a high rate. Such a result is expected in the analysed corpus as it is based on the UCAS guiding instructions, which state that applicants ought to enumerate their academic achievements and professional experiences.

Moving to the U.S corpus, the rhetorical investigation of the PS in the different disciplines showed a slight variation between both cultures. The results showed that

Move 2 is still prominent with a lesser frequency especially in Business and Psychology and with lower average sentences. Further, this move still carries the main weight of the applicant's communicative act, as it occupied, approximately, the 1/3 (34 %) of the PS but with a lower degree when compared to its length in the British corpus. This result confirms what Ding (2007) found in her study conducted on American medical and dental PSs. This suggests that with the high frequency and dominance of M 2 in the native speaker corpora, undergraduate applicants increase their chances of meeting the expectations of the committee members by valuing most of their academic achievements and research qualifications and highlighting their involvement in community services and their high commitment for the field.

The third move mentioned in the used model deals with "relevant life experiences." Indeed, while Move 1 and Move 2 provided the main reasons for applying for the target undergraduate programs and the valuable academic and professional credentials, Move 3 focuses entirely on the most relevant work experiences related to the field and the applicant's community involvement. Furthermore, contrary to the previous moves, which marked high frequency in both cultures, this move seems to be optional, it is present in 50 % of the British and American corpora. In this move, prospective applicants are supposed to describe their previous experiences, and thus their suitability for the field. Starting with the British corpus, the rhetorical findings showed that Move 3 was found only in 60 % of the Physics and Business PSs and 30 % in psychology ones. However, there were some British students who were aware enough of the requirements of the admission process as they opted to include some prior experiences. These findings seem similar to Ding's (2007) results. Indeed, contrary to Move 1 and Move 2, Move 3 featured lower frequency despite the fact that her study is based on graduate PSs. This can be explained by the paucity of the professional experiences in the target discipline among the majority of applicants.

The British findings were similar to the American ones regarding Move 3's frequency. Indeed, the rhetorical investigation indicates that this move appeared only in 50 % of the U.S corpus signifying its optional nature in the three disciplines. This result is expected because both British and American corpora are based on undergraduate statements of purpose. Nonetheless, in the American data we may notice that there is an interdisciplinary gap. In fact, this move occurred in 70 % of Business applications, whereas, it only occurred in 30 % of the physics and 40 % in the psychology applications. This might be attributed to the fact that in the business field there was more reliance on professional experiences, which is highly advocated by the commission members even in the undergraduate level.

The fourth move is "stating goals." This move serves to indicate the applicant's goals in the future study and what s/he aspires to learn. This can serve as a strategy of self-promotion as the candidate is informing the reader that s/he is, to a certain extent, knowledgeable about the field applying for and s/he has well examined the course of the prospective program (Hsaio, 2004). He stated that "the purpose of pointing out the preparation for future study is to make a claim that the applicant has a long term intention and interest in this area and has shown careful consideration in opting for future study (p. 43).” He added also “Through this move the reader can evaluate how well the applicant’s interests and goals correspond to the courses that the program can provide for him (p. 43).” Further, Ding (2007) claimed that this move portrays the applicant’s intended future career after graduation, which stresses the goal-orientedness and strong motivation of the applicant.

This move was optional and it constituted 5 % of the whole British PS as average. It was present in 60 % of the Business PSs, 70 % of Physics ones. However, it appeared in 20 % of the Psychology letters. We may consider the following example selected from the business program.

Extract 1:

> My future aim, after university, is to start up my own successful business and pursue a computing related career with possible geographical/environmental links. (PS 9, Business).

It is clear, from this sample, that in this move, candidates try to show their ability and will to succeed in the target program; hence, persuading the admission committee that they are the applicants they should choose (Callaghan, 2004). Interestingly, in the psychology program, only two out of ten applicants used the *stating goals* move although it is an essential rhetorical element in the statement of purpose according to the UCAS website. This striking paucity of this move in this particular discipline may be explained by, either the students’ unawareness of the genre conventions and requirements, or their uncertainty of their future career in this field.

Regarding the U.S findings, they are not very distinguishable from the British ones, except the psychology field. Indeed, this move appears to be optional in both Business (50 %) and Physics (60 %) but obligatory in psychology (90 %). In fact, this move showed a high frequency in the psychology discipline compared to the other disciplines and compared to all British disciplines. This may suggest that American psychology applicants have clearer future plans and more intentions than their British counter parts.

The last move is “describing personality.” According to Ding’s model (2007), this move serves to explicitly describe or demonstrate the applicant’s unique experience and personality to distinguish him/her from the large pool of applicants. The most striking aspect of this move is that it was totally absent in both the British and American PSs. On the contrary, it was replaced by other moves, which are non-applicable to the employed model in this study. In fact, the British and American PSs seemed to be more preoccupied with explaining the reasons behind application, credentials and their relevant experiences more than stating their personal issues. This move was found in Ding’s (2007) study as an optional move emphasizing the personality strengths.

The variations which were noticed in the American corpus at the rhetorical level may be attributed to the fact that the U.S.A has been a multi-cultural country with multi-cultural minorities speaking various languages. Therefore, it should be expected that these cultural variations would be reflected in the country’s academic genres, which is the case in the genre under investigation. Indeed, it is inferred that the American writing style in this genre is no longer direct and explicit as it is argued in previous studies (Li, 2011; Mauranen, 1993; Yunxia, 2000). On the contrary, with the presence of the cyclical and additional moves and the use of short stories and long sentences, the U.S writing style appears to be more implicit and digressive, the fact that made it more colloquial and interactive than the British one (Biber, 1987). This can be ascribed to its influence by other minorities (like the Asians, Chinese, French, etc. (…)) who were known for their different and digressive rhetorical patterns. In this context, Pokrivack, Hevesiova, Smileskova and Janecova (2010) argued that "the differences that occurred on American rhetoric can be explained by the fact that the U.S.A has been a country of several cultures speaking various languages (p. 9)." Therefore, earlier hypotheses, which overlooked the stylistic, rhetorical and cultural differences between both countries and considered the Anglo-American society as "one block" need to be questioned and even revised, because this will lead automatically to cultural stereotyping.

4.2 Linguistic Analysis

After the identification of the main moves and rhetorical strategies, the researcher analysed the use of the self-promotion strategy. While the moves and sub-moves were manually identified and counted, the most frequently used words and key features were calculated with the help of Ant con 3.2.4 2011 software. The use of this software has revealed the top most frequent recurring linguistic signals in both corpora such as: The personal pronoun "*I*" and its possessive form "*My*" which were used in the self-promotion strategy. This analysis has also shown the presence of other linguistic features such as the boosters and hedges in both corpora.

"*I*" and "*my*" appeared to be the highest-ranking lexical items in the British and American corpora. They were present in all the PSs and in all the disciplines, but with different frequencies. Indeed, it is indicated (Appendix 3) that the first personal pronoun "*I*" tends to be much more frequent than its possessive form "my" within both groups of applicants. In addition, the average of occurrence of "I" in each British PS was approximately 17, whereas it was only 8.3 with "my." Similarly, in the U.S letters the average of occurrence of "I" featured 16.7 per statement compared to 10.4 for the average of "my."

The personal pronoun "I" appeared in both corpora with high frequency in the three disciplines. Starting with the British corpus, the use of "I" ranged from seven occurrences (PS 5 psychology) to 40 (PS 2 psychology). Nonetheless, there is a clear variation at the level of disciplines. In fact, this item was most frequently used in the physics discipline (194) but it was least used in Psychology (130), and in Business (188). However, linguistic findings revealed the presence of another linguistic variation at the move level.

The personal pronoun "I" was the most frequent in the first two moves of the PSs in the three disciplines. This could be attributed to the rhetorical nature of this particular genre. Indeed, some moves such as Move 1: *Explaining reasons* and

Move 2: *Credentials* represent the cornerstone of the statement and they are the space where students try to promote and glorify themselves by emphasizing the "I". Further, despite the variation at the move and discipline levels, the function of "I" did not differ in any of them; In fact, it served to portray the applicants as suitable, competent and well-determined potential students.

Moving to the American corpus, the results showed some statistical variations in the use of "I" in the American PSs. Indeed, this element featured higher presence in the American corpus in the psychology discipline, whereas for the others the average is approximately the same. In addition, similarly to the British PSs, in the American PSs, the personal pronoun "I" was centring in Move 1 and 2 and in some cases in Move 3.

The possessive adjective "*my*" was used by both groups for promoting themselves. It served to demonstrate the applicant's qualifications, valuable experiences and motivating reasons to pursue the target level. The results prove that this element was intensively present all through the different moves of the statement but with a low rate when compared to the personal pronoun "*I*," its frequency ranged from 4 (PS2, Business) to 21(PS4, Business).

With respect to disciplines, the adjective "*my*" was most frequent in British business (102) and least frequent in Physics (73). This may be attributed to, as it was shown in the rhetorical analysis section, the paucity of the relevant experiences and qualifications in the physics discipline. Nonetheless, at the move level, similarly to the previous linguistic signal "*I*", its possessive form "*my*" was highly frequent in the first two moves and with lesser degree in Move 3.

With regards to the American corpus, the linguistic analysis revealed some statistical variations. In fact, it seems clear that "*my*" appeared to be more frequent in the American data than the British one. In addition, contrary to the British corpus, the highest frequency of "*my*" in the U.S PSs was found in Business whereas, the lowest was in psychology. Concerning its frequency of distribution, this feature was essentially present in the first two moves.

The linguistic analysis of the compiled PSs revealed an intensive use of some linguistic devices that consisted mainly in the hedges and boosters. These elements help speakers or writers to express both interpersonal and ideational (or conceptual) information, allowing writers to communicate more precise degrees of accuracy in their truth assessments (Halliday, 1994). Further, Hyland (2000) asserted that hedges and boosters could actually convey the major content of an utterance in carrying authorial judgments. Therefore, it is clear that, in the analysed statements, applicants made use of these linguistic signals for more persuasion of the significance of their qualifications.

The linguistic investigation of the 30 selected PSs from the different disciplines demonstrated that both groups of students showed a strong tendency to use some hedging devices in their writings to convey their intentions appropriately. Nonetheless, the frequency of these devices differed from one statement to another, from one discipline to another, and from one culture to another. Indeed, Swales (1990) assumed that hedges were extensively used in Anglo-American academic writings to project "honesty, modesty, proper caution, and often diplomacy

(p. 174).” The most occurring hedges in both corpora were modal auxiliaries, verbs, adverbs and adjectives. They occurred with different frequency all through the corpus. The modal auxiliaries (*can*, *may*, *would*) appeared to be the most frequent in both corpora. Hedging verbs came in a second position (*feel* and *believe*). Whereas, the adverbs and adjectives were the least frequent devices used by British and American applicants.

The three major modal auxiliaries identified during the process of linguistic analysis of the collected PSs were: “*May*”, “*can*” and “*would*.” The linguistic investigation of the data proved a recurrent use of these three devices with the exception of the modal “*may*” which showed very low frequency in the three disciplines. This can be expected because the modal “*may*” generally expresses uncertainty or probability, which is not the aim of the writers in this particular genre. On the contrary, in the genre of personal statements, students are supposed to show certainty and confidence to portray a positive image of themselves. Similarly, the frequency of the modal “*may*” is still low within the American PSs, with a higher occurrence of “*can*” and “*would*”. Both British and American candidates opted to use these two hedging auxiliaries to promote their candidatures and to glorify their abilities.

Verbs were found to be the second hedging features in the analysed corpora. The most frequent ones were: “*Believe*”, “*seem*” and “*feel*.” As was noticed in the British PSs, “*believe*” and “*feel*” showed high frequency and appeared, for the most part, in all the statements and in all the disciplines. Whereas, the introductory verb “*seem*” appeared only once in the British Physics PSs and it was totally absent in Business and Psychology. This can be explained by the fact that this verb expresses uncertainty, which is inappropriate for this genre.

The same results are almost applicable to the American data. In fact, the verb “*seem*” appeared only once in Physics and it was frequently absent in the others. On the contrary, for the verbs “*believe*” and “*feel*,” although they showed higher frequency, their rate is still low, especially with the verb “*feel*” in Physics and Business. The main two hedging adverbs identified in the English PSs, were: “*Likely*” and “*often*”. In both corpora, they were rarely used especially for the adverb “*likely*” which showed no occurrence in both cultures and in all the disciplines. Regarding the adverb “*often*,” it was scarcely present in the British and American corpus in some disciplines. Its frequency ranged from zero to eight.

Hedging adjectives were found to be the fourth hedging element in the selected English PSs. Similarly, to the previous devices (adverbs), adjectives are quantitatively rare; however, the two main hedging adjectives found were “possible” and “probable”. The adjective “probable” showed no presence in both cultures and in all the disciplines. However, the adjective “possible” was present in some PSs in all the disciplines and in both cultures. Its occurrence ranged from one to four per discipline.

In light of the above description, it could be inferred that the use of the hedging devices (modal auxiliaries, verbs, adverbs and adjectives) was not really preferred by the British and American applicants. This can be expected since students, when dealing with the genre of PSs, are not supposed to express probability or uncertainty. They should, instead, minimize their use of the hedging elements to show a certain degree of certainty and self-confidence.

Boosters were recognized as the most frequent devices used in the present corpus to express the student's valuable candidature, high commitment and future plans. Several boosting features were depicted in the compiled corpus. However, the most occurring ones can be categorized as the following: Modal auxiliaries, adverbs and adjectives. The highest presence of boosters was found in the modal auxiliary "will" in both corpora (25 in British psychology, 34 in American physics and psychology), whereas verbs, adverbs and adjectives were the least boosting devices used by the British and American applicants.

The findings of the analysed corpus revealed that the modal auxiliary "*will*" was the most frequently used in the different disciplines. Besides, the modal auxiliary "*must*" was also identified as another boosting device in the corpus. However, the linguistic analysis revealed a large difference between both auxiliaries in terms of their frequency. In fact, "*will*" seems to dominate all the English letters of application, while the modal "*must*" appeared to be rarely used particularly in the British PSs where it occurred only once. Nonetheless, the modal auxiliary "*will*" appears to be the most frequent booster not only among the boosting modal auxiliaries but also among all the boosting devices. Indeed, this modal was most found in Move 4: *Stating future goals*, where students gave their potential plans and thus giving the impression of maturity, certainty and self-confidence. In the same context, Murphy (2010) pointed out that speakers generally opt for the use of the modal "*will*" since it is "a more direct form which expresses confidence and certainty, unlike the modal forms (p. 139)". The boosting modal auxiliary "*will*" is also found in some additional moves, namely the move of expressing commitment where candidates show their readiness to overcome all the challenges and express their zealous will to excel in the target program.

Nonetheless, unlike its predecessor, the modal auxiliary "must" shows no great presence in the corpora. Indeed, it appeared only once in PS 3 in the British discipline of psychology. In the American data, this modal had few occurrences as well. This noticeable absence might be ascribed to the fact that this modal generally expresses "a strong obligation" (Murphy, 2010, p. 140) which is in sharp contradiction with the requirements of the genre of PSs. Indeed, applicants are supposed to express requests, commitments and polite compliments without showing any imposition or obligation on the reader or the evaluator.

The findings of the analysed data demonstrated that adverbs were intensively used by both groups of applicants as a second boosting strategy. In fact, "*very*" and "*always*" were found to be the most frequent boosting elements in the statements. The adverb "*very*" was the most frequently used adverb in the British and American PSs. It was highly present in all the disciplines. It was always followed by an adjective and it functioned as an intensifier in the different moves of the PS. Nonetheless, the adverb of frequency "always" marked fewer occurrences although it was present in all the disciplines and in both corpora. It did not occur in all the British and American PSs, but it featured higher frequency in the physics discipline, in both cultures, more than the others.

In addition to modal auxiliaries and adverbs, adjectives were employed as a boosting strategy in the British and American PSs. The most frequent boosting adjectives were "*important*" and "*clear*." The element "*important*" was the most frequent. It occurred in some American disciplines such as Business and psychology, while it was absent in physics. Regarding the British corpus, this adjective showed a higher presence in Business (it occurred 11 times). This boosting feature was employed by both groups of applicants as a self-promotion strategy and for more conviction of their relevant credentials.

The adjective "*clear*," however, was found to be rarely apparent in all the PSs. Indeed, it was frequently absent in the American corpus, but it appeared only once in a British physics statement. This may be explained by the fact that this genre lacks the aspect of scientific demonstration that is why such adjectives seem to be generally absent.

5 Conclusions, Implications and Recommendations

The main findings of this study showed that Ding's (2007) five-move structural model for interpreting school application letters was a very useful starting point, but not totally applicable for the British and American statements due to certain structural and rhetorical modifications. The rhetorical analysis of the statements under investigation has revealed some similarities and differences between both corpora. In addition, the linguistic analysis has demonstrated the presence of the two main linguistic strategies namely the self-promotion and the use of boosting and hedging strategies. These strategies served mainly to highlight the candidate's presence and strengthen their positions in the statements. The divergences and convergences between both corpora were attributed to certain socio-cultural and academic factors.

The results of the present study offered valuable implications. From a linguistic perspective, this research is of great importance in understanding the move analysis of the British and American undergraduate university application letters, if not, it provided, at least, some helpful insights about the cognitive structuring of the genre of promotional literature.

Further, the present study may fill in the gap in the rhetorical studies of the British and American academic genres. Indeed, it has shown the main differences and similarities between the British and American writing styles. In addition, it has tackled certain cross cultural and interdisciplinary variations between the statements. These findings may enrich the field of cross-cultural and genre studies.

From a pedagogical point of view, the present study highlights the potential benefits of including the genre of PSs in the classroom activities assessment, particularly in the reading and writing courses for students. Indeed, school teachers need to provide more information of PS writing for both native and non-native undergraduate learners because of the prominence of the PS writing in the student's academic career. In fact, the inclusion of this type of writing activities in the various academic and technical writing courses may help to raise the learners' "rhetorical and genre consciousness" (Swales, 1993, as cited in Bhatia, 2002, p. 14). This would help students who are applying for different universities to know the audiences' expectations and thus enhance their chance of being admitted in the target program. By making clear to students, for instance, what teachers expect and value in their writing tasks, applicants would know the criteria of evaluation and success. This may give them greater motivation and confidence to write appropriately (Hyland, 2007).

This study has some limitations which need to be reconsidered for the development of potential follow-up studies. From a methodological perspective, this study has relied merely on "textual approaches" (Hyland, 2009, p. 20) reflected in the genre analysis and corpus analysis of the selected data. Indeed, unlike Brown (2004) and Smaraj and Monk (2008), this research was not able to include "contextual approaches" (Hyland, 2009, p. 20) such as questionnaires or interviews neither with the students nor with the committee admission members, whose "insider views" can yield valuable information regarding expected features in PS writings (Sibo, 2011). Therefore, further studies are needed to focus on the expectations of the admission committee members on the one hand, and the writers' choices and assumptions on the other.

In addition, there is a noticeable lack of cross-cultural and cross-disciplinary comparison in this research due to the limited sources and the difficulties to get access to some reputable online journals. Hence, further studies should explore in depth the cross-cultural and cross-disciplinary variations in this type of genre. It would also be interesting to investigate PSs from different languages and different cultures and examine the socio-cultural factors being behind the rhetorical and linguistic differences or similarities.

References

Ackland, G. M. (2009). *A discourse analysis of English and French research article abstracts in Linguistics and Economics*. San Diego, California: Montezuma Publishing Press.

Bhatia, V. (2002). Applied genre analysis: A multi-perspective model. *Iberica, 4*, 14–16.

Biber, D. (1987). A textual comparison of British and American writings. *American Speech, 62*, 99–119.

Brown, R. M. (2004). Self-composed rhetoric in psychology personal statements. *Written Communication, 21*(3), 242–260.

Callaghan, G. M. (2004). Writing a winning statement of purpose. Retrieved May 30, from http://www.sjsu.edu/faculty/gcallaghan/graduate/winningstatement.htm

Connor, U. (2002). Contrastive rhetoric and academic writing: Multiple texts, multiple identities. *Forum*: *Applied Linguistics Newsletter*, *23*(1), 1–6.

Ding, H. (2007). Genre analysis of personal statements: Analysis of moves in application essays to medical and dental schools. *English for Specific Purposes, 26*(3), 368–392.

Halliday, M. (1994). *An introduction to functional grammar* (2nd ed.). London: Arnold.

Helal, F. (2013). Genres, styles and discourse communities, in global communicative competition: The case of the Franco-American AIDS War (1983–1987). *Discourse Studies, 16*(1), 47–64.

Hsaio, C. (2003). *Analysis of panel data*. Cambridge, UK: Cambridge University Press.

Hyland, K. (2000). *Disciplinary discourses: Social interactions in academic writing*. London, England: Longman.

Hyland, K. (2007). Genre pedagogy: Language, literacy and L2 writing instruction. *Journal of Second Language Writing, 16*(3), 148–164.

Hyland, K. (2009). *Academic discourses*. London: Continuum.

Kaplan, R. B. (1966). Cultural thought patterns in intercultural education. *Language Learning, 16* (1), 1–20.

Li, Y. (2011). *A genre analysis of English and Chinese research article abstracts in linguistics and chemistry*. (Unpublished master's thesis). San Diego State University, California, The United States of America.

Mauranen, A. (1993). Contrastive ESP rhetoric: Metatext in Finnish-English economic texts. *English for Specific Purposes, 12*, 3–22.

Murphy, B. (2010). *Corpus and sociolinguistics: Investigating age and gender in female talk*. Amsterdam: John Benjamins.

Pokrivack, A., Hevesiova, S., Smileskova, A., & Janecova, E. (2010). *Literature and culture*. Tlač: Vydavateľstvo Michala Vaška, Prešov.

Samraj, B., & Monk, L. (2008). The statement of purpose in graduate program applications: Genre structure and disciplinary variation. *English for Specific Purposes, 27*(2), 193–211.

Sibo, C. (2011). *Genre features of personal statements by Chinese English-as-an additional-language writers: A corpus-driven study*. (Unpublished master's thesis). University of Victoria.

Swales, J. (1990). *Genre analysis: English in academic and research settings Cambridge*. New York: Cambridge University Press.

Taylor, G., & Tinguang, C. (1991). Linguistic, cultural, and sub cultural issues in contrastive discourse analysis: Anglo-American and Chinese Scientific texts. *Applied Linguistics, 12*(3), 319–336.

Yunxia, Z. (2000). Building knowledge structures in teaching cross-cultural sales genres. *Business Communication Quarterly, 63*(4), 49–68.

Part IV
Assessment of Productive Skills

Learner Differences: A Trojan Horse Factor in Task-Based Oral Production Assessment

Mohamed Ridha Ben Maad

Abstract Research efforts to operationalize task difficulty and gauge the extent and direction of its effect on oral performance has been the common denominator between the different models within the purview of task-based assessment (TBA). This area has benefited from an assessment triad conceived by Skehan (1998) wherein oral performance can be evaluated along three areas: Complexity, accuracy, and complexity (CAC). Despite the plethora of empirical findings triggered by TBA models, the methodological schema to standardize difficulty predictors and regulate task demands has remained almost unaltered for almost two decades, and so reached some conventionalism evident in the dependency cross-sectional format. As such, a number of related research projects were invariably committed to establishing an effect structure based on a well-studied range of task design features and/or task conditions. This effect equation seems to be deterministic and speculative in the absence of a real role for individual differences, since measuring task effect against hypothetical learners across assorted experimental contexts can but produce an impressionistic picture of task difficulty.

Keywords Task difficulty · Individual differences · Oral production · Proficiency · Processing

1 Introduction

Based on the prior definitions of Brindley (2009, p. 437), Crookes (1986) and Nunan (1989) defined TBA as "the process of evaluating, in relation to a set of explicitly set criteria, the quality of the communicative performances elicited from learners as part of goal-directed, meaning-focused language use." This definition is anchored in the communicative wave, giving primacy to the communicative goals

M.R. Ben Maad (✉)
Institut Supérieur des Cadres de l'Enfance, University of Carthage, Tunis, Tunisia
e-mail: ridhamaad@hotmail.com

S. Hidri and C. Coombe (eds.), *Evaluation in Foreign Language Education in the Middle East and North Africa*, Second Language Learning and Teaching,
DOI 10.1007/978-3-319-43234-2_9

in syllabus design and assessment with communicative proficiency being operationalized in terms of knowledge and ability to use language appropriately in different social contexts. As task completion has become the ultimate purpose of second language (L2) learners, abilities-oriented proficiency assessment has therefore given way to models that mesh up well with the communicative facet of the target language.

This explains in some way the surge of testing models attendant to spoken language assessment, portraying among other things the communicative event in a more comprehensive way (Bachman & Palmer, 1996). In this vein, a change away from the prevailing standardized, uniform rater-focused assessment towards a more multidimensional, complex learner-focused one as the former has been increasingly criticized for validity issues and biases emanating from assessors/raters and scales (Skehan, 2001).

The main rationale in the present chapter is to address how task-based assessment may capitalize on the role of individual differences as a constructive tributary rather than an impediment to the effort of capturing a principled image of the task-learner interaction. In that, we wish first to review task difficulty which is a core concept deeply engrained in cognitive psychology whose understanding has been consequential for the different research configurations documented in task-based assessment. Among such configurations is the triadic assessment scheme advanced by Peter Skehan and his research associates in the mid-1990s and whose bearing is still felt in most recent literature. Also worth discussing is that this research format has reached the status of orthodoxy evident in its oversight of individual differences as an outstanding intervention factor of a possible disturbing effect on the already established task-constrained take on L2 learners output. Eventually, we will shed light on a research line which attempts to reposition the role of the learner as a fundamental yet approachable component of any task-based assessment equation.

2 Task Difficulty: Detrimental or Instrumental?

It follows from an overview of the TBA literature that task difficulty constitutes the organizing concept around which the prevailing cognitive approach to task (see Brindley, 2009; Ellis, 2009; Révész, 2014) has advanced two models contrasted on whether task difficulty is detrimental or necessary to the learning process. The first model is conceived by Skehan (1998) that difficulty has a taxing effect on L2 performance and the various degrees of task difficulty reflect differentials in the cognitive demands, which in turn manipulate the allocation of attentional resources towards different areas of performance. Skehan (1998) distinguished three types of task difficulty. First, *code complexity* represents the linguistic type of difficulty to be grappled with so as to accomplish any given task. For example, comprehension texts with high frequency vocabulary are much easier to understand than similar texts with low frequency vocabulary. Also, some discourse modes, such as narratives, may be more demanding than other ones. Second, *cognitive complexity* refers to the cognitive demands triggered by the content of processed information (e.g., the amount of computation involved in a task topic or discourse genre). Third, *communicative stress* underlies the performance conditions of tasks. It reflects the "urgency with which a task needs to be completed" (Skehan, 1998, p. 100) under external conditions like time pressure and the number of participants who control conversational exchanges.

Far from the idea of trade-offs triggered by the testing and resource-depleting nature of cognitive demands, the other model describes difficulty as a resource-directing, thus conductive to L2 learning achievements. Robinson (2001), in this regard, distinguishes three types of task-triggered cognitive demands: (i) A *task difficulty* category being associated with individual difference variables whose estimation is hardly predictable (Bachman, 2002; Elder, Iwashita, & McNamara, 2002), (ii) a *task conditions* category that stands for the external factors bearing on task takers (e.g., interlocutor and setting), and (iii) a, *task complexity* category that constitutes the cognitive demands triggered by a given task design feature or sequencing conditions. According to Robinson (2001), increasing task complexity may direct L2 learners' attention to a given area of L2 performance, and hence to L2 learning attainment. For instance, Robinson (2001, p. 317) argued that more cognitively complex task design features (e.g., displaced/past time reference) may push L2 learners to stretch their interlanguage and try out greater morpho-syntactic complexity. It sounds therefore rather plausible to capitalize on predictable and generalizable task complexity conditions (e.g., planning, task design, and topic) which are easy to control and/or assess than on unpredictable learner-related variables (e.g., aptitude, goal orientation, and learning styles).

The resource-directing nature of task difficulty has been further evidenced through the nature of two processing modes: Analysability and formulaicity. Ben Maad (2010), in keeping with findings reported by Foster (2001) and Temple (2000), contended that time pressure as task sequencing variable engages different processing patterns due to the differences in terms of their formulaic repertoires. That is, non-native speakers whose formulaic reservoir is limited perform better in easy tasks due to their resort to such depository. However, when task difficulty is high and so their communication needs, their performance exacerbates. This might be explained by their recourse to the rule-based system (i.e., analysability) after exhausting the use of the formulaic depository. With reference to the model advanced by Levelt (1989), such resource-directing decision denotes that these capacity-stretched task takers' encoding system was unable to develop the complex verbal plans from the conceptualizer with whatever pre-fabricated morphosyntactic and lexical combinations (i.e., formulaicity). Their longer pauses and frequent reformulations being types of communication strategies allowed them to use novel message-form mappings that call for more morpho-syntactic elaboration and hence a restructuring behaviour.

3 The Triadic Assessment Scheme

Skehan (1996, 1998) put forward an assessment scheme constituting a triad of fluency, accuracy, and complexity (CAF). The CAF scheme is well grounded in the cognitive approach which holds that these performance areas stand as “goal areas” to which L2 learners should selectively attend due to the trade-off situation ensuing from task difficulty. This tripartite scheme enables a principled account of the variation occasioned by difficulty differentials in a predictable way. To such thrust for a strong connection between the processing trade-offs and the tripartite scheme, Skehan (1996, p. 50) argued that “there is not sufficient capacity for learners to devote resources to each of them so that they can be met simultaneously, [and so] decisions about prioritization of attentional resources have to be made during communication and learning”. Wolfe-Quintero, Inagaki, and Kim (1998, p. 59) concurred with this statement, positing that “with such a system for estimation of task difficulty, learner performances on carefully sampled tasks can be used to predict future performances on tasks that are constituted by related difficulty components.” The implications of this performance taxonomy for future research and pedagogy were such seeing the huge volume of research conducted in L2 learning and teaching literature where the model functions as an organizing framework to tap difficulty-performance patterns.

Findings in this framework have provided predictions about L2 learners’ task-engagement patterns which would eventually provide a discerning input for pedagogical decisions about task sequencing and evaluation in L2 classes. As a matter of reflection, some meta-studies such as Cumming (2006) and Ellis (2000) asserted that this research line operates within a proficiency-focused pedagogical agenda that seeks to develop standard task properties that guide learners towards predictable performance outcomes. In fact, the tripartite assessment scheme advanced by Peter Skehan and other research associates was an attempt to develop a balanced arrangement that helps expedite the process of L2 learning and enhance its quality. Consistent with this triadic configuration, a number of task-focused researchers (e.g., Kuiken, Mos, & Vedder, 2005) maintain that learners should proportionally (a) work towards a more lexically and syntactically elaborate interlanguage (complexity), and (b) gain more control over the rule-based system (accuracy), and (c) increase the degree of lexicalization to streamline real time processing (fluency).

4 Learner Variation in TBA

To the wide appeal of the tripartite assessment scheme evidenced in the substantial L2 research effort documented so far, the role of individual differences has remained under-represented. An overview of studies inspired by Robinson (2001) and Skehan (1998) would permit the assertion that their assessment frame of reference is in some way deterministic and speculative as task difficulty components were explored against a hypothetical learner whose profile is generalizable across different research contexts and purposes (Ellis, 2000). It should be mentioned though, that the disproportionate focus on task structure and sequencing variables to the detriment of L2 learner-inherent variation has been broached by a number of L2 assessment scholars since the early 2000s. Bachman (2002) called for a revision of the conceptualization of task demands. Bachman (2002) noted that Skehan (1998) treated task demands as detached variables that can be isolated for empirical testing. Bachman claimed that communicative stress and task complexity are fundamentally individual characteristics. Consequently, he argued, task demands "are not inherent in tasks themselves, but are functions of the interactions between a given test-taker and a given test task [and so the] empirical estimates of task difficulty are not estimates of a separate entity, 'difficulty,' but are themselves artifacts of the interaction between the test-taker's ability and the characteristics of the task" (Bachman, 2002, p. 464). Critical of the reductionist, yet established TBA stance where task demands are treated as detached variables, Bachman contends that a tenable assessment of task-triggered outcome should start from containing the psycholinguistic reality of s/he who performs a given task.

In the same vein, Elder et al. (2002) and Iwashita, McNamara, and Elder (2001) proposed a balanced assessment of task difficulty, yet not fully confined to the difficulty emanating from task features and task conditions. A balanced assessment of task difficulty should instead focus on the often mentioned but seldom empirically attested, effects of individual difference variables. This calls for redefining, and thus operationalizing, the construct of difficulty not as a detached task variable, as typically defined in the cognitive approach to task (Robinson, 2001; Skehan, 1998), but rather as a factor that mediates the effect of the individual differences on L2 speaking performance and development. Rather than looking to standardize task difficulty predictors by way of drawing on speculative results of a model that views the human factor as an anomalous component from the assessment equation, Elder et al. (2002, p. 305) argued that "there may be some value in canvassing test-takers' perceptions of task difficulty to determine how influential these are in test performance."

5 Learner Variation Redeemed?

As mentioned earlier, reference to the value of including IDs in TBA is not exclusively lacking in the L2 research literature. What seems more judicious, though, is the necessity to position and/or demarcate an established role for ID variables in the assessment process. Ellis (2009, p. 499) accordingly pointed to the need to "investigate the mediating role played by such ID factors as working memory, language aptitude, willingness to communicate, and risk-taking (…)

[since we] are told so little about the actual learners in many of the studies." The treatment of IDs as a mediating constituent suggests that it is the learner's perception of task difficulty and not the intuitive judgment of researcher and/or external rater that predicts the real effect of tasks on L2 learning outcomes. Similarly, Révész (2014) argued for a kind of "independent evidence" that should be verified rather than assumed to ascertain that task manipulations/constituents have indeed led to the predicted learning outcomes. TBA would hence establish for a principled practice grounded on hands-on view of the task-taking process far from the simplistic intuitions about hypothetical individuals.

Findings reported in Ben Maad (2012) attested to this assessment course through an extensive validation effort to establish difficulty as a matter of Tunisian participants' (*N* = 30, pre-intermediate English proficiency level) immediate accounts before and while performing the tasks. Ben Maad (2012) was carried out based on the assumption that goal orientation, a relatively unexplored ID variable in L2 research, and task conditions would have a joint effect on speaking production and development. The findings revealed that some L2 learners were committed to speech automaticity/formulaicity whereas others focused on restructuring. Incorporating this complex, yet surely revealing joint-effect design, into mainstream task-based agenda would stretch the edge of this ever-growing research area even further. That is, the inclusion of the ID variable of goal orientation in this effect formula would highly substantiate the contention that any L2 learner-originated aspect should not be treated as a Trojan horse factor to be suppressed but rather as an illuminating resource.

6 Conclusion

Researchable ID variables, such as being attendant to L2 learner goal orientations, can be integrated into educational praxis as illuminating parts of TBA only on the stipulation of being subjected to an extensive validation effort so as to consolidate their reliability and validity robustness. Henceforth, task difficulty is seen as a matter of learner perception more than a prerogative of practitioners or task developers. Established ID variables-evidently through extensive validation treatment-may well yield a more realistic and predictable picture of how such variables interact in their effect on L2 performance with task-triggered factors. Instead of relying on a pre-existing taxonomy of difficulty as that of Robinson (2001) or Skehan (1998) , the experimental substantiation of individual differences as in Ben Maad (2012) may enable practitioners to explain the related, yet systematic, variation in the learners' perceptions of task difficulty. It is therefore only through rejecting the view of tasks as "neutral devices for testing," as Iwashita et al. (2001, p. 406) pointed out, that TBA stakeholders can address the scope of learner variation in their projects with more efficiency.

References

Bachman, L. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*, 453–476.

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

Ben Maad, M. R. (2010). Holistic and analytic processing modes in non-native learners' performance of narrative tasks. *System, 38*(4), 591–602.

Ben Maad, M. R. (2012). *The Joint effects of goals and tasks on EFL speaking performance and development: A processing-based perspective*. Unpublished doctoral thesis. University of Manouba, Tunisia.

Brindley, G. (2009). Task-centred language assessment in language learning: The promise and the challenge. In K. van den Branden, M. Bygate, & J. Norris (Eds.), *Task-based language teaching: A reader* (pp. 435–454). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Crookes, G. (1986). *Task classification*: *A cross-disciplinary review*. Technical report 4. Honolulu: University of Hawai'i Press.

Cumming, A. (2006). *Goals for academic writing: ESL students and their instructors*. Amsterdam: John Benjamins.

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? *Language Testing, 19*, 347–368.

Ellis, R. (2000). Task-based research and language pedagogy. *Language Teaching Research, 4*(3), 193–220.

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics, 30*, 474–509.

Foster, P. (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 75–93). Harlow: Longman.

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information processing approach to task design. *Language Learning, 21*, 401–436.

Kuiken, F., Mos, M., & Vedder, I. (2005). Cognitive task complexity and second language writing performance. In S. Foster-Cohen & P. García-Mayo (Eds.), *EUROSLA yearbook* (pp. 195–222). Amsterdam: John Benjamins.

Levelt, W. (1989). *Speaking: From intention to articulation*. Massachusetts: MIT Press.

Nunan, D. (1989). *Designing tasks for the communicative classroom*. Cambridge: Cambridge University Press.

Révész, A. (2014). Towards a fuller assessment of cognitive models of task-based learning: Investigating task-generated cognitive demands and processes. *Applied Linguistics, 35*, 87–92.

Robinson, P. (2001). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 114–137). Cambridge: Cambridge University Press.

Skehan, P. (1996). A framework for the implementation of task based instruction. *Applied Linguistics, 17*, 38–62.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 167–185). Harlow: Longman.

Temple, L. (2000). Second language learner speech production. *Studia Linguistica, 54*, 288–297.

Wolfe-Quintero, K. Inagaki, S., & Kim, H. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity*. Technical report 17. Manoa: University of Hawai'i Press.

Assessing ESL Students' Paraphrasing and Note-Taking

Yasmine Soheim

Abstract ESL university students are expected to be familiarized with paraphrasing and note-taking skills throughout the first few months of their new academic life in order to be prepared to do integrated writing tasks, which are increasingly being used in academic writing. Their teachers continuously highlight the concept of plagiarism and emphasize its seriousness. Therefore, assessing the paraphrasing and note-taking skills on a regular basis is of great importance to the students' academic success. After teaching paraphrasing and note-taking, it is preferable to administer a formal assessment such as a test for two main reasons. First, as a new skill to the majority of undergraduate students, it becomes critical to assess it on a regular basis in order to assure its mastery. Second, this kind of assessment becomes essential as it serves as a diagnostic tool. The researcher describes the different phases of developing (initial planning, test specs, and item writing and moderation) and administering a test assessing both paraphrasing and note-taking skills of Arabic-speaking ESL undergraduate students.

Keywords Paraphrasing · Note-taking · Academic writing · Plagiarism · ESL

1 Introduction

ESL university students are expected to be familiarized with paraphrasing and note-taking skills throughout the first few months of their new academic life. Actually, their teachers highlight the concept of plagiarism and emphasize its seriousness. Therefore, assessing the paraphrasing and note-taking skills on a regular basis is of great importance to the students' academic success. After teaching paraphrasing and note-taking, it is preferable to administer a formal assessment such as a test for two main reasons. First, as a new skill to the majority of undergraduate students, it becomes critical to assess it on a regular basis in order

Y. Soheim (✉)
American University in Cairo, Cairo, Egypt
e-mail: soheim@aucegypt.edu

S. Hidri and C. Coombe (eds.), *Evaluation in Foreign Language Education in the Middle East and North Africa*, Second Language Learning and Teaching,
DOI 10.1007/978-3-319-43234-2_10

to assure its mastery. Second, this kind of assessment becomes essential as it serves as a diagnostic tool.

In integrated tasks, different language skills, such as reading and listening, are combined in one task with writing, where students can summarize or paraphrase from the source text in order to produce their own work. By adopting this approach, students are getting better prepared for what is expected from them during their coming academic courses, such as research papers, case studies, reaction papers and reports (Gebril, 2009; Leki & Carson, 1997; Plakans, 2008; Weigle, 2004).

In the following sections, the researcher describes the different phases of developing (initial planning, test specs, and item writing and moderation) and administering a test assessing both paraphrasing and note-taking skills of Arabic-speaking ESL undergraduate students.

2 Theoretical Background

For decades, independent writing tasks have been prominent in second language learning. However, after being widely criticized (Cho, 2003; Cumming, Kantor, Powers, Santos, & Taylor, 2000; Gebril, 2006; Gebril & Plakans, 2009; Hamp-Lyons & Kroll, 1996; Leki & Carson, 1997; Plakans, 2007; Weigle, 2002, 2004) for not accurately eliciting the writing construct, integrated writing tasks have gained more ground in both high-stakes testing and classroom contexts (Gebril & Plakans, 2009, 2013). In fact, integrated tasks are “more complex and more demanding than traditional stand-alone or independent tasks, in which test-takers draw on their own knowledge or ideas to respond to a question or prompt” (Brown, Iwashita, & McNamara, 2005, p. 1).

On a similar note, writing instructors are required to be familiarized with the issue of copying text referred to as verbatim source use, which has been greatly investigated (Currie, 1998; Johns & Mayes, 1990; Pennycook, 1996; Shi, 2004). Pennycook (1996) argues that the concept of plagiarism is not similarly perceived among different cultures. Therefore, students need to learn how to make correct textual borrowings from a source text, without falling into plagiarism. In their study, John and Mayes (1990) analysed direct copying in 80 writing samples that were divided into high and low proficiency, where participants were required to summarize a text. According to their study, direct copying was more present in the writing samples of the low proficiency level participants. Other studies (Cumming et al., 2005; Gebril & Plakans, 2009) showed that high proficiency students used summarizing as a source of integration, medium proficiency level students employed paraphrasing, whereas low proficiency students used the text sources the least. A possible explanation to these results could be that low proficiency students might fear falling into plagiarism while using the source text. Thus, writing instructors should address teaching paraphrasing and note-taking skills in their classrooms in order to help their students in performing better on their integrated writing tasks.

3 Method

3.1 Context of the Study

This study explores the different phases of developing and administering a test assessing both paraphrasing and note-taking skills of Arabic-speaking ESL undergraduate students. The location setting of the testing was at an American university in Cairo. Freshmen undergraduates, who were enrolled at the Intensive English Program (IEP), represented the subjects of the testing. The IEP had two different levels: English 98 (lower intermediate) and English 99 (upper intermediate). At this initial phase, IEP students were only allowed to take intensive English language courses. These courses included grammar, writing, reading, vocabulary, and study skills. For each section, there were three responsible teachers; the first one taught grammar and writing, the second was assigned to teach reading and vocabulary, and the third was responsible for teaching students the study skills course. Students attended classes four days a week, with an average of five to six hours of classes. The assessment in question was developed for the study skills course. During the study skills course, students were expected to learn the basic skills needed for oral presentations and improve their listening skills. Actually, students were being prepared through their first semester at AUC for a series of English achievement tests that they should pass at the end of the semester in order to fulfil the English language requirements, which were mandatory to pursue the rest of their academic courses through their different majors.

As mentioned earlier, a study skills course should build the skills that the university students need in their various future academic courses. Those skills included paraphrasing and note-taking. Most students were not usually familiar with the rules of paraphrasing and they were not good note-takers. Therefore, it was the teacher's duty to teach the new undergraduates how to paraphrase and how to take notes. After the teaching process of those skills, it was preferable to test students in order to know where they stood. Developing an assessment of paraphrasing and note-taking was the teacher's responsibility. The creation, administration and scoring (according to a developed scoring rubric) of a test on paraphrasing and note-taking was a challenging task. Part of the difficulty of such tests was that those tests were considered integrative. In other words, to assess those students' abilities in paraphrasing and note-taking, reading and listening skills were also included in the test.

3.2 Participants

A class of sixteen IEP students, placed in level 99, received preparation to take the test in question. These students were considered as in upper intermediate level. Their age ranged from eighteen to twenty years old. There were eleven male students and five female students. The majority of the students were Thanawiyya Amma (secondary school) holders, except for two of them (one male and one female student) who were IGCSE holders. One male student was being challenged by visual impairment, which required extra preparation during the test planning. All students were in their first semester at university.

3.3 Test Purpose and Available Resources

According to Brown (2010), the teacher has to determine the purpose and the usefulness of any assessment during the initial planning. After teaching paraphrasing and note-taking, it was preferable to administer a formal assessment such as a test for two main reasons. The first reason was to measure the students' understanding of the past lessons especially that it was almost new to the majority of them. In this case, the test served as an achievement test. The second reason was to diagnose what students lacked in this particular lesson, which needed more work in the future. This test was administered in the middle of the semester. It was formative in the way it diagnosed the students' strengths and weaknesses and allowed the teacher to emphasize on what was needed in order to succeed in the final exam. In other words, students were assessed for learning. Moreover, this test was criterion-referenced because the students were compared to a set of standards, unlike norm-referenced tests (Alderson, Clapham, & Wall, 1995; Brown, 2005; Hughes, 2003).

For a class of sixteen students, there was no need for extra human resources such as proctors and raters. The university allowed the teachers to make use of the photocopiers through an early written order to be submitted in a specific place at least twenty-four hours before the required date. The test room was very well equipped with AC and all the technology needed. The audio tracks could be played on the room PC. As a backup plan, the teacher could make an order from the administration to borrow an audio player in case the room PC fails to work properly. The seating arrangements could be easily done before the test as all chairs and tables had wheels.

Additional arrangements to accommodate the visually challenged student were taken into consideration. Such accommodation included having a reader or installing an application on the student's PC to read aloud to him while using his headphones. Also, the student was being trained on fast touch-typing in order to facilitate the process of taking notes.

3.4 Instruments

As its name indicates, IEP is very intensive and students are taking rigorous English language courses that teach them the different skills they need in order to pursue successfully their academic life in university. Grammar and writing, reading and vocabulary, and study skills are the three different subjects that three instructors are teaching to every section. The testing in question is part of the study skills course. During the study skills course, students usually encounter a large array of new skills that they need to learn. In fact, this specific course is giving the IEP students the opportunity to develop those skills through various learning phases. One major cornerstone skill that most students are struggling with is the paraphrasing skill. Learning to paraphrase is an essential step and therefore requires extensive training. Note-taking is usually associated with paraphrasing because students will definitely need to take paraphrased notes from lectures or readings. It is preferable for teachers to assess the level of their students on a regular basis especially after teaching a new skill such as paraphrasing and note-taking in order to be aware of the learning curve of each student. Paraphrasing and note-taking are two language skills that are integrative, which makes them challenging in order to be tested. Therefore, most of the time, reading and listening skills are also integrated in the testing process. The following test specs are following the same Blueprint format suggested by Bachman and Palmer (2010).

3.5 Test Specs

The paraphrasing and note-taking assessment includes two parts. The first part is designed to assess the paraphrasing ability of the student after working around five to six weeks since the beginning of the semester on acquiring it. The paraphrasing part integrates reading as a second skill in the assessment. The second part of the assessment is concerned with note-taking. After being trained for six weeks on taking notes and deducing the outline of the material in question, students are assessed based on their note-taking skills following a listening.

As for the first part of the test, paraphrasing is assessed according to two tasks. In each task, students are given a short passage that should not be less than 75 words and not more than 110 words, and they are required to produce a paraphrase for the corresponding paragraph.

In the second part of the test, where note-taking is assessed, there are two main tasks. Students listen to a recorded mini-lecture that should not last more than 10 min. During the first task, students should take notes while listening. As for the second task, they are expected to produce an outline, based on their notes, stating the main ideas and mentioning the supporting details of the mini-lecture.

The first part on paraphrasing should come before the second part on note-taking. The rationale behind this sequence is that paraphrasing can be considered as a warm-up to note-taking. In other words, students need to paraphrase their notes in the second part of the test. Concerning the sequence of tasks in the first part, the two short passages that test-takers are required to paraphrase are classified from easier to more challenging in order to accommodate the majority of students. The sequence of the two tasks involved in the second part of assessment of note-taking is evident due to its nature. Test-takers take notes from the listening. Then, they produce the outline according to their notes.

The two parts of the test are of equivalent importance. Paraphrasing and note-taking are two major skills that students should acquire by the end of the semester. Hence, the two parts of the tests have an equal weight of 50 %. Furthermore, the two tasks from part one are also similar in their importance. The weight is divided equally among them because they share the same nature of input and output. Similarly, the weight of the two tasks from part two is equally distributed with 50 % each because taking notes while listening to a mini-lecture is the foundation step towards writing a valid outline. As for the time allotment for each part, the test is designed in a way that each part should take 30 min:

Part one: 30 min

- Task 1: 15 min (paraphrasing first passage)
- Task 2: 15 min (paraphrasing second passage)

Part two: 30 min

- Task 1: 10 min (taking notes while listening to the mini-lecture)
- Task 2: 20 min (writing an outline based on the notes taken in task 1)

Students are given the general following instructions at the beginning of the test:

The test has two distinct parts. Each part is weighted at 50 %. In the first part, you are required to read the two passages carefully and provide a hand-written paraphrase in the specified area on the answer sheet. Each task is worth 25 % of the final grade. In the second part, you are required to take hand-written notes while listening to a 10-minute mini-lecture. Finally, you need to create an outline based on your notes in the final task of the test. The weight of the two tasks of part two is equal (Table 1).

The constructs to be assessed in the two parts of the test are as follows:

Table 1 Specs of input and output of part one of the test

Length	75–110 words
Topics	Mostly academic
Range of vocabulary	Familiar but may include some specific vocabulary related to the topic of the passage
Range of structures	Clear structures that are quite common in academic texts
Readability/difficulty level	Passages should be readable and most students should understand most of it with an increasing level of difficulty in the second passage
Style	Formal
Genre	Academic
Nature of output	Expected response is also visual: handwritten real production
Relationship between input and output	Non-reciprocal because test-takers answer the questions without additional output. Direct relationship because students answer questions

Paraphrasing a short passage: This part assesses the ability of the test takers to produce an accurate and complete paraphrase to a short passage ranging between 75 and 110 words approximately with an increasing level of difficulty in the two tasks. The students are also assessed on their ability to follow the basic foundations of paraphrasing learned in class, such as semantic completeness, lexical and semantic difference, and the overall quality of their paraphrased production (McCarthy, Guess, & McNamara, 2009).

Taking notes while listening to a mini-lecture: This task assesses the ability of the students in identifying and taking notes using abbreviations learned in class about the main ideas and supporting details following an academic mini-lecture that lasts around 10 min. The channel of input for the first task of part two is audio as it involves listening to a mini-lecture. The type of input is for comprehension and identification of main ideas and supporting details. The language will most likely include academic vocabulary such as any academic lecture in a university context. Finally, the vehicle of the input is a recorded track such as MP3.

Creating an outline from the notes taken during task one: Test-takers are assessed on their ability on creating an organized outline based on their notes, where the main ideas and supporting details are identified and organized according to the methods learned in class.

The passing score for this test is 65 %. The rationale behind this passing score is to prepare the test-takers for the final exam. The test is to be administered near the middle of the semester in order to identify the students' weaknesses in paraphrasing and note-taking. It is a formative test after which students receive positive washback and the teacher is able to better diagnose the paraphrasing and note-taking abilities of the students in the class. This test reveals to the teacher the different learning achievements in paraphrasing and note-taking skills. Hence, the teacher is able to obtain a solid picture of what is needed by the students concerning the skills in question. In other words, the results of the test will be used to make instructional decisions, such as more paraphrasing practice.

Scores will be delivered to students using percentage results. Moreover, they will get back their answer sheets with written feedback from the teacher in order to increase the positive washback. Test-takers will get the opportunity to discuss their results with their teacher according to the scoring rubric.

The test is administered in the classroom during the regular class hours. It is a one-hour exam. The class teacher will be proctoring the students. Students receive printed sheets of the test and are required to answer on a separate answer sheet with their own handwriting. Before the assessment administration, the class teacher needs to prepare the class settings. Students should be seated comfortably, with enough space between them. Also, the teacher should be prepared with the audio track of the mini-lecture which is essential for the second part of the test. A back-up plan should be ready with the teacher, who should seek technical assistance whenever needed. In order to help students, succeed in the tests, the testing environment should be supportive. For instance, the A/C should be working properly and noise should be minimized. In order to facilitate the performance of the students, the teacher should read aloud all the different task instructions in front of all

Table 2 Specs of input and output of part two task one

Length	9–11 min
Topics	Academic such as university lectures
Range of vocabulary	Academic with some technical terms
Range of structures	Syntax: Narrow to moderate range of organized structures
Readability/difficulty level	Similar to academic lectures in university context to be as authentic as possible
Style	Formal
Genre	Academic mini-lecture
Speed and accent	The speed is moderate. The accent is the standard English accent
Nature of output	The channel for the expected response is visual. Test takers are expected to produce handwritten notes following the listening
Relationship between input and output	Non-reciprocal because test-takers take notes without additional input Direct relationship because students answer questions

the test-takers. Moreover, the teacher should require the students to keep the test sheet face down until they are told to begin. As a result, all test-takers get the same exact time for the test. After 25 min from the beginning of the exam, the test administrator should remind the test takers that they should get prepared to listen to the mini-lecture in five minutes. On the 30th minute, the audio track is played and students are doing the first task of part two from the test. After the allotted time for the test, the teacher collects all the test sheets and the answer sheets from students and verifies that the number of sheets is correct. Finally, special precautions should be taken concerning the audio track and having a technical back-up plan is necessary (see Table 2).

4 Data Analysis

The scope of this section is to analyse the items of the test that has been developed after being piloted. Ten undergraduate students from the IEP kindly accepted to take the test during their regular class hours. They were all in the same class and shared the same ELIN99 level.

The test consisted of two parts with two tasks under each. The first part assessed paraphrasing skills and the second part evaluated note-taking skills. The test was formative and therefore was administered near the middle of the semester to help in offering useful feedback to students before the final exams, taking place at the end of the semester. The expectations concerning the first part were that students would perform better on the first task than on the second one due to the difference of difficulty between the two passages. In the second part of the test, it was anticipated that students would encounter some challenges due to the topic of the mini-lecture they listened to. Furthermore, creating an outline out of rough notes was estimated to be quite a difficult task.

5 Results and Discussion

Figure 1 shows the scores of each student on the two tasks on Part I.

As Fig. 1 shows, the scores for the two tasks were quite similar. The highest score on this task was 25 and the lowest one was 13. The student who scored the lowest performed similarly on the two tasks. Two students seemed to have grasped the skills of paraphrasing and therefore scored a very high score on both tasks.

Figure 2 shows the scores of the ten test-takers on Part II of each task.

Although students expected not to perform very well on this part, the scores turned out to be quite high for the majority of them. The highest score for both tasks was 25 and the lowest for task 1 was 10, and 13 for task 2.

Figure 3 represents the total scores of the ten participants.

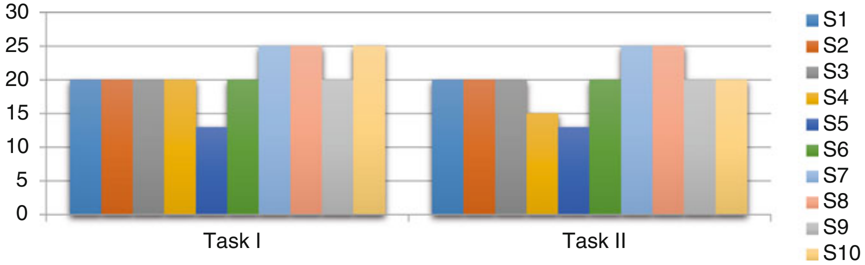


Fig. 1 Part one scores

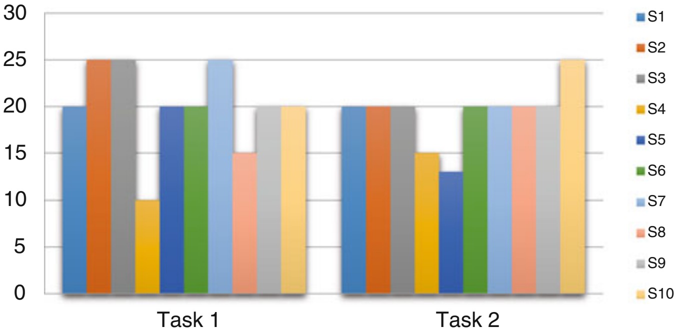


Fig. 2 Part two scores

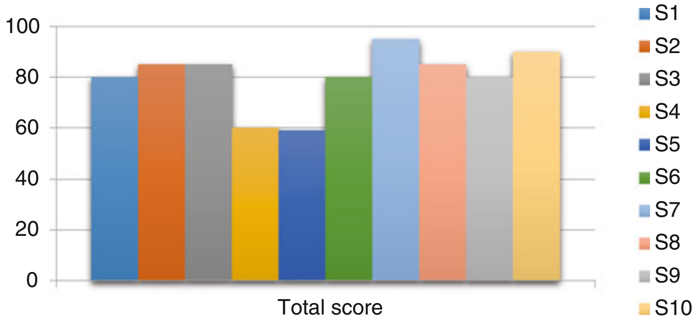


Fig. 3 Total scores (Mean: 79.9; Median: 82.5; SD: 11.14; Highest score: 95; Lowest score: 59)

When analysing the total scores of the ten test-takers, it could be concluded that the performance of the participants was satisfying. The test was not very difficult. However, it served its purposes for formative feedback and for diagnostic purposes.

6 Implications and Limitations

Tests affect both teaching and learning (Brown, 2010). The objective of the teacher is to have positive washback from the assessment. The students have enough time to prepare for their test on paraphrasing and note-taking through a series of in-class and homework practices. This test has a formative nature. Hence, the student expects rich feedback from the teacher. Actually, the teacher can provide individual feedback on the test during office hours. If it is too difficult to meet all students during office hours, the teacher can record the feedback and send it to each student individually on their emails in order to accommodate all students and give them the comments they need to improve their paraphrasing and note-taking abilities. Minimizing or even avoiding negative washback is one of the teacher's great responsibilities. Furthermore, conducting such formative tests help in improving the students' learning skills.

Perhaps one of the greatest limitations of this study was the limited number of participants, which constrained the researcher from making any generalizations. Also, testing paraphrasing and note-taking can be subjective such as essay questions. Therefore, one of the challenges to the teacher is to try as much as possible to minimize subjectivity. This was done through developing a very clear and strict scoring rubric. However, the development of a scoring rubric for paraphrasing and note-taking was indeed the challenge of the whole test development. Another constraint was the preparation of the students to the test. The teacher had to be aware of the progress of his/her students during the semester. In addition, setting the date of the test was the decision of the teacher. S/he will not only determine when the students were ready to take the test but also when the students would be able to perform well. What if the students' performance on the test was outstanding? Does this mean that the teacher had done quite well in teaching paraphrasing and note-taking? Selecting the different parts of the test certainly represented a constraint to the teacher in order to accommodate the weakest student and also challenge the strongest student in the class.

7 Conclusion

It is important to note that in integrated tasks, such as the one administered in the current study, different language skills are combined in one task with writing, where students are expected to summarize and paraphrase from the source text in order to produce their own work without any sort of plagiarism. Students are hence getting better equipped with the needed tools for their future academic courses, which will inevitably include research papers, case studies, and reflection papers.

The test administered by the researcher mainly served for formative feedback and for diagnostic purposes. An interesting finding after piloting this test was that two of the students who usually perform poorly on other tasks, got the highest scores. On the other hand, two of the best performers in class scored a modest total compared to their language abilities. This finding needs to be further researched through piloting other similar tests in the future.

Appendix 1

Name: *Date:*

SID:

Mid-term Test

Paraphrasing and Note-taking

Score: / 100

PART I: Paraphrasing

Task 1: You will have 15 minutes to read the passage and write an accurate and complete paraphrase to it. This task is worth 25 points.

> Of the more than 1000 bicycling deaths each year, three-fourths are caused by head injuries. Half of those killed are school-age children. One study concluded that wearing a bike helmet could reduce the risk of head injury by 85 percent. In an accident, a bike helmet absorbs the shock and cushions the head. From "Bike Helmets: Unused Lifesavers," Consumer Reports (May 1990, p. 348.

………
………………………………………………………………………………………………….

Task 2: You will have 15 minutes to read the passage and write an accurate and complete paraphrase to it. This task is worth 25 points.

> The twenties were the years when drinking was against the law, and the law was a bad joke because everyone knew of a local bar where liquor could be had. They were the years when organized crime ruled the cities, and the police seemed powerless to do anything against it. Classical music was forgotten while jazz spread throughout the land, and men like Bix Beiderbecke, Louis Armstrong, and Count Basie became the heroes of the young. The flapper was born in the twenties, and with her bobbed hair and short skirts, she symbolized, perhaps more than anyone or anything else, America's break with the past. From Kathleen Yancey, English 102 Supplemental Guide (1989): 25.

………
………………………………………………………………………………………………….

PART II: Note-taking

Task 1: Listen to the mini-lecture while taking notes of the main ideas and the supporting details. Please use abbreviations whenever possible. This task is worth 25 points.

http://www.ted.com/talks/angela_lee_duckworth_the_key_to_success_grit.html

………
………………………………………………………………………………………………….

Task 2: You will have 20 minutes to revise your notes in order to create an organized outline stating clearly the main ideas and supporting details in the mini-lecture. This task is worth 25 points.

………
………………………………………………………………………………………………….

Appendix 2

Paraphrasing scoring rubric

4	3	2	1
• Effective paraphrasing strategies • No violation of paraphrasing rules (order, phrasing, ideas) • Development of a smooth and controlled paraphrase of the original text	• Effective paraphrasing strategies • No violation of paraphrasing rules (order, phrasing, ideas) • Paraphrased text is not completely smooth and controlled	• Minor violation of one of the paraphrasing rules (order, phrasing, ideas), but no explicit plagiarism • Paraphrased text is awkward and uncontrolled	• Text is plagiarized due to serious violation of paraphrasing rules

Adapted from: http://8gtaela.weebly.com/uploads/2/4/3/0/2430057/paraphrasing_rubric.doc

Note-taking scoring rubric

Category	4	3	2	1
Keywords versus Copying	Notes are recorded as keywords and phrases	Notes are primarily recorded as keywords and phrases	Notes are primarily copied from the source. Some evidence of keywords and phrases	Notes are copied directly from the source. Notes are not present, missing; no attempt shown
Relevance	Notes relate to the topic	Notes primarily relate to the topic	Some notes relate to the topic, but many don't	Notes are not related to the topic
Organization	Has heading, topic/subtopics listed, and search terms are listed	Heading, topic, and search terms are mainly complete	Some evidence that notes are organized, bulleted, and neat. Heading, topic, and search terms are incomplete	No heading, topic, or search terms listed
Citations	All notes refer to their source	Notes primarily refer to their source	Some evidence that notes refer to their source	No evidence that notes refer to their source
Quantity	More than enough notes are taken to create the product	A sufficient number of notes are taken to create the product	Nearly enough notes are taken to create the product	Not enough notes are taken to create a product Not present, missing; no attempt shown

Adapted from: http://ckjh.cksd.wednet.edu/school/lmc/note%20taking%20rubric.pdf

Appendix 3

General notes on sources used to create the test:

The two tasks of paraphrasing were inspired from the worksheets prepared by Dr M. Rayan at the IEP in the AUC.
It can be found on the Study Skills Google site created by the IEP.
Retrieved June 12, 2008, from http://owl.english.purdue.edu/owl/resource/619/02/

The mini-lecture entitled "Angela Lee Duckworth: The key to success? Grit" can be found on TED talks on the following link:

http://www.ted.com/talks/angela_lee_duckworth_the_key_to_success_grit.html
The transcript is also available on the same webpage.

References

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Brown, H. D. (2010). *Language assessment: Principles and classroom practices*. New York: Pearson ESL.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to english language assessment*. New York: McGraw.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on english-for-academic-purposes speaking tasks (TOEFL Monographs Series #MS29)*. Princeton, NJ: Educational Testing Services.

Cho, Y. (2003). Assessing writing: Are we bound by only one method? *Assessing Writing*, *8*, 165–191.

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). TOEFL 2000 writing framework: A working paper (TOEFL Monograph Series Report No. 18). Princeton, NJ: Educational Testing Service.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing, 10*, 5–43.

Currie, P. (1998). Staying out of trouble: Apparent plagiarism and academic survival. *Journal of Second Language Writing*, *7*(1), 1–18.

Gebril, A. (2006). *Independent and integrated academic writing tasks: A study in generalizability and test method.* Unpublished doctoral dissertation, The University of Iowa, Iowa City, USA.

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing, 26*(4), 507–531. doi:10.1177/0265532209340188

Gebril, A., & Plakans, L. (2009). Investigating source use, discourse features, and process in integrated writing tests. In *Spaan fellow working papers in second/foreign language assessment* (Vol. 7, pp. 47–84). Ann Arbor: The University of Michigan.

Gebril, A., & Plakans, L. (2013). Towards a transparent construct of reading-to-write assessment tasks: The interface between discourse features and proficiency. *Language Assessment Quarterly, 10*(1), 1–19. http://dx.doi.org/10.1080/15434303.2011.642040

Hamp-Lyons, L., & Kroll, B. (1996). Issues in ESL writing assessment: An overview. *College ESL, 6*(1), 52–72.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

Johns, A. M., & Mayes, P. (1990). An analysis of summary protocols of university ESL students. *Applied Linguistics, 11*, 253–271.

Leki, I., & Carson, J. (1997). "Completely different worlds": EAP and the writing experiences of ESL students in university courses. *TESOL Quarterly, 31*, 39–69.

McCarthy, P., Guess, R., & McNamara, D. (2009). The components of paraphrase evaluations. *Behaviour Research Methods., 41*(3), 682–690.

Pennycook, A. (1996). Borrowing others' word: Text, ownership, memory, and plagiarism. *TESOL Quarterly, 30*(2), 201–230.

Plakans, L. (2007). *Second language writing and reading-to-write assessment tasks: A process study*. Unpublished Doctoral Dissertation, The University of Iowa, Iowa, City: IA, USA.

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing, 13*(2), 111–129.

Shi, L. (2004). Textual borrowing in second-language writing. *Written Communication, 21*, 171–200.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing, 9*, 27–55.

Criteria for Assessing EFL Writing at Majma'ah University

El-Sadig Yahya Ezza

Abstract Recent applications of quality assurance (QA) standards have required academic programmes in Saudi universities to base teaching on a variety of learning outcomes along with relevant teaching and assessment methods that are subject to regular reviews both internally and externally. In line with these educational developments, the present study attempted to find out if the repeated check-ups of what to assess in the students' writing have helped to standardize the assessment criteria for EFL writing courses at Majma'ah University (MU). A questionnaire and follow-up interviews were used to collect data to answer four research questions pertaining to the use of marking rubrics and whether or not their use is affected by three variables: Instructional experience of the faculty, academic levels at which writing courses are offered and the type of writing being assessed. The study findings indicated that participants used the ten marking rubrics included in the questionnaire. They also showed no statistically significant differences in the use of marking rubrics based on the study variables.

Keywords Assessment · Criteria · Writing · Rubrics · Validity · Reliability · Standardization

1 Introduction

The current decade has witnessed production of a huge body of research on EFL writing in Saudi Arabia. Most studies investigated aspects of writing such as proficiency, performance, errors and strategies (Al-Harthi, 2011; Aljafen, 2013; Al-Nufaie & Grenfell, 2012; Jahen & Idrees, 2013). Despite the vital role that assessment can play in the development of students' writing and instructional practices, extended web search for writing assessment research in the Saudi EFL context has resulted in a few studies that had addressed peripheral assessment issues

E.-S.Y. Ezza (✉)
Majma'ah University, Al Majma'ah, Saudi Arabia
e-mail: e.ezza@mu.edu.sa

S. Hidri and C. Coombe (eds.), *Evaluation in Foreign Language Education in the Middle East and North Africa*, Second Language Learning and Teaching,
DOI 10.1007/978-3-319-43234-2_11

such as peer feedback, peer-review and teacher's corrective feedback (Al-Shahrani, 2014; Grami, 2010; Jahen, 2012). The tendency to investigate these aspects of students' writing receives strong justification from curricular practices underlying writing instruction in many tertiary institutions. In other words, most institutions do not provide enough writing practice, thus, causing students to produce low quality prose that gives rise to the bulk of research indicated earlier. In fact, it is a general tendency in Arab tertiary institutions to undervalue the teaching of communication skills in that an examination of the skills courses in a number of Arab universities (in Colleges of Arts per se) showed they represented about 16 % of the English syllabus at King Saud University, 23 % at Cairo University, 15 % at Petra University (Jordan), 16 % at Beirut Arab University, 23 % at Damascus University, and 14 % at the University of Khartoum (Abdalla, 2007). Abdalla also observed that EFL faculty tended to classify themselves as linguists or literature specialists to support their specialty claims to teach linguistics and literature courses, leaving skills courses to be taught by junior faculty or teaching assistants. Implicit in such curricular practices is the fact that writing instruction is not a well-informed activity since it is assigned to EFL faculty who do not have appropriate instructional experience nor do they know the theories underlying writing instruction and assessment. So, they literally follow the textbooks prescribed for the teaching of writing courses and then impressionistically assess their students' performance.

However, recent applications of quality assurance (QA) standards have required academic programmes in Saudi universities to base teaching on a variety of learning-outcomes along with relevant teaching and assessment methods that are subject to regular reviews both internally and externally. In line with these educational developments, the present study sets out to decide if repeated check-ups of what to assess in the students' writing have helped to standardize the assessment criteria for EFL writing courses at Majma'ah University. Thus, there are four questions to answer in this connection:

a. What marking criteria do EFL faculty use to assess their students' writing?
b. Are there assessment differences attributable to the faculty's academic status?
c. Do the academic levels at which writing courses are offered require different types of assessment criteria?
d. Do different types of writing require different assessment criteria?

2 Theoretical Background

Since the early 1870s, academic circles in English-speaking countries, notably the United States, have been attaching heightened significance to writing instruction. Part of the reason for such an interest in writing was the need to address the literacy crisis that had afflicted the American college education (Bazerman et al., 2005; Connor, 1996). A century or so later writing became a subject of academic enquiry with higher status as evidenced by the fact that it was taught by full-time faculty instead of the previous practice when it had been taught by part-time instructors and teaching assistants. Equally important, a number of journals were devoted to research on writing, including "*College Composition and Communication*, *Written Communication*, *Rhetoric Review*, *Journal of Basic Writing*, *Journal of Teaching Writing*, and *Journal of Second Language Writing*" (Connor, 1996, p. 59). Also, writing has recently become an essential requirement for young Americans seeking secondary and post-secondary education to the extent that failure to master it prevents them from "completing high school, obtaining a post-secondary degree, acquiring a job that pays a living wage, and participating fully in community and civic life" (Graham, Harris, & Hebert, 2011, p. 10).

English-speaking academia has also paid special attention to the development of theories, models and criteria for the assessment of students' writing. The general practice indicates that assessment criteria are institutional in nature. That is, it is the academic institutions or educational authorities or both that set the assessment criteria to be implemented by writing instructors. Generally speaking, these official bodies differ as to the type and number of criteria to assess the students' writing. For instance, the Somerset Local Educational Authority (LEA) in UK bases the assessment of the students' writing on eight criteria: Originality, vocabulary, elaboration, organization, syntactic agreement, spelling, handwriting and layout (Wilkinson, 1983, p. 68). The Australian Curriculum, Assessment and Reporting Authority (ACARA) employs a system consisting of ten marking criteria: Audience, text, structure, ideas, persuasive devices, vocabulary, cohesion, paragraphing, sentence structure and spelling (ACARA, 2012, p. 6). In the US, the City University of New York uses a scoring guide consisting of five categories: Critical response to a writing task, development of ideas, structure of the response, sentence and word choice, and grammar, usage and mechanics. The last two categories subsume scoring items that are treated as independent marking criteria in the British and Australian assessment models.

Technically speaking, these criteria are known as scoring rubrics that could be used holistically or analytically. Where holistic scoring is concerned, it is conceived to take the whole text into account to assign an overall score to it (Becker, 2011, p. 116). Also, Brenland (1983) details that in holistic scoring raters score students' written performance for prominence of certain features important to that kind of writing or assign it a letter grade. By contrast, analytic scoring takes into account a number of text components such as accuracy, cohesion and content with each one being scored separately (Becker, 2011). Becker lists many advantages and disadvantages inherent in the use of each type of scoring. For instance, holistic rating emphasizes what writers do well and ensures greater score validity because of the authentic, personal reaction of the rater; yet, it does not give precise diagnostic information about the students' writing ability. Also, the scores mainly depend on the rater rather than on the characteristics of the text. As to the analytic scoring, it shows high score reliability and identifies writers' strengths and weaknesses; however, analytic rating of one scale may influence the rating of another. Moreover, analytic scoring is also time consuming.

Both holistic and analytic scoring systems are widely conceived as traditional methods of writing assessment. As such, they have for decades been criticized for a variety of reasons. For instance, Huot (1996) contends that traditional assessment is preoccupied with technical aspects of assessment such as reliability and validity. He details that agreement between independent raters to establish inter-rater reliability is not a crucial issue since both raters are not "equally good judges for all courses (p. 555)." He also reports that "more reliable measures like multiple choice are less valid ways to evaluate student writing;" as such, the argument goes, "the properties of a test which establish its reliability do not necessarily contribute to its validity (p. 558)." Thus, dissatisfied with traditional focus on technical aspects of writing assessment procedure, Huot proposed a new writing assessment whose "epistemological basis honours local standards, includes specific context for both the composing and reading of student writing, and allows for the communal interpretation of written communication (p. 561)." Also, unlike a traditional reliance on statistical validation and standardization of writing assessment, "more qualitative and ethnographic validation procedures like interviews, observation, etc." are employed to validate and standardize the proposed writing assessment procedures. Needless to say that the introduction of context is crucial in establishing a writing culture in the institution where writing is taught. Overall, in such a writing context, institutions can control and enhance instructional practices so that at least writing courses are entrusted to instructors who are cognizant of writing theory and practice. Also, Huot's proposal might particularly benefit Arab academia to reconsider its current trio of writing courses, i.e., "Writing 1," "Writing 2," and "Advanced/or Essay Writing" not only in terms of the number of writing courses and quality of instructional practices, but also in terms of the standardization and validation of assessment procedures.

Criticism of traditional assessment also comes from the Writing Across the Curriculum movement (WAC); WAC "challenged traditional assessment on general skills in undifferentiated testing situations" (Bazerman et al., 2005, p. 120). In other words, WAC theory maintained that writing practices take different forms in different disciplines; therefore, they could not be assessed using the same rubric-based procedures since good writing in, say, literature cannot be so conceived in physics, and vice versa. This reasoning receives support from research findings. For instance, Showeglar and Shamoon (1991), as cited in Bazerman et al. (2005, p. 123), report that sociology teachers rejected students' writing reflecting lines of reasoning pertaining to related fields such as anthropology and psychology despite their thematic and methodological relevance to sociology.

Contrary to Huot's argument, classroom practitioners do not necessarily assess students writing in pursuit of numerical consistency. Practice informs that students writing is usually rated by (individual) course instructors. What is more, writing instructors in a number of Arab universities reflected that instances of exam answers re-scoring by independent committees scarcely resulted in significant result differences. In fact, writing assessment is intended to serve a number of pedagogical purposes. For instance, as an integral component of teaching and learning, assessment is widely conceived to provide "up-to-date information or feedback about students' progress, allowing teachers and students or both to adjust what they are doing" (Graham, Harris & Hebert, 2011, p. 12). So, it is plausible to maintain that critical views of rubric-based assessment do not detract from the value of traditional assessment particularly in situations that base writing practice on the acquisition and application of rubrics to writing assignments. This is the exact case of English departments in Arab universities where an average textbook prescribed for a writing course emphasizes such writing components as cohesion, vocabulary, sentence structure and mechanics. Apparently, then, scoring rubrics "can be used to indicate how well a student has achieved mastery of aspects of L_2 writing" (Becker, 2011, p. 114).

3 Method

3.1 Context of the Study

The study was conducted at MU where five English departments were established in its five campuses to teach two types of EFL syllabus: Educational syllabus and literature-oriented syllabus. The Educational syllabus equips the students with literacies needed for their future teaching profession while the literature-oriented syllabus focuses on general purpose literacies to produce e.g., future translators, teaching assistants, school teachers, language researchers, creative writers, etc. However, both types of syllabus include the same number and types of writing courses: Writing I, Writing II, Advanced or Essay Writing.

3.2 Participants

The study participants were EFL faculty who were in the service of Majma'ah University. They teach at the five MU campuses: Community College, College of Education and Preparatory Year Deanship in Majma'ah City, College of Education in Zulfi City, Colleges of Science and Humanities in Ghat, Hotat Sudair and Rumah Cities. Questionnaire copies were both emailed and distributed manually to 80 EFL faculty but only 38 responses were received. Table 2 shows that the participants belong to four distinct academic ranks: Associate Professor(s) (1), Assistant Professors (10), Lecturers (23), and Teaching Assistants (4). In addition to their general tendency to use marking rubrics to assess students' writing, the study also aims to decide whether or not they differ in such a use based on their academic ranks, i.e., research question 2.

3.3 Instruments

Initially, a criterion-referenced marking guide developed by the Australian National Assessment Program was used to collect data for this study. It consisted of ten marking rubrics: Audience, text structure, ideas, persuasive devices, vocabulary, cohesion, paragraphing, sentence structure, punctuation, and spelling. Because the copyright allows that it could be used in whole or in part for non-commercial purposes, it was both emailed and distributed manually to the EFL faculty teaching at different MU campuses to state the frequency of their use of its marking criteria along a five-point Likert-type scale of frequencies, ranging from always to never (i.e., always = 5, often = 4, sometimes = 3, rarely = 2 and never = 1). Although the instrument was originally reliable and valid, the addition of new variables and frequency of use necessitated the re-calculation of its reliability. Cronbach's Alpha was used for this purpose, resulting in the co-efficient of 0.86 which indicates a high consistency. A follow-up interview was also used to elicit qualitative data that could not be detailed in the participants' responses to the questionnaire. It took about a month to distribute and collect the participants' responses to the questionnaire. In some campuses faculty required permission from their department heads to be able to answer the questionnaire.

3.4 Data Analysis

Three different types of statistics were used to analyse the study data. Data elicited to answer the first research question was analysed in percentage terms. Indeed, the calculation of frequency of use could best be analysed to compare and contrast the significance attached to different types of marking rubrics. Data pertaining to the second and third research questions was analysed using a Kruskal-Wallis test. The decision to apply this test rested on its rank-based, non-parametric nature that could be used to decide if there were significant differences between two or more groups of an independent test. A Mann-Whitney test was used to analyse the data collected to answer the fourth research question owing to its non-parametric nature, which is specific to the comparison of two groups within a variable.

4 Results and Discussion

Question 1: What marking criteria do EFL faculty use to assess their students' writing?

The frequency of use of each marking criterion has been calculated in percentage terms. Table 1 details the participants' responses.

Table 1 Frequency of use of marking rubrics

Nb	Criterion	Frequency				
		Always (%)	Often (%)	Sometimes (%)	Rarely (%)	Never (%)
1	Audience	21.1	26.3	36.8	10.5	5.3
2	Text structure	47.4	26.3	23.7	2.6	0.00
3	Ideas	55.3	21.1	18.4	5.3	0.00
4	Persuasive devices	21.1	34.2	31.6	13.2	0.00
5	Vocabulary	39.5	31.6	21.1	7.9	0.00
6	Cohesion	34.2	34.2	18.4	13.2	0.00
7	Paragraphing	26.3	28.9	28.9	15.8	0.00
8	Sentence structure	60.5	18.4	10.5	10.5	0.00
9	Punctuation	50.0	18.4	13.2	13.2	5.3
10	Spelling	55.3	18.4	13.2	13.2	0.00

Apart from "audience" and "punctuation"—both are "never" used at 5.3 %—all the other criteria were frequently used though with different percentages as indicated by the multiple "zero" occurrences of the "never" frequency. An equally important observation is that most participants would "always" focus on the "sentence structure" in assessing the students' writing.

Other top percentages also "always" occurred in the examination of "ideas," "spelling", "punctuation", and "text structure" in the students' writing. Further elaboration on these issues will be attempted in the result interpretation below.

Question 2: Are there assessment differences attributable to the faculty's academic status?

This question is intended to investigate the effect of participants' academic status on their use of marking criteria in assessing students' writing. Table 2 classifies the study participants based on their academic statuses while Table 3 presents a statistical analysis using the Kruskal-Wallis test to decide if these statuses could affect the assessment of the students' writing.

Table 2 also shows that the study participants represent all the academic ranks currently in the service of MU. Generally speaking, instructors of the two lower academic ranks, i.e., Lecturers and Teaching Assistants are expected to teach writing courses. However, faculty of the two higher ranks, i.e., Assistant Professors and Associate Professors, also assess essays written as answers to exam questions as detailed in Tables 4 and 6. Test statistics in Table 3 shows that the asymptotic significance is 0.614, which is greater than the test significance level, i.e., 0.05; thus, indicating no significant differences in participants' responses based on their academic ranks.

Question 3: Do the academic levels at which writing courses are offered require different types of assessment criteria?

Table 2 Participants' ranks

Ranks			
	Status	N	Mean rank
AVERESLIK	Associate professor	1	29.50
	Assistant professor	10	20.95
	Lecturer	23	18.04
	Teaching assistant	4	21.75
	Total	38	

Table 3 Test statistics based on participants' ranks

Test statistics[a,b]	
	AVERESLIK
Chi-square	1.805
d.f.	3
Asymp. sig.	0.614

Note a = Kruskal-Wallis test, b = Grouping variable: Faculty's Academic Status; Asymp. sig.—asymptotic significance; in this study asymp. sig. ≤0.05

Table 4 Academic levels of courses

Ranks			
	Level	N	Mean rank
AVERESLIK	Early	26	19.21
	Intermediate	12	20.13
	Total	38	

Table 5 Test statistics for the academic levels of courses

Test statistics[a,b]	
	AVERESLIK
Chi-square	0.065
d.f.	1
Asymp. sig.	0.799

Note a = Kruskal-Wallis Test, b = Grouping variable: Academic Levels of Courses; Asymp. sig.—asymptotic significance; in this study asymp. sig. ≤0.05

Data pertaining to this question is presented in Tables 4 and 5.

Table 4 indicates that only courses offered at early and intermediate years, i.e., 1–2 and 3–4 consecutively were included in this investigation. In Table 5 the Kruskal-Wallis Test was employed to analyze the participants' use of the marking rubrics based on the academic level at which the writing courses are offered. Test statistics show no significant difference in the participant's use of the marking rubrics despite the difference of the academic levels at which the students study these courses.

Question 4: Do different types of writing require different assessment criteria?

Table 6 Types of writing

Types of writing				
	Type	N	Mean rank	Sum of ranks
AVERESLIK	Writing courses	26	20.38	530.00
	Linguistics and literature	12	17.58	211.00
	Total	38		

Table 7 Test statistics for types of courses assessed

Test statistics^{a}	
	AVERESLIK
Mann–Whitney U	133.000
Wilcoxon W	211.000
Z	−0.782
Asymp. sig. (2-tailed)	0.434
Exact sig. [2 * (1-tailed Sig.)]	0.485^{b}

Note a = Mann-Whitney Test, b = Grouping variable: Types of Writing; Asymp. sig.—asymptotic significance; in this study asymp. sig. in either case, one- and two-tailed, ≤0.05

This question draws on the idea that the assessment of different types of writing require different criteria. For instance, a text written as an exam answer to a question in linguistics or literature might not be assessed in terms of criteria similar to those used to assess a paragraph in a writing course. Tables 6 and 7 summarize the data collected to answer this question.

Table 6 reveals that participants who teach writing courses are more than twice the number of their counterparts who teach linguistics and literature. In fact, they correspond to the number of participants who teach courses offered at early years of university education reported in Table 4.

A Mann-Whitney test was used to analyse the participants’ responses to the questionnaire based on the type of writing being assessed, i.e., paragraphs and essays in linguistics, literature, and writing courses. Test statistics in Table 5 reveal that asymptotic significance is greater than 0.05, indicating no difference in the participants’ employment of marking rubrics for all types of writing.

In an attempt to interpret the findings just reported, it is apparent that there are no significant differences in the participants’ employment of the marking rubrics. The participants, that is to say, use almost the same or similar rubrics to assess the students’ writing despite their different instructional experiences as well as the differences in academic levels at which courses are offered and the types of writing being assessed. In fact, the statistical analysis of data pertaining to the first research question, i.e., whether or not the participants use rubrics, was a prelude to the nature of findings generated by the subsequent research questions. In other words, as Table 1 indicates, most participants reported high, average and zero uses of marking rubrics included in the questionnaire. More specifically, despite the differences in their instructional experiences, participants employed the same marking criteria in assessing essays and paragraphs written in linguistics, literature, and writing courses in violation to the universal principle that features of good writing differ from discipline to discipline (Bazerman et al., 2005).

However, participants might wish to argue that different components of the English syllabus, e.g., linguistics and literature, do not constitute mutually exclusive disciplines that can be assessed using different marking rubrics. Such an argument seems to receive support from the emergence of "literary linguistics", i.e., an interdisciplinary activity "in which a linguistic approach is used to analyse fiction" (Azevedo, 2009, p. 2). This argument can be refuted on two grounds. First, it is generally conceived that

> Literary critics have railed against the cold scientific approach used by scholars of languages in their analyses of literary texts, whilst linguists have accused their literary colleagues of being too vague and subjective in the analyses they produced (McIntyre, 2012, p. 1).

Implicit in this observation is the fact that linguistics and literature are not only thematically different, but also cannot be appropriately handled in the same genre. Indeed, textbooks prescribed for teaching linguistics and literature courses focus on concepts that are entirely different. In other words, while an introductory textbook in linguistics might abound in such concepts as *language, linguistics, dialect, accent, phonetics, phoneme, and phonotactics*, a similar textbook in literature would centre upon concepts such as *character, climax, conflict, hyperbole, imagery, irony, metaphor, paradox and plot.* Second, the reference to the interdisciplinary activity connecting linguistics and literature is irrelevant since linguistics equally relates other branches of knowledge, e.g., geolinguistics, psycholinguistics, sociolinguistics, and forensic linguistics; it cannot, therefore, be assumed to qualify for similar assessment procedures as these disciplines owing to the interdisciplinary relationship between them.

The follow-up interview with some participants provided a plausible interpretation for the assessment dilemma emanating from the use of similar rubrics to assess students' written production in linguistics, literature and writing courses. It was agreed that students start studying these subjects in their second year of college education. At this stage, it was reflected that most students suffer from acute lexical deficiency and production of grammatically ill-formed sentences. So, rather than equipping students with skills needed for writing in linguistics and literature, the EFL faculty at Majma'ah University focus on teaching the basic sentence structure and writing mechanics, among others. There is evidence from previous research findings at Majma'ah University that partially supports this reasoning. A first year reading class failed to recognize the meaning of the word "small" but after it was written twice on the board as "small" and "SMALL" that most students smiled in relief (Ezza, 2013, p. 22). These same students were expected to enrol in linguistics and literature classes the following academic year. Apparently, these subjects were beyond the students' competence and, as such, participants were justified in focusing on writing components that the students could handle.

The follow-up interview also revealed that participants did not know much about theories of writing instruction and assessment. Writing theories were neither part of their previous training, nor did they read about them throughout their professional life. As a result, they were not aware of the theoretical principles underlying materials prescribed to teach writing courses and assess the students' performance accordingly. In such a situation, it was maintained that writing instructors "may be susceptible to basing their evaluations of NNS writing more on sentence-level concerns than on content or rhetorical concerns" (Sweedler-Brown, 1993; as cited in Weigle, 2007, p. 201). This was the exact case of the present study participants who prioritized the assessment of sentence structure as evidenced by the highest frequency of use in Table 1.

5 Conclusion

This study was intended to investigate EFL faculty's use of a marking guide consisting of ten rubrics to assess students' written performance at MU. The need for the study emanated from a general tendency to standardize the teaching and learning processes in Saudi academia. This policy is expected to benefit English language skills in general and writing skills in particular owing to their limited space in the EFL syllabus in many Arab universities. Overall, academic quality assurance stipulates that assessment procedures should be applied to all courses based on many learning outcomes that are determined prior to instruction and assessment. As such, it is expected to eliminate instructional practices that once put skills courses at a disadvantage in the manner reported in the first section of the study. In light of these facts, the present study set out to answer four questions pertaining to whether or not EFL faculty at Majma'ah University use marking criteria to assess the student's writing along with the effects of the three variables on such a use: Instructional experience, academic levels at which courses are offered and the types of writing being assessed.

The study findings indicated that the participants used the ten marking rubrics included in the questionnaire. They also showed no statistically significant difference in the use of marking rubrics judged by the study variables. These findings were argued to reflect a violation of a universal principle that writing standards differ from discipline to discipline, which, in turn, jeopardizes the efforts made to standardize the teaching and learning processes at MU. Emphatically, writing instruction and assessment did not receive due attention from EFL faculty at MU. Therefore, this paper recommends a number of procedures to enhance the teaching and assessment of writing given the enculturating role that writing plays in academia. First, it is highly recommended that the Deanship of Quality and Skills Development in collaboration with the English departments provide special training programmes to be implemented by local and international writing experts. Second, writing courses should be entrusted with more experienced English faculty; it is assumed that their advanced academic training and extended experience will enhance both instructional practices and assessment procedures. Third, there is currently an urgent institutional need to publicize a writing culture through the establishment of writing centres and special chairs on writing research.

To conclude, there are at least two limitations in the study that should be addressed in future research. First, the marking rubrics investigated in this study pertain for the most part to the skills that could be encountered in writing courses. While it is true that they could be considered in the assessment of literary and linguistic essays, they do not by themselves define good writing in literature and linguistics. Therefore, this paper recommends the inclusion of writing skills that are more specific to these disciplines, e.g., argument structure, demonstration of knowledge, etc. in future research.

Second, the instruments used in this study do not provide information about whether the EFL faculty employed the marking rubrics holistically or analytically. Thus, further research should examine marking samples to decide on faculty's preference for holistic and analytic use of marking rubrics.

Appendix 1

Dear EFL/TESOL Faculty at Majma'ah University;

The following questionnaire is intended to collect data to answer a number of research questions regarding the use of marking criteria in assessing students' writing at Majma'ah University. You are kindly requested to spare some of your valuable time to answer it.

Regards;
El-Sadig Yahya Ezza
Associate Professor, Community College

Criteria for Assessing EFL Writing at Majma'ah University

Questionnaire

A. *Demographic Information*

1. Academic Status:

   i. Professor: [ ]
   ii. Associate Professor: [ ]
   iii. Assistant Professor: [ ]
   iv. Lecturer: [ ]
   v. Teaching Assistant: [ ]

2. Type of Writing Assessed:

   i. Essays and paragraphs in writing courses: [ ]
   ii. Essays and paragraphs in linguistics and literature: [ ]

3. Levels at which courses are offered:

   i. Early levels (1–2): [ ]
   ii. Intermediate levels (3–6): [ ]
   iii. Advanced Levels (7–8): [ ]

B. Tick the frequency of marking criteria you use in assessing students' writing; where:
5 = always, 4 = often, 3 = sometimes, 2 = rarely, 1 = never

	Marking criteria	Description	Frequency of use				
1	Audience	The writer's capacity to orient, engage and persuade the reader	5	4	3	2	1
2	Text structure	The organization of the structural components of a persuasive text i.e., introduction, body and conclusion					
3	Ideas	The selection, relevance and elaboration of ideas for an argument					
4	Persuasive devices	The use of a range of persuasive devices to enhance the writer's position and persuade the reader					
5	Vocabulary	The range and precision of contextually appropriate language choices					
6	Cohesion	The control of multiple threads and relationships across the text, achieved through the use of referring words, ellipsis, text connectives, substitutions and word associations					
7	Paragraphing	The segmenting of text into paragraphs that assists the reader to follow the line of argument					
8	Sentence structure	The production of grammatically correct, structurally sound and meaningful sentences					
9	Punctuation	The use of correct and appropriate punctuation to aid the reading of the text					
10	Spelling	The accuracy of spelling and the difficulty of the words used					

Appendix 2

Dear EFL/TESOL Faculty at Majma'ah University;

The following questionnaire is intended to collect data to answer a number of research questions regarding the use of marking criteria in assessing students' writing at Majma'ah University. You are kindly requested to spare some of your valuable time to answer it.

A. *Demographic Information*

1. Academic Status:

   i. Professor: [ ]
   ii. Associate Professor: [ ]
   iii. Assistant Professor: [ ]
   iv. Lecturer: [ ]
   v. Teaching Assistant: [ ]

2. Type of Writing Assessed:

   i. Essays and paragraphs in writing courses: [ ]
   ii. Essays and paragraphs in linguistics and literature: [ ]

3. Levels at which courses are offered:

   i. Early levels (1–2): [ ]
   ii. Intermediate levels (3–6): [ ]
   iii. Advanced Levels (7–8): [ ]

B. Tick the frequency of marking criteria you use in assessing students' writing; where:
5 = always, 4 = often, 3 = sometimes, 2 = rarely, 1 = never

Criteria for Assessing EFL Writing at Majma'ah University

	Marking criteria	Description	Frequency of use				
1	Audience	The writer's capacity to orient, engage and persuade the reader	5	4	3	2	1
2	Text structure	The organization of the structural components of a persuasive text i.e., introduction, body and conclusion					
3	Ideas	The selection, relevance and elaboration of ideas for an argument					
4	Persuasive devices	The use of a range of persuasive devices to enhance the writer's position and persuade the reader					
5	Vocabulary	The range and precision of contextually appropriate language choices					

(continued)

(continued)

	Marking criteria	Description	Frequency of use				
6	Cohesion	The control of multiple threads and relationships across the text, achieved through the use of referring words, ellipsis, text connectives, substitutions and word associations					
7	Paragraphing	The segmenting of text into paragraphs that assists the reader to follow the line of argument					
8	Sentence structure	The production of grammatically correct, structurally sound and meaningful sentences					
9	Punctuation	The use of correct and appropriate punctuation to aid the reading of the text					
10	Spelling	The accuracy of spelling and the difficulty of the words used					

References

Abdalla, A. Y. (2007). The significance of incorporating language skills into EFL syllabus. *Khartoum University Journal of ADAB, 25*, 17–29.

Al-Harthi, K. (2011). *The impact of writing strategies on the written product of EFL Saudi male students at King Abdul-Aziz University*. A thesis submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Applied Linguistics. New Castle University, UK.

Aljafen, B. S. (2013). *Writing anxiety among EFL Saudi students in science colleges and departments at a Saudi university*. A thesis submitted to the School of Graduate Studies and Research in partial fulfillment of the requirements for the degree Master of Arts. Indiana University of Pennsylvania, USA.

Al-Nufaie, M., & Grenfell, M. (2012). EFL students' writing strategies in Saudi Arabian ESP writing classes: Perspectives on learning strategies in self-access language learning. *Studies in Self-Access Learning Journal, 3*(4), 407–422.

Al-Shahrani, A. (2014). Investigating teachers' written corrective feedback practices in a Saudi EFL context: How do they align with their beliefs, institutional guidelines, and students' preferences? *Australian Journal of Teacher Education, 37*(2), 101–122.

Australian Curriculum, Assessment and Reporting Authority. (2012). *Persuasive writing marking guide*. Sydney: ACARA.

Azevedo, M. M. (2009). Literary linguistics in the context of a literature department. In J. Collentine, M. Garcia, B. Lafford, & F. M. Marín (Eds.), *Selected proceedings of the 11th Hispanic linguistics symposium* (pp. 1–8). Somerville: Cascadilla Proceedings Project.

Bazerman, C., Little, J., Bethel, L., Chavkin, T., Fouquette, D., & Garufis, J. (2005). *Reference guide to writing across the curriculum*. West Lafayette, Indiana: Parlor Press and the WAC Clearinghouse.

Becker, A. (2011). Examining rubrics used to measure writing performance in U.S. intensive English programs. *The CATESOL Journal*, *22*(1), 113–130.

Brenland, H. M. (1983). *The direct assessment of writing skills: A measurement review*. College Board Report No. 83-6. New York: College Entrance Examination Board.

Connor, U. (1996). *Contrastive rhetoric: Cross-cultural aspects of second language writing*. Cambridge: Cambridge University Press.

Ezza, E. Y. (2013). Intervention strategies in a Saudi EFL classroom. *Journal of Arts and Humanities, 2*(2), 17–24.

Graham, S., Harris, K. & Hebert, M. (2011). *The benefits of formative assessment: A report from Carnegie corporation*. New York: Alliance for Excellent Education.

Grami, G. M. A. (2010). *The effects of integrating peer feedback into university-level ESL writing curriculum: A comparative study in a Saudi context.* Unpublished doctoral dissertation, Newcastle University. Accessed March 15, 2015 from http://www.kau.edu.sa/files/0005407/researches/57369_27610.pdf

Huot, B. (1996). Toward a new theory of writing assessment. *College Composition and Communication, 47*(4), 549–566.

Jahen, J. H. (2012). The effect of peer reviewing on writing apprehension and essay writing ability of prospective EFL teachers. *Australian Journal of Teacher Education, 37*(4), 60–84.

Jahen, J. H., & Idrees, M. W. (2013). EFL major student teachers' writing proficiency and attitudes towards learning English. *Umm Al-Qura University Journal of Educational and Psychological Sciences, 4*(1), 10–72.

McIntyre, M. (2012). Linguistics and literature: Stylistics as a tool for the literary critic. *SRC Working Papers*, *1*, 1–11.

Weigle, S. C. (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing, 16*, 194–209.

Wilkinson, A. (1983). Assessing language development: The Crediton project. In A. Freedman, I. Pringle, & J. Yalden (Eds.), *Learning to write: First language/second language* (pp. 67–86). New York: Longman.

Part V
Textbook and ICT Evaluation

An Evaluation of the Place of Culture in English Education in Tunisia

Tarek Hermessi

Abstract This study, situated within cultural content evaluation and analysis research, uses the content analysis technique to investigate the cultural dimension of English education in a Muslim EFL setting, namely Tunisia. It specifically scrutinizes the place of culture in official curricular documents and locally-produced teaching materials. The materials analyzed and evaluated, in this study, are two official curricular documents as well as eight textbooks and seven teacher guides used in Tunisian basic and high schools. The study revealed that, unlike in many other Muslim EFL settings, Tunisian curriculum designers and textbook writers do not have any a priori ideological objection to the inclusion of culture in the English program. However, they seem to approach the cultural dimension of L2 education in a nonsystematic, unprincipled way. The study informed also that there exists a "latent cultural curriculum" that is subservient to the presentation of the structural and functional aspects of language. Such latent cultural content is hypothesized to only lead to a basic form of factual cultural knowledge as students are never engaged in intra-cultural or intercultural dialogic, interpretive experiences. The implication of this study is that although culture is present in L2 education by choice or by coercion, the treatment of cultural content remains a function of two variables, namely (1) the philosophical and ideological foundations of language curricula and (2) the distance between the culture(s) representing the source language and that/those representing the target one.

Keywords Evaluation · EFL context · Teaching materials · Curriculum · Document survey · Content analysis · Intercultural communicative competence · Cultural distance

T. Hermessi (✉)
Institut Supérieur des Langues de Tunis, Tunis, Tunisia
e-mail: hermestic@yahoo.com

S. Hidri and C. Coombe (eds.), *Evaluation in Foreign Language Education in the Middle East and North Africa*, Second Language Learning and Teaching,
DOI 10.1007/978-3-319-43234-2_12

1 Introduction

The emergence of the concept of intercultural communication and the subsequent elaboration of the concept of intercultural communicative competence (henceforth, ICC) by Michael Byram and Claire Kramsch has resulted in the so-called "cultural turn" of foreign language teaching. During the 1990s, research on the place of culture in L2 education has flourished and cultural content evaluation and analysis studies varied both theoretically and methodologically. Theoretically, these studies adopted either the modernist, structuralist definition of culture as a static, homogeneous, nation-based concept or the post-modernist conceptualization of culture as a heterogeneous, dynamic, non-essentialist notion (see Kramsch, 2009). Methodologically speaking, cultural content evaluation and analysis studies embraced three approaches, namely content analysis, critical discourse analysis and semiotic analysis (see Weninger & Kiss, 2014). Weninger and Kiss (2014), after surveying several studies on the cultural content of L2 textbooks, contended that most of these studies lack principled criteria for analysis and evaluation and a clear theoretical framework. Besides, researchers remain vague about the way "textual and visual data (i.e., content) is unitized and coded (p. 6)."

The place given to culture in L2 education, mainly English education, remains an ideological and socio-political, rather than a purely linguistic issue. It is a function of the philosophy adopted in education in general and foreign language education in particular. In fact, foreign language teaching/learning, though accepted to be indispensable to integrate global economy and benefit from technological progress, is perceived by many to represent a potential threat to local cultures (See Adaskou, Britten, & Fahsi, 1990; Phillipson, 1992; Pennycook, 1994; Canagarajah, 1999, among others). A plethora of culture evaluation studies have been carried out in contexts where the cultural dimension of L2 education is both accepted and approached systematically (See Risager, 2011; Sercu et al., 2005, for instance). A few studies, however, evaluated the place of culture in curricular settings where culture is not deemed to be an essential component of L2 teaching. Scrutinizing such settings is assumed to be central to complete the picture of the cultural dimension of L2 education.

The consideration of culture in foreign language curricula, in many L2 contexts, is neither taken-for granted nor systematic. Among these contexts is the Islamic FL context, which pedagogical assumptions and practices might not necessarily be in line with the prevalent views on the link between L2 education and culture. As a matter of fact, curriculum developers, in this context, approach FL as a set of structural codes that can be taught/learnt with either minimum or no reference to culture (see Cortizzi & Jin, 1999). They, actually, strive to either nullify the cultural load of foreign languages or keep to a minimum. In this study, it is assumed that the relationship between culture and L2 education is elusive as it might manifest itself in the form of incidental, latent curriculum. Such curriculum is believed to transpire from the cultural potential of language (materials) and the impossibility of disentangling language from culture regardless of whether ICC is explicitly considered as an educational objective or not. In this vein, the study specifically evaluates how culture is approached in the English curriculum in Tunisia by examining official curricular documents and teaching materials (teacher guides and student textbooks).

2 Theoretical Background and Study Framework

Theorizing and problematizing the place of culture in language learning has been paradigmatically grounded within anthropology, cultural studies, and linguistics. Traditionally, culture has been approached from the structuralist, nation/state-based, essentialist, and normative perspective. In applied linguistics, such approach has been challenged by an “alternative,” “nonstandard”, post-modernist approach that is grounded within the Neo-Vygotskyan socio-cultural theory (see Kramsch, 1993, Atkinson, 1999, Holliday, 1999, among others). Byram (2009) identified five types of “savoirs” that make up ICC, namely (1) “savoir être” or intercultural attitudes (2) “savoirs” or knowledge, (3) “savoir comprendre” or skills, (4) “savoir apprendre/faire” or skills of discovery and interaction, (5) savoir s’engager or critical cultural awareness (see Table 1).

Culture, in L2 curricula, has been approached from three different perspectives: the mono-cultural approach, the intercultural approach and the transcultural approach. The mono-cultural approach is essentially factual knowledge-oriented. Galloway (1985, as cited in Omaggio-Hadley, 2001, p. 348) identified four variants

Table 1 The components of intercultural competence and their definition

Types of savoirs	Definition
Savoir être or attitudes-curiosity/openness	Curiosity and openness, readiness to suspend disbelief about other cultures and belief about one’s own and ability to decenter and relativize one’s own values, beliefs and behaviors
Savoirs or knowledge	knowledge of social groups and their products and practices in one’s own and the target language culture along with general processes of societal and individual interaction
Savoir comprendre or skills of interpreting/relating	Ability to put ideas, events and documents from two or more cultures side by side and relate, compare and interpret them in a way that minimizes misunderstanding what people say, write or do
Savoir apprendre/faire or skills of discovery and interaction	Ability to optimally use intercultural knowledge and skills in everyday life cultural encounters
Savoirs’ engager or critical cultural awareness	Ability to discern how one’s culture can lead to the rejection of the perspectives, practices and behaviors of the other on the one hand and to critically evaluate the other on explicit criteria on the other. It has connotations with political engagement related to “education for intercultural citizenship”

Based on Byram (2009, pp. 337–340)

of this approach, namely "(…) the Frankenstein approach (a taco from here, a flamenco dancer from there), a Gaucho from here, a bullfight from there), the 4-F approach (Folk dance, festivals, fairs and food), the Tour-Guide approach (Monuments, rivers, cities, etc.), and the By-The-Way approach (sporadic lectures or bits of behavior selected indiscriminately to emphasize sharp differences)". The mono-cultural approach can only result in critical, interpretive cultural awareness if learners are brought to discern the link between the products, practices, and perspectives within source or target culture. The intercultural or comparative approach is, by definition, critical and interpretive as it presents aspects of both source culture and target culture and invites learners to mediate between the two, giving ICC all its sense. In contrast, the transcultural approach considers the aspects of global middle class and youth culture that are basically the outcome of technological, economic and media globalization. It critically and interpretively relates these aspects with both source culture and target culture in the aim of nurturing a sense of being a 'citizen of the world' in L2 learners (see Byram, 2014; Kramsch, 2010). The notion of global culture; in the present study, was defined in terms of the emergence of a global culture or a 'culture franca' that is not tied to any L2, youth culture as a case in point. Such global culture has resulted in the advent of a "global consciousness" or "awareness of the world as a single space" (see Robertson, 1992).

Several evaluation studies have been conducted in Islamic EFL settings wherein the cultural dimension of L2 curricula, in general, and English curricula, in particular, are approached in terms of compatibility with Islamic values and local cultures. Al-Issa (2005), for instance, critically investigated the ideology and philosophy of English language teaching, in Oman, by surveying official curricula documents and teaching materials. He found a conflict between the National English Language Plan/Policy, which underscores the importance of considering culture in L2 teaching, and the content of teaching materials. In Iran, AliAkbari (2004, p. 5) evaluated the "state of culture" in four Iranian high school ELT textbooks and concluded that the content of these textbooks can in no way develop intercultural competence in Iranian learners of English. In Saudi Arabia, Cortizzi and Jinn (1999) noted that in the textbook "English for Saudi Arabia" (Al-Qureishi, Watson, Hafseth & Hickman, 1988)

> (…) virtually every setting is located in the source culture. When the textbook characters greet one another, talk about professions, make Arabian coffee, or talk about going on pilgrimage to Mecca, they are predominantly Saudi Arabians performing culturally familiar activities in their own country with their own citizens (in English) (p. 205).

In a study conducted in a context that is very close (both geographically and socio-culturally) to the context of the present study (Tunisia), Adaskou, Britten, and Fahsi (1990) investigated design decisions on the cultural content of a secondary school English course in Morocco. They, surprisingly, recommended that such decisions not be based on theories of intercultural communicative competence, but rather on "(…) prevailing attitudes towards foreign culture among teachers of English (p. 3)." The aforementioned studies revealed that, in Omani, Iranian, Saudi Arabian and Moroccan Islamic EFL settings, L2 curriculum designers seem exhibit a priori ideological objections to the inclusion of foreign cultural content in language programs. They seem also to ambivalently treat cultural content as they proclaim the importance of culture in curricular documents while de-emphasizing it in textbooks and teacher guides as well as in actual teaching.

This study evaluated the place of culture in English curriculum in Tunisia, a North African, Islamic country with complex socio-cultural and linguistic characteristics. It found its rationale in the necessity to investigate the place of culture in L2 education in contexts in which source culture is distant from target culture. In such contexts, curriculum designers are hypothesized to be less prone to include a cultural component in language teaching and consequently either overlook culture or approached in a non-systematic way. The study scrutinized official curricular documents and teaching materials (student textbooks and teacher guides) to highlight the place of culture in the English program, in Tunisia. It specifically addressed the following questions:

(1) Do Tunisian policy-makers have a priori ideological objections to the inclusion of culture in English education?
(2) Is ICC an explicit educational objective in the Tunisian curriculum of English?
(3) Is there a "latent" cultural content in English teaching materials?
(4) What type of intercultural competence can such "latent" cultural content develop?

3 The Study

3.1 Context of the Study

Socio-culturally speaking, Tunisia has represented throughout its three millennia history a melting pot for different peoples and cultures. The original culture of Tunisia is the Berber one; such culture has been influenced by all the peoples and cultures that marked the history of the Mediterranean region, namely the Carthaginian, Phoenician culture, the Roman culture, the Greek culture, the Arab-Islamic culture, the Moorish culture, the Turkish culture, the Italian culture and the French culture. Two factors have shaped the current socio-cultural landscape in Tunisia. The first is the impact of Habib Bourguiba, the first president of the country, who, influenced by Turkey's Kamel Ataturk, managed to found a modern nation by "betting on" women's liberation and education. The second is the geographical and economic ties of Tunisia with Europe, in general, and the Mediterranean European countries, in particular.

The native language of Tunisia is Tunisian Arabic (TA), which can be considered as a language rather than a dialect or a variety of "Modern Standard Arabic" (MSA) or of "Classical Arabic" (CA). MSA and CA are both, technically speaking, dead languages, i.e., languages without native speakers. From a psycholinguistic point of view, MSA can be considered as the first foreign language of Tunisia and French as its second foreign language. Tunisia, a former French colony, has always given the French language an important place in its educational system. French has served as a foreign language as well as the medium of instruction for the major part of school subjects, mainly scientific, technical and vocational ones since the early years of primary education (from year 6 to 12) and secondary education (from 13 to 19). At university most scientific and technical specialties are delivered in the French language.

In the late 1960s, Tunisia witnessed a campaign of Arabicization which relegated French to the status of a foreign language and upgraded Modern Standard Arabic to the status of medium of education mainly in basic and secondary education. In the early 1970s, English was introduced in the Tunisian educational system as a school subject starting from the age of 16. The educational reform of 1990 reorganized schooling into 9 years of basic education (from 6 to 15) and 4 years of secondary education (from 15 to 19). It also reinforced the place of English as a foreign language in the school system by lowering the age at which students start learning it to 12. In the year 2009, the Tunisian Ministry of education and training in collaboration with the British Council launched an ELT reform project in the aim of bringing English education in Tunisia ‘up to the Common European Framework standards’ (De Lotbinière, 2009). In 2014, the Tunisian government decided to lower the age at which French is learnt at school to the age of 8 and English to the age of 9.

The English language program in Tunisia is regulated by the 2002 framework law of education and supervised by the “Department of Pedagogy and Norms of Basic Education and Secondary Education” in the Ministry of Education and Training. This official board collaborates with basic school and high school ELT inspectors and teachers to define the guidelines for the English curriculum at the levels of objectives, syllabus, teaching materials, and assessment. Such curriculum is spelt out in two regulation documents issued by the “Department of Programs and Norms” of the Ministry of Education and Training. These two documents are called “official programs” and entitled respectively “English Programs for Basic Education” and “English Programs for English Education in Secondary Education years 1, 2, 3, 4”. The teaching materials are locally produced and published by a government body, the “Centre National Pédagogique” in conformity with the specifications of the so-called “Official Programs.” The ELT materials used in Tunisian schools include a student book, a workbook and a teacher guide for each grade.

3.2 Materials and Procedure

This study relied on the document survey technique and the content analysis technique to survey the curricular documents and teaching materials. The curricula documents were the 2002 framework law of Foreign Languages, the English

Programs for Basic Education (2006) and the English Programs for English for Secondary Education (2008). The examination of official language curricular documents was intended to gain insight into the philosophy of the English curriculum in Tunisia in relation to the place of culture in language teaching. The teaching materials were the textbooks and teacher guides used in grades 6–9 of basic school and grades 1–4 in high school. The textbooks and teacher guides were numbered from 1 to 8 (TB1 through TB8 and TG1 through TG8), except for grade 8 of basic school which does not have a teacher guide.

As regards the procedure followed in this study, the researcher firstly appraised official curricular documents regulating English education in Tunisia to see how culture is approached. Secondly, Teacher Guides have been surveyed, in their turn, from cover to cover in search for any reference to the cultural dimension of English language teaching. Thirdly textbooks have been in their turn examined from cover to cover at three different levels. The first level of analysis was concerned with form, layout and organization (whether texts are presented in series or not, the types of illustrations and texts used), as well as segments of society represented. The English textbooks and teacher guides, used in the eighth grades of basic school and high school education, have been scrutinized to identify any instances of cultural content. Such scrutiny was based on Joiner's (1974) checklist for the evaluation of the cultural content of language textbooks and Sheldon's (1988) checklist for the evaluation of the overall content of language textbooks.

The second level of analysis consisted in a content analysis of the amount and the type of cultural references. Cultural references have been classified according to whether they are textual or non-textual and direct (referring to a particular aspect of culture) or indirect (referring to an English-speaking country as the birth place of a person in a biography for instance). They have also been classified as touristic (if they refer to the touristic highlights of an English speaking country), factual (if they refer to a geographical, social or historical feature pertaining to an English speaking country) or analytical/critical (if the reference engages students in comparative, interpretive discussion of culture or highlights the negative, controversial aspects of culture). The third level of analysis consisted in closely surveying textbooks to classify cultural references according to whether they pertain to target culture, source culture or global/international culture. This level of analysis was based on the "National standards in Foreign Language Education Project" (1996) classification of culture into products, behavioral practices and 'philosophical' perspectives, as well as Moran's (2001) notion of people. The same framework has been used by Yuen (2011) to investigate the representation of culture in English textbooks in Hong Kong. The cultural references in English textbooks have accordingly been classified into four components: products, perspectives, people and icons and practices/lifestyle in the Tunisian/Arab culture, The British/US culture, and global culture. The choice of the British and US cultures was motivated by the results of the second part of the document survey which revealed their predominance in English textbooks (see Table 2 on p. 21).

Table 2 Cultural reference by english speaking countries

Textbook	Great Britain	USA	Australia	Canada	New Zealand	South Africa
TB1	2		1			
TB2	3	1	1			
TB3	17					
TB4	6					
TB5	5	1				
TB6	13	1				
TB7	1	20				
TB8	22	6	1	3	1	1
Total	69	29	3	3	1	1

4 Results and Discussion

4.1 Culture in Curricular Documents

The examination of official curricular documents revealed that language policy-makers in Tunisia do not to have any ideological a priori objection to the inclusion of culture in teaching English as a foreign language. Curriculum designers seem also aware of the educational breadth of foreign languages, which represent an opportunity for cultural encounter. In fact, the "2002 education framework law" stipulates that students must be proficient in "Arabic, the national language" and in two foreign languages at least (Article 1). Article 51, of the same law, stipulates that foreign languages represent a "means of communication" and an "instrument allowing for direct access to the products of universal thought along with the techniques, scientific theories and civilizational values it conveys". In addition, foreign languages prepare youth to be up-to-date with developments in the aforementioned areas and contribute to this development "(…) in a way that enriches national culture and ensures interaction with universal culture" [Translation mine and bold added] (Leclerc, 2012).

In English Programs for Basic Education (2006) and English Programs for Secondary Education (2008), the expressions "target culture" and "Anglophone context" have been used without any further elaboration or comments about what they refer to or mean in relation to the culture or the cultures representative of the English language. The use of "Anglophone culture" can be taken as an indication that, English curriculum designers in Tunisia, assume all English-speaking countries to represent the English culture. In addition, official curricular documents seem to focus on the linguistic structural aspects as well as the socio-pragmatic and socio-semantic dimensions of communicative competence [conventions of use, level of formality, appropriateness (…)] more than ICC as defined in the literature review section. Actually, "culture" has been explicitly mentioned twice in the two documents mentioned above under the heading of "the status of English as a subject matter and its contribution to the achievement of cross-curricular learning goals" in the following words:

> As a means of communication English will foster learner self-expression as well as appropriate interaction with peers and other interlocutors, which in turn, will ensure *access to universal culture through Anglophone context* [Italics added].
>
> The learner needs to understand how the language system works and how language conventions can vary according to purpose, audience, context and culture and apply this knowledge in speech and writing in both formal and informal situations. (English programmes for Basic Education, 2006, p. 4 and English Programmes for Secondary Education, 2008, pp. 4–5)

Culture is also mentioned in broad terms under a section labelled "Aims of Reading/Listening skills" in English Programmes for Secondary Education (2008):

> Expand one's knowledge of the world
>
> Develop awareness of aspects of the target culture
>
> Compare one's culture to that conveyed in texts
>
> Develop appreciation of self, environment and culture (pp. 11, 25, 45, 51, 63)

None of these goals has been further elaborated in the "official programs". The use of such vague, generic expressions as "aspects of target culture", "appreciation of self, environment and culture" shows the lack of systematicity and absence of clear vision regarding the place of culture in teaching English as a foreign language in Tunisia.

In the Tunisian context, the philosophy of English education is not clearly and consistently formulated given that there is no unified framework for teaching foreign languages. Although the 2012 framework law of education specifies the place and rationale of foreign languages in the Tunisian educational system, official documents regulating English language education in Basic and High schools do neither explicitly nor extensively spell out actual or estimated learner needs or learning goals and objectives. In addition, they do neither specify descriptors of proficiency/competencies nor standards of evaluation and testing. As for the place of culture in English education, the 2012 framework is more specific than the "official programs". As a matter of fact, it stipulates that foreign language education should allow students to have access to "the products of universal thought", "interact with universal culture" and use such access to "enrich national culture". Tunisian language policy-makers seem, therefore to adopt a non-systematic approach to culture. Although they do not have a priori objections to the inclusion of culture in English curriculum, they treat culture in a nonsystematic, inconsistent, sometimes frivolous way.

The statement of philosophy, premised on ideological, educational, and socio-economic assumptions, represents the cornerstone of education, in general, and language curriculum, in particular. Such statement determines decisions pertaining to teaching method, syllabus, objectives, and teaching materials. The philosophy of a given curriculum states the skills, competencies, beliefs and attitudes to be nurtured in the student and allows for pinpointing the profile of the desired product of any educational program. The statement of philosophy and ideology of English education in Tunisia is almost absent in official curricular documents.

4.2 Culture in Teacher Guides

In the prefaces of TGs, culture has been explicitly mentioned 6 times. The most explicit reference to culture is found in TG2:

> (…) the book is similar to a story (…) The story is about a cultural exchange whereby a British teenager comes to stay with a Tunisian family as a token of *intercultural learning*, which facilitates *access to universal culture* (official programmes) (…) the learner both discovers the main characters and learns language pertaining to *everyday life situations* (official programmes) as well as *the moral attitudes* they entail (…) The aim of this book is therefore two-fold: to teach and consolidate language and language skills and also to breed a new generation of *tolerant* responsible and autonomous youth. (p. 5–6) [Emphasis added]

The scrutiny of this excerpt reveals that the primary goal of English teaching is to master the structural aspects of language along with the "language" used in "everyday life situations". However, there is another goal tacked to the first, which is "the moral attitudes" the language of "everyday life situations" entails. These so-called "moral attitudes" are neither explained nor translated into learning or teaching goals. In the end of the passage, there is also a clear reference to "intercultural learning" and to one of the components of intercultural competence as defined by Byram (2009), namely tolerance of "otherness". Such a reference seems, nevertheless, to be more of a hollow "slogan" than a principled educational objective.

TG6, in a "lesson" about "Travel is fun and Broadens the mind", refers to "develop awareness of aspects of target culture" as a learning objective without any further details or instructions (p. 21). In TG8, in lesson 5, which deals with "Comparing Educational Systems", the writers state that "finding out about other people and countries is fundamental to expand one's knowledge of the world and evaluate one's own context, if time allows [sic] explore in depth some of the details mentioned in various [educational] systems" (p, 25). Though, comparing educational systems can only lead to a factual form of intercultural competence, it remains a secondary goal, for Tunisian curriculum designers, to be pursued only "if time allows". In a similar vein, TG8 refers to "Cultural notes" under the section of "what to insert in portfolios" without any explanation about the rationale and purpose of using such notes. The only explicit instance of a comparative approach to culture is found in TG4. Under the unit of "Sharing Family Responsibility", we read the following note to teachers: "the idea is to get your pupils to understand that Mark is a modern father and that he is sharing family responsibilities with his wife. This will provide the pupils with a good start to compare between this family and their own (p. 11)" [Emphasis added]. It would be very interesting to see how teachers proceed with this goal.

4.3 (Latent) Culture in English Textbooks

Although the cultural dimension of language teaching is neither explicitly nor systematically approached in English curriculum in Tunisia, there exists a latent substantial cultural content presented essentially under reading and listening materials and activities. The latent cultural content of English textbooks was analyzed at the levels of form and organization, amount, type and nature. At the levels of form and organization, the document survey revealed that English textbooks are not presented in a series; rather each textbook is written by a team of inspectors and teachers. The illustrations are mainly drawings in TB1/TB2/TB3 and pictures in TB4 through TB8, with a few instances of cartoons and comic strips. The texts are dialogues, letters, and newspaper articles in all textbooks with occasional use of jokes and anecdotes in TB3/TB7 and songs and poems in TB1/TB2/TB3. Although textbook writers claim to use authentic materials, it is difficult to verify such a claim given that the sources of the materials used for teaching are either not available or cited in nonconventional way. In addition, no specific information is provided about the extent to which the texts have been contrived. There are, nevertheless, a few references to online sources (www.eslmonkey.com, for instance), 'unknown' newspapers and magazines, or to texts taken from Non-Governmental Organizations or financial institutions documents [UN, UNICEF, WB, IMF (…)].

The predominant setting is the city; the characters are mostly young, the social category represented is middle class. Woman is positively portrayed; the characters portrayed in textbooks have both Tunisian and Western names; however they are all dressed in the western fashion and there are no local traditional Tunisian dresses. In addition, although the Islamic code of dressing is not compulsory in Tunisia, still many women are veiled and this code of dressing does not appear in any of the 8 textbooks. It is difficult to know whether the non-representation of veiled women is a conscious (political or ideological) decision or a "naïf" omission.

At the level of amount, all instances of cultural references were counted and classified according to the English-speaking countries they refer to (see Table 2). The English-speaking countries represented in Textbooks are Britain and USA essentially with a few references to Canada, New Zealand, Australia and South Africa. Tunisian textbook writers seem, however, to essentialize the English culture to the British culture and to a lesser extent to the American one. As Table 2 informs, the culture of Great Britain has been referred to 69 times while the American culture 29.

The latent cultural content has been also classified according to whether it refers to British/US culture, Tunisian/Arab culture or global/international culture. Table 3 presents the cultural content of English textbooks pertaining to the British and Americans. It can be noted that although such content does neither reinforce stereotypes nor enshrine negative views of USA and Britain, it does not represent uncomfortable realities [unemployment, poverty, family breakdowns, racism (…)], in the two countries either. The people and icons represented in textbooks include inventors and scientists as well as musicians, writers (Shakespeare, for instance)

Table 3 British/US culture in Tunisian EFL textbooks

Perspectives	Easter, friendship, family responsibilities between men and woman, generation gap and parents/teens relationships, teenagers and money, violence at school, love, neighbors
Practices and lifestyle	School system, a day in a park in London, public transport in London, accommodation in London, shopping in Edinburgh, pets (…)
People and icons	British: • Musicians/singers: Sade, Sting, Paul McCartney, Cat Stevens, Britney Spears, David Beckham • Other figures: William Henry Porter, William Shakespeare, William Wordsworth, Michael Foster, Lady Diana US: • Musicians/singers: Louis Armstrong, Steve Wonder, C, Santana, Lionel Richie, Mariah Carey, Akon, Beyoncé Knowles, Rihanna, Justin Timberlake • Other US/BRITISH figures: Martin Luther King, Frederick Douglass, Sonia Sanchez, Alexander Graham Bell, James Hillier, Dorothy West, Quentin Reynolds, Isaac Asimov
Products	Music/songs: 'What a wonderful World', 'Englishman in New York', 'Say you, Say Me', 'Songs of Freedom', 'Hotel California', 'Father and Son', "Ebony and Ivory', Hero, 'Bridge Over Troubled Water' Poetry: 'Catch the fire' 'the Bard's Sonnet 18' 'The richer, the poor' 'The daffodils' Fiction: 'True Love', 'Later', 'The Winter's Tale', Lord of the Rings • Other: The touristic season in Scotland and Edinburgh, Harry Potter, Scottish Traditional Dress, The Museum of London, Holidays in USA [Hawaii, New York, White Water magic, Disney Land, Texas scenery and wildlife (…)]

and singers, who belong predominantly to "high culture". The types of music that textbooks contain include Pop Rock, Folk Rock, Soul, Hip Hop, R&B, and Jazz, with songs dating from the 70s, 80s and early 90s. The singers include Santana, Lionel Richie, Steve Wonder, Cat Stevens, Sting and Mariah Carey. The extensive amount of "high culture" content in English textbooks seems to correspond more to textbook writers tastes than to learners tastes.

Tables 4 and 5 respectively present the perspectives, products, people/icons and practices/lifestyle in the Tunisian/Arab culture and the global/international culture. Table 4 informs that English textbooks focus on the touristic highlights of Tunisia. It also sheds light on people and icons that belong to 'Arab culture' such as the historian and sociologist Ibn Khuldun, the love story icons of classical pre-Islamic poetry Qais/Layla and Antar/Abla. Table 5, in contrast, presents what can be labeled global cultural 'topics' such as views of people around the world on school, generation gap, environment, science and technology, charity, consumerism, internet addiction, nutrition, philanthropy (…).

At the level of nature and type of cultural references, Table 6 indicates that the English textbooks with the largest amount of cultural references are TB3, TB6, TB7

Table 4 Tunisian/Arab culture in Tunisian EFL textbooks

Perspectives	Decision-making in family, boys/girls friendship, 'civility', equality between man and women
Practices and lifestyle	The weekly Market in Houmet Souk
People and icons	Aziza Othmana, Ibn Khuldun, Qais/Layla, Antar/Abla, Wissam Hmem
Products	The Tunisian flag, Touristic highlights of Tunisia, Tunisian Cuisine

Table 5 Global/international culture in tunisian EFL textbooks

Perspectives	How people around the world feel about school?, generation gap and pushy parents, health and environment, consumerism, love, human rights, music, violence, child labor, nutrition, jobs and success stories, science and technology [genetic engineering, cloning, in vitro fertilization (…)], philanthropy, charity and volunteering (…)
Practices and lifestyle	Valentine's day, Mother's day, Entertainment and hobbies, Health, Internet addiction (…)
People and icons	Alfred Nobel winners with a text on Ahmed, H. Zewail, Maria Montessori, king of Flaminco, Joakin cortes, Celine Dion, Zine Eddine Ziden (…)
Products	• TV channels: Teletoon, Spacetoon and Boomerang, Food and Cuisine, MTV, CDs, Tourism in Malaysia (Island of Peneng), Tourism in Malta, Immortality (song) • UN and its institutions: UNICEF, UNESCO, WFO, World Bank (…) • NGOs: OXFAM, HRW, Red Cross, Red Crescent, Islamic Relief (…)

Table 6 Nature and type of cultural references in English textbooks

Textbook	Direct reference		Indirect reference	Type of reference
	Textual reference	Nontextual reference		
TB1	2		1	Touristic: NA Factual: NA
TB2	1		4	Touristic: NA Factual: NA
TB3	13	3	2	Touristic: 2 Factual: 12
TB4	3	1	1	Touristic:2 Factual: 2 Analytical: 1
TB5	4		2	Touristic: 1 Factual: 3
TB6	7	4	4	Touristic: 8 Factual: 3
TB7	12		8	Touristic: 1 Factual: 11
TB8	19	5	10	Touristic: 11 Factual: 13

and TB8. It also shows that all the cultural references are either touristic or factual. The important amount of cultural content of the English program can, in best cases, lead to an incidental, and basic form of factual, cultural knowledge. As a matter of fact, students are never engaged in a dialogic, interactive and interpretive use cultural content. There is a single explicit instance of critical interpretive culture learning in TB4 (p. 46) (see Table 3); it was based on a text on violence in a British school where students make fun of the accent of a schoolboy from Kingsbury. Tunisian learners were invited to "say whether the same things happen the same way in your school or not."

The latent culture curriculum of English textbooks is presented sporadically and non-systematically in the "By the way approach" and the "tour guide" approach. These two approaches to culture make ideal people meet in idyllic places. In TB2, for instance, Peter, a British Citizen visiting Tunisia, discovers Houmet Essouk (an area in the touristic island of Djerba, in the south of Tunisia) Market "where all people are kind and smile to him while he takes photos of them." In TB3, Imen, a Tunisian student visiting England, discovers the monuments of London [Westminster Palace, Buckingham palace, Trafalgar square, London Tower, Camden market by tube, River Thames (…)].

The cultural content presented in English textbooks is rather bland. It involves no evaluative comments, no comparative frame of reference, and no invitation to critical thinking. Actually, the textbooks used in Tunisian schools reveal a few feelings or opinions; students are almost never invited to relate practices to perspectives within the English culture or compare practices and perspectives in Tunisian culture and "English" culture in spite of the opportunities available. In fact, textbooks 4 through 8 are full of culturally-loaded topics and themes that are likely to generate critical, interpretive cultural reflection. These topics include generation gap, violence at school, living without parents, attitudes and values, rights and duties [equal chances and roles for men/women, boys/girls; human rights, equality, tolerance and respect for others (…)]. They can be optimally used to develop intercultural communicative competence and engage students in discussing the link between practices and perspectives both intra-culturally and inter-culturally.

Intra-culturally, such topics would reveal different views among Tunisian students and possibly teachers themselves. In fact, the Tunisian society, considered by many to be one of the most 'westernized' Arab-Muslim countries, is deeply divided between modernists, secularists and conservative, Islamists. Perspectives concerning the aforementioned issues might also generate attitudes that vary as a function of dwelling setting, i.e., rural areas vs. urban areas and probably social class as well. Intra-cultural variation in perspectives cannot be discerned by curriculum developers unless they become aware of the notions of subculture (variations within the same culture) and co-culture (cross-cultural overlaps among distinct cultural groups). In addition, they can consider global middle class culture and global youth culture as manifested in social media, video games, music, sports, cuisine, cartoons, TV programs, and fashion, in language programs. The presentation of such cultural content can actually pave the way to the inclusion of culture-specific practices, products and perspectives. The non-critical, non-interpretive exposure to (foreign) culture can in no way prepare students for situations they might find themselves in the future. On the contrary, it can nurture misunderstanding, stereotyping and stigmatization.

5 Conclusion

This study investigated the place of culture in English education in Tunisia. It revealed that Tunisian language policy-makers, unlike policy-makers in many other EFL Islamic settings, do not have any ideologically-motivated or religiously-driven a priori objections to the consideration of culture in English education. However, it showed that these policy-makers do not approach culture teaching in a principled, systematic way. In fact, the examination of official curricular documents showed that there exists no clear theoretical or pedagogical frame of reference to teaching culture in Tunisian schools. However, the scrutiny of teaching materials indicated also that in spite of the fact that culture is not explicitly recognized as an educational objective, there exists a substantial "incidental" or "latent" cultural content in English textbooks. In light of the results of this study, theoretical, pedagogical and methodological implications can be drawn.

Theoretically speaking, the inclusion of culture in language curricula can be argued to be a function of the philosophy and ideology adopted in curricular documents. The Tunisian English curriculum, exam-oriented (having as ultimate goal to prepare students for the Baccalaureate exam), lacks such philosophy and ideology. English, in Tunisia, seems to be taught for no obvious reason to use Abott's (1987) term. Culture in the Tunisian curriculum of English is approached incidentally and sporadically and ICC is often confused communicative, socio-pragmatic and socio-cultural competence. In this respect, Byram (2014) contended that there exists a lack of understanding of the significance of intercultural competence and its relationship to linguistic competence. The place of culture in L2 curriculum remains, mainly in settings where source culture and target culture are distant from each other, a complex issue that cannot be squared within deterministic theoretical frameworks of intercultural communicative competence. In language curricula, culture, however, will always be there, awaiting around the corner for the first opportunity to come to the surface by choice or by coercion. It would manifest itself as a 'latent curriculum', even if it is not set as an educational or teaching goal. Curriculum design and teaching materials are, generally speaking, a function of expertise and authority as contended by Cortizzi and Jinn (1999). Both statement of philosophy and materials writing should therefore be entrusted to language teaching experts.

Pedagogically speaking, in EFL contexts, four curricular positions in relation to the place of culture in L2 education can be identified: (1) the no need for culture position, (2) the red-tape position, (3) the nonsystematic position and (4) the systematic position. In the first position, L2 education is considered to be a necessary evil and curriculum developers exhibit open ideological and political objections against the inclusion of culture in language programs; they tend to treat L2 as a structural code that can be taught with the source culture as a frame of reference i.e., presenting foreign language in local situations, with local characters, local perspectives, behaviors, and products. The cultural dimension of L2 is therefore either inexistent or maintained to a strict minimum. In the second position, curriculum developers think that culture can be included in language programs but impose red-tape restrictions on cultural content. They include only the 'neutral' cultural aspects and censor all the aspects that are deemed morally, religiously or even politically incompatible with the source culture. In the third curricular position, although curriculum developers have no ideological or political objections to teaching culture, they approach it in a nonsystematic way and tend not to have a clear idea about the state of the art theoretical and pedagogical views on the cultural dimension of foreign language education. Teaching materials might contain cultural content of which curriculum developers and practitioners might not even be aware. In addition, culture might be undermined because of timetable pressure or exam orientation. In the fourth position, curriculum developers have no ideological and political objections against the inclusion of culture in L2 program, see culture and language as inseparable entities and consider L2 learning to be an opportunity for nurturing empathy tolerance of and openness to others; they tend to treat the cultural component in a systematic, rigorous way at the levels of approach, design and teaching techniques.

Methodologically speaking, this study is an analysis of the content of curricular documents and the cultural load of English textbooks in Tunisia. It is also an evaluation of the extent to which the teaching goals pertaining to the cultural dimension of L2 proclaimed in curricular documents and teaching materials are met or not. Traditionally, the term 'evaluation' has been used to refer to research concerned with examining language teaching materials in terms of content and appropriateness. In addition, L2 textbook research has relied on checklists which might prove inappropriate for examining such a complex issue as the cultural dimension of L2 education. In this respect, Weninger and Kiss (2014) noted, that, in L2 textbook research, a distinction should been established between evaluation and analysis. Evaluation is inevitably subjective as it measures the potential or actual effects of materials on their users (Tomlinson, 2003, p. 16). Analysis, a "more principled and theoretical approach to the examination of language teaching materials"; in addition, it is more objective as it assesses what materials contain using different theories and frameworks; it "(…) points further than judging the appropriateness of a particular book in a given educational context (…) with specific students in mind (Weninger & Kiss, 2014, p. 3)." A distinction must therefore be established between analysis studies and evaluation studies with analysis studies trying to assess the content of teaching materials and evaluation studies trying to gauge the effect of such content on learners and the extent to which language curricular achieve their proclaimed goals.

References

Abott, G. (1987). EFL as education. *System, 15*(1), 47–53. doi:10.1016/0346-251X(87)900479

AliAkbari, M. (2004, August). *The place of culture in the Iranian EFL textbooks in high school level.* Paper Presented at the 9th Conference of Pan-Pacific Association of Applied Linguistics, Seoul, Korea.

Al-Issa, A. (2005). The role of english language culture in the Omani language education system: An ideological perspective. *Language, Culture and Curriculum, 18*(3), 258–270. doi:10.1080/07908310508668746

Atkinson, D. (1999). TESOL and culture. *TESOL Quarterly, 33*(4), 765–786. doi:10.2307/3587880

Brooks, N. (1964). *Language and language teaching*. Orlando, Florida: Harcourt Brace Janovich Inc.

Brooks, N. (1986). Culture in the classroom. In J. M. Valdes (Ed.), *Culture-bound: Bridging the gap in language teaching* (pp. 123–129). Cambridge: Cambridge University Press.

Byram, M. (2009). The intercultural speaker and the pedagogy of foreign language education. In D. K. Deardorff (Ed.), *The Sage handbook of intercultural competence* (pp. 333–349). London: Sage Publications Ltd.

Byram, M. (2014). Twenty-five years on-from cultural studies to intercultural citizenship. *Language, Culture and Curriculum, 27*(3), 209–255. doi:10.1080/07908318.2014.974329

Chastain, K. (1988). *Developing second language skills: Theory and practice*. Orlando, Florida: Harcourt Brace Janovich Publishers.

Cheng, K. K. Y., & Beigi, A. B. (2012). Education and religion in Iran: The inclusiveness of EFL (english as a foreign language) textbooks. *International Journal of Educational Development (32)*, 310–315. doi:10.1016/j.ijedudev.2011.05.006

Cortizzi, M., & Jin, L. (1999). Cultural mirrors: Materials and methods in the classroom. In E. Hinkel (Ed.), *Culture in language teaching and learning* (pp. 196–219). Cambridge: Cambridge University Press.

De Lotbinière, M. (2009, February 6). Tunisia turns to a new language partner. Retrieved from: http://www.theguardian.com/education/2009/feb/06/tunisia-tefl

Galloway, D. (1985). *Motivating the difficult to teach*. Boston: Addison-Wesley Publications.

Holliday, A. (1999). Small cultures. *Applied Linguistics, 20*(2), 237–264. doi:10.1093/applin/20.2.237

Hughes, G. H. (1986). Culture in the classroom. In J. M. Valdes (Ed.), *Culture-bound: Bridging the gap in language teaching* (pp. 162–168). Cambridge: Cambridge University Press.

Joiner, E. G. (1974). Evaluating the cultural content of foreign-language texts. *The Modern Language Journal, 58*, 242–244. doi:10.1111/j.1540-4781.1974.tb05106.x

Kiss, T, & Weninger, C. (2013). A semiotic exploration of cultural potential in EFL textbooks. *Malaysian Journal of ELT Research*, *9*(1), 19–28. Retrieved from: http://www.melta.org.my/majer/vol9(1)/majer%20kiss%20weninger.pdf

Kramsch, C. (1993). *Context and culture in language teaching*. Oxford: Oxford University Press.

Kramsch, C. (2009). Cultural perspectives on language learning and teaching. In K. Knapp, B. Seidlhofer & H. G. Widdowson. (Eds.), *Handbook of foreign language communication and learning* (pp. 219–246). New York: Morton de Gruyter.

Kramsch, C. (2010). Theorizing translingual/transcultural competence. In G. S. Levine & A. Phipps (Eds.), *Critical and intercultural theory and language pedagogy* (pp. 15–31). Boston, MA: Heinle Cengage Learning.

Leclerc, J. (2012). Tunisie. Loi d’orientation de l’éducation et de l’enseignement scolaire 2002 dans L’aménagement linguistique dans le monde, Québec: TLFQ, Université Laval. Retrieved from http://www.axl.cefan.ulaval.ca/afrique/tunisie-loi-2002-educ.htm

Moran, P. R. (2001). *Teaching culture: Perspectives in practice*. Heinle and Heinle publications.

Omaggio-Hadley, A. (2001). *Teaching language in context*. Boston: Heinle & Heinle.

Risager, K. (2011). The cultural dimensions of language teaching and learning. *Language Teaching, 44*(4), 485–499. doi:10.1017/S0261444811000280

Robertson, R. (1992). *Globalization, social theory, and global culture*. Thousand Oaks, CA: Sage.

Seelye, H. (1993). *Teaching culture: Strategies for inter-cultural communication* (3rd ed.). Lincolnwood, IL: National Textbook Company.

Sercu, L., Bandura, E., Castro, P., Davcheva, L., Laskaridou, C., Lundgren, U., Ryan, P. (2005). *Foreign language teachers and intercultural competence: An international investigation*. Clevedon: Multilingual Matters.

Shah, K. S., Afsra, A., Fazel e Haq, M. H. & Khan, A. Z. (2012). Course contents of english language textbooks and their relevance to learners' culture in an Islamic context. *Journal of Education and Practice, 3*(12), 165–180. ISSN 2222-1735 (paper) ISSN 2222-288X (online).

Sheldon, L. E. (1988). Evaluating ELT textbooks and materials. *ELT Journal, 42*(4), 237–247. doi:10.1093/elt/42.4.237

Stern, H. (1992). *Issues in language teaching*. Oxford: Oxford University Press.

The National Standards in Foreign Language Education Project. (1996). *The national standards for foreign language learning: Preparing for the 21st century*. New York: The National Standards in Foreign Language Education Project.

Tomalin, B., & Stempelski, S. (1993). *Cultural awareness*. Oxford: Oxford University Press.

Tomlinson, B. (2003). Materials evaluation. In B. Tomlinson (Ed.), *Developing materials for language teaching* (pp. 15–36). London: Continuum.

Tomlinson, B. (2012). Materials development for language learning and teaching. *Language Teaching, 45*(2), 143–179. doi:10.1017/S0261444811000528

Tunisian Ministry of Education. (2006). *English programmes for basic education*. Retrieved from: http://www.edunet.tn/ressources/pedagogie/programmes/nouveaux_programme2011/preparatoire/langues/anglais_college.pdf

Tunisian ministry of education. (2008). *English programmes for secondary education*. Retrieved from: http://www.edunet.tn/ressources/pedagogie/programmes/nouveaux_programme2011/secondaire/anglais.pdf

Weninger, C., & Kiss. T. (2014). *Analyzing culture in foreign/second language textbooks: Methodological and conceptual issues*. Retrieved from: http://www.academia.edu/6707905/Analyzing_Culture_in_Foreign_Second_Language_Textbooks_Methodological_and_Conceptual_Issues

Evaluation of ICT Use in Language Education: Why Evaluate, Where to Look and with What Means?

Faiza Derbel

Abstract The aim of this chapter is to discuss how best to go about evaluating the use of Information and Communication Technologies (ICTs) in language education and to propose guidelines for future research in this area. It begins by deconstructing "evaluation" as a concept and reviewing the status of evaluation studies in English language teaching (ELT) drawing first on meta-analyses and reviews of research on Computer-Assisted Language Learning (CALL) and distance education, and then focusing on the shift in evaluation practice in light of emerging areas of ICT application facilitated by mobile devices and Web 2.0. Evaluation of ICT use will be examined in terms of its foci and emerging methodologies to highlight the need for taking into account the complexity of learning and teaching with ICTs to measure and confirm the "alleged" benefits of ICT use and also to develop knowledge and understanding of the use of ICTs in language education in various contexts. The paper ends by pointing to directions for evaluation studies of relevance to researchers in countries of the Middle East and North Africa (MENA) that this Handbook is targeting.

Keywords Evaluation · Frameworks · Research methodologies · Incorporation of ICT · Web 2.0 · MENA region · Language education

1 Introduction

This chapter delves into the issue of evaluation of the use of Information and Communication Technologies (ICTs) in education with specific reference to English language education. It discusses what methodology to adopt to carry out evaluation studies involving the use of ICTs. That is, how to frame the questions,

F. Derbel (✉)
Faulty of Letters, Arts and Humanities, University of Manouba, Manouba, Tunisia
e-mail: fderbel26@gmail.com

S. Hidri and C. Coombe (eds.), *Evaluation in Foreign Language Education in the Middle East and North Africa*, Second Language Learning and Teaching,
DOI 10.1007/978-3-319-43234-2_13

how to determine their purpose, and what research designs to adopt in light of current technological advances. For a start, what is meant by ICTs today (2015) and how is evaluation defined? According to the World Bank Group (2000), ICTs include any form of "hardware, software, storage, processing, transmission, and presentation of information (voice, data, text, images) (n.p.)." Recognized ICTs include, and the list is not exhaustive, multimedia authoring tools, various forms of distance learning platforms, video conferencing systems, mobile technology, social network platforms (Facebook, My Space, Skype, etc.), and various interactive and engaging devices. Educators are understandably attracted to engaging in what is now known as "ICT-enhanced," "ICT-supported" or "ICT-based" instruction. Judging from the meta-analyses and research reviews of technology use available (Bax, 2003; Burston, 2015; Chapelle, 1998, 2003; Cox & Marshall, 2007; Debski, 2003; Liu, Moore, Graham, & Lee, 2003), teachers have been experimenting for more than half a century with hardware, software, learning systems, networks, mobile devices, downloadable applications and the abundant resources available on the web and in the cloud (Dudeney & Hockly, 2012). Educators are pressed to evaluate the new tools and learning spaces their students are using but are not certain whether the current methodologies are adequate for the task (Goodwin-Jones, 2005a, b, 2011; Greenhow, Robelia, & Hughes, 2009).

Technology-using teachers are generally attracted to using ICTs with their students and to accessing the wealth of digital resources and authoring tools available to them to introduce new input or to design innovative learning activities (Greenhow et al., 2009). Indeed, some teachers found that individualized Skype sessions with team members in the process of completing project work can be integrated within their courses to facilitate communication with their students (Goodwin-Jones, 2005b). Tasks such as reserving an air ticket, checking a bank account, reading news online, and applying for a job can become more authentic thanks to ICT use (Smyth & Mainka, 2006).

However, while teachers may embrace technology, they also assume the responsibility for any decisions they make in their own classrooms in terms of what can be revealed as the impact/outcome effect of ICT use on their learners' school results. What if the taken-for-granted "beneficial" aspects of ICTs do not reflect on their performance? What if the virtues of the application cannot be appreciated by the learners? Is there a way to measure the educational potential of ICT and gauge the degree of its success/failure? One way of resolving the uncertainties is to evaluate the use of ICT in context. That is, investigate whether the ICT-supported instructional mode is leading to improvement based on "proof" from the learners themselves and/or following an appraisal of the situation by applying a given set of measurable criteria. Ideally, evaluators (whoever they may be) will try to establish a relationship between ICT use and learning effects (conceptualized as gains in achievement, improved motivation or changed attitude). Evaluation can be tied up to assessing compatibility with "desired" learning objectives, curriculum specs, and performance targets. This chapter delves into the issue of evaluating ICT use in language education which is, on hindsight, a straight forward task but has become complex in light of the advances in technology and the widening of the scope of teaching/learning with ICTs. It will be argued in this paper that data-driven context-specific evaluation studies carried out by teachers-as-researchers is the most promising approach.

2 Defining Evaluation

It has become difficult to define the term evaluation or delimit its scope. As Harland (1996) rightly points out, "the term itself has been stretched and stretched to encompass an ever-widening range of activities, undertaken for an ever-increasing range of purposes (p. 91)." Simonson, Smaldino, Albright, and Zvacek (2009) suggest "evaluation (…) is the systematic investigation of the worth of an on-going or continuing distance education activity (p. 349)" which highlights the merit of an online program in terms of strengths, weaknesses, benefits and drawbacks. Therefore, setting criteria and benchmarks are important for the practice of evaluation as well as deciding who is well-placed to carry out the evaluation. As the title of this paper suggests, there is first a need to consider the reason for the evaluation, define the nature of the task and decide what (methodology) can be used as proof of "good use" or value-added to instruction and learning. Evaluators need to look for evidence (sources of data) and select the strategies to collect the needed evidence and reach results by using the necessary (and viable) analytical tools. By way of illustration, suppose that the teaching of a particular course is to be done through the use of a software "extolled" by its developers and marketers for its ability to make learning and teaching effortless, enjoyable and manageable within record time. In this case, a clear method of evaluation and fixed criteria for acceptance (features and properties) can help the decision maker (teacher or administrator) decide whether such software is relevant to the curricular goals in her specific context (Reiser & Kegelmann, 1994). Evaluation in that sense consists of determining the quality of the software beforehand but leaves open the question of who will do the evaluating, what information is needed and what procedures can be used to collect it.

There is certainly an expert side to evaluating instructional software as the process will entail knowledge of the technical and teaching skills required to exploit the useful aspects of this software. One way of carrying out the evaluation is to experiment with the software with a small number of students and determine its worth. In a scenario described by Reiser and Kegelmann (1994), the learners can be invited to use the software, the teachers observe them as they use it, and then ask them to express their opinions about the software. Indeed, many teachers I know engage in this off-the-cuff evaluation working with a trial version. As reflective practitioners (Nunan & Lamb, 1996), teachers are constantly confronted with situations that require evaluation of new tools (e.g., interactive whiteboard) or new content (e.g., a reading from an online magazine or a conversation thread on a blog) in order to reach decisions about the applicability of the ideas. Therefore, evaluation of ICTs is part of the day-to-day activities of technology-using teachers while planning, delivering and assessing the degree of success of their attempts.

Besides engaging in informal evaluation as potential users of ICTs in their work, teachers can find themselves subjected to evaluation as part of an institutional plan to introduce ICTs. In this case, the evaluation consists in appraising the progress of an intervention by an outsider looking for data evidence (e.g., coming from observation of a teacher in action) by an "evaluation expert" to determine whether and to what extent the teaching strategies the teachers are using match the "desirable" practice promoted by the intervention (Hedberg, 2011, pp. 9–12). Evaluation of interest to policy makers and funding bodies generally revolves around macro-level concerns related to cost effectiveness, programme fidelity, availability of infrastructures, specific applications, access, enrolment and drop out levels, etc. The evaluation methods in these circles are somewhat canonized considering the long history of what is now established practice within funding organizations (e.g., The World Bank, UNESCO, and OECD). Funding bodies tend to generalize their own methods of information gathering, discursive practices, measurement techniques and interpretative frames of reference (see Wagner et al. 2005). This type of evaluation work serves the purpose of *monitoring and evaluation* whereby the constructs are broken down into *intended outcomes* (teachers' technical and pedagogical skills and learners' desired outcomes, information skill, attitudes, and so on) and *Moderating factors* (level of community support, access to ICTs, and availability in the home or community) (Wagner et al., 2005, p. 8). Survey data can be collected and submitted to statistical analysis to obtain results about student outcomes by way of indications of increased knowledge of school subjects and/or improved attitudes about learning. *Monitoring and evaluation* is a frequent occurrence in the evaluation discourse of international funders because the function of the evaluation, for ill or for good, is to warrant the good management of the projects they are funding. Evaluation in this case serves a control function (Harland, 1996) and is not necessarily conceived to explain practice or entangle intervening factors.

This chapter is written with the teacher-as-researcher and research student in mind and not the "expert" evaluator working for an international body. As academic and researcher functioning in a technology-challenged educational context, am often challenged by teachers in the audience questioning whether technology makes a difference and cynically ask: What kind of language are the learners likely to "pick up" from peers during online collaborative work? A more encouraging question came from a learning system provider who asked: "What can be the best way to evaluate what we are doing in Tunisia?" Of interest to me and my students is the evaluation of ICT use to investigate "the potential of technology for language learning" (Chapelle, 2003, p. 36). Evaluation can focus on learners working with ICT to pinpoint, for instance, what strategies they use to complete a task to provide evidence for the quality of the occurring exchanges which can, in turn, serve as evidence for the learning taking place as a process or as product. Evaluation of teaching with ICTs means, in this paper, any activity involving the exploration of specific situations to see whether the use of ICTs is anchored in criteria of "good practice," "good results" and assumptions about quality teaching/learning (Simonson, Smaldino, Albright, & Zvacek, 2009). The sense making to result from the exploration and thorough the examination of a software, courseware or teacher-developed materials and the learning arising during implementation are taken to be context-specific and personalized by the teachers in-action (Schön, 1983). In this type of evaluation, there is as much to gain for the teachers who will see for themselves how they fared as users of ICT. In the title of this paper I ask: *Why evaluate, where to look and with what means?* That is, the evaluator is required to determine the purpose of the evaluation as practice, fix targets for the evaluation (where to look) and determine and deploy the research instruments and analytical tools needed. By doing so, it is hoped that this chapter gives the inspiration, conceptual understanding, and practical guidance for prospective evaluation researchers to proceed on scientific grounds and explore with confidence situations of interest in their own context.

3 Conceptualizing Evaluation

Most teaching resource books used in teacher education and training include a section at least on evaluation. Authors (e.g., Reece & Walker, 1992; Simonson et al., 2009; Smyth & Mainka, 2006) encourage teachers to engage in evaluation of their learners and their own teaching by providing them with easy-to-use checklists, ready-made questionnaires, and sample interview protocols for them to apply step by step. These teacher educators/trainers are keen on evaluation because they see in the practice an opportunity for teachers to see for themselves how they are faring as teachers and how their learners are progressing. These resource books are recommendable as starting points for novice researchers interested in evaluation studies in school contexts. More specialized evaluation books (e.g., Hopkins, 1989) are also helpful in explaining the distinctions between different approaches to evaluation.

As far as evaluating ICT use, researchers would need to distinguish between formative and summative evaluation. Summative evaluation means that the software is evaluated while used by teachers and learners in order to determine what the outcomes are as revealed in test scores while formative evaluation means trying out software before designing and delivering instruction so that a decision is reached about its potential for learning (Lan, Sung, & Chang, 2013; Levy, 1999; Shaughnessy, 2003). Alternatively, Chapelle (2003, p. 81) argues that the evaluation can be equally focused on the *product* by measuring *what* has been learned as specified in the objectives of the instruction or *process* of learning by collecting information about *how* the software was used by learners and to do what learning activities. She adds that these distinct evaluation types can, when combined, contribute to a more exhaustive description of the situation of ICT use and help sharpen views of the issues surrounding the practice of teaching with technology. Furthermore, Ruhe and Zumbo (2009) propose with reference to distance education, focusing on the course's *underlying values* and *unintended consequences* as additional features of the implementation to uncover the "hidden" aspects of a distance course to complete the picture. The underlying values of a course, they suggest, can be inferred from the language used to describe it (found in course outlines, descriptions of the goals, standards, and assessment principles). Labels like "innovative" and "cutting-edge" can be used to infer the intent and contrast it with the outcomes to see if there is a match. Unintended instructional consequences can be technical flaws, high dropout rates, or witnessed discrepancy among the course components. As for the unintended social consequences, these include, for instance, isolation of the distance learner and the need for the tutor to play the role of coach and facilitator (Derbel, 2013).

Judging from developments in the field of CALL (Chapelle, 2001, 2003; Levy, 1997), evaluation practice focusing on the measurement of *effects*, *impact*, *outcomes*, *efficiency*, was considered the legitimate way of reaching answers about ICT use which was at that time needed to make the case for technology or to convince administrators, parents and teachers to generalize the experiences of high profile projects (Chapelle, 2003, pp. 70–73). Indeed, in the early days of CALL (1960s) high profile projects in the US and UK were evaluated with comparisons with traditional teaching situations. The main purpose of the evaluation was to establish whether the PLATO learning system, for instance, had a positive or negative effect on learners' achievements and attrition rates compared with traditional classes. The field moved on to develop more contextualized and sophisticated methodologies. The approach changed in the 1980s and 1990s as teachers began to play a bigger role in programming and developing their own materials, taking advantage of the availability of user-friendly authoring tools (Levy, 1997, p. 43). The objective of evaluation is no longer the justification of spending but rather exploring and understanding the learning processes as they unfolded during implementation (Garcia-Villada, 2009). As a result, evaluation criteria and methodological solutions were needed to help teachers make sense of the situations they were documenting and analysing from learning/teaching frames of reference. Chapelle (1994), for instance, illustrated how CALL activities can be researched to determine the quality of the communication resulting from their use, drawing in this case on an interaction analysis framework to produce a description of the learner-computer interactions, the language produced and the learner's level of engagement with the materials.

In her 2003 book, *English Language Learning and Technology*, Chapelle succinctly summarizes evaluation research in CALL and demonstrates how researchers have successfully used frameworks from allied disciplines (Second Language Acquisition, Discourse Analysis, Testing and so on) and analytical tools fit to the focus and purpose of the evaluation (learner, teacher, task, the interactions, or output). By way of illustration, Chapelle (2003, pp. 82–108) explains that researchers can investigate how learners use the options in the software, observe and document whether learners use the help functions (e.g., subtitling or annotations) and the feedback built into the software. Results of this evaluation can be used by software developers to improve the design and increase its learning potential. Similarly, she points out that content analysis has been "successfully" used to analyse the language of learners as they were performing computer-mediated communication (CMC) tasks and describe the characteristics of the register used, the written interactive discourse and other moves. Inferences can then be made about the value of the task and the language abilities of the learners.

The evaluation work helped with its cumulative effect to establish the field of CALL and the value of teaching/learning with technology. The studies overviewed in Chapelle's (2003) book can inspire novice evaluators when looking for an area of focus (learner, teacher, communication tasks, CALL materials, and multimedia activities), types of evaluation (formative/summative or product/process), research designs (experimental, case study or action research), framework for analysis and interpretation (content analysis, interaction analysis, discourse analysis) and positioning of the evaluator(s) (insider or outsider).

Most importantly, the shift starts by asking the question differently (Cox & Marshall, 2007). Instead of asking whether teaching with technology is "better than" teaching face-to-face or "more effective" in improving learners' test scores, it is more informative to ask questions about how the experience of learning L2 with CALL went and find evidence for its contribution to learning language. The evaluator can examine the situation from multiple perspectives and collect quantitative as well as qualitative data about the on-going implementation of technology-supported lessons. Jamieson et al.'s (2005) study illustrates a possible design they used to evaluate the *Longman English Online* (*LEO 3*) during implementation. Chapelle's (2001) six criteria of CALL evaluation (language learning potential, learning focus, learner fit, authenticity, positive impact, and practicality) were used to examine the materials from the perspectives of the developer, teacher and the learners. Their scheme of data elicitation included questionnaires, reflections, and interviews based on specific evaluation questions for each criterion. The "numeric summaries (p. 32)" obtained from test scores and questionnaire analysis were supplemented with interview and guided reflection data. Conclusions were then reached about how *LEO 3* matched the six criteria and how it was appreciated by this group of learners and the teacher. To evaluate teacher developed reading materials while used by learners, Derbel (2001) resorted to a screen capturing software (*Camtasia*) to observe the learners' "navigation paths" and "look-up behaviour" which were later corroborated with the self-reporting data obtained from learners by means of a stimulated recall protocol. Strengths and weaknesses of the materials and design options were revealed and juxtaposed with the teacher's pedagogical intent.

Another evaluation study reflecting the shift is Lan et al.'s (2013) action research study focusing on a proposed mobile-supported cooperative reading system. Their design reflected an attempt to bridge evaluation and pedagogical concerns by incorporating formative and summative evaluation within a two-loop action research cycle. They collected data before and while teaching, followed by repair cycles of the system's functions and interface. The teachers taught two units and administered pre- and post-tests to measure their students' attitudes examined along five dimensions of the interface and functions of the system.

Evaluation work of teacher practices is also a possible focus. To trace and construct the learning processes of teachers making the transition from traditional teaching to teaching with interactive whiteboards, learning objects and network links, Hedberg (2011), for example, developed the "concerns-based adoption model" (p. 4) by drawing on the literature on teacher acceptance of innovation and was able to trace the teachers' learning curves and changes in their pedagogy with the whiteboard from the narratives he collected over the semester. He was also able to determine the factors shaping the quality and pace of their learning to teach with the interactive whiteboard and learning objects. This study design reflects a shift in focus from the technology itself to the teacher implementing it. Similarly, Starkey (2011) developed "the digital age learning matrix" (p. 24) as a research tool to evaluate how digitally-able beginning teachers taught lessons using digital technologies. He used the matrix as a reference to evaluate whether the teachers' practice is aligned with the principles of constructivism (taken to be the learning theory compatible with the requirements of 21st century digital learning).

The studies reviewed so far indicate that there are generally two perspectives: The psychometric type and the more interpretative type. A note of caution is due, however. It has been pointed out earlier that experimental designs fell out of favour in the 1980s and 1990s among teachers interested in process research but meta-analyses and reviews of research focusing on impact/effect of ICT on learning indicate that this research tradition has not been abandoned (Cox & Marshall, 2007; Liu et al., 2003; Ting, 2005). Despite the detected methodological "flaws" and shortcomings in a big number of studies which did not qualify for inclusion, these authors merely call for the sharpening of the research practices by collecting more in-depth data and increasing the length of time spent in the field (i.e., carry longitudinal studies). Felix (2008) praises Allum (2004) and Nutta et al. (2002) for incorporating delayed post-tests and self-reporting data to measure learners' attitudes towards computers and school subjects. Liu et al. (2003) mention Brett (1997) who compared multimedia technology and simple audio and video equivalents to prove the superiority of multimedia CALL. There must be new concerns for researchers who find themselves working with the Web 2.0 generation and facing new challenges. I outline the following observed trends.

Greenhow et al. (2009) note that the most prominent change consists in the broadening of the conceptualization of "classroom" and that "[t]oday learners have more choices about how and where to spend their learning time (e.g., in online settings or in private, public, or home school options) than they did 10 years ago" (p. 247). Therefore, to recall part of the question I am asking in this paper ("*where to look?*"), it seems that any attempt to evaluate the use of Web 2.0 tools (e.g., social networks, media sharing, blogs, etc.) will imply data collection from different electronic and physical spaces. More innovative data collection methodologies (also outlined by Goodwin-Jones, 2011) are emerging to evaluate learning in Web 2.0. like web surveys, digital photography and movies, voice recognition tools, and more powerful data aggregation tools.

Evaluating learning in Web 2.0 can be guided by questions about the learning opportunities (e.g., creation of content), the role(s) teachers play in the process considering the open-endedness of virtual spaces, the education value of learners' participation in Web 2.0 and how they bridge the learning opportunities in different spaces (in and out-of-school and on social networks). Currently, researchers started to dispute the impact of mobile technologies and constant internet connectivity on academic performance. Researchers are asking whether new working styles are developing and whether these are conflicting with the learning objectives proposed by teachers. For example, Rosen, Carrier, and Cheever (2013) attempted to assess the level of distraction during task performance and Lepp, Barkley, and Karpinski (2015) whether learners are able to self-regulate to resist the distraction. Thus, they are seeking through their work to reach answers and clarifications about the opportunities and the risks for academic performance. Other researchers are more positive in their outlook to the use of mobile phones in L2 learning. Burston (2015), for instance, singles out Chen, Chang, Lin, and Yu (2009), Liu (2009) and Tai (2012) as exemplifying the use of mobile phones in communicatively-oriented activities. Therefore, it appears that more up-to-date methodologies are not necessarily being used with recent technological innovations.

Evaluation work can help make the case for technology use in language education; i.e., establish benefits, gains in achievement, learning potential associated with a particular situation of ICT-supported teaching. Evaluation can lead to better understanding of the processes facilitated by the introduction of an ICT tool and how it contributes to language learning. More informative evaluations are the ones which yield thorough descriptions of the tools used, level of learners involved, the relevance for the curriculum or lesson, the support mechanisms, and the learning theory embedded in the materials and implemented by the teacher. Based on the discussion in the previous sections, what is known today has been made possible by revising the methodologies used and especially by theorizing from the ground naturally occurring teaching (Guba, 1981) with the involvement of the main actors in the research process. The ultimate goal of evaluation is to provide answers and give clarifications about the use of ICTs where "mixed feelings" about their worth are shared among funders, administrators and/or teachers and parents. Evaluation studies can also reach unintended, unexpected or conflicting results (Felix, 2008).

4 Recommendations for Future Research

In the beginning of this paper, I had hoped that the analysis would lead to providing guidelines for novice teachers they can use as points of departure for the design of evaluation studies that tap into the implementation of ICT-supported teaching/learning in the MENA region. I will start from my understanding of my own context in Tunisia and hope that readers in the region can establish parallels relevant to their own contexts. What has been done on the issue of ICT use in education in Tunisia is somewhat patchy. Much of it has not been documented even though a good number of teachers I know have been trying ideas for a number of years and some are quite advanced technology users. Moreover, little information is available on the official policy related to ICT use in education. A report by Hamdy (2007) included a listing of the state-led projects and policy decisions from 2001 upwards focusing on the introduction of ICT as a compulsory subject in schools. Another book by Chabchoub and Bouraoui (2004) included elements of the story of state policy such as the Family Computer project in 2000 or background to the creation of the Virtual University of Tunis (UVT) in 2002 and subsequent plan(s) to "digitize" 20 % of the courses at university by 2006–2007. UVT has now made strides into course development according to the information on their websites (number of courses online) and has put in place a teacher training program for tutors. Moreover, Master's level research studies are relatively few (Ben Gayed, 2007; Ben Hammed, 2012; Charfi, 2010; Klibi, 2014; Lachheb, 2013) and generally focused on "effects," "motivation" and "hurdles and obstacles". Therefore, there is a lot left to explore about actual implementation.

As this chapter is meant to guide researchers interested in evaluating ICT use in educational contexts, it is important to remind them that research methodology is a set of options and solutions to assist them in finding answers to their questions. Therefore, if research students (or practitioners for that matter) are interested in exploring the "worth" or "outcome" of applying a software, using an authoring tool, or adopting a social network platform in their own school, they can find guidance in these reviews. Other possible studies of the evaluation type can be, for instance, conceived to answer a question like: "*What happens when learners perform a writing task on a class blog*?" The researcher can resort to digital data recording (using an outside camera focusing on learners and a "tracking" software focusing on what happens inside the computer) collected during the performance of the writing task. The operational data can be transcribed and coded according to a system which is likely to serve the researcher's purpose (see Chapelle, 2003; Derbel, 2001; Jamieson & Chapelle, 1987). Such rich data can help uncover on-task/off-task behaviour, strategies for task completion, the teacher's role/presence (if the teacher is involved), and online/offline dialogue with peers and teacher (whether instances of negotiation of meaning and request for help do occur). Taken together, these details can help researchers in the process of interpretation (relying on an appropriate theoretical framework) in order to make inferences and reach conclusions about the quality of the learning experiences (the aim of the evaluation). In the end, the results should contribute to clarifying how blogging can be considered a "good" or "beneficial" way of teaching writing supported by evidence from the data. The following sections are suggestions for possible studies to evaluate ICT use in language education.

I begin by suggesting that evaluation studies focus on ICT-supported teaching of the traditional language skills (listening, speaking, reading and writing). The research design should be powerful enough to capture the intricacies of the situation. A mixture of process and product data and examine them drawing on theoretical models, (e.g., Goodman's (1968) interactive reading model), to design and determine the quality of the instruction and learning experiences. Focus can be on task variation, diversification of the reading resources, individual learning strategies, or group dynamics. Second, the object of the evaluation can be a specific ICT tool and its application to teach learners of different ages or of different specializations (e.g., medical doctors). A third area can be based on Greenhow et al.'s (2009, p. 250) suggestion to study modes of self-expression (e.g., when learners produce digital content by means of a specific authoring tool) and using it to communicate in various learning spaces in both formal and informal settings. I have learnt when invited to a Tech Age Teachers' meeting in 2014 (see https://www.irex.org/projects/tech-age-teachers-tunisia) that teachers across sectors and levels are being trained to integrate technology in their teaching and are out there creating teaching/testing materials, class websites, blogs and wikis. A multiple case study of these teachers will be of interest and a good starting point for the appraisal of the philosophy of empowering teachers with technological know-how and their experience with incorporating technology in their teaching in natural settings (Guba, 1981). Finally, but not exclusively, researchers/teachers can evaluate attempts to blend face-to-face teaching with online collaboration projects on social networks with a special emphasis on the process and product. Chapelle's (2003, p. 110) advice about generating CALL text from CMC exchanges can be extended to the new electronic tools and spaces teachers are using today.

5 Conclusion

I have, in this chapter, discussed evaluation as a concept and illustrated how it was used to appraise attempts to implement ICT-supported teaching. I have argued that evaluation can be conceptualized differently and that methodologies have to be compatible with the questions asked and explained that to this day researchers continue to look for answers about the value of technology use for learning but methodologies evolved over the decades as researchers grappled with the questions (Bax, 2003; Chapelle, 2003; Dudeney & Hockly, 2012; Levy, 1997). Theories, conceptual frameworks and analytical tools from already established fields have proven useful (Chapelle, 2003; Liu et al., 2003). However, completely innovative practices may be needed to match the nature of evolving uses and practices and concerns (Greenhow et al., 2009; Goodwin-Jones 2011; Starkey, 2011). I will end with Levy's remark about CALL research in 1990s: "(…), evaluation is crucial if CALL is not to be entirely technology-led, and if we are to identify and build upon prior successes" (p. 41). In Tunisia, for instance, teacher educators and practitioners are cast in the role of recipients of training in technological skills with no research activity to follow up implementation and teacher mediation. The research I am suggesting can help bring into focus the role of teachers in ICT use in language education.

References

Allum, P. (2004). Evaluation of CALL: Initial vocabulary learning. *ReCALL Journal, 16*(2), 488–501.

Bax, S. (2003). CALL—Past, present and future. *System, 31*(1), 13–28.

Ben Gayed, L. (2007). *Effects of CALL on motivation of Tunisian learners of English for specific purposes* (Unpublished Master's thesis). University of Sfax.

Ben Hammed, R. (2012). *Normalisation: Obstacles and solutions* (Unpublished Master's thesis). University of Manouba.

Brett, P. (1997). A comparative study of the effects of the use of multimedia on listening comprehension. *System, 25*(1), 39–53.

Burston, J. (2015). Twenty years of MALL project implementation: A meta-analysis of learning outcomes. *ReCALL, 27*(1), 4–20.

Chabchoub, A., & Bouraoui, K. (2004). *Introduction à la pédagogie numérique* [*An Introduction to digital pedagogy*]. Publication de l'ATURED.

Chapelle, C. A. (1994). CALL activities: Are they all the same? *System, 22*(1), 33–45.

Chapelle, C. A. (1998). Multimedia CALL: Lessons to be learned from research on instructed SLA. *Language Learning & Technology, 2*, 22–34. Retrieved 09 December 2015, from http://www.llt.msu.edu/vol2num1/article1/index.html

Chapelle, C. A. (2001). *Computer applications in second language acquisition*. Cambridge, UK: Cambridge University Press.

Chapelle, C. A. (2003). *English language learning and technology*. Philadelphia, USA: John Benjamins Publishing Co.

Charfi, H. (2010). *A comparison of the metacognitive reading strategies applied by self-regulated EAP students in print and online environments* (Unpublished Master's thesis). University of Sunderland, UK.

Chen, T.-S., Chang, C.-S., Lin, J.-S., & Yu, H.-L. (2009). Context-aware writing in ubiquitous learning environments. *Research and Practice in Technology Enhanced Learning, 4*(1), 61–82.

Cox, M. J., & Marshall, G. (2007). Effect of ICT: Do we know what we should know? *Education and Information Technology, 12*, 59–70. doi:10.1007/s10639-007-9032-x

Debski, R. (2003). Analysis of CALL (1980–2000) with a reflection on CALL as an academic discipline. *ReCALL, 15*(2), 177–189.

Derbel, F. (2013). Facilitation of learning in electronic environments: Reconfiguring the teacher's role. In M. Ciussi & M. Angier (Eds.), *Proceedings of the 12th European Conference on E-Learning, ECEL* (pp. 94–100). Academic Conferences & Publishing Int., Ltd.

Derbel, F. (2001). *The integration of CALL in the ESL classroom: Reconciling agendas* (Unpublished Master's thesis). Iowa State University, Ames, Iowa, USA.

Dudeney, G., & Hockly, N. (2012). ICT in ELT: How did we get here and where are we going? *English Language Teaching Journal, 66*(4), 533–542.

Felix, U. (2008). The unreasonable effectiveness of CALL: What have we learnt in two decades of research? *ReCALL, 20*(2), 141–161.

Garcia-Villada, E. (2009). CALL evaluation for early foreign language learning: A review of the literature and a framework for evaluation. *CALICO Journal, 26*(2), 363–389.

Goodman, K. S. (1968). The psycholinguistic nature of the reading process. In K. S. Goodman (Ed.), *The psycholinguistic nature of the reading process* (pp. 15–26). Detroit, MI: Wayne State University Press.

Goodwin-Jones, R. (2005a). Emerging technologies, messaging, gaming, peer-to-peer sharing: Language learning strategies & tools for the Millennial Generation. *Language Learning & Technology, 9*(1), 17–22. Retrieved 01 February 2015, from http://llt.msu.edu/vol9num1/pdf/emerging.pdf

Goodwin-Jones, R. (2005b). Emerging technologies, Skype and podcasting: Disruptive technologies for language learning. *Language Learning & Technology, 9*(1), 17–22. Retrieved 01 April 2015 from http://llt.msu.edu/vol9num3/pdf/emerging.pdf

Goodwin-Jones, R. (2011). Emerging technologies: Mobile apps for language learning. *Language Learning & Technology, 15*(2), 2–11. Retrieved 01 February 2015, from http://llt.msu.edu/issues/june2011/emerging.pdf

Greenhow, C., Robelia, B., & Hughes, J. (2009). Learning, teaching, and scholarship in a digital age, Web 2.0 and classroom research: What path should we take now? *Educational Researcher, 38*(4), 246–259.

Guba, E. G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *ECTJ, 29*, 57–91.

Hamdy, A. (2007). Survey of ICT and education in Africa: Tunisia country report. The World Factbook, 2007. Retrieved 05 July 2015, from https://cia.gov/cia//publications/factbook/geos/ts.html

Harland, J. (1996). Evaluation as Realpolitik. In D. Scott & R. Usher (Eds.), *Understanding educational research* (pp. 91–105). London, UK: Routledge.

Hedberg, J. G. (2011). Towards a disruptive pedagogy: Changing classroom practice with technologies and digital content. *Educational Media International, 48*(1), 1–16. Retrieved 01 February 2015, from http://dx.doi.org/10.1080/09523987.2011.549673

Hopkins, D. (1989). *Evaluation for school development*. Buckingham, UK: Open University Press.

Jamieson, J., & Chapelle, C. A. (1987). Working styles on computers as evidence of second language learning strategies. *Language Learning, 37*, 523–544.

Jamieson, J., Chapelle, C. A., & Preiss, S. (2005). CALL evaluation by developers, a teacher, and students. *CALICAO Journal, 23*(1), 1–44. http://www.tesl-ej.org/pdf/ej64/a2.pdf

Klibi, A. (2014). *Implementation of computer assisted language learning in Tunisia: EFL teachers' perceptions and perspectives* (Unpublished Master's thesis). University of Manouba.

Lachheb, A. (2013). *Information technology effects on Tunisian college students, Tunisian English majors as a case study* (Unpublished Master's thesis). Grand Valley State University, USA.

Lan, Y.-J., Sung, Y.-T., & Chang, K.-E. (2013). From particular to popular: Facilitating EFL mobile-supported cooperative reading. *Language Learning & Technology, 17*(3), 23–38. Retrieved 01 April 2015, from http://llt.msu.edu/issues/october2013/action.pdf

Lepp, A., Barkley, J. E., & Karpinski, A. C. (2015). *The relationship between cell phone and academic performance in a sample of US college students* (pp. 1–9). Jan-March: Sage Open.

Levy, M. (1999). Design processes in CALL: Integrating theory, research and evaluation. In K. Cameron (Ed.), *CALL: Media, design, and application* (pp. 459–469). Exton, PA: Swets & Zeitlinger.

Levy, M. (1997). *Computer-assisted language learning: Context and contextualization*. Oxford, UK: Oxford University Press.

Liu, T.-Y. (2009). A context-aware ubiquitous learning environment for language listening and speaking. *Journal of Computer Assisted learning, 25*(6), 515–527.

Liu, L., Moore, Z., Graham, L., & Lee, S. (2003). A look at the research on computer-based technology use in second language learning: A review of the literature from 1990–2000. *Journal of Research on Technology in Education, 34*(3), 250–273.

Nunan, D., & Lamb, C. (1996). *The self-directed teacher: Managing the learning process*. Cambridge, UK: Cambridge University Press.

Nutta, J. W., Feyten, C. M., Norwood, A. L., Meros, J. N., Yoshii, M., & Ducher, J. (2002). Exploring new frontiers: What do computers contribute to teaching foreign languages in elementary school? *Foreign Language Annals*, *35*(3), 293–306.

Reiser, R., & Kegelmann, H. (1994). Evaluating instructional software: A review and critique of current methods. *Educational Technology Research and Development, 42*(3), 63–69.

Reece, I., & Walker, C. (1992). *A practical guide to teaching, training and learning: A practical guide*. Tyne and Wear, UK: Business Education Publishers Ltd.

Rosen, L. D., Carrier, M., & Cheever, N. A. (2013). Facebook and texting made me do it: Media-induced task switching while studying. *Computers in Human Behaviour, 29*, 948–958.

Ruhe, V., & Zumbo, B. D. (2009). *Evaluation in distance education and e-learning*. New York: The Guilford Press.

Schön, D. (1983). *The reflective practitioner: How professionals think in action*. New York, USA: Basic Books.

Shaughnessy, M. (2003). CALL, commercialism and culture: Inherent software design conflicts and their results. *ReCALL, 15*(2), 251–268.

Simonson, M., Smaldino, S., Albright, M., & Zvacek, S. (2009). *Teaching and learning at a distance: Foundations of distance education* (4th ed.). Pearson Education, Inc.

Smyth, K. & Mainka, K. (2006). *Pedagogy and learning technology: A practical guide*. Edingburgh Napier University. Published online by Creative Commons, 2010. Retrieved 01 February 2015, from http://www.creativecommons.org/licensing/by-nc-Sa/2.5

Starkey, L. (2011). Evaluating learning in the 21st Century: A digital age learning matrix. *Technology, Pedagogy and Education, 20*(1), 19–39.

Tai, Y. (2012). Contextualizing a MALL: Practice design and evaluation. *Educational Technology & Society, 15*(2), 220–230.

Ting, S. E. (2005). The impact of ICT on learning: A review of research. *International Educational Journal, 6*, 635–650.

Wagner, D. A., Day, B., James, T., Kozma, R., Miller, J., & Unwin, T. (2005). *Monitoring and evaluation of ICT in education projects: A handbook for developing countries.* Washington, DC: InfoDev/World Bank. Retrieved 01 December 2014, from http://www.infodev.org/en/Publication.9.html

World Bank. (2000). *Monitoring and evaluation: Some tools, methods and approaches.* Washington, D.C.: World Bank Group. Retrieved 01 December 2014, from http://www.worldbank.org/oed/ecd/

Part VI
Evaluation of ELT Certificates and Programs

Evaluating the Certificate of Teaching English as a Foreign Language (CTEFL): A Way to Quality

Hala Salih and Abuelgasim Sabah Elsaid Mohammed

Abstract This study aims to evaluate the *Certificate of Teaching English as a Foreign Language (CTEFL)* programme provided by the Postgraduate Unit, English Language Institute, University of Khartoum, by identifying the students' opinion on the four modules, discovering the students' perception of the teaching, and their own performance. In addition, it tries to judge the success of the programme as perceived by the students and ways for improvement. To this end, the study used a questionnaire. The participants were 13 students who studied in cohort 5 of the CTEFL. The results showed that the students found the programme successful since it contained interesting and useful modules. The students were also satisfied with their instructors' and their own performance.

Keywords SHEQUIP · CTEFL · Programme evaluation · Students' perception · Improvement · Certification

1 Introduction

In recent times, the English language has become the first language of the world. It is estimated that the number of people in the world that use the English language to communicate on a regular basis is 2 billion. As the English language has gone beyond its natural boarders; non-native speakers of English outnumber native speakers three to one as asserted by Crystal (1997). English became the dominant business language and it has become almost a necessity for people to speak English if they are to enter the global workforce. Research from all over the world shows

H. Salih (✉) · A.S.E. Mohammed
University of Khartoum, Khartoum, Sudan
e-mail: halasalih64@gmail.com

A.S.E. Mohammed
e-mail: abuelgasims@gmail.com

S. Hidri and C. Coombe (eds.), *Evaluation in Foreign Language Education in the Middle East and North Africa*, Second Language Learning and Teaching,
DOI 10.1007/978-3-319-43234-2_14

that cross-border business communication is most often conducted in English. In addition, the importance of the English language in the global world of research and publication cannot be ignored.

In Sudan, English language teaching was deeply rooted in the educational system. Its history can be traced with the beginning of colonization of Sudan by the British in 1898, as it became the language of education and civil service. After the end of colonization in the sixties, a long process of Arabicization started whereby Arabic became the medium of instruction in the schools and English language became a subject among other subjects. In 1990, a decision was taken to extend Arabicization to higher education and thus Arabic replaced English as the medium of instruction in institutions of higher education.

> In 1990, a decision was made by the Ministry of Higher Education to teach first year students at all Sudanese Universities in Arabic. By the end of 1994 all Sudanese universities (16 universities and university colleges (6 colleges) are expected to teach all subjects in Arabic. (Wagi'alla 1996, p. 347)

The step to Arabicize teaching in higher education was accompanied by other decisions such as changing English language textbooks, taking extensive reading out of the curricula of schools and closing down the English language teacher training institutes. These changes in the status of the English language in Sudan led to a sharp deterioration of English language proficiency levels among university graduates.

A need for English language grew in Sudan due to political and economic reasons. In 2005, Sudan signed the *Comprehensive Peace Treaty* ending 45 years of civil war between the southern and northern part of the country. Language was a main article in the peace treaty. The English language was alleviated to the position of a second language, so it should be used in civil service and teaching in institutions of higher education. With the cession of the southern part of Sudan into the Republic of Southern Sudan in 2011, English became even more important as it was chosen as the formal language of the new country. As the north needed to communicate with its new neighbor, thus, the English language continued to play an important role in the relationship between the two countries. Economically, despite sanctions on Sudan the local economy continued to grow with the growth of petrol, telecom and construction industries. These industries among others needed the English language for their businesses. They wanted to employ graduates with better English language proficiency. Unfortunately, universities were graduating students with poor English language proficiency levels. Thus, the need for quality English language programmes has grown and programme evaluation has become a necessity to improve the current situation.

Because of the deterioration in English language services and the absence of benchmarks and standards, Sudan did not develop a history of programme evaluation. Programme evaluation can provide information to stakeholders and sponsors such as the effects, potential limitations, or apparent strengths of the programme and thus lead to improvements in the quality of existing language services. It can also indicate the programme's impact on participants and discover problems or needs early on to prevent more serious problems later. It can also recommend improvements for the future to ensure quality and inform staff about the programme.

Hoping to provide quality English language programmes, the English Language Institute (ELI) started a project to evaluate all its current programmes. This paper highlights the results of the evaluation of the *Certificate of Teaching English as Foreign Language* (CTEFL) programme offered by the ELI. The evaluation of the CTEFL programme attempts to respond to some questions regarding the achievement of objectives of the programme, teachers' performance, learners' attitudes toward the programme, and the relation between the programme and students' needs.

2 Theoretical Background

Evaluation is an intrinsic part of teaching and learning of languages. Programme evaluation is a systematic method for collecting, analyzing, and using information to answer questions about a programme's effectiveness and efficiency. According to Brown (2005), evaluation is one of the components of any language curriculum. It refers to the organised accumulation and analysis of data to enhance the curriculum and to measure its effectiveness in a specific setting.

A number of approaches can be followed in evaluation. The first is a product-oriented approach, which concentrates on measuring the achievement of the programme's learning outcomes. The second is a static-characteristic approach, which deals with the available resources such as the proportion of teachers compared with students, the number of books, the degrees available in the institution among others. Outsider experts usually administer this approach. The third approach is a process- oriented approach that attempts to examine the process of learning. It includes both formative and summative evaluation. The last approach is a decision-facilitating approach. It focuses on collecting data to enable programme's managers to make necessary decisions (Brown, 2005).

Richards (2001) proposes that programme evaluation attempts to respond to some questions regarding the achievement of objectives of the programme, teachers' performance, learners' attitudes toward the programme, and the relation between the programme and students' needs. Answers to these questions enable programme administrators to make various types of decisions. Sanders (1992) and Weir and Roberts (1994) as cited in Richards (2001) suggest that evaluation may concentrate on a variety of aspects such as programme planning and organisation, programme content, teachers and teaching, materials, and students.

Course and material evaluation is frequently covered in the literature. However, there are few studies in programme evaluation. Barazaq (2007) conducted a study whose aim was to identify the effectiveness of Student-Teacher Training Programme (STTP) in the Gaza Strip, Palestine. The participants were 200 student teachers at Gaza Islamic University, Aqsa University, and Azhar University. Barazaq used a questionnaire to collect data for the study. She found that the programme was quite good and well organised. It also equipped students with the necessary skills required for teaching the English language.

Edwards and Owen (2002) evaluated the MA TESL/TEFL Open/Distance Learning Programme at Birmingham University, UK. The aim was to evaluate the programme impact on students', who were in-service teachers, performance. The subjects were 148 programme graduates. Edwards and Owen used a questionnaire for data collection. They revealed that the programme had a strong impact on the students since 90 % of the subjects remembered what they studied in the MA. They also found that the programme was successful because it prepared the students to teach English.

Biyik (2007) studied the Distance English Language Teacher Training Programme (DELTTP) in Anadul University, Turkey. His objective was to ascertain the programme's adequacy. The participants were 26 graduate students of the academic year 2004–2005. Biyik used a questionnaire for the students and an interview with the administrators, instructors, and three students. He concluded that the programme was not able to train the desired number of teachers in that short time period despite the fact that it was successful because it met students' needs and expectations.

Regmi (2008) evaluated the ELT Programme at Kathmandu University, Nepal. His participants were eight students enrolled in the academic year 2007. He used an interview to collect data and found that the ELT programme had partially achieved its aims.

Fordden (1997) assessed the ELT Graduate Programme at the University of Antioquia, Colombia. She aimed at identifying students' feeling towards the programme and to solve any problems therein. To collect data, Fordden (1997) employed classroom observations and an interview. The results showed that the ELT Graduate Programme teachers were well qualified and up-to-date. Some students complained about their teachers' poor selection of reading lists. Students also regarded teachers as facilitators. In addition, the study revealed that all the courses were relevant and their content was useful. The programme suffered from the time factor because it was short. A major finding was that students wanted to have formative assessment. Fordden (1997) suggested some developments regarding content, timing, methodology, and assessment.

3 CTEFL Background

In 2011 the English Language Institute (ELI) was established by the University of Khartoum to promote the teaching and learning of the English language. With this objective, the ELI started working on localizing international degrees and certificates. One of these certificates is a three-month certificate called the *Certificate of Teaching English as a Foreign Language* (CTEFL). It was designed in joint collaboration with Reading University in a project funded by a grant from the British Council under the Sudan Higher Education Quality Improvement Project (SHEQUIP). The main idea behind the CTEFL is to localize a certificate that resembles Cambridge CELTA that will allow non-English language graduates to specialize in teaching English to widen the base of English language teachers and provide certification for those who were already teaching the English language with no certification. The CTEFL was launched in 2012 and was run twice a year. It was made up of four modules. Module 1 is called '*Language Proficiency*' and the aim of the module is to raise the students' proficiency level from intermediate level to an upper-intermediate level. Module 2, 3 and 4 were '*Teaching Language Skills*', '*Core Issues*' and '*Language Analysis*', which were run simultaneously after the students successfully passed module 1. They were coded as follow:

M1 CTLP: Language Proficiency
M2 CTTS: Teaching Language Skills
M3 CTCI: Core Issues in ELT
M4 CTLA: Language Analysis.

The CTEFL was continuously evaluated and changes were made to improve the quality of the teaching. In 2014, the number of students grew and they were from very diverse backgrounds (retired army generals, pharmacists, journalists, veterinary doctors, Islamic Sharia professors). The administration of the ELI decided to carry out a more in-depth programme evaluation, a practice which was not very regular and rare in Sudanese institutions of higher education.

The ELI, with the diverse students' profile, wanted to identify the students' opinion on the four modules, discovering the students' perception of the teaching, and their own performance. In addition, the programme evaluation wanted to judge the success of the programme as perceived by the students. The ELI considered the students as one of the most important stakeholders as they were adult learners already successful in their careers, but they wanted to acquire new skills of teaching the English language. Also due to the ELI limited resources the programme evaluation was seen as a way to assist in prioritizing resources by identifying the programme components that are most effective or critical for students' successful learning.

Thus, the programme evaluation in the study aimed at:

- Identifying students' opinion about the CTEFL four modules of content.
- Discovering students' perception of the CTEFL teaching methodology.
- Finding students' self-evaluation of their performance in the programme.
- Judging the programme's success.

4 Methodology

4.1 Participants

The participants of this study were cohort 5 of the CTEFL programme. The total number of the students in cohort 5 was 17. Out of these 17, only 13 (76.9 %) participated in this study. There were 10 (76.9 %) males and 3 (23.1 %) females.

Their age range was 24–64. All of them were university graduates with diverse degrees. Two of them were master's holders and two others had higher diplomas. The others were B.A. and B.Sc. holders. They were specialised in different fields such as English language, pharmacy, engineering, and commerce among others. They studied at various Sudanese universities such as SUST, Cairo University (Khartoum Branch) and Juba University. Five of them graduated from the University of Khartoum, and one from Manchester University, UK.

4.2 Instrument

A questionnaire was used to collect data. It was adapted from a questionnaire designed by the School of English Language and Literature, Aristotle University of Thessaloniki, Greece. The questionnaire contained four sections. Section A collected personal information about the students. The second section, B, sought to identify students' opinion about the CTEFL's four modules regarding their learning outcomes, content, and sessions. It covered items 1–10. Items 11–12 concentrated on written and/or oral presentation and quizzes, respectively. Section C (items 13–17) attempted to collect data concerning the teaching of the modules. Questions were designed to evaluate the instructors of the four modules. The last section, D, (items 18–22) was about the students' self-evaluation of their performance in the CTEFL. Question 23 asked students to assign a percentage for the success of the CTEFL and question 24 required them to suggest any ideas to develop the programme. Students were provided with five-Likert scale options that ranged between strongly agree to strongly disagree and they were asked to tick their appropriate choice.

4.3 Procedure

The questionnaire was distributed to the participants during their last lectures. Seventeen copies were handed out to the students. The total of returned copies was 15. Two copies were excluded because the subjects answered only two or three questions. The final number of copies was 13 (76.5 %). Statistical Package for Social Science (SPSS) version 21 was used to analyse the data.

4.4 Validity and Reliability

Before the actual implementation of the questionnaire, it was sent to two assistant professors specialised in the English language to evaluate its content in terms of relevance and appropriateness to the study objectives. They commented on the content and suggested some changes to make it more suitable to the objectives of the study. Their comments were incorporated in the final version of the questionnaire. The questionnaire was distributed to three former CTEFL students for piloting. The three students answered the questionnaire smoothly and without finding any difficulties. Cronbach Alpha was used to measure the reliability of the questionnaire and the value was 0.94, which was quite appropriate for the questionnaire to be administered.

5 Results and Discussion

The questionnaire was analysed using SPSS. The results were grouped according to the objectives of the study.

5.1 Results of the First Study Objective

The first objective was to identify the students' opinion on the four CTEFL modules and their content. It is covered in questions 1–10 on the questionnaire. For M1 CTLP, the results are shown in Table 1.

Table 1 shows that 12 (92.3 %) of the participants *agreed* that M1 learning outcomes were clear and the additional material were useful. The participants also regarded the module as useful since 11 (91.7 %) of them *agreed* on that. Similarly, 11 (84.6 %) of the subjects *agreed* that the material and sessions were well organised. Out of 13, 10 (76.9 %) of the participants *agreed* that the material used was relevant to the learning outcomes, and it was interesting. Seven (53.8 %) of the

Table 1 Students' opinion on M1 CTLP (language proficiency)

No.	Question	Agree		Not sure		Disagree	
		No.	%	No.	%	No.	%
1	The module learning outcomes were clear	12	92.3			1	7.7
2	The module material was relevant to the learning outcomes	10	76.9	1	7.7	2	15.4
3	The material taught was well organized	11	84.6			2	15.4
4	Each session was well organized	11	84.6	1	7.7	1	7.7
5	The additional material used (videos, slides, photocopies, etc.) were helpful	12	92.3			1	7.7
6	The module learning outcomes were achieved	10	76.9	3	23.1		
7	The module was interesting	10	76.9	1	7.7	1	7.7
8	The module was difficult	5	38.5	1	7.7	7	53.8
9	The module was useful	11	91.7			1	8.3
10	The module required lots of study	8	61.5	1	1 7.7	4	30.8

subjects *disagreed* that the subject was difficult. However, 8 (61.5 %) *agreed* that it required lots of study. These results indicate that the students were satisfied with M1 CTLP. They found it well organised in terms of content. This may be attributed to the use of the *Link Up* Upper Intermediate course book in teaching this module. The book is organised into 20 units. It is an integrated course that covers topics, skills, structure, and vocabulary. Additionally, the results imply that the students were able to cope with the content of the book, and they did not find it difficult. The module's aim was to improve learners' proficiency, so the students found it useful especially since most of them had been away from learning the English language for a long time. Moreover, it appears that the students were aware of the module-learning outcome and they felt that they achieved those objectives. In conclusion, the students were satisfied with M1 CTLP.

Concerning M2 CTTS, the results are displayed in Table 2.

According to the results displayed in Table 2, almost all the students 12 (92.3 %) believed that the additional material used was helpful. Eleven students (84.6 %) also *agreed* that the learning outcomes were clear, the material was relevant to the learning outcomes, the material and sessions were well organised. Ten (76.9 %) of them *agreed* that the module learning outcomes were achieved, the module was interesting, and it was useful. As for the difficulty of the module, 6 (46.2 %) of the subject *disagreed*, but 5 (38.5 %) found it difficult. The module required lots of study as seen by 6 (46.2 %) of the participants. However, 5 (38.5 %) were *not sure*. It seems that they could not judge the difficulty of the module since the distribution of cases is quite similar. The results indicate that M2 CTTS was to the students' expectations. It can also be stated that the module was interesting and useful for the students.

Table 3 summarises students' opinions on M3 CTCI. The vast majority of the subjects 10 (90.9 %) *agreed* that the material was well organised and the additional material used was useful. Similarly, 10 (83.3 %) agreed that the material was

Table 2 Students' opinion on M2 CTTS (teaching the language skills)

No.	Question	Agree		Not sure		Disagree	
		No.	%	No.	%	No.	%
1	The module learning outcomes were clear	11	84.6	1	7.7	1	7.7
2	The module material was relevant to the learning outcomes	11	84.6	1	7.7	1	7.7
3	The material taught was well organized	11	84.6	1	7.7	1	7.7
4	Each session was well organized	11	84.6			2	15.4
5	The additional material used (videos, slides, photocopies, etc.) were helpful	12	92.3			1	7.7
6	The module learning outcomes were achieved	10	76.9	1	7.7	2	15.4
7	The module was interesting	10	76.9	1	7.7	2	15.4
8	The module was difficult	5	38.5	2	15.4	6	46.2
9	The module was useful	10	83.3	1	8.3	1	8.3
10	The module required lots of study	6	46.2	5	38.5	2	15.4

Table 3 Students' opinion on M3 CTCI (Core Issues in ELT)

No.	Question	Agree		Not sure		Disagree	
		No.	%	No.	%	No.	%
1	The module learning outcomes were clear	9	75.0	3	25.0		
2	The module material was relevant to the learning outcomes	10	83.3	2	16.7		
3	The material taught was well organized	10	83.3	1	8.3	1	8.3
4	Each session was well organized	10	90.9			1	9.1
5	The additional material used (videos, slides, photocopies, etc.) were helpful	10	90.9			1	9.1
6	The module learning outcomes were achieved	9	81.8	2	18.2		
7	The module was interesting	7	70.0	2	20.0	1	10.0
8	The module was difficult	3	25.0	2	16.7	7	58.3
9	The module was useful	9	75.0	2	16.7	1	8.3
10	The module required lots of study	7	58.3	4	33.3	1	8.3

Table 4 Students' opinion on M4 CTLA (language awareness)

No.	Question	Agree		Not sure		Disagree	
		No.	%	No.	%	No.	%
1	The module learning outcomes were clear	12	100.0				
2	The module material was relevant to the learning outcomes	11	91.7	1	8.3		
3	The material taught was well organized	12	100.0				
4	Each session was well organized	11	91.7	1	8.3		
5	The additional material used (videos, slides, photocopies, etc.) were helpful	12	100.0				
6	The module learning outcomes were achieved	11	91.7	1	8.3		
7	The module was interesting	11	100.0				
8	The module was difficult	4	33.3	2	16.7	6	50.0
9	The module was useful	10	81.8	1	9.1	1	9.1
10	The module required lots of study	9	75.0	1	8.3	2	16.7

relevant to the learning outcomes and the material was well organised. Nine (75.0 %) of the students found the module learning outcomes were clear, 9 (81.8 %) of the subjects *agreed* that the learning outcomes of the module were achieved. Also 9 (75.0 %) of the students found the module useful. The difference in the percentage was due to some missing answers regarding these items. A number of students 7 (58.3 %) believed that the module required lots of study but it was not difficult. Seven (70.0 %) of the subjects stated that it was interesting.

The participants also expressed their opinion on M4 CTLA. The results are shown in Table 4. From the table, it can be seen that all the participants 12 (100 %) *agreed* that the module learning outcomes were clear, the material was well organised, and the additional material used was useful. Almost all of them 11 (91.7 %) found that the material was relevant to the learning outcomes, each session was well organised, the learning outcomes were achieved, and the module was interesting. Ten (81.8 %) of the subjects *agreed* that the module was useful, but they found it demanding since 9 (75.0 %) of them stated that it required lots of study. Nevertheless, only 4 (33.3 %) found it difficult and 6 (50.0 %) found it easy.

It seems that the students were satisfied with the CTEFL programme modules. This implies that the programme meets students' needs for teaching English language. The results also indicate that making students aware of the learning outcomes is an integral part to guide their learning. It is worth mentioning that at the beginning of the programme, the students were provided a leaflet informing them about the CTEFL programme and its learning outcomes. Regarding the content, students found it relevant to the learning outcomes and their level. In the planning stage of the CTEFL, the content was catered to be at a certificate level to suit those who are interested in the programme and who were not specialised in English language. The results also show that students acquired the skills needed to teach the English language. These results agree with what was revealed by Barazaq (2007). She found that the Teachers' Training Programme at the Islamic University of Gaza, Palestine, was quite good and it met students' needs. She also found that the programme equipped students with the necessary skills to teach English. The results also are in agreement with Edwards and Owen's (2002) findings. They concluded that their participants found the MA TESL/TEFL useful and interesting. Similarly, Biyik (2007) revealed that the DELTTP at Anadul University, Turkey, met students' needs.

Questions 11 and 12 tried to elicit students' opinions on the assignments and quizzes in the four modules. The results are presented in Tables 5 and 6.

It is clear from Table 5 that 12 (92.3 %) of the participants *agreed* that the assignments helped them understand the particular subject matter. Eleven students (84.6 %) also agreed that the assignments topics were given on time and there was guidance from the instructors. Ten of them (76.9 %) *agreed* that the deadline for submission/presentation was reasonable and the instructors' feedback was helpful and detailed.

Table 5 Students' opinion on written and oral presentations

11	In case where there were written and/or oral presentations	No.	%	No.	%	No.	%
11a	The topic (s) was given in time	11	84.6	1	7.7	1	7.7
11b	The deadline for submission/presentation was reasonable	10	76.9	1	7.7	2	15.4
11c	There was guidance from the instructors	11	84.6			2	15.4
11d	The instructors' feedback was helpful and detailed	10	76.9	1	7.7	2	15.4
11e	The assignments helped you to understand the particular subject matter	12	92.3			1	7.7

Table 6 Students' opinion on the modules quizzes

12	In case there were quizzes	No.	%	No.	%	No.	%
12a	You were informed well in advance	9	75.0	2	15.4	1	7.7
12b	The errors and the corrections were explained	11	91.7			1	7.7
12c	The quizzes helped you to understand the particular subject matter	12	92.3			1	7.7

As for the quizzes, Table 6 shows that 12 (92.3 %) of the subjects *agreed* that the quizzes were helpful in understanding the particular subject matter. Eleven (91.7 %) of them also believed that errors and corrections were explained. However, it seems that the students were not satisfied with the quiz dates since 9 (75.0 %) of them *agreed* that they were not informed in advance.

From Table 6, it seems that both the instructors and students are aware of the importance of assignments and quizzes to aid student learning. This type of formative assessment can help students understand the module content. Furthermore, it informs students about their performance and instructors about their teaching. The students were satisfied with their instructors' guidance on the assignments. These results disagree with what was revealed by Regmi (2008). He found that instructors in the ELT programme of Kathmandu University assigned homework to students without sufficient practice in the classroom and the students were not satisfied with this.

5.2 Results of the Second Study Objective

The second objective of the study was to discover the students' perception of the teaching process in the CTEFL. This objective was covered in questions 13–17 in the questionnaire. Table 7 sums up the results.

The table shows that 12 (92.3 %) of the participants *agreed* that the instructors were receptive/open to students' questions, consistent in keeping class hours, and their performance was very good. Eleven (84.6 %) of the students also *agreed* that the instructors provided.

Table 7 Students' perception of the teaching in the CTEFL

No.	Question	Agree		Not sure		Disagree	
		No.	%	No.	%	No.	%
13	The instructors were committed to the modules	11	84.6	1	7.7	1	7.7
14	They were receptive/open to students' questions	12	92.3			1	7.7
15	They were consistent in keeping class hours	12	92.3			1	7.7
16	They provided you with additional bibliography	11	84.6	1	7.7	1	7.7
17	The overall performance of the instructors was very good	12	92.3			1	7.7

The fourth is a part time lecturer who is a foreigner but has taught M1 CTLP before. All of them are well qualified, and more importantly, they are dedicated to their modules. They were closely supervised by the programme coordinator. These results agree with what Fordden (1997) found. She concluded that her subjects were satisfied with their instructors' performance since they were all well qualified.

5.3 Results of the Third Study Objective

The third objective of this study was to discover students' evaluation of their own performance in the programme. The results are presented in Fig. 1.

Figure 1 shows that all 13 students (100.0 %) *agreed* that they had no difficulty understanding the modules. Twelve (92.3 %) of them stated that they were satisfied with their performance, they were rarely absent from classes, and they always understood the instructors. Almost all the participants 11 (84.6 %) agreed that they participated in the class discussions. It seems that the total number of students in the programme (17) was advantageous to students. It enabled them to follow up and discuss during sessions. The results also indicate that students were enjoying the modules since they found them useful and interesting, so they were rarely absent from classes. It can be stated that the well qualified instructors can attract students to any programme and make it interactive. In addition, students took responsibility for their own learning. It can therefore, be stated that choosing good teachers is an

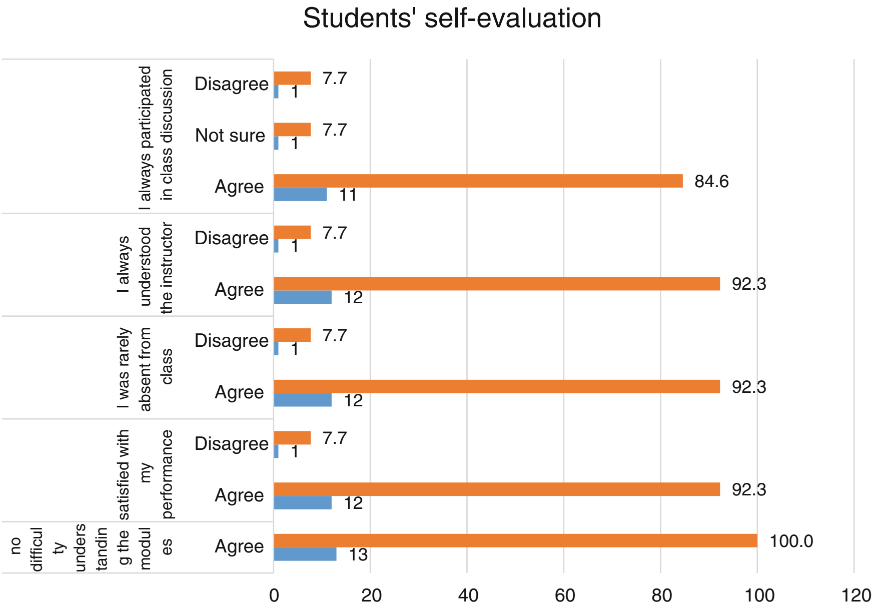


Fig. 1 Students' evaluation of their own performance

integral part of the programme's success. These results are consistent with those of Barazaq (2007), Edwards and Owen (2002), and Fordden (1997) who found that their subjects were satisfied with their performance in the programmes they were studying in.

5.4 Results of the Fourth Study Objective

The fourth objective was to identify how successful they considered the CTEFL programme to be. This objective was covered in item 23 of the questionnaire. It asked students to assign a percentage regarding programme success. Table 8 sums up the results.

It is clear that 6 (50.0 %) of the students suggested that the CTEFL was 90–100 % successful. Nevertheless, the other 3 (25.0 %) regarded the suitable percentage of the success of the CTEFL as 80–89 %, and the other 3 (65–75 %). This is an indication that the CTEFL was successful (90–100 %). The results also indicate that the students were satisfied with the programme. This can be attributed to the interesting, useful, and well-organised modules in addition to the good instructors' performance. The results are in accordance to what was revealed by Edwards and Owen (2002) who found that the MA TESL/TEFL in Birmingham University was successful. Biyik (2007) concluded that the DELTTP in Anadul University was successful in quality, but it suffered from a shortage of time.

Question 24 in the questionnaire asked students to propose any suggestions to develop the CTEFL. Out of 13, 12 (92.3 %) responded to this question. The following is a summary of their suggestions:

- Increase programme time
- More teaching practice
- M1 should be taught by a native speaker
- M2 and M3 should be joined in one module
- There should be real teaching in real classes
- Introducing field visits
- M3 needs more time.

It is clear that students were not satisfied with the microteaching period, which was part of the Core Issues module, which was only two weeks. One week was for mentoring and the second week was for assessment. Students also suggested joining M2 CTTS and M3 CTCI in one module because they are complementary. However,

Table 8 Students' perception of CTEFL success

Suggested percentage	No.	%
90–100	6	50.0
80–89	3	25.0
65–75	3	25.0
Total	12	100.0

the ELI administration could find it illogical to join M2 and M3 as M2 details the teaching of the four skills while M3 focuses on theoretical aspects and methodology of teaching that cannot be included in M2. The suggestions of the students to increase the time for teaching practice and have native speakers as teachers for M1 were all found to be valid by the ELI administration and were put into action.

6 Conclusions and Recommendations

In conclusion, the CTEFL Programme provided by the ELI, University of Khartoum, as the results suggest, is successful. It meets learners' needs and expectations. It contains modules with clear achieved learning outcomes. These make it useful and interesting to students. The teaching process in the CTEFL is very good since it was performed by well-qualified, dedicated, and committed instructors. The students are satisfied with the programme and their performance in it.

However, there is no programme that is absolutely successful. Thus, to further develop the CTEFL, the study recommends the following:

- The microteaching time span should be increased.
- A native speaker is needed to teach M1 CTLP.
- Real teaching situations are required to be incorporated in the programme either inside the ELI or outside in the community.

References

Barazaq, M. Y. (2007). *Students-teachers' training programme evaluation in English language teaching colleges of education in Gaza Strip universities* (Unpublished PhD thesis). Islamic University of Gaza. Online at www.library.iugaza.edu.ps. Accessed December 9, 2014.

Brown, J. D. (2005). *Evaluation*. Oxford: Oxford University Press.

Biyik, C. O. (2007). A preliminary evaluation of the distance English language teachers training programme (DELTTP) in Anadolu University. *Turkish Online Journal of Distance Education-TOJDE*, *8*(1), 143–162. Online at www.tojde.edu.tr. Accessed December 9, 2014.

Crystal, D. (1997). *English as a global language*. Cambridge: Cambridge University Press.

Edwards, C., & Owen, C. (2002). What should go into an MA TEFL programme? Teachers' evaluation of the taught components of a sample programme. *ELTED*, 17. Online at www.bravestar.csv.warwick.ac.uk. Accessed December 9, 2014.

Fordden, C. (1997). Curriculum evaluation: The case of the ELT graduate programme at the UDEA. *KALA Revista de lenguaje y cultura Medelin, 2*(1–2), 18–42.

Regmi, K. D. (2008). Evaluation of ELT programme at Kathmandu University. Online at www.eric.ed.gov/ED529044. Accessed December 10, 2014.

Richards, J. C. (2001). *Curriculum development in language teaching*. Cambridge: Cambridge University Press.

Wagi'alla, A. (1996). English in Sudan. In J. Fishman, A. Cohen, & A. Rubal-Lopez (Eds.), *Post-imperial English: Status change in former British and American Colonies, 1940–1990* (pp. 339–356). New York: Mouton de Gruyter.

An Evaluation of the Challenges of Sudanese Linguistics and English Language-Related Studies' Ph.D. Candidates: An Exploratory Qualitative Study

Awad Alhassan and Holi Ibrahim Holi Ali

Abstract This qualitative evaluative study is based on semi-structured interviews with two Sudanese professor supervisors and three Ph.D. candidates who have recently completed their Ph.D. study in linguistics and English language-related studies at the Graduate College, University of Khartoum, Sudan. The study explored and evaluated the problems and challenges Ph.D. candidates encountered during their candidature from the perspectives of both supervisors and candidates themselves: How do these challenges impact on their study? How do they cope with these challenges? How can these challenges be overcome? The study provided new insights into doctoral education in Sudan, specifically in the University of Khartoum. The study adopted a qualitative methodology with semi-structured face-to-face interviews being the principal method of data collection along with the collection of some institutional documents, some of which are being used during interviews in a discourse-based format. Five tape-recorded interviews were conducted with both candidates and supervisors. Interview data were coded and analyzed inductively. Results of data analysis revealed that there were many problems and challenges doctoral students experienced throughout their Ph.D. candidature, such as supervision-related challenges, resources-related and organizational challenges. In addition, there were a number of strategies candidates reportedly used to deal with these challenges and both candidates and supervisors reportedly held varied perceptions about what makes a good quality Ph.D.. The study recommendations, implications along with its limitations and suggestions for further research were presented and discussed.

Keywords Linguistics · English Language-related studies · Ph.D. candidates · Problems · Challenges · Qualitative · Evaluative

A. Alhassan (✉)
University of Khartoum, Khartoum, Sudan
e-mail: awad_alhassan@hotmail.com

H.I.H. Ali
University of Huddersfield, Huddersfield, UK
e-mail: howlli2@yahoo.com

S. Hidri and C. Coombe (eds.), *Evaluation in Foreign Language Education in the Middle East and North Africa*, Second Language Learning and Teaching,
DOI 10.1007/978-3-319-43234-2_15

1 Introduction

Doctoral education is the highest level of academic qualification someone can attain in higher education and it is the core and fundamental degree of academic practice (Pyhältö, Toom, Stubb, & Lonka, 2012). Engaging in postgraduate research not only entails undertaking the research but also developing research skills in order to become an independent researcher (Brydon & Fleming, 2011, p. 996). The fundamental goal of Ph.D. research in higher education degree programs is to produce independent researchers who are able to adapt to diverse contexts in both the academia and industry (Manathunga & Lant, 2006). Ph.D.s, unlike other less challenging postgraduate degrees in academia, require high quality standards and benchmarking.

A successful Ph.D. candidate should be in full command of the subject area of their research and its current trends of knowledge and debate and they should also be able to extend the debate and contribute to the existing knowledge. They should show originality in their produced Ph.D. thesis. Given such a high status of doctoral degree in academia, the produced Ph.D. thesis should therefore be of a high quality and should meet the required standards and benchmarking. Sudanese linguistics and English-language related studies, Ph.D. candidates, however, seem to have experienced a range of challenges that hinder them from meeting the required high standards and benchmarking often expected to be met in the produced Ph.D. theses.

Previous studies (e.g., Ayiro & Sang, 2011; Dysthe, Samara, & Westrheim, 2006; Edwards, 2002; Hasrati, 2005; Mackinnon, 2004 cited in Gunnarsson, Jonasson, & Billhult, 2013; Löfström & Pyhältö, 2014; Winter, Griffiths, & Green, 2000) have highlighted a range of challenges and problems Ph.D. candidates encounter with their Ph.D. study. The most frequent challenges range from supervision, assessment of high quality Ph.D. these, lack of resources, lack of focus, poor research design to inadequate conceptualization of research questions, inadequate research background, lack of training in methodological and writing skills, and lack of research facilities. The present study attempts to explore these challenges and problems Sudanese Ph.D. candidates encounter during their Ph.D. candidature. The overarching objective of the study is to provide some pedagogical implications to inform the doctoral education research training and development programmes in the Sudanese higher education institutions. The study adopted a qualitative methodology whereby semi-structured intervening was used as the main method for data collection along with the collection of some institutional documents to enhance the study both methodological and analytical triangulation. The study is part of a large project with multiple phases of investigation covering a number of Sudanese universities but the current study reported in this chapter is the first phase and was only confined to the University of Khartoum The implications for the development and sustainment of Ph.D. research training programs will be discussed and recommendations for the development of criteria for high quality Ph.D. theses will be presented.

2 Theoretical Background

In the last decade of the twentieth century, there was a major influx and expansion of doctoral studies undertaken at the higher education institutions in the UK, USA and other countries (Morley, Leonard, & David, 2003). Doctoral education is the highest education level of academic qualification and the foundation for research and development. The key part of undertaking a Ph.D. is to become an independent researcher. Engaging in postgraduate research does not only entail undertaking the research but also developing research skills in order to become a researcher (Brydon & Fleming, 2011, p. 996). The fundamental goal of research in higher education degree programs, particularly doctoral degrees, is to develop independent researchers who are able to adapt to diverse workplace contexts in academe, industry and the profession (Manathunga & Lant, 2006).

PhDs, unlike other less challenging postgraduate degrees in academia, requires high quality standards and benchmarking. A successful Ph.D. candidate should be in full command of the subject area of their research and its current trends of knowledge and debates and they should also be able to extend the debates and contribute to the existing knowledge. However, the Ph.D. degree candidates in Sudanese higher education institutions seem to have fallen short of these standards and requirements. They seem to have experienced a range of challenges and difficulties during the course of their Ph.D. candidature that hinder them from attaining the required standards. Firstly, supervision-related challenges such as lack of expertise and experience on the part of supervisors. Secondly, resources-related challenges which include a lack of sufficient resources and research facilities. Thirdly, there are also some significant other challenges such as the lack of funding, lack of research training and development programs on both the methodology and academic writing levels, candidates' inadequate research background knowledge, and the uncontrolled growth of the number of Sudanese postgraduate students wishing to pursue Ph.D. study, etc.

Producing a high quality Ph.D. thesis is the core element in doctoral education. However, there seems to be a practical problem facing Ph.D. candidates and supervisors in higher education institutions as to how to produce and judge Ph.D. research quality (Winter, Griffiths, & Green, 2000). Quality research can be defined operationally as the research that is completed on time and have a rigorous research design which is internally and externally valid, based on a reliable data sources, using appropriate analytical methods which are meaningful (Mahmood, 2011). The issue of monitoring and benchmarking the quality of Ph.D. theses has been studied widely and it needs to be addressed (Kyvik & Thune, 2014). Measuring the quality of scientific output is traditionally done by using peer review and scientific methods. Judging the quality of a Ph.D. thesis and finding the appropriate and explicit criteria for assessment is not an easy task. For example, Gulbrandsen (2000) argued that the concept of research quality should be divided into quality elements which demonstrate different criteria of good research and which could be extended to cover Ph.D. theses and other pieces of research. In the same line Marsh, Rowe, and

Martin (2002) noted that establishing appropriate benchmarking criteria or framework for measuring the Ph.D. effectiveness is not an easy task. However, there are several main indicators which can be used to measure Ph.D. effectiveness and quality. The criteria should include the following: Originality, solidity, scholarly/scientific relevance and practical/societal utility. From a research quality perspective, originality is a common concept which implies novelty in relation to current existing knowledge and theory.

Original research contributes to new perspectives, data or methods. However, what exactly constitutes originality in doctoral theses is open to different interpretation. Additionally, the solidity element includes the idea of stringency, validity, reliability, correctness, truthfulness and consistency. This element is closely related to the mastery of a body of scientific knowledge and appropriate methodologies. In doctoral theses, elements such as the structure of the arguments and the manner of the thesis presentation are also seen as important quality criteria when assessing theses (e.g., Mullins & Kiley, 2002, cited in Kyvik & Thune, 2014). Moreover, scholarly relevance comes as part of research quality. This may include that the research problem, theory, methodology or results must be interesting to other researchers in the same or similar fields. Practical utility is a fourth element of research quality which deals with the external or extra-scientific relevance. This means the research should be of interest, importance or utility not only to the scientific community, but also to specific users or society in general. All these elements are mutually important for measuring research quality.

3 Ph.D. Supervision

Ph.D. education is the core and fundamental degree of academic practice (Pyhältö et al., 2012). Supervision is a key element for the successful of Ph.D. journey. It is, however, a pedagogical challenge in higher education (Gunnarsson et al., 2013). The success of the Ph.D. degree depends on supervisors. They must provide expertise, time, and support to foster in the candidate the skills of and attitudes towards research, and to ensure the production of a thesis is of an acceptable standard (Heath, 2002). It is widely believed that supervision should be approached from pedagogical perspectives (Zeegers & Barron, 2012). Ph.D. supervision has two major dimensions: The involvement of the supervisors in the provision of intellectual expertise to students, and their involvement in counselling students and boosting their confidence and morale (Hockey, 1994). Moreover, doctoral supervision provides a potential arena for learning to identify problems that arise during the study and solve them in an ethically and sustainable manner (Löfström & Pyhältö, 2014). Supervision has been identified as one of the most important determinants of doctoral studies and good doctoral supervision is viewed as central to the achievement of positive outcomes from research education (Halse & Malfory, 2010).

Lee (2008) offers five dimensions for the role of supervisors in dissertation writing: Identifying functional aspects (project management), enculturation (encouraging the student to become a member of the academic community), critical thinking (encouraging the students to question and analyze their work), emancipation (getting the students to question and develop themselves) and developing a quality-relationship whereby the student is inspired, nurtured and cared for. The features or elements constitute a part of an apprenticeship model for supervision. The supervisors' role is to bring down the students from their professional pedestal, as a process of status 'deconstruction', so that they can progress as researchers (Watts, 2009).

However, there are some supervision-related problems which can hinder the Ph.D. progress and completion. Problems, such as lack of supervision, overdependence on supervisors, and being at cross-purposes with supervisors are reported in the literature as the main causes to problems such as prolongation of studies, lower level of well-being and dropping out (e.g., Dysthe, Samara, & Westrheim, 2006; Edwards, 2002; Hasrati, 2005; Mackinnon, 2004 cited in Löfström & Pyhältö, 2014). Other problems which have been reported are supervisory relationships, including lack of supervision or interpersonal friction.

Further Gunnarsson et al. (2013) have reported some supervision-related problems such as an inadequately low supervisory meeting frequency resulting in a stressful and lonely walk Ph.D. education journey. Many academics have reportedly said that study for a Ph.D. and supervising it is a really complex task (Denicolo, 2003). The most fundamental problem which is repeatedly discussed in the literature and which is encountered by supervisors is the lack of motivation among some of their Ph.D. students (Hockey, 1996). Additionally, previous studies on doctoral students' experience and difficulties report that attrition rate, distress and disengagement are the most encountered challenges (Mahmood, 2011; Pyhältö et al., 2012). Moreover, Ayiro and Sang (2011) conducted a qualitative study based on 52 Ph.D. candidates and 60 academics in Kenyan public universities to explore the challenges that Kenyan Ph.D. candidates experienced during the course of their studies and the sources of these challenges and difficulties. The study was primarily aimed to enhance *quality assurance processes* in the award of PhDs by Kenyan universities. Some of the major sources of the problems reported by participants are: Lack of focus, poor research design, inadequate conceptualization of research questions, inadequate research background, lack of training in methodological and writing skills, and lack of research facilities. In the same line, McCarthy, Hegarty, Savage, and Fitzpatrick (2010) noted that Ph.D. candidates may experience challenges in establishing their conceptual frameworks, methodological issues, ethical dilemmas and even accessing their study participants.

4 Context of the Study

The study was conducted in Sudan. The Republic of the Sudan (henceforth Sudan) is an African sub-Saharan country situated in the North-east of Africa and bordered by seven countries. Standard Arabic is the official language of the country while

English is the second official language. Besides these two languages, there are numerous indigenous languages spoken in Sudan. English is a foreign language in Sudan, but it is used as a second official language after Arabic in the official transactions of the governmental institutions. It is also taught as a subject in schools and universities and it is used in some Sudanese higher education institutions as the medium of instruction and assessment in certain disciplines.

English was used as the medium of instruction in higher education institutions until 1990 when new policies of Arabicisation[1] were introduced whereby English was replaced by Arabic as a medium of instruction in higher education. These policies stipulated that the higher education curricula must be taught and/or translated into Arabic and Arabic must be used as a medium of instruction and assessment instead of English in all undergraduate and postgraduate programs in the higher education institutions. Despite this English has remained the medium of instruction and assessment in a number of postgraduate programs in these institutions. Due to its status and historical international links with English speaking universities, especially in Britain, the University of Khartoum in particular has retained the use of English as the medium of instruction and assessment in a number of postgraduate programs run in different faculties and institutes of the university.

Comments: I guess all the previous paragraphs under the heading of “context of the study” should compacted into two short paragraphs. I guess no need for all these sub-titles.

The graduate school at the University of Khartoum was established in 1973. Since its inception, it has been offering Ph.D. degree programmes in Linguistics and English-language related studies in three departments housed in two faculties at the University of Khartoum. Since 1978 to 1993, seventy-five Ph.D. candidates have completed their Ph.D. study at these three departments. The theses are divided into three broad areas namely, linguistics, English Language and educational studies. There are about 15 theses in linguistics, 12 in education and 48 in English language. The first Ph.D. thesis in English was completed in 1978. From 1993 to present there seems to have been an upsurge in the number of Ph.D. candidates as there are seventy-four Ph.D. theses that have been completed since then.

5 Significance of the Study

This study as investigated the challenges and needs of the Sudanese Ph.D. candidates doing research in linguistics and English Language-related areas of study in the Sudanese higher education institutions. It is a large project with multiple phases of investigation covering a number of Sudanese universities but the current study

[1] Arabicisation is the use of Arabic as a sole medium of instruction in the higher education intuitions. The policies were introduced in the 1970s but practically came into effect in the 1990s.

reported in this chapter is the first phase and will only be confined to the University of Khartoum. The implications of the findings of the study are intended to better inform the Ph.D. supervision and research training and development programs to maintain high quality of Ph.D. theses that meet the international quality assurance and benchmarks. The implications have an ecological significance for the doctoral education in the context of the study as, to the best of our knowledge, there has not been any study so far conducted on this topic in the context. The study is, therefore, pioneering and would open up potential avenues for more future search in the Sudanese context and across the region.

6 Methodology

In this section we explain the research design of the study and the qualitative methodological approach adopted and justify this methodological choice. We also explain the process used for the recruitment of the study participants both supervisors and candidates. The section will also include the description and discussion of the methods used for the data collection. We will conclude the section by discussing the process applied to the data transcription, coding and analysis procedures including inter-rater reliability checks. Five participants (two supervisors and three candidates) took part in the study. Supervisors had at least ten years of experience in supervising Ph.D. degree and candidates have recently completed their Ph.D. theses (less than two years).

The project was explained in writing to the participants and they were informed that their participation was voluntary, and they were freely able to withdraw from the study at any time. A written informant consent form was obtained from all participants. As a characteristic of being exploratory, the study used semi-structured interviews as a principal method of data collection along with some documentary data to enhance triangulation. To gain the insider participants' *emic* perspectives and thereby lessen the outsider researcher's *etic*/outsider perspectives on the issues under investigation, the study adopted a face-to-face interview method, which would help the researcher gain the participants' "views, understandings, interpretations [as well as] experiences" (Mason, 2002, p. 63). Qualitative interviews can be semi-structured or open depending on the purpose of the research (Dörnyei, 2007; Kvale, 1996), and the present study adopted the semi structured interview format since the purpose was to explore and discover as many issues as possible from both supervisors and candidates regarding the issues under investigation in the context of the study. Besides the use of interviews, the study also used some documents to enrich data collection. A range of documents were collected and analyzed. The documents helped triangulate the data as they were used in the interviews in a discourse-based format with the study participants (for further discussion of discourse-based interviewing, see e.g., Lillis, 2001; Odell, Goswami, & Herrington, 1983). The documents included the graduate school higher degree handbook and regulations, rules, policies and some statistics on the number of Ph.D. degrees awarded from 1970s to date.

6.1 Coding of Interview Data Transcripts

We adopted an exploratory open strategy to code our data in the sense that we coded everything so that we could discover as many potential issues as possible from the data. We began by reading the interview transcripts, summarizing them, or 'discover[ing] particular events, key words, processes, or characters that capture the essence of the piece' (Coffey & Atkinson, 1996, p. 31). We had two sets of interview transcripts: Two supervisor participants' transcripts and three candidate participants' transcripts. We transcribed all interview recordings verbatim. Having finished transcription, we chose two representative interview transcripts, one from each set of data, and we read through and summarized the topics addressed by the informants while making crude codes. This summary was done manually in the margins of the text using Microsoft Word's *add new comment* function. For example, the following summary code was initially made for the chunk of text below from one of the supervisors' representative interview transcripts: *Supervisors*' *views on the types of challenges and problems*

> They are not themselves trained to research. They never wrote in their undergraduate studies. They never wrote long essays or short essays in the term, in the time of four years' time they, they, they study. So when they come to write research they don't know how to do it. People don't know how to compile or how to develop a paragraph. They don't know how to develop a paragraph. They know nothing about the simple idea of opening and ending that paragraph (…) (S1).

We then read again through the remaining interview transcripts of both supervisor and candidate participants and applied these summary codes. After adding the summary codes to the remaining interview transcripts, we went back and read again carefully and closely through the two representative interview transcripts for both supervisor and candidate participants and using Microsoft Word's *add new comment* function once again, we added besides the summary/crude codes new refined codes in order to make the codes/themes more representative and more accurate. So, for instance, to the same above chunk of text, we added the following refined code: *Challenges and problems*. Such new codes were again applied to the remaining supervisor and candidate participants' interview transcripts. *The challenges and problems* code then became the main code for all types of challenges and problems including a number of sub-codes representing the range of the types of these challenges and problems which were adequately covered in the analysis section.

As we were planning to conduct inter-reliability checks with a second coder, we again refined and rewrote our codes to be more transparent and more reader friendly, and added shorthand codes for convenience. Instead of asking, for instance, the second rater to assign a code of challenges and problems to the bits of text in the transcripts when supervisors talk about the challenges and problems of candidates, this became a simple shorthand and user-friendly code, *CHALLPROB*, with the code definition clearly provided.

Having finalized the two lists of codes for both supervisor and candidate participants, we then sent two typical and representative interview transcripts along with the two lists of codes to a second coder and asked them to try the codes on these two transcripts. The second coder was a Ph.D. student of applied linguistics who was familiar with qualitative research as they themselves were using qualitative research methodology for their study. After the second coder/rater had finished coding, we met with them to calculate the percentage of inter-rater reliability and secondly to discuss and resolve our coding disagreements.

Simple percentages of agreement and disagreement were calculated by dividing the number of coding agreements over the total number of coding episodes multiplied by hundred (number of agreements/total number of coding episodes x 100). The disagreements were counted and documented so that we would later discuss and resolve them. The percentage of our agreement on the teacher supervisor' interview transcript was 90 % and on candidates participants' interview transcript 88 %.

While some researchers (e.g., Coffey & Atkinson, 1996) in the literature of qualitative research warned against using coding as synonymous with analysis other researchers (e.g., Miles & Huberman, 1994) have considered the coding of qualitative data to be analysis. In this study we differentiated between the two processes. We first coded our data into categories and themes and we then moved on to further analysing these themes by establishing more linkages and connections between them by comparing and contrasting participants' views on the issues under investigation. To put it simply, coding brings interview data on an idea or theme together; analysis 'lies in establishing and thinking about such linkages (…) how we use the coding and concepts' (Coffey & Atkinson, 1996, p. 27).

Having explained the coding and the analytical procedures in the previous section whereby the final themes and categories in the data were identified for further analysis, in this section we will interpret and explain the data by establishing comparisons, contrasts, and linkages with reference to our research questions. The analysis will be presented thematically according to the study reach questions. Data transformation is done manually and undertaken through the identification of emerging themes and codes. Thematic analysis is employed because it has the potential to produce diverse interpretations of the data and offers more insightful interpretations to the data in question (Braun & Clarke, 2006).

7 Results

This section will only be devoted to the data analysis and a separate section will follow where the results will be discussed and connected with the relevant literature.

RQ1. What are the challenges Sudanese Linguistics and English-related studies Ph.D. candidates encounter throughout their Ph.D. candidature and how do these challenges affect the quality of the produced Ph.D. thesis?

Both supervisors and candidates reported a range of challenges and difficulties that Ph.D. candidates experienced with their Ph.D. study. The section below summarises these challenges, their sources, their negative impact on the candidates' performance to produce good quality Ph.D. thesis, and the coping strategies students used to try and overcome these problems. The main challenges which were reported are: Supervision-related challenges, resources-related challenges and organizational-related challenges. The results suggest that there was a variety of challenges that were encountered by Ph.D. students during their candidature. The candidates highlighted a number of problems throughout their candidature which would affect the quality of their produced theses. *Supervision problems were most frequently mentioned including the lack of guidance*:

> *We don't have guidance.* You keep reading and reading a lot [in the literature]. There is *no guidance,* there is no one to ask, there is no one to refer to, just trying to *find your own way,* this is my biggest problem from the beginning (C1).

This candidate carried on to report problems with academic writing Ph.D. candidates encountered and that the Ph.D. study did not train them to be good writers to develop beyond the Ph.D. study due to lack of feedback on writing:

> (…) even people, when after finishing their thesis, they *haven't mastered techniques* of dissertation writing (…) Why? Because we don't have plans. We don't have someone to say, this is wrong. This is right. This is supposed to be done like that because of that and so on (…) You have to go and read and find your own way (…) You can find it hard to be (…) but at last you (…) the [lack of training on] techniques [of academic writing] is our problem (C1).

Candidate's (C1) comment shows that lack of guidance and lack of planning are some of the challenges that they personally encountered during their Ph.D. study and it is clear that C1 struggled with their Ph.D. and this would clearly have a negative impact on their study. It can be seen from the above quotations that the lack of guidance and planning was a major challenge for them along with the lack of academic writing skills and techniques which were needed not only for successful Ph.D. writing up but also for carrying on and developing as a publishing researcher beyond the Ph.D..

Similarly, candidate (C2) reported problems regarding lack of subject specialist supervision and thus lack of guidance:

> [My supervisor told me], I personally cannot help you. I just help you in the *technical way* of doing the research but you will be *responsible for everything* regarding the topic of your Ph.D.. I seek no help from anyone but I had to dig for my own [way]. That's one. And, the second challenge was how to find the sources for writing (C2).

Moreover, this candidate went on and reported that even the external examiner of their thesis did not have the relevant expertise on the subject area of their Ph.D. area of research:*My external examiner* just dealt with it [my Ph.D. thesis] as research so that is just like my,

> my supervisor. *They look at the structure* of research and whether I followed the methodology of research or not that is because *they don't have idea about the content* (C2).

Furthermore, in the same line, candidate C2 reported supervisor's lack of expertise in the research area of their Ph.D. to the extent that supervisor suggested change of the first chosen topic:

> The first challenge was when I wrote this [my research proposal], the supervisor said, "*This is a good field and you might be the pioneer in this field and you will never find anybody here who will help you in this field because as far as I know no one has supervised this area before.*" So, this is an altogether new area. But when I came after a year the supervisor changed their mind and said that, "*No, I don't want you to* (…)" they said they didn't want me to write on this field and have to choose any other field (C2).

So apparently, this lack of expertise and thus instability in supervision would have negative impact on the candidates' motivation and focus which would in turn result in poor quality produced Ph.D. theses.

Similarly, C1 reported that this lack of expertise and subject specialism on the part of supervisors resulted in having no content-related comments and feedback on their work:

> There is no feedback. Supervisors *only give comments on the language and language correction. Nothing has to do with the subject* or on research nor writing as how to develop discussion and arguments and all these kinds of things (C1).

This lack of expertise at the part of some supervisors was also confirmed by the Supervisors themselves:

> Unfortunately, some people now are supervising students who *are not qualified themselves.* You see a number of supervisors who *never care to write research anywhere.* They didn't have any [research] article in any periodical. They *never attended any international or local conference* and they *supervise PhDs*. This is a problem. *Those people who you call professors*, or whom we call professors, are *not qualified.* A professor should be qualified, *should be a researcher, they should be research-oriented themselves* (S2).

Students also reported supervision instability and sometimes they had to change supervisors which might be problematic for them due to the difference in supervision styles among supervisors:

> I felt that it's going to be *very difficult actually to change the supervisor* because as you know *every supervisor has their own way [of supervision]* (…) and their own programme [and schedule] to look at the work and so on. I had already finished most of the part with Dr X [supervisor] when, Dr Z [a new supervisor] was nominated actually to be my supervisor. It was just to ensure that I finish and then write the final report to the Graduate College (C3).

C3 candidate illustrates that there was instability in their supervisors because they kept changing their supervisors and this could be one of the challenges that encountered by during their study. Candidates also reported facing some challenges in getting cooperative, supportive, close and timely supervision:

> I remember I was planning to write up a chapter. That chapter took three years to get finalized. *I was trying to establish some conceptual framework for my methodological choice* and found it difficult. [Before I start working on it]. *I went to see my supervisor to just discuss the general plan with them for writing up the chapter on what to be included*

> *and so on. They told me 'don't ask me, you have to go and write the chapter and then bring it to me and after that we can see. I was just asking for discussion* (C1).

Comparing Ph.D. candidates' and supervisors' views we might gain a clear picture about these challenges and how to deal with them. A supervisor (S1) illustrates that there are many challenges that Ph.D. encounter with the methodological issues:

> One of the problems, I'm lucky enough to have most of my students were competent in English, and they don't have language problems, but the first problem you could observe, that they lack knowledge on research methodology as well as skills, they come with very little research experience as well as knowledge. They face many problems, regarding where they understand the basic preliminary words and drafting techniques of research proposals (S1).

However, another supervisor's (S2) response contradicts the above mentioned supervisor's views in many ways such as some candidates' low level of proficiency in English Language is one of the challenges which were highlighted by supervisors in addition to the lack of background knowledge in certain areas which could otherwise help in writing their theses successfully:

> They [candidates] come from *very poor backgrounds* (…) *they don't have good English proficiency*. They can't write, they can't speak English. If you ask a candidate how many books you have read in your life time in English, they would say a couple of books, you see. They didn't have this, the real guts to do it. Their English is broken, they cannot express themselves in English, spoken or written. They don't have access to the English culture, to the British or American culture. I was wondering how someone [for instance] could hand in a Ph.D. in literature [when] they come from a very poor, geographical background. They never gone to the cinema, they have never seen a film before; he/she never listened to a song. How come would they be accepted for a Ph.D. study, would they be able, to write about the literature of those people? So they will depend on two or three books of criticism and they will start to copy (S2).

This supervisor went on to comment that candidates themselves are part of the problem because they commit themselves to other duties and they do not have time for their study.

> They don't have problems. The problems are themselves. They are not, they don't have *free time for research*, they are all teaching or working, you see, they don't devote any time for research. They are working all day from 8 to 5. *When are they going to sit to study and research?* (S2).

Clearly, this lack of language proficiency coupled with lack of general knowledge and lack of devotion to and commitment to study on the part of the candidates could negatively impact on the quality of Ph.D. research produced by those candidates.

Another supervisor (S1) agreed with these comments but they seem to put that in a sympathetic way:

> The main problem is that they are not full-time students. They are working, supporting themselves, their families, and even the pressing economic circumstances have [have an impact] on them (S1).

Lack of academic reading and writing skills on the part of candidates was also highlighted by supervisors:

> *They never read periodicals*. They never read (…) They depend on books and you know how difficult it is to make a Ph.D. out of books. That is one thing. Secondly, they are not themselves trained to research. They never wrote in their undergraduate studies. *They never wrote long essays or short essays in the term, in the time offour years' time they, they, they study. So when they come to write research they don't know how to do it* (…) *people don't know how to compile or how to develop a paragraph* (…) *they don't know how to develop a paragraph.* They know nothing about the simple idea of opening and ending that paragraph. They have never been to any course of criticism, neither theoretical nor practical. *But how can you write [a Ph.D. thesis] if you have never been to this kind of courses* (S2).

RQ2. How do they respond to these challenges?

Candidates used a range of survival strategies to cope with these challenges and had their theses successfully completed. One of the strategies was seeking help from pervious Ph.D. candidates and/or colleagues:

> While I was in my office, one of my colleagues came in. He just paid me a short visit and I told him that, I start for Ph.D. but unfortunately I met a lot of challenges and I don't want to give up or surrender. So, I just want to find a way to find access to books. He said, "What was your problem?" I told him that my problem was a problem of resources, and he said that wasn't a problem and that he would help me. And he asked if I knew the books that I want to use. I said, "Yes, I have about five books that I need." Then he invited me to his house and we went there and within half an hour he downloaded four out of the five books that I had ordered through the British Council. And he showed me a website for free books download (C2).

Another survival and coping strategy reported by candidates is the persistence and perseverance to obtain guidance and close supervision from supervisors:

> *I'm chasing [my supervisor] every week* like that call them and say "I have something ready have you finished reading the first one?" Sometimes they say, "No." Sometimes they just can't bring it to me. But I keep on asking them like that. So, *I go to them at their office. Sometimes I meet them at their home* (C2).

The candidates also reported that they were reading previous theses and following similar styles and formats and generally try to find similar theses to help them curve out their topics and arguments:

> I try to overcome these problems [lack of guidance]. I tried to choose one of the theses that have been done outside Sudan for guidance. I spent one complete year just reading and trying to find my way (C1).

Additionally, they use the internet to find similar theses and they start curving out their topics and arguments. In contrast, supervisors reported that they followed many strategies to help their candidates to complete their theses successfully by providing them with books, references and other materials as well as close supervision throughout from the very early stages of the Ph.D. candidature:

> Before conducting their research and before reading [the] literature, and them processing to the actual writing. *I direct them to focus on literature review first,* like six months, so they're reading, and then they come up, give me a summary, oral presentation of what they

> have done regarding and [reviewing] the literature, and then proceeding to actual writing. When they write, *I take the whole work, read it carefully, have all my comments written,* and then I ask the student to meet me and we sit together for two to three hours, going over the work page by page, sentence by sentence, highlighting the problems and showing them how to solve them (S1).

Supervisor (S2) went on in the same line reportedly highlighted that they help their supervisees in formulating their research topics:

> [One of my Ph.D. students said to me] "I don't know but I like phonology but I don't know how to do it. What shall I do, I don't, I don't find a [topic] for my Ph.D.." I said, "Do you have a recorder?" [They] said, "Yes." "Do you know a nearby nursery school?" [They] said, "Yes." I said, "Go, please, to that nursery school, hide your recorder, play with the children in Arabic and speak to them in Arabic and let the teacher speak to them in Arabic and decide the age. Take from 3 to 5 year-olds. When you finish two hours of recording, come back to me." [They] said, "But-" I said, "No. Can you do that?" [They] did it. I said, "Now this is your Ph.D.." "What shall I do with it?" "Go and listen. Put your theme at this. You want know at what age those children can pronounce the Arabic sounds" (S2).

All these problems and challenges described above seem to have their impact on the quality of the produced Ph.D. theses as we will see in the following section.

RQ3: *What constitutes a good Ph.D.*?

This question attempts to explore both Ph.D. candidates' and supervisors' views about what makes a good and high quality Ph.D. according to their own understanding. A supervisor (S1) illustrates that a good Ph.D. is the one which has:

> For me, a good quality Ph.D. thesis, first of all, depends on, *finding an original topic, originality is very important,* and, when I have a new student, the first thing I ask them is what do you want to do? and then, the other question would be, *what are you going to add to knowledge?* what is your expected *contribution?* And then I don't normally accept the answers they give, I just ask them, go and read literature thoroughly, and then come back, *see what people have done, and what is left for you to add* (S1).

It can be seen from the above that this supervisor's views about what makes a good Ph.D. are originality and contribution. It appears that establishing originality in a thesis is something of a paramount importance but Ph.D. candidates face many challenges which may impede them to meet the standards and criteria to achieve originality. Another supervisor (S2) states their views about the deterioration of Ph. D. education due to the above mentioned challenges. They reported that the lenient admission criteria for the Ph.D. degree in turn resulted in low-quality produced Ph. D. theses. They also reported an alarming recent upsurge in the number of people who wish to do Ph.D.:

> There is a big problem in the Sudan. Not all of those students who are accepted to do Ph.D.s are qualified to do them, especially in English language. Their English is below the standard. If I had the authority I would not let them do even a postgraduate degree. Doing a Ph. D. in the Sudan is easy now. Anyone can apply for Ph.D. and they will get into it, just like that. If you go to the different universities and you pick any thesis, you will see how things are below the standard and let me say this is a scam, it's not English at all (S2).

Supervisors also called for rethinking and revolutionising the existing research methodology paradigm in the context of the study which reportedly seems to be contributing to the production of low quality Ph.D. these:

> These descriptions [PhDs] are not good and also this *fashion of questionnaire, questionnaire, questionnaire*, all researchers in this country [Sudan] now, in all fields, use questionnaires. They give the questionnaire blindly to the students or to teachers and they answer them. *They produce them out [questionnaire data] in very beautiful diagrams and graphs and tables and that's it, it's a Ph.D.* (S2).

To maintain high quality Ph.D. thesis supervisors also called for limiting Ph.D. education to the main three universities in the country [Khartoum university is one of them] and these universities should make resources and facilities available to postgraduate students besides tightening their Ph.D. admission criteria. They also warmed about the duplication of Ph.D. topics due to the lack of databases shared among universities.

> *Not all universities have the right to offer postgraduate degrees*. Postgraduate studies should be limited to very few universities. For example, *the University of Khartoum, University of Al Jazeera and University of Sudan for Science and technology, full stop.* But these universities including the University of Khartoum should have a postgraduate residence, a postgraduate library and they should be up to date. They should have library (…). *They should not accept anybody for Ph.D. research.* They should accept distinguished students, and therefore *they should write at least twenty pages of research proposal and has to be examined by a panel of professors and they should also have what we call research bank or thesis dissertation titles bank.* Because now this research is repeated, you know, because now by writing only two or three pages of research proposal you can get accepted into a Ph.D. programme and you start off your study. Such *PhDs could be duplicated anywhere. So we have now duplications, plenty of PhDs, the same topic is dealt with in different universities and they have [different] PhDs* (S2).

The same supervisor went on to highlight their own strict way of upholding the standards of high quality Ph.D. regardless of the institutional policies by taking only the best and few number of Ph.D. candidates at a time:

> *I don't accept* (…) *anybody to do Ph.D.. While [some] people [may] handle twenty candidates,* I may have three or four at a time but they are selected. If I select somebody and I feel that [they are] not up to the standard of Ph.D., I will cross [their] name out and this happened. After [they] spent two years with me, I said to [them], "You can't do it" (S2).

Additionally, students reported that maintaining the quality of Ph.D. research entails cooperative, supportive and high quality supervision:

> To raise the quality of the [Ph.D.] research here, *I think we have to appoint a foreign supervisor to be a co-supervisor.* It has to be someone *from outside Sudan.* You know from experience, *foreign supervisors are kind* with the *information* they give you and *their time* and they just *give everything* (C1).

Students also suggested that training Ph.D. students on academic writing conventions could also help improve the quality of Ph.D. research:

> If we want to improve *our higher degree studies, we have to focus on providing courses on research writing.* Actual writing, how to write, *how to develop discussion and argument*, just all these things. *We do not have such kinds of things and experiences* (C1).

Clearly, from the above quotations of both supervisors and candidates, it seems that there are concerns about the quality of Ph.D. research in the context of the study. Participants, however, reported different perceptions regarding the elements of quality in Ph.D. research.

8 Discussion

The findings on the need for supervisors with expertise and subject specialism suggest a significant correlation between good supervision and success and sustainability of high quality Ph.D. thesis. The findings resonate with the views and findings of many researchers (e.g., Halse & Malfory, 2010; Heath, 2002; Löfström & Pyhältö, 2014) in that supervision and the role of subject-specialist supervisors are highly important as they can be the guardians of quality assurance of the produced Ph.D. theses.

The findings on that candidates are left without guidance to struggle with their Ph.D. candidature corroborates with similar findings in the literature (e.g., Fergie, Beek, McKenna, & Creme, 2011; Krase, 2007). One of Fergie et al.'s (2011, p. 236), participants reported lack of guidance and support with their Ph.D. candidature "You're on your own, and it [Ph.D.] requires a great deal of diligence and discipline, and it's a lonely walk." Our participants in the current study reported similar challenges but their challenges are compound in nature. Fegrie et al.,'s (2011) study was conducted to explore experiences of five Ph.D. students at the University College London taking a writing module on 'developing a literature review'. The study focused on writing rather than supervision challenges since the issue of good and adequate specialist supervision is taken for granted in the context of their study. In contrast, the findings in the context of the current study suggest that challenges encountered by the Ph.D. candidates are of a compound nature as they involve both lack of supervision with expertise and subject specialism on the chosen area of the Ph.D. research and the lack of support and advice on the academic writing skills and conventions. Moreover, the call by candidates for the appointment of foreign co-supervisors who are perceived to be more cooperative and supportive than the national/local supervisors suggest some socio-cultural implications to be considered in any plans for professional training of national/local supervisors.

To sum up, the main findings were summarized and discussed in relation to the relevant literature. Several challenges were reported the candidates such as supervision-related challenges, resources related challenges and other significant ones. Some key issues such as technical supervision or supervising by general experience of research process versus supervising by expertise and specialist knowledge of the Ph.D. topic were elaborated.

9 Conclusions, Recommendations, Implications and Limitations

The present study was able to provide insights and deep understanding to Ph.D. students' perceptions about some of the challenges that they encountered during their candidature. Based on the study findings, the study concludes that Ph.D. candidates in questions have encountered several challenges and difficulties during their course of the study and these challenges seem to have negative impact on the quality of the candidates' produced Ph.D. theses and Ph.D. education in general.

The most common challenges which emerged from the participants' responses are: First, intellectual challenges such as difficulty on how to choose the study area and finding a topic, difficulty in data collection, lack of research culture and training, topics rejection by supervisors, ability to reflect on, report on one's progress,, constructing a theoretical/conceptual framework that guide the research, getting finished on time: Poor submission and completion rates, and maintaining high-level academic and scholarly writing. Second, supervision-related challenges such as ineffective supervision, insufficient guidance, lack of constructive feedback, difficulty in finding supervisors, changing supervisors/instability in supervisors, inadequate supervision, supervisors' lack of expertise and specialist subject knowledge in the chosen Ph.D. research area, lack of effective joint supervision schemes, unsuccessful supervision cycle, busy supervisors, lack of progress, fragmented communication with supervisors, lack of timely feedback, lack of motivation as well as lack of appropriate means of communication with supervisors.

Third, organizational challenges which are: Lack of planning and focus, heavy workloads, problems in attending supervisory meetings, timetabling and time management, disengagement with research and research community, balancing research commitments with other commitments, and lack of time for reading. Finally, resources-related challenges such as lack of references, lack of access to research databases, lack of editing and proof-reading skills, and challenges in finding relevant literature. Regarding participants' perceptions about what makes a good quality Ph.D.; this can be summarized as follows: Originality and novelty, should be based on a genuine problem, can answer real life problems, easy to read and understand, has contribution to the existing knowledge of the field, well-organized, written in plain English, based on rich data and guided by sound theoretical framework, supportive and cooperative supervision, foreign co-supervision, and training on academic writing skills and conventions. As for candidates coping strategies in dealing with the reported challenges, a number of strategies were reported such as seeking help and resources from others such as previous Ph.D. students and other staff members. Furthermore, they tend to copy writing styles from previous Ph.D. theses, contacting previous Ph.D. survivals, etc.

Based on the findings of the study number of implications and recommendations can be presented and discussed. The results suggest that Ph.D. education in the context of the study needs a reform in a range of aspects (e.g., admission regulations, supervisors' and candidates' training, revising supervision procedures, making resources available and accessible, and using the internationally-recognized benchmarking standards and criteria in Ph.D. award and education in general). As we have highlighted above, the challenges seem to have negative impact on the candidates' Ph.D. theses quality and Ph.D. education in general. The results indicate that these challenges can be overcome by modifying and updating Ph.D. award rules and regulations and using the internationally-recognized benchmarking standards. Moreover, the results suggest the need for research training and development programs to be introduced to help Ph.D. candidates with both research process (methods and methodology) and research product (academic writing and best practices of writing up Ph.D. thesis). Ph.D. candidates should be encouraged to publish throughout their candidature in peer-reviewed journals and co-authoring with their colleagues and supervisors should also be encouraged. Additionally, transferrable skills related to career development should be introduced and emphasized. Further, plagiarism detection software should be used to help students to learn about plagiarism in order to avoid the practice. Supervisory boards should be set up to regularly monitor candidates' progress.

These supervisory boards should include the main supervisor and other two supervisors, one acting as an advisor and the other as a head for the board. Supervisors should be nominated and assigned according to their areas of expertise and specialist knowledge in the area of Ph.D. research topics and the number of supervisees should not exceed more than five at a time. Graduate college of the University of Khartoum should organize annual postgraduate conferences, workshops, seminars, symposia to offer a platform for postgraduate students where they can share their ideas and report their preliminary findings at different stages of progress, and to get timely feedback about what they have done in their ongoing Ph.D. projects. Funds should be made available to help Ph.D. candidates to finish their degrees on time. Finally, resources should be made available and subscription to international peer-reviewed journals and famous database should be made to guarantee access to help candidates keep abreast with the debates and latest developments in their relevant fields of researched.

The study is an explanatory in nature. It adopted a qualitative methodology with interviewing being the method of data collection. Being a small scale study including only five participants from only one discipline, the findings of the current study are not intended to be generalizable. The implications of the study, however, have an ecological significance for the doctoral education in the context of the study as, to the best of our knowledge, there has not been any study so far been conducted on this topic in the context. We therefore, believe this study would be pioneering and would open up potential avenues for more future search in the Sudanese context and beyond across the region. The implications could also be transferable to and applicable in other similar contexts. To gain more insights into and understanding of the challenges of Ph.D. education in Sudan and thus suggesting further recommendations for Ph.D. research training and development, future research could expand the current investigation by including more universities and participants as well as disciplines. Methodologically, future studies could combine both text and context by analyzing the text of Ph.D. theses themselves and the context in which they have been produced. Corpus-based textual analysis methods combined with ethnographic contextual analysis methods would aptly fit the purpose.

References

Ayiro, L. P., & Sang, J. K. (2011). The awards of the PhD in Kenyan universities: A quality assurance perspective. *Quality in Higher Education, 17*(2), 163–178. doi:10.1080/13538322.2011.582794

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101.

Brydon, K., & Fleming, J. (2011). The journey around my PhD: Pitfalls, insights and diamonds. *Social Work Education: The International Journal, 30*(8), 995–1011.

Coffey, A., & Atkinson, P. (1996). *Making sense of qualitative data*. London: Sage Publications.

Denicolo, P. (2003). Assessing the PhD: A constructive view of criteria. *Quality Assurance in Education, 11*(2), 84–91. doi:10.1108/09684880310471506

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative and mixed methods*. Oxford: Oxford University Press.

Dysthe, O., Samara, A., & Westrheim, K. (2006). Multivoiced supervision of master's students: A case study of alternative supervision practices in higher education. *Studies in Higher Education, 31*(3), 299–318.

Edwards, B. (2002). *Postgraduate supervision: Is having a PhD enough? In paper presented to the Australian Association for research in Education Conference*. Australia: Brisbane.

Fergie, G., Beek, S., McKenna, C., & Creme, P. (2011). It's a lonely walk: Supporting postgraduate research through writing. *International Journal of Teaching and Learning in Higher education*, *23*(2), 236–245, available online at http://www.isetl.org/ijtlhe/

Gulbrandsen, J. M. (2000). *Research quality and organisational factors. An investigation of the relationship*. Trondheim: Department of industrial economics and technology management. Norwegian University of Science and Technology.

Gunnarsson, R., Jonasson, G., & Billhult, A. (2013). The experience of disagreement between students and supervisors in PhD education: A qualitative study. *BMC Medical Education*, *13* (134), 1–8. Available on http://www.biomedcentral.com/1472-6920/13/134

Halse, C., & Malfroy, J. (2010). Retheorizing doctoral supervision as professional work. *Studies in Higher Education, 35*(1), 79–92.

Hasrati, M. (2005). Legitimate peripheral participation and supervising PhD students. *Studies in Higher Education, 30*(5), 557–570.

Heath, T. (2002). A quantitative analysis of PhD students' views of supervision. *Higher Education Research & Development*, *21*(1), 41–53. doi:10.1080/07294360220124648.

Hockey, J. (1996). A contractual solution to problems in the supervision of PhD degrees in the UK. *Studies in Higher Education*, *21*(3), 359–371. Available on doi: 10.1080/03075079612331381271

Hockey, J. (1994). Establishing boundaries: Problems and solutions in managing the PhD supervisor's role. *Cambridge Journal of Education, 24*(2), 293–305.

Krase, E. (2007). May be the communication between us was not enough: Inside a dysfunctional advisor/L2 advisee relationship. *Journal of English for Academic Purposes, 6*(1), 55–70.

Kvale, S. (1996). *Interviews: An introduction to qualitative research interviewing*. London: Sage.

Kyvik, S., & Thune, T. (2014). Assessing the quality of PhD dissertations: A survey of external committee members. *Assessment & Evaluation in Higher Education, 40*(5), 768–782. doi:10.1080/02602938.2014.956283

Lee, A. (2008). How are doctoral students supervised? Concepts of doctoral research supervision. *Studies in Higher Education, 33*(3), 267–281.

Lillis, T. M. (2001). *Student writing: Access, regulation, desire*. London: Routledge.

Löfström, E., & Pyhältö, K. (2014). Ethical issues in doctoral supervision: The perspectives of PhD students in the natural and behavioral sciences. *Ethics & Behaviour*, *24*(3), 195–214. Available on doi:10.1080/10508422.20133.830574

Mahmood, S. T. (2011). Factors affecting the quality of research in education: Student's perceptions. *Journal of Education & Practice*, *2*(11), 34–40. Available on www. iiste.org.

Mackinnon, J. (2004). Academic supervision: Seeking metaphors and models for quality. *Journal of Further and Higher Education, 28*(4), 397–405.

Manathunga, C., & Lant, P. (2006). How do we ensure good PhD student outcomes? *Education for Chemical Engineers, 1*(72), 72–81. doi:10.1205/ece.05003

Mason, J. (2002). *Qualitative researching* (2nd ed.). London: Sage.

Marsh, H. W., Rowe, K. J., & Martin, A. (2002). PhD students' evaluations of research supervision: Issues, complexities, and challenges in a nationwide Australian experiment in benchmarking universities. *The Journal of Higher Education*, *73*(3), 313–348. Available on http://www.org/stable/1558460

McCarthy, G., Hegarty, J., Savage, E., & Fitzpatrick, J. (2010). PhD away days: A component of PhD supervision. *International Nursing Review*, *57*(4), 415–418. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1466-7657.2010.00828.x/full

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. London: Sage Publications.

Morley, L., Leonard, D., & David, M. (2003). Quality and equality in British PhD assessment. *Quality Assurance in Education, 11*(2), 64–72. doi:10.1108/09684880310471489

Mullins, G., & Kiley, M. (2002). It's a PhD, not a Nobel Prize: How experienced examiners assess research theses. *Studies in Higher Education, 27*(4), 369–386.

Odell, L., Goswami, D., & Herrington, A. (1983). The discourse-based interview: A procedure for exploring the tacit knowledge of writers in nonacademic settings. In P. Mosenthal, L. Tamor, & S. A. Walmsley (Eds.), *Research on writing: Principles and methods*. (pp. 221–236). New York, NY: Longman.

Pyhältö, K., Toom, A., Stubb, J., & Lonka, K. (2012). Challenges of becoming a scholar: A study of doctoral students' problems and well-being. *International Scholarly Research Network*, 1–12. Doi:1.5402/2012/934941.

Watts, J. H. (2009). From professional to PhD student: Challenges of status transition. *Teaching in Higher Education*, *14*(6), 687–691. Available on doi:10.1080/13562510903315357

Winter, R., Griffiths, M., & Green, K. (2000). The academic qualities of practice: What are the criteria for a practice-based PhD? *Studies in Higher Education, 25*(1), 25–37.

Zeegers, M. & Barron, D. (2012). Pedagogical concerns in doctoral supervision: A challenge for pedagogy. *Quality Assurance in Education*, *20*(1), 20–30. Available on doi:10.1108/09684881211198211

https://www.google.com/?gws_rd=ssl#q=university+of+khartoum

Part VII
Quality Assurance, ESP Needs Analysis

Quality Assurance and Foreign Language Programme Evaluation

Donald F. Staub

Abstract The global higher education market is changing quickly as increasing numbers of institutions are entering the playing field. Higher education is no longer solely the province of universities funded by governments. The result of this present condition is that higher education institutions are seeking ways to distance themselves from their competitors. Many are doing so by offering English-medium instruction, which often entails the establishment of an EFL programme for all incoming students who will eventually study in the University's English-medium programmes. These EFL programmes are increasingly under pressure to demonstrate their value to external and internal stakeholders. Thus, quality assurance and evaluation are becoming critical activities for EFL programmes wishing to demonstrate their worth. This paper examines the design and implementation of a quality assurance initiative at an English-medium university in Istanbul, Turkey. Qualitative and quantitative data is used to evidence the success and challenges of establishing this effort.

Keywords Quality assurance · Evaluation · Foreign language program · Qualitative/quantitative data

1 Introduction

As the higher education marketplace expands rapidly in many countries, students and their families are becoming more discerning when selecting a post-secondary school. Access and financial considerations are not the sole criteria driving their decision-making. Demonstrated quality and value-added are becoming the distinguishing factors when making this critical choice; what Gallagher (2015, para. 2) refers to as the "marketization of higher education due to a focus on value and transparency." Increasingly, universities, faculties, and instructional programmes

D.F. Staub (✉)
Isik University Sile, Istanbul, Turkey
e-mail: staubdonald2@yahoo.com

S. Hidri and C. Coombe (eds.), *Evaluation in Foreign Language Education in the Middle East and North Africa*, Second Language Learning and Teaching,
DOI 10.1007/978-3-319-43234-2_16

are seeking means to validate institutional quality. While Engineering and Business programmes have set the pace for quality assurance, primarily through accreditation, there is little doubt that Foreign Language programmes, particularly EFL programmes, should be at the forefront as well. As the gateway to many, if not all programmes in universities, EFL programmes may be the largest instructional units in a university. Consequently, EFL programmes are quite often the face of the university when it comes to attracting students. A strong grasp of English—provided through a preparatory or foundation program-opens the door to a quality higher education. As well, a quality EFL programme is also a proxy for less time (and money) spent at this early stage, meaning a quicker path to a degree. Thus, in order to continuously affirm the value of an EFL programme, ongoing evaluation and quality assurance is becoming a mandatory activity carried out in parallel with the curriculum delivery.

Quality of EFL programmes have not always garnered the current level of attention. Indeed, for many programmes the need for evaluation still warrants justification, which explains the relevance of this chapter. In most higher education institutions, the EFL programme is not viewed as an academic unit (i.e., resulting in research by faculty or degrees for students). Rather, they are outside the perimeter of the institutional core, acting as a service unit, playing the role of a loosely coupled (Weick, 1976) entity. As such, while loosely coupled units by definition may enjoy greater flexibility in how they operate, they are often marginalized in terms of resources and status. This is often evidenced via contrasting academic designations-where the EFL programme may be administered by a director (not a chair or dean), and content is delivered by instructors and teachers (not assistant, associate, or full professors). As such, and for better or worse, EFL programmes often operate unchecked internally and externally.

Yet, there are a number of global trends in post-secondary education that are pushing universities to reconsider the amount of attention given to the quality of the EFL programme. With the steady growth of higher education institutions delivering so-called English-Medium Instruction (EMI), or Content and Language Integrated Learning (CLIL) (e.g., Coleman, 2006; Maiworm & Wächter, 2002), schools are finding themselves in a position whereby evaluation and quality assurance are unavoidable. Practically speaking, in many such institutions, the EFL Programme may be the gateway and gatekeeper for a significant proportion of students. Ensuring the successful, timely completion of the EFL Programme for the greatest number of students is essential, if for nothing other than to prevent a bottleneck in the EFL programme. Additionally, the growing privatization of higher education has resulted in a context where institutions compete for student-consumers, thus quality and reputation become value-added features that provide a competitive advantage, particularly in a crowded marketplace. This has led to an accreditation movement across the field. It is also important to note that a lack of attention to the quality of the EFL programme may result in issues regarding student persistence and retention. The EFL Programme may serve as an essential stepping-stone toward a student's academic and career goals. Students that struggle with the language learning, or feel dissatisfied with the effectiveness of the programme, may choose to leave the institution. While the effectiveness of the EFL programme is closely associated with the efficient delivery of the curriculum, there are numerous additional factors across the programme that comprise the broader definition of its quality.

From this context of increasing need for attention to the quality of EFL programmes, a number of pertinent research questions emerge. First, what would one expect to find in the design of a comprehensive, rigorous evaluation of an EFL programme? Second, what considerations should be made in order to design and implement a sustainable quality assurance effort within an EFL programme? This qualitative case study carried out at an EMI higher education institution in Istanbul, Turkey will seek to provide clarity to these two important questions.

2 Theoretical Background

The foundation of an effective and sustainable quality assurance programme is the belief and actions of the organization, demonstrating a “collective commitment” (Maki, 2004, p. 11) to this endeavour. The organization may be the larger institution in which the EFL programme is housed, or may be the unit in which the programme resides; e.g., a faculty or a school of foreign languages. The point to emphasize is that quality assurance cannot be viewed as a desktop exercise, carried out by a few sequestered individuals who are charged with examining data and creating reports. After all, the goal of assessment is not just to gather evidence, but also to make evidence-informed changes (Banta & Blaich, 2011, p. 25). Thus, the sustainable quality assurance enterprise must be “driven by the internal curiosity” of a learning organization as opposed to external forces. It should be envisioned, planned, implemented, and “reflected in structures, processes, and practices,” with the intent of becoming institutionalized and leading to the broad pursuit of knowledge for improvement. The task for leaders then, is to foster the learning environment by supporting and taking responsibility for quality assurance initiatives, particularly at the unit level where learning takes place (Banta, Jones, & Black, 2009, p. 12). That is, a context must be created where evaluation and assurance are “purposefully planned and intentionally reflective” (Bresciani, 2006, p. 23), and where a feeling of trust is continually reinforced; that is, where data and results are utilized in the manner for which they were specified (Bresciani, 2012, p. 418). Thus, there is the critical groundwork that must be laid in order for a healthy, productive environment that sees evaluation and assurance not as perfunctory activities, rather opportunities for the organization, or unit, to learn and grow.

The literature on programme assessment is quite clear regarding the establishment of a productive learning environment within the unit. In the design and implementation of a quality assurance programme, the first important step is to develop a plan that gives prominence to evidence and makes it a “consequential factor” in programme planning and review processes (New Leadership Alliance, 2012, p. 7). A critical piece of the plan is to deliberately consider the weight of time required of faculty and staff to participate in evaluation and assurance initiatives. By those developing the plan, it should be acknowledged that impact on learning cannot take place without increased engagement from faculty and staff (Banta & Blaich, 2011, p. 23) and that evaluation and assurance-related activities are time-intensive, given the expectations for analysis and reflection (Baker, Jankowski, Provezis, & Kinzie, 2012, p. 10). Additionally, such activities are often perceived as bureaucratic and externally-driven and needlessly encroach on time that should be dedicated to students (Banta & Blaich, 2011, p. 25).

Beyond planning, when it comes to implementation, it is essential that logical connections are consistently reinforced to all involved in the enterprise between the perceived busy work of data collection and analysis and the goal of improved quality of learning. This can occur through opportunities at various points in the assessment cycle provided for faculty and staff to develop an understanding of assessment and advance their knowledge and experience with it (Banta et al., 2009, p. 15). Collecting and analysing programme data must be viewed as a means to an end; it should exist to provide stakeholders with the information to satisfy their own natural curiosity about the results of their work (Bresciani, 2012, p. 15), which, in turn, results in discussions around learning and the curriculum. Blaich and Wise (2011) make the case that “assessment data only has legs (p. 12)” if there is a feedback loop with the data speaking to questions that staff have about learning and the programme. If the message of data supporting programme improvement becomes blurred, assessment runs the risk of becoming marginalized (Maki, 2004, p. 15). As Banta et al. (2009, p. 5) stress, assessment that “spins in its own orbit” will not succeed.

Evaluation and quality assurance efforts are a double-edged sword. On one side, Banta and Blaich (2011, p. 27) point out that assessment is a subversive activity, raising questions about learning and quality, and creating discomfort regarding established habits within a programme. On the other side, when implemented effectively, a focus on quality assurance leads to consistent programme improvement and growth. For the latter condition to prevail, how the quality assurance effort is implemented is critical in both the short and the long term. In the short term, it is about building trust; demonstrating that data is for improvement and not punishment. Tagg (2007, p. 37), arguing for why organizations have such difficulty with change, believes that the most fundamental problem of colleges is that the people within them do not learn very well. Therefore, the long-term goal for the quality assurance effort should be to foster a learning organization that comes from consistent, productive conversations around data and quality, as well as the reinforcement of trust, through conscientious and equitable use of data. Bresciani (2006) offers “criteria for good practices” to guide the design stage of an outcomes-based assessment programme. In the case of the institution examined in this study, her list of criteria was adapted and has served as a valuable point of reference for their own quality assurance programme development. Her nine criteria outline an instructive cycle of design and implementation from clarity of goals and expectations, to collaboration and recognition of effort, to coordination and flexibility of implementation, to ongoing evaluation of the process. The following case study explores the design and implementation of a quality assurance initiative in a university-level EFL programme, and the degree to which Bresciani's criteria may serve as a framework for other such programmes in the initiation of quality assurance efforts.

3 Method

This research was carried out using the Case Study approach, which explores a "bounded system, such as a process, activity, event (…)" (Creswell, 1998, p. 112). The case is bounded within a particular context, which is the School of Foreign Languages (SFL) at a private university in Istanbul, Turkey. Furthermore, this can be viewed as an instrumental case study (Stake, 1995), whereby it is used to illustrate an issue, which is the design and implementation of a Quality Assurance initiative of an EFL programme. In Turkey, there are currently 185 universities; 41 % (n = 76) are private institutions. According to data provided by the website of the Turkish Higher Education Council, 84 % (n = 155) of the 185 universities have EFL preparatory programmes. As of Summer 2015, fewer than ten EFL programmes in the country had received accreditation, with the site of this case study being one of them.

The case study site has five Faculties-Arts and Sciences, Economics and Administrative Sciences, Architecture and Design, Engineering, and Fine Arts. English is taught as the medium of instruction in all faculties and academic programmes, with the exception of the Faculty of Fine Arts and a recently established Psychology programme in the Faculty of Arts & Sciences. The school has an enrolment of approximately 3000 undergraduate and 500 graduate students; nearly half of the undergraduates live on campus. Approximately 80 % of all incoming freshman enrol in a one-year, intensive academic English preparatory programme at the school; the remaining 20 % are exempt either through performance on the EFL proficiency exam or standardized exams such as the TOEFL or IELTS. An *Exit Exam* is developed and administered at the end of the academic year by the EFL preparatory programme. Students who successfully complete the EFL programme move along to their academic programmes as freshmen. Those who do not successfully complete the programme may enrol in an intensive summer programme and enter an Exit exam at its completion. If they are once again unsuccessful, they must re-enrol in the EFL programme for a second academic year. Between the 2011–2012 and 2014–2015 academic years, the average percentage of EFL students required to re-enrol for a second year in the programme was 48 %.

The organizational structure for the EFL programme is that there is one director, who reports directly to the rector of the university, and an administrative layer, comprised of the *Coordinators* team: The Academic coordinator, the Administrative coordinator, the Curriculum coordinator, the Testing and Assessment coordinator, the Integrated Technology coordinator, and the Student

Learning Centre coordinator. In 2014, prior to pursuing accreditation, positions were created for a Quality Assurance Coordinator and a Continuous Professional Development Coordinator.

Prior to the 2014–2015 academic year, attention to quality could be characterized as unstructured and unsystematic. Data analysis was conducted primarily through end-of-year Exit Exam results and annual survey data gathered from students and staff. Most data could be classified as output-based (e.g., number of students passing or failing, number of students visiting the Student Learning Centre, number of teachers receiving professional development during the year). While broad learning outcomes were identified (e.g., "students are able to write a 600-word informational essay"), a comprehensive list of learning outcomes and objectives was lacking, as was an accompanying outcomes assessment process. Little data found its way into a feedback cycle that informed adjustments to the curriculum or staff development. In addition, data on incoming students was not analysed to serve as predictive analytics, nor was there an attempt to gather and analyse data on those students who had left the EFL programme prior to successful completion. In sum, the EFL programme lacked a systematic means for aggregating, disaggregating, and analysing data. Therefore, to some degree, life was easy. If students were unsuccessful, the responsibility was their own. With only a broad light cast upon the programme, there were few sharp beams that focused on specific issues. Potential solutions remained in the shadows.

In the 2014–2015 academic year, the director of the SFL determined that the EFL programme would pursue accreditation, which would serve two purposes. First, in the crowded university marketplace for student attention, a stamp of accreditation provides schools with a competitive advantage. Second, the director determined that evaluation of quality-of processes and instruction-were to be a major priority from that point forward. The first step was the establishment of an SFL-wide quality assurance system, embodied in the creation of a Quality Assurance Unit (QAU), which included a place on the organizational chart as well as the hiring of a coordinator and one part-time staff member, who is also a full-time instructor. The first priority for the QAU was to facilitate the accreditation application process. However, accreditation was not seen as the sole intent of the QAU. The mission of the QAU is to ensure continuous improvement across all aspects of the SFL, and the accreditation process is viewed as an important way to periodically affirm ongoing quality assurance activities. This study is an evaluation of the design, development and implementation of a QA effort within an EFL programme.

The design and development stage of the QA process took place over a seven-month period (i.e., May–November) as the EFL programme conducted a self-study and revised or established policies, procedures, and processes that would result in an effective, sustainable QA system. Implementation and evaluation are ongoing. From the outset, there has been a consistent focus on data collection and analysis for evaluative and planning purposes. Through surveys, focus groups and interviews, data has been gathered in order to formatively evaluate the effectiveness of the design implementation, as well as assist with planning. Surveys have been used on a frequent basis as they help to acquire a "quantitative or numeric description" of attitudes or opinions of a population through a sampling of that population (Creswell, 2012, p. 155). Despite the risk of playing a role in the modern phenomenon of "overload" (Sue & Ritter, 2012) from too many digital surveys, brief online surveys (e.g., Google Forms or Survey Monkey) have proven to be an effective way to collect data. Finally, whereas surveys are able to provide rapid feedback on activities, that is, to "learn about what you cannot see" (Glesne, 1999), interviews and focus groups have also served as useful means for the collection of evaluative data. Interviews and focus groups are generally semi-structured in order to guide the conversation so that if there are multiple interviewers, the integrity of the process is maintained (Fontana & Frey, 1998, p. 52).

4 Results

The design and development phase of the quality assurance initiative took place over a seven-month period, from May to November 2014. The first step in the process was the establishment of a Quality Assurance Unit (QAU). A coordinator was hired and an office was dedicated to the Unit. The initial task of the QAU was to lead the first-time accreditation application process. A working team was assembled from current EFL instructors and staff, and relevant tasks were allocated; e.g., a curriculum team leader assumed responsibility for addressing standards related to learning, the Assessment Coordinator addressed standards related to testing and assessment. The thread woven through all standards in the application, however, was quality-establishment of policies, procedures, and a system that would focus on continuous improvement across all aspects of the EFL programme. Therefore, while the QAU kept one eye focused on the accreditation application, the other eye was trained on the broader goal of an effective and sustainable Quality Assurance system.

In terms of Quality Assurance, the EFL programme, and the SFL broadly, was a blank slate. Prior to the decision to pursue accreditation, no systematic, collaborative, results-oriented process of targeted data collection and analysis had existed. Data was often anecdotal. Reports were sporadic, often created on an as-needed basis. For curricular purposes, aggregation and analysis of class-level data, such as attendance and assessments, did not occur. Programme-level student learning outcomes and objectives existed as disparate lists, with little emphasis placed on the alignment of these outcomes and objectives with assessments. For administrative purposes, broad sweeping, end-of-year surveys were administered, yet purposeful, task-specific surveys were not; in neither case were there follow-up focus groups, interviews, or action plans. Likewise, processes and procedures (e.g., recruitment, staff induction, student database management, student orientation) were not subject to analysis for effectiveness and efficiency. "Quality Improvement" was not a bullet in anyone's job description.

Thus, in parallel with addressing the individual standards in the accreditation application, the QAU devised a framework that has guided the design, development, and implementation of quality improvement policies and processes within the EFL programme and across the SFL. The framework divided the establishment of a quality assurance programme into two major components: Organization and Monitoring & Adjusting. Under each of which, a list of sub-categories was created. Each of these sub-categories would be comprised of a foundational piece (i.e., definitions, descriptions, policies, procedures) and an implementation piece (i.e., collecting, analysing, and acting upon relevant data, or what is commonly known as *closing the loop*).

The first crucial step in establishing a quality assurance programme was to create the Quality Assurance Unit. By adding this Unit to the organizational chart, assurance was being institutionalized and thus legitimized, particularly among the EFL staff. Prominent office space within the EFL building was allocated to the Unit, as was funding for two position-a full-time coordinator and a part-time *officer*. Subsequently, the Continuous Professional Development Unit (CPD) was established, and subsumed under the QAU. The rationale for doing so was that in the cycle of quality improvement, continuous and broad-ranging training for the staff results in an improved programme; the QAU identifies training needs and the CPD develops and delivers. Similar to the QAU, CPD was assigned a part-time coordinator, a part-time officer, and office space. This structural move also raised the profile and the importance of CPD among the EFL staff.

The QAU began its existence by creating its own vision, mission, and Quality Assurance policy for the EFL programme, specifying that the QAU would be responsible for monitoring, reporting, and guiding the EFL in ongoing quality improvement, as well as facilitating professional development opportunities. The policy also specifies the formation of a Quality Assurance advisory committee, comprised of SFL stakeholders, and which meets semi-annually for the purpose of providing direction to the QAU. Functionally, the QAU at this time published a *Quality Assurance Manual*, which articulated the role and responsibilities of the QAU, with descriptions of all significant functions (e.g., administrative and student learning outcomes assessment, programme review), as well as an annual calendar identifying a timeline for data collection, analysis, action plan development, and reporting. For accreditation purposes, the QAU also oversaw the revision of the *EFL Student Manual* and the *Staff Manual*, as well as the creation of an *EFL Policy Manual*.

The next step for the QAU was the development of a Data Warehouse for the SFL. Prior to the establishment of the QAU, little attention had been paid to the array of EFL-relevant data that could be analysed for quality purposes. A centralized repository for information on student demographics, student activity, and student performance had not existed previously. The QAU identified a list of sub-fields from all three of these broad areas for the data warehouse. The student demographics are intended for use as predictive analytics for retention and student performance in the EFL. Fields under this category include: Student high school GPA, does the student have a scholarship, did the student attend a private or public high school, the location of the high school, the parents' educational background, whether the student has siblings attending a university, whether the student lives on campus. Under student activity, does the student participate in the student government or any clubs, does the student have work-study, to what extent does the student utilize the EFL's Student Learning Centre. Student performance data includes in-class skills assessments, mid-term and final scores (sub-divided into the skills sections), and attendance records. The academic year is divided into four, seven-week modules. The performance data is entered at the end of each module. One reason for the warehouse is to help identify students who are struggling, either before they walk through the door for the first time, or as early as possible once they have started the EFL programme. Another reason for the warehouse is that the data can be analysed for trends across levels, skills, and time within the EFL programme.

The final step in establishing the QAU was the development of a Communications Plan. The Plan covers a range of communication objectives with the overall aim of increased transparency, clarity, and collaboration throughout the EFL. For collaborative activities (e.g., addressing accreditation standards or writing policy), a shared repository was created. The free cloud storage, Google Drive was deemed most suitable for the needs of the project. In addition to storage, it allows collaborators to synchronously edit documents. A file system was arranged according to the major objectives and standards for the accreditation application, and all working members of the accreditation team were granted authority to upload or edit relevant files. This permitted the QAU coordinator to take inventory at any point in time of current progress toward the completion of the application. Beyond accreditation, the Google Drive remains a repository for QAU and CPD staff to store and collaborate on work.

Furthermore, an additional aspect of communication was the need to archive and distribute pertinent information to internal and external stakeholders through an easy-to-manage means. The desire of the QAU was to remain transparent, yet avoid burdening the staff by flooding in-boxes with non-essential information. The solution was the creation of a website/blog utilizing an open source website creation platform. The *SFL Quality* website became a dynamic space for QAU/CPD team members to announce relevant internal and external activities (e.g., workshops and conferences), to share information about research or webinars, to report and reflect upon internally organized activities, and to serve as a publically accessible repository of slide presentations and non-sensitive materials and reports created by these two units (e.g., manuals and workshop handouts). The website also serves as a venue to report on evaluation and research projects carried out by the two units.

Finally, in an effort to establish effective communication throughout the EFL programme, it was important to move in concert with the broader communications network of the SFL, and not at crossed-purposes. In terms of the organizational culture, this was critical. The QAU and CPD were striving to gain acceptance and legitimacy. If they were viewed as contributing to a perceived over-burdening of staff, they would struggle to build a base of support. Organizing trainings, workshops and meetings required a good deal of advance planning to ensure that conflicting events (e.g., assessments, coordinator-level meetings, staff meetings) were not scheduled for that time, nor that the planned QAU activity became part of a cluster of other activities arranged by other EFL units, thus giving credence to the belief that the staff were over-worked. During the first academic year of its existence, much angst was experienced by the QAU and CPD units as many scheduling conflicts arose due to the lack of a common working calendar. While certain dates for Curriculum and Testing were announced well in advance, these units would frequently organize last-moment meetings to address pressing issues. These meetings would sometimes clash with QAU or CPD meetings, which generated conflict. The solution in this case was for the Curriculum and Testing units to create a master calendar of their own events at the beginning of the academic year, and the QAU and CPD units would target their own activities for dates that resulted in the fewest conflicts manageable.

In addition to the foundational pieces necessary to establish the QAU, the other major initiative was the design of an integrated monitoring and adjustment system for student learning and administration of the EFL. This was implemented through two primary activities: Assessment of Student Learning Outcomes (SLOs) and Programme Reviews. Regarding SLOs, the QAU and the Curriculum Unit established curriculum maps for each skill area comprised of objectives to be addressed during each of the four modules in the academic year. Instructors are required, twice per module, to submit maps indicating which objectives they had achieved in the previous three weeks. The maps, along with assessment results, permit the QAU to perform frequent analyses of curriculum delivery to determine the pace of progress toward end-of-year SLOs. The QAU has also implemented a separate approach to evaluation through Programme Reviews. Generally speaking, the programme review is a self-study conducted by an academic unit as a means to evaluate the effectiveness and efficiency of the programme. Often, such reviews are carried out by the programmes themselves, with guidance and support from an institutional effectiveness unit, such as the QAU. The programme review takes a wider view of quality than the narrower focus of student learning outcomes assessment, examining both quantitative and qualitative data, and looking more broadly at issues such as student success and persistence, instructor preparation and professional development, and instructor and student perceptions of the programme.

In terms of defining *programme* for the purpose of evaluation, the QAU determined that in order to more effectively analyse the quality of the EFL programme broadly, it would be useful to examine it according to its three separate levels, or tracks, which are distinguished by the language abilities of the learners in those tracks (i.e., track-one students are near-beginners whereas track-three students may be able to complete their requirements after one semester). While there is certainly much integration across the EFL curriculum, each of the three tracks has their own unique features that allow them to be viewed as separate entities. The rationale for conducting the reviews is that rather than examining student performance at the aggregate, or EFL programme level, it is easier to identify issues and solutions, at a micro level.

Therefore, each programme, or track is reviewed on an annual basis. Data is gathered and analysed during the first semester of the academic year, with analysis and planning taking place in the second semester. The logic behind this timeline is that if the decision is made to change textbooks, there is sufficient time to select and order them for the next academic year. The data is gathered and organized in a report by the QAU. This includes student performance data (including SLO data), student success and persistence data, information about instructors (e.g., demographics, hours of instruction, office hours), student survey and focus group data, and instructor survey and focus group data. In the second semester of the academic year, the QAU facilitates an analysis discussion with a working team comprised of members from the Curriculum Unit and instructors from that track. The outcome of the meeting is a set of concrete action plans to be developed during the summer, and implemented the following academic year.

The QAU is not solely focused on the quality of learning. There are numerous administrative functions that must be evaluated and monitored for quality on a continuous basis. Just as there are Student Learning Outcomes, there are Administrative Outcomes (AOs) as well to provide a comprehensive evaluation of the EFL programme. AOs are used to analyse the effectiveness and efficiency of routine processes and non-instructional activities. In large part, AOs are identified by the accreditation standards (e.g., staff recruitment and induction, grievance procedures). There are also a number of AOs that have been identified based on local needs and initiatives, such as student retention, the Student Learning Centre, or the development and implementation of the Continuous Professional Development programme. Additionally, there is a steady stream of initiatives and pilot projects across the EFL programme, which require close observation to determine their effectiveness. Examples of recent projects include the implementation of an Early Alert system, Individual Learning Plans for low-performing students, and Peer Observation programme for instructors, as well as pilots of a First Year Experience and Early Alert programmes. Data collection and analysis for administrative outcomes is both qualitative and quantitative, and continues throughout the year.

5 Discussion

After 1.5 years of design and implementation of a quality assurance system for the EFL programme, the primary question to be addressed is *how successful is the effort?* To what degree has the QAU impacted quality, and what is the potential for sustainability of the quality assurance initiative. The data gathered for this study has revealed that, in a broad sense, the initiative appears to be on the right track. Organizationally, the foundation for quality has been laid and the evidence suggests a positive impact on administrative outcomes and the establishment of a culture of assessment within the EFL programme. This is not to say that it has been an effortless process; some initial issues needed to be resolved, and various challenges persist-typical to any change effort. Nevertheless, at this stage in the maturation process of the quality assurance initiative, a number of distinct themes and *take-aways* have emerged that warrant discussion, from the need for quality assurance, to the need for rigorous planning and execution, to the need for effective leadership.

EFL programmes often find themselves as marginalized units on campus. They may assume a visible role on campus if one views them from an enrolment numbers (and revenue generation) perspective. Unfortunately, this is not how it usually works. As EFL programmes generally lack the prominent researchers and publications output of academic programmes on campus, they are often viewed as less prestigious within the university caste system. This is an ironic situation. On the one hand, the loosely-coupled (Weick, 1976) nature of EFL programmes allows them a certain degree of latitude in, for example, hiring practices and enrolment management. Yet, this may also suggest that the institution does not hold the programme to the same standards as other units. In many institutions, the EFL programme may be the gateway and gatekeeper for a significant proportion of students. It is here that students begin their university lives, and if the programme does not meet their needs or expectations, it may also be where they end their university lives (at least at that school). As the global higher education sector continues its expansion, with more and more institutions offering English Medium Instruction, or Content and Language Integrated Learning, schools are seeking a competitive advantage that will help them stand out. One such way is for universities to assure the quality of their foreign language programmes, which is reflected in the recent growth in the EFL programme accreditation market. As mentioned previously, this was a driving factor behind the EFL programme at the institution, where this case study has been conducted, being the first unit on campus (and one of the first EFL programmes in the country) to seek accreditation. This accomplishment has not escaped the attention of the university administration, which has actively promoted this achievement on the university website.

An important consideration in establishing a sustainable quality assurance programme is to examine the current organizational structure. Where are the supports and where are the barriers to success? In a meta-analysis of institutions implementing assessment programmes, Baker (2012, p. 10) and colleagues found that successful schools were those that “worked diligently over time to create structures, processes, and an atmosphere conducive to the use of assessment to improve student learning.” One simple example of structure, as seen at the institution where this case study was carried out, is the establishment of an office that oversees Quality Assurance. By doing so, quality improvement has been legitimized; assuring stakeholders that there is a long-term vision for the process. Additionally, the physical space that the QAU occupies is prominent, on the main floor of the EFL building, set visibly between instructor and administrative offices. With the creation of this Unit, staff have demonstrated a greater interest in quality. The increased interest is quantifiable, *vis a vis* the number of instructors who have visited the QAU since opening its doors, with many noting, “it’s about time.”

An additional example of how organizational structure becomes a barrier to engagement is through workload. On one hand, the structure of the EFL programme is viewed as favourable as instructors are required to only teach four days a week. However, this also meant that teachers had only four days per week to teach, prepare for classes, hold tutorials for their students, and attend meetings. In surveys and interviews, teachers have consistently raised concerns about being "overworked", and for this reason, quality assurance and professional development activities are perceived as burdensome. This is not the image that the QAU is hoping to portray as it attempts to build engagement. Therefore, as opposed to working at odds with an existing structure by trying to layer on traditional forms of training and workshops, the QAU, from the beginning, has worked to foster engagement by altering its delivery methods. For example, in-house webinars have been created and distributed to the staff. Face-to-face sessions are shorter with smaller working groups. One successful endeavour has been the "Nano Conference," with its three separate five-minute presentations. 65 % of respondents to the follow-up survey indicated that the timing was "just right". The important point here is that when creating a new entity within an existing structure, it is critical to understand the limits that the structure imposes on the organization and its employees. To gain acceptance, the new entity must carefully insert itself into the current environment.

Furthermore, the effective quality assurance programme does not magically appear-as much as we would like it to. It requires planning (Lennon et al., 2014; New Leadership Alliance, 2012). As the author Lewis Carroll noted wisely, "If you don't know where you are going, any road will get you there." In other words, effective planning leads to a roadmap that articulates a comprehensive, manageable list of quality-related activities, their importance, and an assessment plan for each. There are generally two sets of forces, external and internal, driving quality assurance. The former could be generally viewed as more directive, coming from the top-down, whereas the latter may be seen as more deliberate and aimed at continuous improvement. One reason why external calls for quality assurance are perceived as invasive is that they are usually accompanied by non-negotiable, unrealistic deadlines. There is little appreciation given to careful, strategic planning that is critical for effective, sustained design and implementation. Therefore, effectively planned quality assurance initiatives are characterized by two important requirements: Inclusion and time.

Planning requires inclusion. Effective planning requires input and participation from a breadth of stakeholders. *Ineffective* planning is a small group working in isolation in order to complete a task with a given deadline. Just like good assessment itself, it is important to receive multiple perspectives in planning. At the very least, the list of stakeholders should include the faculty conducting the instruction and the assessments, the students who will receive the instruction, and relevant administrators. Bresciani (2009, p. 87) concluded that a significant threat to successful implementation of programmes had been the absence of instructors in the "planning and delivery of programs, the assessment, or in the discussion of results." Planning must also result in shared consensus. There must be general agreement on those key qualities and quantities that define a successful student and an effective programme.

Planning requires time. As can be imagined, the processes described earlier take time. The simple logistics of gathering and analysing the data generated by stakeholders is time consuming. The additional layer of ensuring that consensus is achieved means that more time must be dedicated to the deliberate process of seeking input and achieving buy-in. One important caveat is that over-planning can be toxic too. There is nothing more frustrating and de-motivating to busy instructors than talk that appears hollow and appears to lead nowhere.

Thus, an important component of quality assurance is the assessment plan-what data will be collected, when will it be collected, and by whom will it be collected and analysed. There is a wide range of outcomes-both learning and administrative-that the QAU is responsible for identifying and monitoring. This required the establishment of a detailed assessment plan, which includes a yearly assessment calendar that specifies what data will be collected, when it will be collected, when it will be analysed, when action plans will be devised, and who will ensure that these steps take place on time. In order to establish continuity in the wording and design across all assessments, the QAU utilized the SMART acronym as a guide.

Specific: The wording used for the assessment will specify the knowledge, skill, or behaviour that is being assessed, the expectation of acquisition-i.e., awareness, understanding, or application, the population that is being assessed, and the assessment instrument. An illustration of this tenet can be found in one Student Learning Outcome from this case study:

> EFL students [population] will demonstrate the ability to write [expectation of acquisition] an effective response essay on the end-of-year Final Exam [assessment instrument].

Measurable: The assessment should articulate the expected percentage change (from baseline to target) in performance as a result of instruction, clinical experience, etc.
Achievable: The established target should be achievable within the stated timeframe. We may want 70 % of EFL students to demonstrate proficiency in writing an effective response essay, but if our baseline is 50 %, it may be ambitious to expect a 20 % increase within the stated timeframe.
Relevant: The assessment must be aligned with established Standards, such as the Common European Framework, as well as with the needs and mission of the unit, which will include those of the academic programmes in which the EFL students will one day study.
Timeframe: A beginning and ending point for learning and assessment must be stated. Is the target long or short-term? For what time period will the population be assessed? When will the assessment take place?

The acronym has proven quite helpful when explaining assessment planning to instructors who are unfamiliar with the exercise. Specifically, for each identified outcome and its corresponding assessment, a simple table was created that contains the following columns: Outcome, Assessment, Target, Results, Use of Results, and Person Responsible. The Outcome is identified by the QAU along with relevant stakeholders (e.g., Curriculum Unit, Administrative Coordinator, Programme Director), as is the Assessment—with wording guided by the SMART method. Baseline data informs a realistic Target. Results are collected once the specified assessment is given, and the QAU facilitates assessment meetings where teams analyse the data and develop action plans.

The QAU encourages the collection of data from multiple sources. In situations where an outcome can be assessed using, for example, both quantitative and qualitative data (e.g., the effectiveness of the Student Learning Centre), confidence in the findings increases. Likewise, achievement of a specific learning objective (e.g., determining the meaning of vocabulary in context) is analysed with data from in-class and end-of-module assessments. The QAU also encourages instructors to utilize performance-based assessments in order to measure the amount of deep-learning (as opposed to rote memorization) that is taking place among the EFL students.

The commonly accepted phrase for completing an assessment cycle is *closing the loop*. This means that the steps in the assessment cycle are effectively accomplished-identifying the outcome, assessing it, analysing the data, and acting upon the data. Completion of at least one full cycle could be called a *success*. Completion of cycles on a continuous basis, particularly resulting in curricular change and improvement, is referred to as *sustainability*. The steps described previously in establishing the assessment plan may lead to one successful completion of the assessment cycle. Yet, to realize sustainability, the EFL programme must strive to establish a culture of assessment and a learning culture. To do so takes time, planning, and inclusiveness.

If a culture of assessment does not already exist, the process to establish such a culture takes time. Change takes time. The QAU has been very deliberate in its data collection and analysis approach, focusing more on programme-level data, and very little on classroom-level data. In a Culture where failure generally results in admonishment rather than support, it is little wonder that instructors are extremely reluctant to examine class-level data, for fear that their class will be singled out, or that their capability as an instructor will come into question. The QAU and the CPD unit piloted a Peer-Observation initiative in the 2014–2015 academic year, and it was received very coolly, with the conclusion being that instructors were very apprehensive about being under the microscope in their own classrooms-despite the non-critical nature of the design. Thus, regarding the data on student learning, the approach of the QAU has been to only facilitate conversations around programme level data, and in turn, emphasize problem solving over fault-assignment. Again, the transformation is taking time, through extended conversations about "what people hunger to know about their teaching and learning environments and how the assessment evidence speaks to those questions" (Blaich & Wise, 2011, p. 12). The approach taken by the QAU is that while programme-level data may be useful in stimulating conversation, eventually the curiosity of the instructors themselves will encourage them to begin asking questions that require classroom level data, which is a positive step toward what Bresciani calls "forming 'habits' of assessment" (2009).

In addition to time, careful planning is essential. It is important to create a roadmap that leads the way to the goal of a sustained and effective quality assurance initiative (New Leadership Alliance, 2012). This is where the standards in the accreditation self-study were useful as a guide to creating an initial plan. The plan for sustainability articulates clear direction for assessments (i.e., what will be assessed and when), responsibilities (i.e., who is responsible for collecting, analysing, and reporting data), and communication (i.e., who will write reports, and when will they be submitted). Across the EFL programme, responsibility for reporting data is widely distributed across coordinators-e.g., the Curriculum coordinator, the Testing coordinator, the Administrative coordinator-and funnelled to the QAU to be used in reports for broader data analysis. For instance, the Student Learning Centre coordinator creates an activity report at the end of each semester. That data is fed into the QAU data warehouse and becomes part of broader analysis to determine success predictors for EFL students.

Inclusiveness is the glue that holds the successful assessment plan together. The greater the breadth of engagement from EFL stakeholders, the greater the likelihood of sustainability. As Hersh and Keeling (2013, p. 9) argue, “too often, assessment is orphaned to the province of a small group of dedicated faculty and staff,” which significantly hinders the growth of engagement in the process. Ensuring that the EFL instructors who implement the assessment plan have a voice in designing and administering the program is essential; “significantly growing and deepening faculty involvement” is where Hutchings argues that the “real promise of assessment” lies (2010, p. 6). It is important that those who are committed to the programme feel that they are making a contribution to continuous improvement, and that their contributions are recognized. Which means that public acknowledgement of instructors who lend time to the quality assurance initiative is highly valuable. This is a practice that the QAU is well aware of and employs at presentations and trainings.

At the same time, there must be a level of accountability. Establishing an atmosphere of collaboration and collegiality does not necessarily mean an environment free of conflict. Just as action is recognized, so is inaction. When issues regarding the quality assurance process arise with individuals, the QAU coordinator sets up a meeting with the instructor to seek out a resolution. If that does not result in a positive change, the SFL director has a conversation with them. To retain its accreditation status, the EFL programme must ensure that all staff are actively engaged in quality improvement. In cases where instructors have shown reluctance to contribute, an honest conversation about the value of total participation in the process has yielded positive results.

Finally, broad, active engagement requires a high level of trust and what is known as psychological safety (e.g., Edmondson, 1999). In such a context, individuals feel a sense of security in the reporting of results, as well as in taking risks to suggest changes in instruction. If there is a sense of support, not fear of punishment or retribution, then instructors are more willing to collaborate and experiment. If it is clear that conversations focus on improvement, not punishment, then the door swings open to greater collaboration for the sake of quality improvement. Thus, as the Unit works to develop a culture of assessment within the EFL programme, it is very careful in the early stages to focus on results at the programme level, and avoid the isolation of data at the classroom level. As mentioned previously, eventually, instructors will begin asking for classroom level data for the sake of improvement. That occurrence will signal an environment of psychological safety and a learning culture.

From the outset of the quality assurance initiative, communication and transparency have been explicit objectives of the QAU. The need for clear, consistent communication cannot be overstated (Bresciani, 2009). Thus, it is important that the quality assurance team share relevant information with all stakeholders (Baker et al., 2012; Bresciani, 2012) and that information is presented in an "actionable" way (Banta & Blaich, 2011, p. 27). This includes intent, rationale, expectations, milestones, outcomes, and a proposed timeline. Transparency is essential for sustainability, and yet often ignored (Miller, 2012; New Leadership Alliance, 2012), as the perception is that it invites criticism. On the contrary, transparency contributes to confidence in a unit by demonstrating a commitment to improvement. As part of the communication plan, transparency in terms of activity and progress ensures that a larger body of stakeholders is aware and informed. Transparency and communication assist quality assurance planners in guarding against the perception that progress is not being made, or that only a select group is taking responsibility, which means that those not involved can ignore the process. Therefore, multiple channels of communication are advised—from staff meetings, to emails, to social media (tweets, Facebook posts, blogs), to conversations in the corridor or over coffee. The important point is that communication must be continuous *and* relevant. Busy stakeholders will quickly tune out from agenda items and emails that do not have a connection to their daily work lives. To this end, the QAU pushes out updates on quality assurance initiatives via its website. EFL staff members choose to subscribe to the website, and receive email announcements when new material (e.g., reports, initiative updates, conference announcements) is posted to the website. The QAU coordinator actively participates in as many meetings as possible-among coordinators or instructors-so that information is frequently gathered and information is shared. Establishing a visible, active presence across the EFL programme is a critical objective toward engaging EFL stakeholders.

Quality assurance is not a single event, carried out for the purpose of satisfying an accreditation visit. QA is one component in an integrated process that includes curricular design, teaching, assessment, professional development, and programme administration. In this way, the quality assurance initiative must be developed in such a way that is both integrated and systematic. And, equally important, it must also be perceived as such.

Even if a school does not have the resources (i.e., money, time, or people) to establish and staff an office, an attempt should be made to create a modest system that regularly monitors a manageable number of administrative and learning outcomes. Quality assurance should be approached from the perspective that relevant activities (i.e., outcomes and assessment identification, data collection and analysis) are formative and not summative; these activities should be ongoing and consistent. Without regular monitoring, whether semi-annually, annually, or even biannually, an EFL programme will simply fall into a reactive, rather than proactive mode of operation.

Moreover, if the programme is genuinely aiming for improvement, there will be a constant cycle of new initiatives that are designed, piloted, and assessed. Thus, in parallel with the continuous assessment rotation of core outcomes, there will also be a secondary level of evaluation involving ongoing research and development of new initiatives. For instance, since the 2014–2015 academic year, in addition to monitoring primary student learning and administrative outcomes, the QAU was involved in the development, implementation, and evaluation of such initiatives as Peer Observation for instructors, Individual Learning Plans for students, a First Year Experience, an Early Alert System, as well as a variety of professional development activities.

Crucially, all of the actions discussed earlier do not take place without the vision and support of leadership. Indeed, quality assurance and accreditation in the context of this case study are still more or less an institutional choice; they are not driven by external mandates. Leadership (i.e., the Director of the SFL) has been instrumental in initiating accreditation and structural change, as well as garnering support and resources from the university administration in order to bring these ideas to fruition. Additionally, Distributed Leadership (Spillane & Sherer, 2004; Spillane, Halverson, & Diamond, 2001, 2004) has played a role in the CPD unit's ability to gain a foothold in the SFL organizational culture. Distributed leadership is often misperceived as the distribution of power, when in fact it suggests the distribution of cognition-the spread of vision and values that sustains organizational efficiency and effectiveness. Although the SFL Director paved the way for the QAU by creating an office and a place on the organization chart, she has also stepped aside and allowed the QAU to set an agenda and run its own course. This is not to say that the QAU was simply a case of plug-and-play. There have been moments of anxiety as the existing structure and communications system adjust to this new entity. Again, however, leadership has played the critical role of stepping in on occasion to validate the Unit and reinforce the notion that the QAU is and will remain an integral part of the SFL structure.

6 Conclusions and Recommendations

Foreign language programmes, and EFL programmes in particular, are not often the focus of rigorous programme evaluation. Due to their nature as non-academic programmes, they are often viewed as external to the core operation of the higher education institution. This is unfortunate. As universities strive to differentiate themselves from the growing crowd, foreign language programmes have proliferated. And, as increasing numbers of schools provide instruction through English or another foreign language, such programmes are finding themselves as the gatekeeper to the university's core faculties and departments. Thus, it is logical that such programmes, which are often the largest units on campus, and the first stopping place for substantial numbers of students, should be required to ensure consumers of higher education that they are enrolling in a programme that is quality assured. Yet, there are relatively few EFL programmes that have structured quality assurance into their existing organization and operations, and fewer still that have sought accreditation.

This study has been a case study of an EFL programme in one university in Istanbul, Turkey that decided to pursue accreditation while also establishing a Quality Assurance Unit. The study described the development and implementation of a QAU and its quality assurance system. It also analysed the current state of the initiative, enumerating major themes that have emerged from the data collected over a 15-month period. Foreign language programmes that are considering the establishment of a quality assurance initiative, if not a Unit, should be able to glean lessons from the analysis provided within this study. At the same time, it is accepted that this study is limited in that the data is only for a 15-month period. A longitudinal study must be carried out in order to determine the sustainability of the quality assurance initiative and its long-term impact on the culture of the organization-will it transform into a culture of assurance and inquiry, or will status quo prevail.

Finally, this grounded study may contribute to the field of EFL programme evaluation research by providing a framework *vis a vis* the list of major findings reported in the analysis section. Future research may look to tighten this framework to a more succinct list of factors that affect success and sustainability. Bolman and Deal's Four Frames (2003) may serve as a means for analysing the successful quality assurance start-up, but it is arguably too broad for the specific intent of determining the critical pieces necessary for an effective quality assurance effort. EFL programmes are often too large and too instrumental to not be taken seriously. Regardless of whether institutional leadership elects to evaluate the quality of its EFL programme on a continuous basis, the leadership of the EFL programme should realize the importance of quality assurance, and strive to establish at least a system, if not an entity, that continuously asks the questions- How are we doing? Can we be doing better? If so, how?

References

Baker, G. R., Jankowski, N. A., Provezis, S., & Kinzie, J. (2012). Using assessment results: Promising practices of institutions that do it well. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA). Retrieved from: http://www.learningoutcomesassessment.org/UsingAssessmentResults.htm

Banta, T. W., & Blaich, C. (2011). Closing the assessment loop. *Change: The Magazine for Higher Learning*, *43*(1), 22–27. doi:10.1080/00091383.2011.538642

Banta, T. W., Jones, E. A., & Black, K. E. (2009). *Designing effective assessment: Principles and profiles of good practice*. San Francisco, CA: John Wiley & Sons.

Blaich, C. F., & Wise, K. S. (2011). From gathering to using assessment results: Lessons from the Wabash national study. *NILOA Occasional Paper*, *8*. Retrieved from: http://www.learningoutcomesassessment.org/occasionalpapereight.htm

Bolman, L. G. & Deal, T. E. (2003). *Reframing organizations: Artistry, choice, and leadership* (3rd ed.). San Francisco, CA: Jossey-Bass.

Bresciani, M. (2006). *Outcomes-based academic and co-curricular program review*. Sterling, VA: Stylus.

Bresciani, M. J. (2009) How to build and sustain a culture of assessment in your college and department. Slide presentation. Retrieved from: https://manoa.hawaii.edu/assessment/workshops/pdf/Bresciani%20MAC%202009-09-09.pdf

Bresciani, M. J. (2012). Recommendations for implementing an effective, efficient, and enduring outcomes-based assessment program. *Community College Journal of Research and Practice, 36*, 411–421. doi:10.1080/10668920902852160.

Coleman, J. A. (2006). English-medium teaching in European higher education. *Language Teaching, 39*(01), 1–14. doi:10.1017/S026144480600320X.

Creswell, J. W. (1998). *Qualitative inquiry and research design: Choosing among five approaches*. Thousand Oaks, CA: Sage.

Creswell, J. W. (2012). *Qualitative inquiry and research design: Choosing among five approaches* (3rd ed.). Thousand Oaks, CA: Sage.

Edmondson, A. (1999). Psychological safety and learning behaviour in work teams. *Administrative Science Quarterly, 44*, 350–383.

Fontana, A., & Frey, J. (1998). Interviewing: The art of science. In N. K. Denzin & Y. S. Lincoln (Eds.), *Collecting and interpreting qualitative materials* (pp. 361–367). Thousand Oaks, CA: Sage.

Gallagher, S. (2015). Competition can generate innovation and change. *University World News (23 January, 2015), 351*. Retrieved from: http://www.universityworldnews.com/article.php?story=20150121093519115

Glesne, C. (1999). *Becoming qualitative researchers: An introduction* (2nd ed.). New York: Longman.

Hersh, R. H. & Keeling, R. P. (2013). Changing institutional culture to promote assessment of higher learning. *NILOA Occasional Paper No.17*. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA).

Hutchings, P. (2010). Opening doors to faculty involvement in assessment. *NILOA Occasional Paper No.4*. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA).

Lennon, M. C., Frank, B., Humphreys, J., Lenton, R., Madsen, K., Omri, A., & Turner, R. (2014). *Tuning: Identifying and measuring sector-based learning outcomes in postsecondary education*. Toronto: Higher Education Quality Council of Ontario.

Maiworm, F., & Wächter, B. (2002). *English-language-taught: Degree programmes in European higher education*. Bonn: Lemmens.

Maki, P. L. (2004). *Assessing for learning: Building a sustainable commitment across the institution*. Sterling, VA: Stylus.

Miller, M. A. (2012). From denial to acceptance: The stages of assessment. *NILOA Occasional Paper*, *13*. Retrieved from: http://www.learningoutcomesassessment.org

New Leadership Alliance for Student Learning and Accountability. (2012). *Committing to quality: Guidelines for assessment and accountability in higher education.* Washington, DC: Author. Retrieved from: http://www.chea.org/pdf/Committing%20to%20Quality.pdf

Spillane, J. P., Halverson, R., & Diamond, J. B. (2001). Investigating school leadership practice: A distributed perspective. *Research News and Comment*, 23–28.

Spillane, J. P., Halverson, R., & Diamond, J. B. (2004). Distributed leadership: Toward a theory of school leadership practice. *Journal of Curriculum Studies, 36*(1), 3–34.

Spillane, J. P. & Sherer, J. Z. (2004). *A distributed perspective on school leadership: Leadership practice as stretched over people and place*. AERA San Diego.

Stake, R. E. (1995). *The art of case study research*. Thousand Oaks, CA: Sage.

Sue, V. M., & Ritter, L. A. (2012). *Conducting online surveys*. Thousand Oaks, CA: Sage.

Tagg, J. (2007). Double loop learning in higher education. *Change: The Magazine of Higher Learning, 39*(4), 36–41. doi:10.3200/CHNG.39.4.36-41

Turkish Higher Education Council. Retrieved from: http://www.yok.gov.tr/web/guest/universitelerimiz

Weick, K. E. (1976). Educational organizations as loosely coupled systems. *Administrative Science Quarterly*, 1–19.

Evaluation in Tunisia: The Case of Engineering Students

Rym Jamly

Abstract This chapter reports on one aspect of a research study conducted within the framework of needs analysis in English for Specific Purposes (ESP). A triangulation of both methods and resources was employed for data collection. While a questionnaire was addressed to all the participants in the present study, a structured-interview was further conducted with the ESP teachers. The content of the adopted instruments was, in most cases, duplicated among informants to ensure the reliability and consistency of results. Though the study did not draw a sound conclusion about the relevance of the Test of English for International Communication (TOEIC) to the learning and target needs of the students, it revealed that more than half of the former graduates who are also actual employees did not attach great importance to the test as a valuable tool to evaluate their communication needs as required by their potential employers. This chapter offers a window of opportunity to build the bridge between learning objectives and learner evaluation in ESP which is a widely neglected area of research especially within the Tunisian context. The real shift resides in recognizing all ESP stakeholders' perceptions of evaluation including learners themselves and the implementation of outcome-based measuring tools reconciling standardized output-oriented assessment with test reliability, authenticity, and impact.

Keywords English for specific purposes (ESP) · Needs analysis · Language testing

1 Introduction

The teaching of ESP in Tunisia has long been characterized by the lack of professional standards and formal training on the part of the teacher who is generally deemed responsible for the whole learning and teaching process. Local and regional

R. Jamly (✉)
Higher Institute of Legal and Political Studies of Kairouan, Tunis, Tunisia
e-mail: rym.jamly@gmail.com

S. Hidri and C. Coombe (eds.), *Evaluation in Foreign Language Education in the Middle East and North Africa*, Second Language Learning and Teaching,
DOI 10.1007/978-3-319-43234-2_17

research studies of ESP in the Arab world where English is taught as either a foreign or a second language pervaded the scene of English language Teaching (ELT) and continued to prosper in response to the ever-changing demands of national and international businesses. Local studies conducted within the framework of M.A. and Ph.D. research projects into ESP learning and teaching revealed that ESP students need to acquire specific target skills, most notably the receptive reading skill, and that ESP courses should be geared towards their specific learning needs with respect to both learning and working environments.

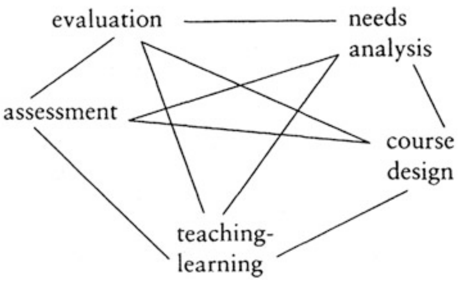
Despite the government's commitment to successive educational reforms, English competence remained beyond the desired level and no formal decisions have been made to revise and enhance the ESP situation. Yet, the future of ESP in Tunisia seemed to follow a wrong path despite its application on a wide range of academic domains at an advanced level namely business, economics, sciences, medicine, and engineering. Notwithstanding the bulk of local research on needs analysis, the fact remains that none of them has been extended to inform the development of ESP test content and task types. The focus of this chapter addresses this research gap in the Tunisian context.

2 Theoretical Background

The term evaluation has always been associated with any language teaching course, though rarely in the general English context (Hutchinson & Waters, 1987). Yet, in ESP as a goal-oriented undertaking par excellence, there exists two layers of evaluation which relate to both the learner and the course. Dudley-Evans and St. John (1998) advocated that evaluation plays a prominent role within the ESP process as is shown in Fig. 1.

Figure 1 suggests that the design of the ESP course depends mainly upon needs analysis which in turn relates subtly to evaluation and assessment. These two latter are conceived to mimic the effectiveness or otherwise of the course at large. Hutchinson and Waters (1987) maintain that assessing the learner performance and evaluating the ESP course helps establish the existing educational cracks and gives feedback on how to repair them especially when it comes to course objectives and

Fig. 1 Stages in the ESP process (Dudley-Evans & St. John, 1998, p. 121)



educational needs. Course evaluation is generally conducted by means of needs analysis which involves all the ESP stakeholders including the learner. Learner assessment in ESP concerns itself with the level of language proficiency needed to perform successfully in the target situation of language use whether academic, professional, or vocational.

Specific purpose testing is generally viewed as criterion-referenced unlike testing for general purposes, which is considered to be norm-referenced. While a norm-referenced test relates to the test taker's "rank" in relation to others, a criterion-referenced test informs the test taker's "level of performance" (Bachman, 1989, p. 248). This would typically imply in a way the theoretical orientation of general English testing and the practical consideration of ESP testing. This implication, however, does not deny that general English test developers eschew the practical outcomes of the test scores, nor does it signify that ESP tests lack the required theoretical basis. The fact remains that ESP test content and methods should be mainly derived from an analysis of a target language use (TLU) situation in terms of the tasks that test takers would typically perform.

In fact, the process of test development and validation in an ESP context is fundamentally a purpose-based of the ilk. Accordingly, an ESP teacher might devise a test to place the learner's background knowledge in accordance with the ESP course (placement test), to ascertain that what is being taught is actually being learnt (achievement test), and to determine the degree of the learner's proficiency level with reference to the required target language tasks (Hutchinson & Waters, 1987). The assessment criteria for the latter, however, are rarely based on a thorough analysis of the target language use situation. Instead, they appear to be primarily language theory-driven (Jacoby & McNamara, 1999), neglecting the specific situations in which language is to be used.

Administering language tests for a specific group of learners might present test developers with controversial issues, including test content's specificity as well as test reliability and authenticity. Bachman and Palmer (1996) summarized and translated these issues into a model, which includes a set of test qualities accompanied by three principles believed to be the basis for the usefulness of a test. These principles apply mainly to the specificity of the purpose of the test, test takers, and the testing situation. Douglas (2000) advocated that there exists a continuum of specificity that is dictated by the intended purpose of a given language test. The question arises as to how specific the test should be, and to what extent it is applicable to all language use situations.

Another issue concerns the question of the interplay between assessing subject matter content knowledge and/or language background knowledge. With reference to a test for trainee doctors, Bachman and Palmer (1996) opined that in order to maintain the balance between language proficiency and content knowledge, a 'knowledge test' is in order. Cognitively speaking, it was believed that language and non-language ability account for the two sides of the same coin to the extent that they should be viewed as "inextricably linked" (Douglas, 2000, p. 39).

The relation between the language course and assessment has long been raised as a major concern in LSP classrooms. In an attempt to justify the lack of evidence for testing in ESP, Alderson and Waters (1983) referred to the influence that tests exert on teaching as the "washback" effect. They suggested that a test, whether good or bad, is intended to have a definite effect on the teaching process. Hughes (1988, p. 145) further claimed that because ESP is fundamentally needs-based, "teaching for the test (…) became teaching towards the proper objectives of the course". The washback not only affects the learning and teaching process; but also gives feedback to teachers, course developers, and even learners on how to proceed within the ESP classroom.

Task authenticity has also been discussed among the major factors affecting the good quality or otherwise of an ESP test. Bachman and Palmer (1996) defined the concept as the extent to which TLU tasks correspond with the test tasks. In other words, had the language test been authentic, test takers would have found themselves engaged in a set of tasks that perfectly mirror the kind of tasks they would perform in the real-life language use situation. This is to highlight the critical relevance of needs analysis to evaluation in ESP. A useful test is by definition that which abides by the requirements of specific TLU domains which in turn is investigated by means of target situation needs analysis. Similar to authenticity is test task iterativeness which reflects the degree of correspondence between the test taker's personal characteristics and both the test tasks and the TLU tasks. Those characteristics which relate to "the test taker's language ability (language knowledge and strategic competence, or metacognitive strategies), topical knowledge, and affective schemata are in turn intended to interact with each other to serve the iterativeness of the task" (Bachman & Palmer, 1996, p. 25).

In short, evaluation in ESP is by no means done at random. Several dimensions contribute, whether directly or indirectly, to the effectiveness and utility of the test. In addition to reliability, authenticity, interactiveness, and impact; Bachman and Palmer (1996) identified the notion of construct validity as a defining feature of useful tests. The construct or the ability to perform specific tasks in the TLU domain should be evidenced by the valid interpretation of test scores and not by the validity of test scores themselves. The study will demonstrate how this quality is of paramount importance especially when it comes to ESP proficiency tests.

The main rationale behind the present research project lies in the established fact that no formal needs analysis study has been conducted on the ESP course of the target population. More precisely, it aims to identify and describe the second-year engineering students' needs and attitudes toward the ESP course content in an attempt to set forth insightful recommendations for a more promising ESP course that best meets both the students' academic and professional 'expectations'. This goal-oriented study tries to raise the target students' awareness about their current ESP learning situation and sensitize them to the language flaws and help them formulate a conclusive conception about what they really need to learn in English with regard to the four language skills.

The fundamental research problem upon which the study hinges pertains to the argument that if learners' needs are not analysed and assessed, they may miss the goal behind the whole teaching process. Very much in the same vein, they may come to realize their failure to be operational in their field of work. Given the fact that there has been no empirical investigation of the students' present and target needs, the present research intends to probe the following research questions:

a. What are the students' ESP learning needs and target needs?
b. Do the students' needs match the content of the ESP course?
c. What are the implications of analyzing the needs of these learners?

3 Method

3.1 Context of the Study

The present study was carried out at the Higher School of Communication of Tunis (SUP' COM), an engineering school founded in 1998 and placed under the tutelage of both the Ministry of Higher Education and the Ministry of Communication Technologies. As far as the distribution of ESP courses along the school educational cycle is concerned, they take place over the three first semesters with a total number of 42 h each semester. During the first week preceding the fourth semester, all the students get into an English language-training course that encompasses thirty hours of practice for the evaluation of students' communication skills in their future professions.

With this end in view, the main assessment tests adopted to measure English proficiency at the school were the Test of English as a Foreign Language (TOEFL) and the Test of English for International Communication (TOEIC). SUP'COM was selected among many institutional settings since it is the first Tunisian higher education institution that took the initiative to incorporate such standardized tests in response to the important role allocated to the English language for the academic and professional success of its students. This perceived priority could also be manifested by the collective decision to render the TOEFL previously and the TOEIC currently an essential prerequisite for the graduates to receive their diploma.

3.2 Participants

Multiple sources were selected as the main participants in the present study: The target students, the ESP teachers, the subject-matter teachers, and the former students. Both questionnaires and structured interviews were given to a random sample with the aim of building as representative a picture as possible. The target students' questionnaire was filled out by 153 respondents representing 85 % of the total number of second-year Telecom engineering students (180 students). As for the ESP teachers, the totality of the three female English teachers successfully responded to the questionnaire and structured interview items. A random sample of former graduates working for international companies in Tunisia and abroad were also invited to participate in the present study. The sample is composed of 50 employees aged between 25 and 29 years all of them holding the National Diploma in Telecom Engineering. As regards the subject-matter teachers, a questionnaire was sent by e-mail and only 28 out of 53 replied.

3.3 Instruments

The choice of data-gathering instruments is fundamentally influenced by Mackay (1978) who stipulated that there are two basic research tools for the collection of data, namely the questionnaire and the structured interview. These methods were described by Long (2005) as “deductive” measures whereby the researcher logically arrives at reliable conclusions about learners’ needs. They proved highly beneficial to elicit both qualitative and quantitative information for both the Target Situation Analysis and the Present Situation Analysis. Decisions about the design and the content of the questionnaire rested mainly upon Hutchinson and Waters (1987) theoretical framework for analysing both learning and target needs which highly relates to the Munbian model (1978) together with the McDonough and McDonough (1997) taxonomy of question types. The formulation of questions fell mainly under six main sets, namely factual, yes/no, multiple-choice, ranked, scaled, and open-ended. Drawing on the general purposes of the study, each question and sub-question concerned itself with a specific purpose on the basis of different “variables” under assessment (Oppenheim, 1992).

4 Data Analysis

For the purposes of triangulation, a combination of qualitative and quantitative research methods was used to glean and analyse data about the sample population. The idea behind collecting qualitative data relates to the search for patterns about the respondents’ perspectives, perceptions, opinions and explanations towards specific issues. The data was mainly gathered through open-ended questions the answers of which came up in a textual format that lent itself to thematic categorization. This kind of data was present in both questionnaires and structured interviews and was analysed using inductive reasoning whereby results were observed and presented in graphs and tables. Quantitative data was also used to serve statistical purposes.

The data which can be measured and counted was mainly collected by means of close-ended questions in questionnaire and structured interview formats with the aim of producing reliable and generalizable results. The latter were deductively analysed using descriptive and inferential statistics. The analysis process started with manually coding the responses to each questionnaire and interview item with the help of the statistical package (SPSS 20) which was also used to generate frequencies and percentages for each coded item among the multiple human sources involved in this study.

5 Results and Discussion

Overall, the analysis of the findings related to the perceived level of English supported the view that the second-year Telecommunications engineering students at SUP'COM need to develop their general proficiency in English. This conclusion was reached in view of the apparent discrepancy between the actual level of students as perceived by their ESP teachers and themselves as opposed to the target level as perceived by the graduates. Not only was there a dispute between the students (59 %) who reported that they are "good" at English and all ESP teachers who indicated that the vast majority of the learners have a "medium" level of English, but also between ESP teachers and graduates (72 %) who believed they enjoy a "good" level. What strengthened their belief was their reported TOEIC/TOEFL scores. This would imply that a "good" level of proficiency in English is at least required by potential employers in the target working situation. Thus, it suggested that students need to improve their current level of English which would in turn maximize their chances to be appointed to their potential jobs.

In general terms, there was a consensus among all the participants in the present study on the need for a combination of both specific and general English. However, it is interesting to notice here that 80 % of the graduates reported that they learnt "general English" in addition to one of the ESP teacher's comment that her choice about the teaching of both Englishes applied only to the first-year English program and that second-year TOEIC preparation sessions could not be described as either general or specific. These two findings would suggest, as it was concluded before, that the ESP course is not specific enough or is too general considering the requirements of the target situation English use.

When asked about whether the ESP course met their needs as future engineers, half of the graduates disagreed. Again, it is worth noting that 34.4 % of the students expressed their dissatisfaction to the fact that the course meets their needs as future engineers. It follows then that both students and graduates are aware that their classroom learning needs in English differ from their needs at the work place and both seem to be crucial to the successful completion of their studies and future professional career. In response to the question pertaining to the degree of importance of English for their current job, this claim was largely illustrated by 78 % of the graduates who reported that the language was "very important".

Bearing in mind that the ESP course in focus is a classroom-based English language training, the ESP teachers took the responsibility for the implementation of the TOEIC as one of the popular assessment tools to measure the students' required proficiency in the target situation. In this respect, it seemed odd to find that all the participants in the present study approved the necessity and importance of the test as one of the learners' job requirements except for 64 % of the graduates who indicated that the test did not constitute a necessary step towards the successful achievements in their current job. It is thus urgent to confirm the graduates' opinion through recourse to the future potential employers' corresponding point of view, which presented one of the big limitations that faced the researcher during this study.

In conclusion, the findings revealed that the target engineering students were expected to be engaged in real-world job chores where both accuracy and fluency are highly required. This conclusion was further emphasized when both students and graduates indicated that they mostly expected more "speaking" in the first place and "specific vocabulary" in the second place in response to their expectations vis-à-vis the ESP course. It follows then that the telecommunications engineering profession instrumentally requires English as an employability skill and this was obviously illustrated by the subject matter and ESP teachers' emphasis on "practice" and "communication" when asked about what students should do in order to be operational in their field of work.

6 Implications and Recommendations

Similar to the centrality of NA is the importance of evaluating learners' performance and the effectiveness or otherwise of the ESP course. Learners' assessment and course evaluation are considered two sides of the same coin. Both were meant to contribute to the process of satisfying students' needs and meeting the course objectives. As for learners' assessment, Hutchinson and Waters (1987) distinguished between three different types of learners' assessment characteristics of ESP, viz. placement, achievement, and proficiency tests. The ESP course offered to the second-year telecommunications students at SUP'COM opens a window of opportunity for them to receive an intensive language training course with the aim of obtaining the TOEIC certificate as a passport to success in their professional life. This test can be classified under both proficiency and achievement tests since it aims to measure the learner's progress until s/he demonstrates his/her ability to meet the demands of the work place by obtaining the required score.

Yet, what seemed to be neglected by ESP teachers was the importance of the placement test as a measurement tool that helps specify the learners' initial background knowledge. This negligence could be justified by the ESP teachers' suggestion to work with homogeneous mini-groups instead of dividing the total number of students by the available three teachers. Yet, course evaluation was by far the most important form that caught most of the attention of the ESP teachers. As the present study revealed, ESP teachers were aware of their students' learning and target needs. This awareness generated their collective decision to substitute the Test of English as a Foreign Language (TOEFL) with the more purpose-oriented

TOEIC. In general, specific purpose language testing has been a subject of considerable debate particularly in relation to general purpose language testing. Notwithstanding these misconceptions, the fact remains that ESP tests differ from those of General English (GE) in that they should be language-specific, content-related, and most importantly, purpose-based.

In an attempt to promote the practice of ESP testing at SUP'COM in particular and in any educational institution in general, a number of suggestions and recommendations are offered based on the main conclusions drawn from this large-scale research project. Firstly, it is highly recommended that ESP testing be based first and foremost on well-founded needs analysis. In view of the significant role English plays at the target work situation, the learners' real world needs especially communication needs should be given more prominence and priority. The testing of ESP should reflect the way English is likely to be used in terms of activities, tasks, modes, channels, identity of communicators, setting of communication and the like.

Second, as far as proficiency tests are concerned, there should be crystal clear purposes for which test developers administer the test and test takers sit for the same test. Raising awareness of the intended purpose of the test helps maximize its benefits both on the learners' involvement and responsibility towards their own learning and the teachers' appreciation of test scores. This would also affect the content, the required level of proficiency, and the value of the test.

Finally, yet importantly, testing in ESP should be viewed as a two-fold procedure, i.e., assessing the learners' ability and providing feedback accordingly. Test scores or outcomes should be analysed regularly and conventionally in order to ensure test validity and reliability. This needs in turn be set in view of the changing nature of the international job market demands as well as the individual value judgments of the key stakeholders of the ESP practice.

7 Conclusion

The purpose of this chapter is twofold. It seeks to establish the link between ESP evaluation and needs analysis on the one hand, as well as language testing and language teaching on the other. The main concern of ESP has always been the use of language and not the language per se. The teaching of ESP proved to involve much more than adapting the course content to the specific purposes and the communicative needs of the learners. Rather, it entails a variety of evaluation procedures to ascertain the learners' ability to perform predetermined communicative tasks in the target situation. Language testing in ESP is also a matching process, which is meant to offer feedback on both the performance of the learners and the effectiveness of the course. The main argument relates to the fact that ESP assessment abides by specific criteria that should be derived from an analysis of the TLU context, most notably authenticity and reliability.

References

Alderson, J. C., & Waters, A. (1983). A course in testing and evaluation for ESP teachers. In *Lancaster Practical Papers in English Language Education* Vol. 5. Oxford: Pergamon Press.

Bachman, L. F. (1989). The development and use of criterion-referenced tests of language ability in language program evaluation. In R. K. Johnson (Ed.), *The second language curriculum* (pp. 242–258). Cambridge: Cambridge University Press.

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Douglas, D. (2000). *Assessing language for specific purposes*. Cambridge: Cambridge University Press.

Dudley-Evans, T., & St John, M. J. (1998). *Developments in ESP: A multi-disciplinary approach*. Cambridge: Cambridge University Press.

Hughes, A. (1988). Introducing a needs-based test of English language proficiency into an English medium university in Turkey. In A. Hughes (Ed.), *Testing English for university study* (pp. 134–146). (ELT Documents#127). London: Modem English Publications in association with the British Council.

Hutchinson, T., & Waters, A. (1987). *English for specific purposes: A learning-centred approach*. Cambridge: Cambridge University Press.

Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes, 18*(3), 213–241.

Long, M. H. (2005). Methodological issues in learner needs analysis. In M. H. Long (Ed.), *Second language needs analysis*. Cambridge: Cambridge University Press.

McDonough, J., & McDonough, S. (1997). *Research methods for English language teachers*. Great Britain: Arnold.

Mackay, R. (1978). *Identifying the nature of learners' needs*. In Mackay & Mountford (Eds.), *English for specific purposes* (pp. 21–42). London: Longman.

Munby, J. (1978). *Communicative syllabus design*. Cambridge: Cambridge University Press.

Oppenheim, A. N. (1992). *Questionnaire design, interviewing and attitude measurement*. London: Pinter Publishers.

Part VIII
Assessment Literacy and Dynamic Assessment

Why Should the Assessment of Literacy in Morocco Be Revisited?

Abdelmajid Bouziane

Abstract This chapter surveys the assessment of reading and writing in Morocco. Despite the scarcity of research in these domains, some interesting findings and suggestions are analyzed. The assessment has been found to be mostly unfair and to harm students both during and after school. Also, teachers are only fairly satisfied with the types of assessment in place. In reading, the studies show that high-order questions are less frequent and that such shortage has negative effect on students' performance in this skill. In writing, the findings of research reported in this chapter show varying degrees of inconsistencies in scoring essays. Instead, the suggestions include adding objective tests to free composition format at lower levels, adopting the analytic method, pre-determining the features to be scored (i.e., setting scales), describing the scales as clearly as possible, considering the appropriate weight for scales in accordance with the students' level and the purpose of evaluation. Alternatively, some recommendations are put forward, such as boosting research in different areas of language education, teacher training at all levels in how to design, administer, and interpret tests' outcomes pedagogically and statistically.

Keywords Assessment literacy · Reading · Writing · Rating inconsistency · ELT · Teacher training · Research · Morocco

1 Introduction

Assessment has always been an essential component of the teaching process. Teachers are generally not only involved in designing, administering, and scoring tests but also in interpreting outcomes, albeit they may not be always reasonably

A. Bouziane (✉)
Faculty of Letters and Humanities Ben Msik, Hassan II University of Casablanca,
Casablanca, Morocco
e-mail: abdelmajid.bouziane@gmail.com

S. Hidri and C. Coombe (eds.), *Evaluation in Foreign Language Education in the Middle East and North Africa*, Second Language Learning and Teaching,
DOI 10.1007/978-3-319-43234-2_18

knowledgeable about all these matters. The ways testing is carried out have been criticized everywhere (cf., Linn, 2000 on testing in the US). In Morocco, voices have articulated their dissatisfaction regarding testing and have suggested revisiting it. Testimonials to such dissatisfactions are reported in this paper which focuses on assessing reading and writing. The paper also provides recommendations in an attempt to remedy some of the flaws.

2 Theoretical Background

2.1 Overview of Assessment System in Morocco

In the primary and secondary levels, Morocco has three high-stake exams. The first one takes place in the last year of the primary level in which young learners sit for an exam in the core subjects of languages (Arabic and French), literacy and numeracy. The second is in the end of middle school during which only a quota of students that high schools can accommodate passes. The third one is the Baccalaureate which is based on a pass/fail system regardless of the quota. Students who score a mean of 10/20 (50 % in the international system) pass this exam. The breakdown of the score calculation comes to 25 % obtained from the regional exam in which students take minor subjects a year before, 25 % from continuous assessment, and 50 % from the core subjects that are assigned depending on the area of specialty. English happens to be among these core subjects but with a low coefficient in science and technical streams. The exam consists of a test that contains a reading text and comprehension question followed by exercises of language and vocabulary and then by writing. In the tertiary level, the implementation of the LMD (*Licence, Master, and Doctorat*) system in Morocco has resulted in two semesters a year with three major types of exams in each semester: Continuous assessment, the end of the semester exam and the retake exam for those who do pass all exams in the semester. The scoring system is compensatory both across modules and across semesters. Masters follow the same system and doctorates follow the international system of the supervisor giving the greenlight to defend the thesis and giving the manuscript to members of the jury who write their reports and, in most cases, examine the work in a public viva.

It is worth mentioning that speaking and listening are excluded from the assessment of English in the Baccalaureate exams despite the fact that teachers devote a big proportion of the classroom time to speaking, and by extension to listening, and young adults and adolescents work for long hours on the Internet and social networks watching and/or listening. Therefore, the youth do not seem to be fairly rewarded in the skills they have developed by themselves and which are as important as other skills.

2.2 Research into Testing in Morocco

Despite the scarcity of research on testing in Morocco, the available body of research tends to question existing assessment practices. El Mazgualdi (1996) conducted an empirical investigation on the psychological impacts of testing on adolescents' attitudes. He found that in case of success, adolescents formulate positive attitudes towards learning and believe that success in studying entails success in life. Similarly, failure leads to negative attitudes towards self and school and to a strong belief that failure will accompany students even after school.

In a comprehensive Ph.D. research study, Ouakrime (1986) evaluated the purposes, the practices and the effects of English in higher education in Morocco. To do so, he administered questionnaires to students from different levels of the Department of English in Fez, teachers from five departments of English throughout the country, and interviews 8 of his colleagues in the department of English. He came to the conclusion that the examination system "increases the risk of exam results being perceived more as subjective judgements made on student performance than a 'fair' evaluation of the range of their knowledge and abilities (p. 238)". He added that this system needed more inspiration from the existing literature, more democracy and transparency as it lacked visibility and feedback provisions. Although, there have been three major reforms, in 2003 when the LMD was introduced and in 2009 with the reform of master and doctoral programs and in 2014 when new official handbooks of standards for each track were introduced, the situation has not improved substantially compared to what Ouakrime described. These new standards stipulated that all the exams must be written. However, there are subjects that need to be tested orally such as Public Speaking or Quran recitations.

Specifically, in ELT, the available body of research, though small, suggests that this area is not immunized from flaws. Hammani (1995) analyzed the scores allotted by teachers in both continuous assessment and academy exams. He found that not only were the marks of continuous assessment inflated but also did not correlate significantly with the academy exams ($r = 0.38$, not significant). Melouk (2001) reported on a national survey involving all EFL test designers in different academies and 563 teachers representing all the regions in the country. The results show that these parties were only partly satisfied with the outcomes of item-bank system as it was implemented in Morocco. The teachers in the survey reported that there was a high degree of adequacy between textbooks and official exams (44.2 %) whereas 33 % thought there was a moderate match and only 5.5 % claimed a perfect match. Particularly, teachers felt marginalized in the processes of item-bank (or test) construction. The survey also reported the lack of teacher training in testing as most teachers underwent short training periods of two to five days. In his concluding remarks, Melouk claimed that evaluation in Morocco "is still in its embryonic stage" because of "the lack of training of both teachers and supervisors in this area (…) Scarcity of field research in this area in Morocco" (ibid. p. 63).

3 Assessing Reading

The small body of research conducted on assessing reading in the Moroccan educational system, at least to my knowledge, suggests that the textbooks used in language education and, probably, the teaching practices to train students in answering thought-demanding questions partly prepare students to become effective readers. Ezzaki (1986) investigated the types of questions posed in some Arabic and French textbooks used in the different levels of language education (primary and secondary schools) in Morocco, and some Baccalaureate tests of English. He found that high-order questions lacked in the investigated data compared to foreign textbooks. A similar design was used by Melouk (1992) and Boubekri (1997) to investigate, respectively, questions posed to test various subjects (including English) in the Baccalaureate exam and a textbook in use then titled *Bridges*. Both researchers confirmed the scarcity of thought-provoking questions. The more frequent questions happen to be the bottom levels of Bloom et al.'s taxonomy (1956). In brief, these research studies show that the types of reading tests as well as cognitive abilities to which they appeal touch only partly upon the overall entirety of comprehension. That is, the assigned activities tend to neglect the thought-demanding levels that enable to develop critical thinking and higher-order reading skills.

It is not surprising, then, that students poorly perform the neglected levels of reading skills. Bouziane (1993) contended that secondary school students were unable to answer high-ordered questions both in English and in French. He has also found that teachers of English and those of French show more inconsistencies in correcting higher-order questions than in correcting low-order ones. The reasons for such poor performances may be mainly the students' imperfect (use of) reading strategies. Also, the Moroccan students' performance in reading is reported to be poor. The Superior Council of Education and Training and Scientific Research (2009) conducted a large-scale study, labeled the National Programme for Education Achievement Testing, on a sample of 26,520 learners and concluded that they showed modest performance in all evaluated subjects. The overall mean of the scores hardly reached 50 % of what the students were supposed to attain by the end of the year in all subjects. A study by the RTI International (2012), titled *Early Grade Reading Assessment*, shows that students (n = 773) in Grade 2 (G2) and Grade 3 (G3) in 40 schools in a region in Morocco reported that 50 % were unable to respond correctly to a single comprehension question. The overall performance by students did not go beyond 18 % of correct answers. Internationally, literacy in Morocco needs far more improvements to gain a better position. In the PIRLS (the Progress in International Reading Literacy Study) 2001, Morocco was classified next to last among the 35 countries that participated in International Student Achievement in Reading (Mullis et al., 2003). In the PIRLS 2006 (Mullis et al., 2007), Morocco regressed and so it did in PIRLS 2011 (Mullis et al., 2012). Ibourk (2013) and Gustafsson et al. (2013) show close relations between the PIRLS poor results and socio-economic and school-related factors whereas the RTI

International's research (2015) points to the negative effect of textbooks on reading at early ages and to teachers who have shown few significant relations between their perceptions and their practices (RTI International, 2014a, b).

This body of research shows that Moroccan students need more adequate training in reading effectively and, inevitably, their performances in reading are far below those of proficient readers. Teachers cannot blame other parties for such a deficiency. The *Official Guidelines* (Ministry of Education 1994, 1996a, b, 2007) have reiterated assigning inferential questions in various open-ended types of questions taking into account the variety of levels of inferencing. Some reading models in the Guidelines do contain high-order questions (e.g. Ministry of Education 1996a); however, more challenging tasks which call for analysis, synthesis, and evaluation should be added. Generally, the teaching of reading in EFL classes in Morocco needs more improvements, particularly in the teaching and testing of higher-order and thought-provoking abilities (Bouziane 1997).

4 Assessing Writing

In writing, the available studies confirm that "no news is good news." Bouchouk (1987) investigated the way teachers of Arabic scored compositions. Although he provided detailed scales and rubrics based on the analytic method of scoring, the raters came up with significantly inconsistent scores. This implies that scoring is likely to be arbitrary. Similarly, Dahbi and Britten (1989a, b) referred to scoring written discourse in high schools as being a staircase phenomenon. That is, compositions are thrown on stairs marked with figures and each composition is awarded the score of the stair on which it lands. This metaphor describes inconsistent scoring. As a remedy, the two researchers conducted a four-phase study whose optimal outcome was to train a group of teachers in using an adapted analytic method stipulated with prior training in understanding and applying its scales. Here is how their experiments unfolded. In phase 1, Dahbi and Britten (1989a) compared the reliability of six composition scoring methods: Three are analytic (descriptive categories, rhetorical-analytic, analytic) and the other three are global (traditional, double traditional, ranking plus traditional). Despite the Traditional and Double Traditional methods being ranked top followed by Ranking plus Traditional because of teachers' familiarity with them, no single method was found to yield a significantly high inter-rater reliability coefficient. As a result, training in methods with explicit scoring criteria was an alternative.

In phase 2, Dahbi and Britten (1989b) compared the reliability of the three analytic scoring methods. The analytic method yielded better results perhaps. A modified version containing such criteria was developed. In phase 3, they compared the reliability of the modified analytic method which contained language and communication criteria. However, this modified method revealed that coherence and completeness posed more problems to teachers than the other criteria, especially language-related ones. The outcomes showed higher inter-scorer reliability, though statistically insignificant because of insufficient training. In the last phase, 4, they checked the reliability of the modified analytic method after teachers' adequate training in using it. This time the results yielded more consistent scoring with the highest inter-scorer agreement. Some factors that resulted in the above successful results were adequate school-based training over time, the use of mixed ability compositions for the training sessions, and the discussion of participants' scoring. Apart from a few hints, very little information is known about the training package these two researchers suggested and probably only little training has been provided based on their recommendations.

Tamek (1989) looked into whether three scorers' (tutors) judgements using the analytic method would discriminate 19 advanced university students' essays. He also compared students' written and spoken products and checked the effect of orality features on rhetorics. To do so, he resorted to quantity, syntactic complexity, and error variables together with occurrence of six orality features. His findings showed that formal analysis did not discriminate significantly between proficient and poor writers. They also confirmed a degree of maturity through the absence of transfer from spoken to written mode. However, the permeation of orality features in functional level discriminated levels and correlated significantly with the scorers' marks.

Naciri (1995) analyzed teachers' (n = 103) responding to students' (n = 618) writing and found that there was a big discrepancy between teachers' beliefs and their actual responses to students' drafts. This implies that these teachers did not refer to stable and collectively shared standards when responding to drafts and, by extension, when scoring final drafts. Particularly in scoring, she found that the same compositions were granted better scores by secondary school teachers than by university teachers, hence inconsistent benchmarking reference. In responding to student drafts, she reported that teachers used a combination of old and new responding techniques but tended to respond to final, rather than evolving, compositions. She also reported the predominance of form-related and negative comments. The predictors of writing quality, in a descending order, came as follows: Vocabulary, organization, grammar, content, and mechanics. Similarly, the scorers in Nadri's (2005) study (n = 6) agreed more on language aspects. He found that inter-scorer reliability was more noticeable in language than in content features. However, this was driven by the scorers' views of the nature of writing as he says: "It seems that the participants [teachers] do not share the same view about writing ability. Whereas some view writing mainly as a linguistic ability, others consider it to be a rhetorical activity (p. 128)." Also, the weighing and priorities granted to criteria played an important role in scorers' (dis)agreements. Interestingly, he concluded that discrepancies occurred more in average ability performances than in high or low ability ones.

Ennaji (1987) scaled up testing of writing across levels and suggested an approach to scoring. He claimed that testing should reflect the teaching purposes and, thus, controlled, directed, and free compositions should be evaluated differently. He suggested that teachers should adopt the analytic method and they should give more weight to form at low levels and gradually shift to content in free composition.

Sadiqi (1986) cited some problems of composition teaching and learning and provided some tentative solutions. She explained that the teaching problems lied in unclear objectives, absence of teamwork, large classes, and subjective marking methods and the students' problems in L1 and L2 interference, little link between the courses they took, infrequent writing, micro/macro structure problems. Alternatively, she suggested evaluating the teaching, encouraging reading, adopting systematic correction of errors, sensitizing students to written discourse and genres. Bellout (1990) discussed assessing writing in tertiary low levels and classroom and exam assessing techniques. She suggested that in-class assessment should address accuracy, fluency, and rhetorical features and that besides free composition, objective tests should be assigned to students at low level. Bouziane (1991) suggested adopting more effective ways of teachers' and peers' feedback to student writing and changing attitudes towards errors. He encouraged self- and peer-correction but stipulated it with careful planning and understanding. He also called for adopting positive attitudes towards errors and favoring the correction of global errors over local ones.

The above state-of-the art of testing provides enough evidence to reconsider testing in Morocco in general and in ELT in particular. All the documented studies call for further teacher training. The common areas of training are test design, writing items, using statistics, using computers, conducting research in testing, and testing matters in general. El Mazgualdi (1996) testifies to this when he writes: "We have not found a study devoted specifically to educational evaluation of the Moroccan system of education (p. 4; my translation)."

5 Recommendations

Some recommendations for a better practice of testing in Morocco seem necessary:

(a) State-of-the art articles should be publicized to share the findings of research, bearing in mind that this latter is in general conducted for academic purposes.
(b) Surveys, like the one conducted by Melouk (2001), and reports are highly recommended to identify teachers' and test designers' attitudes and needs. These will also help evaluate the existing testing practices.
(c) Researchers should be encouraged to conduct research on testing. Scientific productions, such as articles, books, and technical reports and are highly welcome because the available books are textbooks about testing and assessment in general rather than specifically dealing with assessment in Morocco (cf. Madi, 1990; Fatihi, 1995, to cite only a few). Experimentations and studies that are well-informed by theory will result in good practices.
(d) Teacher training should be given priority, both pre- and in-service training. Many of the above studies point to teachers to be partly responsible for poor

quality of assessment. Also training teachers will result in diversifying testing practices. It is worth mentioning that visual interpretation, i.e., how to interpret images and videos, should be included in testing to cater for the students' practices as they tend to be big consumers of online visuals.

(e) There should be a national board of examination which will look after the quality of tests, the design of test rubrics, the training of testers and teachers, and the study of test washback effect. Such a board was envisaged within the new reform of education and training.

(f) Further research in how to introduce criterion-referenced assessment with clear rubrics and descriptors in order to make testing in Morocco more effective has long been awaited. Also, in the tertiary, course descriptions with learning outcomes and grids of evaluation should be public for students to know what is expected from them and therefore work accordingly.

(g) The launch of master and doctoral trainings in assessment in the hope of creating synergy and solid research that will identify flaws and suggest better practices that are inspired from international body of research.

(h) The last, but not the least, recommendation has to do with considering testing from a global perspective, rather than from the bits and pieces of different subjects as is the case now. That is, testing, and teaching thereof, policies should be complementary across disciplines and levels.

6 Conclusion

This article has shown why testing in Morocco should be revisited. The documented research not only raises the drawbacks of the existing testing practices but also calls for major changes. It also provides some recommendations. It should be noted that one of the limitations of this article is that it does not provide alternatives which will lead to better practices.

References

Bellout, Z. (1990). Assessing composition at university first cycle. In *Proceedings of the XIth MATE Annual Conference*, pp. 227–237.

Bloom, B., Englehart, M., Furst, E., Hill, W., & Krathwohl, D. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. New York, Toronto: Longmans.

Boubekri, El H. (1997). Types of questions in *Bridges*. A paper delivered at the MATE 17th Annual Conference in Erfoud.

Bouchouk, El M. (1987). Evaluation of literature in the Baccalaureate exams: From subjectivity to objectivity. *Attadriss*, *10*, 5–27 (in Arabic, my translation).

Bouziane, A. (1991). Feedback in the learner-centred approach. In *Proceedings of the XIIth MATE Annual Conference*, pp. 77–88.

Bouziane, A. (1993). Towards an effective use of reading texts: An investigation. In *Proceedings of the XIIIth MATE Annual Conference*, pp. 83–97.

Bouziane, A. (1997). What research tells us about reading in Morocco. *MATE Newsletter, 18*(2), 8–10.

Dahbi, M., & Britten, D. (1989a). Beyond the staircase: A comparison of the reliability of six composition scoring Methods. In *Proceedings of the IXth Annual Conference of MATE*, pp. 36–46.

Dahbi, M., & Britten, D. (1989b). Improving the reliability of composition scoring: A research-based recommendation. In *Proceedings of the Xth National MATE Conference*, pp. 43–50.

El Mazgualdi, A. (1996). *Evaluation in the Moroccan educational system: A psycho-pedagogical study on the impact of school exams on the pupils' formulation of some psychological attitudes*. Mohammadia: Fdala Publishing House (in Arabic, my translation).

Ennaji, M. (1987). Strategies for testing and scoring Composition. In *Proceedings of the XIth MATE Annual Conference*, pp. 51–61.

Ezzaki, A. (1986). Questioning in language education. In *Proceedings of the Sixth Annual Conference of MATE*, pp. 10–18.

Fatihi, M. (1995). *Methods of measurement and techniques of evaluation: Test and exam construction and treatment of outcomes*. Casablanca: Ennajah Eljadida (in Arabic, my translation).

Gustafsson, J. E., Hansen, K. Y., & Rosén, M. (2013). Effects of home background on student achievement in reading, mathematics, and science at the fourth grade. In M. O. Martin & I. V. S. Mullis (Eds.), *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade-implications for early learning*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA), pp. 181–287.

Hammani, M. (1995). Designing and moderating test items. In *Proceedings of the XVth Annual Conference of MATE*, Casablanca, pp. 91–101.

Ibourk, A. (2013). Determinants of educational achievement in Morocco: A micro-econometric analysis applied to the PIRLS Study. *Regional and Sectoral Economic Studies, 13*(2), 179–190.

Linn, R. L. (2000). Assessment and accountability. *Educational Researcher*, *29*(2), 4–16. url: http://www.aera.net/pubs/er/arts/29-02/linn15.htm (www document).

Madi, L. (1990). *Objectives and evaluation in education*. Rabat: Babel (in Arabic, my translation).

Melouk, M. (1992). *Towards a qualitative analysis to the baccalaureate exams*. A paper presented at the MATE 13th Annual Conference in Ouarzazate.

Melouk, M. (2001). The State of EFL evaluation in Morocco: The testers' and teachers' opinions. *Proceedings of the 21st Annual Conference of MATE in Essaouira*, pp. 55–63.

Ministry of Education. (1994). *Rencontres pédagogiques concernant les professeurs dans l'enseignement secondaire: Documents pédagogiques pour l'anglais*. Sale: DEDICO.

Ministry of Education. (1996a). *Rencontres pédagogiques concernant les professeurs d'anglais exerçant dans l'enseignement secondaire: Documents relatifs à la 3ème année secondaire*. Casablanca: Somagram.

Ministry of Education. (1996b). *Rencontres pédagogiques concernant les professeurs dans l'enseignement secondaire*. Casablanca: Les Editions Maghrebines.

Ministry of Education. (2007). *English language guidelines for secondary schools: Common core, first year, and second year Baccalaureate* (n. p.).

Mullis, I. V. S., Martin, M. O., Foy, P., & Drucker, K. T. (2012). *PIRLS 2011* International results in reading. IEA: Lynch School of Education, Boston College. Available at [Retrieved on 20th April, 2016]: http://timssandpirls.bc.edu/pirls2011/downloads/P11_IR_FullBook.pdf

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Kennedy, A. M. (2003). *PIRLS 2001* International report: IEA's progress in international reading: Literacy study in primary schools in 35 countries. IEA: Lynch School of Education, Boston College. Available at [Retrieved on 20th April, 2016]: http://timss.bc.edu/pirls2001i/pdf/p1_ir_book.pdf

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). PIRLS 2006 International report: IEA's progress in international reading: Literacy study in primary schools in 40 countries. IEA: Lynch School of Education, Boston College. Available at [Retrieved on 20th April, 2016]: http://timss.bc.edu/PDF/PIRLS2006_international_report.pdf

Naciri, L. A. (1995). *Response to students' writing in the Moroccan EFL context*. Unpublished D.E.S. Dissertation. Rabat: Faculty of Education.

Nadri, Y. (2005). *Investigating discrepancies in scoring compositions among university teachers*. Unpublished DESA degree in Education. Rabat: Faculty of Education.

Ouakrime, M. (1986). *English language teaching in higher education in Morocco: An evaluation of the Fez experience*. Unpublished PhD Thesis. London: Institute of Education.

RTI International. (2012) *Student performance in reading and mathematics, pedagogic practice, and school management in Doukkala Abda, Morocco*. EdData II, (Task Order 7). Washington, DC: USAID.

RTI International. (2014a). *Research on reading in Morocco: Analysis of teachers' perceptions and practices*. EdData II, (Task Order 15). Washington, DC: USAID.

RTI International. (2014b). *Research on reading in Morocco: Analysis of initial teacher training*. EdData II, (Task Order 15). Washington, DC: USAID.

RTI International. (2015). *Research on reading in Morocco: Analysis of textbook procurement chain and market for supplemental reading materials*. EdData II, (Task Order 15). Washington, DC: USAID.

Sadiqi, F. (1986). The teaching and learning of composition. In *Proceedings of the Sixth Annual Conference of MATE*, pp. 55–63.

Superior Council of Education and Training and Scientific Research. (2009). *Programme National d'Evaluation des Acquis: PNEA 2008* [The National Programme for Evaluating Achievements Testing].

Tamek, M. S. (1989). *Some aspects of orality in Moroccan university students' written English*. Unpublished D.E.S. Thesis. Rabat: Faculty of Letters: Rabat.

Specs Validation of a Dynamic Reading Comprehension Test for EAP Learners in an EFL Context

Sahbi Hidri

Abstract Validation is carried out to explore the different facets that come into play to design useful dynamic tests. This chapter underscored a theoretical and practical overview of test specs (specs) validation of a dynamic assessment (DA) of a reading comprehension test for learners of English in an EAP program. A special focus was attended to qualitative and quantitative data analyses of interactions between two mediators and test-takers in three testing phases: Input, interaction and output and a Multi-Faceted *RASCH* Measurement, *FACETS* analysis, of test scores. Results showed that test-takers' performance significantly improved with the support of a mediator; thus resulting in a more relevant output. The FACETS quantitative results also confirmed the results of the qualitative analysis. The study suggested a list of specs for designing dynamic reading tests for learners of English in a similar-related context. Limitations and recommendations were also discussed.

Keywords Dynamic assessment · Dynamic reading · Zone of Proximal Development (ZPD) · Mediated Learning Experience (MLE) · Specs validation · Socio-Cultural Theory (SCT) · Classroom-based assessment (CBA) · Cognitive & metacognitive strategies

1 Introduction

In many educational contexts, testing has always been informed by a theory rooted in a classical way of testing with candidates performing solely on the test. Research on assessment has addressed the classical mode of assessment; however, DA, being used in cognitive assessment (de Beer, 2010), has not been given its due importance as a complex enterprise even though it has room in integrating learning and testing

S. Hidri (✉)
English Language Institute, University of Jeddah, Jeddah, Kingdom of Saudi Arabia
e-mail: sahbihidri@gmail.com

S. Hidri
Faculty of Human and Social Sciences of Tunis, Tunis, Tunisia

S. Hidri and C. Coombe (eds.), *Evaluation in Foreign Language Education in the Middle East and North Africa*, Second Language Learning and Teaching,
DOI 10.1007/978-3-319-43234-2_19

and making decisions on the curriculum. de Beer (2010, p. 241) contends that “DA refers to an assessment approach that includes a learning opportunity during assessment in order to provide information on the current as well as the potential future performance levels of the individual being assessed-typically by means of a test-train-retest process.” In implementing DA, there exist some learning handicaps reflected in the test-takers’ poor background knowledge who most often fail to relate actual knowledge of the test to their background knowledge. Classical assessments overlooked the assessment of such types of knowledge, skills, sub-skills and language abilities (Macrine & Sabbatino, 2008). Because of such learning deficiencies, using DA becomes a necessary and complimentary tool to develop the learners’ cognitive and meta-cognitive reading strategies. Traditionally, addressing specs of language skills, such as reading, has gained momentum in language testing and assessment (Alderson, 2000; McNamara, 1996) and it has largely depended on the use of, for instance, Item Response Theory (IRT) in analyzing test scores (Bachman, 2004). Contrary to mainstream assessment, specs validation of dynamic reading comprehension tests has received scant attention in research.

2 Theoretical Background

This chapter probed the relevance of specs validation of a dynamic reading comprehension test. Test specs are defined as an intersection between method and purpose linked to a well-defined criterion that reflects the course objectives and the actual language ability of the learner. Most obviously, specs are a design document needed by many people, such as examinees, test designers, textbook and course developers, teachers, policy makers, institutions and governments. An overriding consideration is that the way specs are designed affects test scores and that the kind of test tasks has an impact on the test takers’ modifiable performance (Hamp-Lyons, 1997). DA, anchored in Vygotsky Sociocultural Theory of mind (SCT) (1986), encourages the development of learning autonomy by regulating attention and engaging the test-takers in joint activities of more than one test-taker with the support of a mediator. DA measures the learning abilities from a process-based perspective (Jeltova et al., 2007). In this regard, contrary to the traditional modes of assessment, DA proposes the development of growth in unveiling the cognitive and meta-cognitive reading abilities in learning or acquiring a second language i.e., cognitive modifiability (Vygotsky, 1981). DA has been implemented in different contexts and it is officially recognized as a testing mode to mediate learners to overcome their learning problems. Engaging mediators and learners in joint activities functions as an important enterprise in implementing DA and what is needed at this level is mastery of the mediation strategies. Research (e.g., Lantolf & Poehner, 2011) has shown that DA has proved to be more reliable than mainstream assessment, especially in unveiling the learners’ potential and cognitive strategies to yield appropriate answers to test items in language skills, such as reading.

As a construct, reading has to be theoretically defined and then operationalized into test items that could be the evidence of the actual language ability of the learner and the course objectives. Such an overriding trend would be adequate for defining the context of the target language use domain and for facilitating learning. Alderson (2000, p. 5) highlights the need for reading as a social activity that is rooted in a given social setting. He contends that

> [R]eading is not an isolated activity that takes place in some vacuum. Reading is usually undertaken for some purpose, in a social context, and that social context itself contributes to a reader's notion of what it means to read, or, as recent thinkers tend to put it, to be literate.

Stated another way, capitalizing on the awareness and relevance of the social context of the classroom settings, the comprehension process should be the ultimate goal in a DA reading task. This trend has been given momentum in DA. Alderson (2000) overtly stresses an overview of intertwined variables to provide a clear track of the act of reading, such as the reader herself, background knowledge, motivation, skills and sub-skills, text genre awareness, readers' comprehension ability, metacognition, content schemata, purpose of reading, language of questions, types of questions, testing skills, grammar, vocabulary, etc. Nonetheless, what is uncertain at this level is the mediators' ability to master these strategies given the eventuality that mediators are confined by their views of language learning and teaching. Cohen (1994, 2007) states that reading is not a passive process. On the contrary, there is a wide variety of types of knowledge required for comprehension and it is the role of the reader to activate such types to comprehend the reading input. Comprehension failure may emerge as a result of distortions in background knowledge; hence the relevance of background knowledge in test design. A good reading ability increasingly culminates in learning maturity, which may further be indicative of successful future development (Valsiner & van der Veer, 1993). Additionally, Cohen (1994) labels background knowledge as a kind of schemata that he classifies according to language, content and textual aspects. When the three types of knowledge are substantively activated, successful reading can be elicited and can easily take place. This kind of reading is called top-down process (Carrell, 1988). A "text-based or data-driven," approach (Cohen, 1994), however, consists of the reader focusing on the textual elements for comprehension purposes, such as morphemes, words, phrases, sentences and then discourse aspects. In defining the construct of reading, Alderson (2000, p. 118) states that:

> Every test is intended to measure one or more constructs. A construct is a psychological concept, which derives from a theory of the ability to be tested. Constructs are the main components of a theory, and the relationship between these components is also specified by the theory.

Constructs are founded on theoretical underpinnings. The theoretical and operational definitions of the construct provide an authoritative view of test specs. For instance, Bachman and Palmer's framework (1996) of test specs embraces an overt definition of the construct, since any score inferences on the language ability should serve as the basis for target language use domain, such as language in context. Assessing language in context should be the target of educational assessment in that it has to evaluate the curriculum (Reynold, Whedall, & Madelaine, 2009). This study fits in the context of evaluating reading from a curriculum-based perspective, since DA tasks and activities of this study "departed" from the reading program of the curriculum.

Many studies (e.g., Ableeva, 2008; Feuerstein, Rand, & Rynder, 1988; Haywood & Lidz, 2007; Kozulin & Gindis, 2007; Poehner, 2007, 2008; Sternberg & Grigorenko, 2002) have approached DA differently and its practitioners have endeavored to assess learners' developmental capabilities in a process-oriented and well-defined cultural context where they are engaged in scaffolding to negotiate meaning. Sternberg and Grigorenko (2002, p. vii) contend that in DA:

> The examiner teaches the examinee how to perform better on individual items or on the test as a whole. The final score may be a learning score representing the difference between pretest (before learning) and posttest (after learning) scores, or it may be the score on the posttest considered alone.

Generating an array of facets on the interconnection between dynamic reading and SCT should be monitored from a particular perspective. Vygotsky's theory of language learning (1981, 1986) may be complimentary in this regard. Dynamic reading conveys the presence of another person to facilitate comprehension. Assuredly, compelling cultural tools, such as language, form the basis for an initial interaction where the environment and the learner become inseparable. In a social setting, the individual's cognitive ability can be malleably developed in the presence of a more competent mediator, and, therefore, success on reading future performances can be predicted on the basis of moment-to-moment interaction. Lidz and Gindis (2003) highlight this growing trend in DA by stating that mainstream assessment focuses on "the child's cognitive performance to the point of "failure" in independent functioning, whereas DA in the Vygotskian tradition leads the child to the point of achievement of success in joint or shared activity" (p. 103).

One way of highlighting the strands of how learning takes place can be explored through Gass's model of input, interaction and output (IIO) (1997). To comment on the model, apperception, comprehended input, intake, integration and output constitute its five parts. In the apperception stage, the reader notices the input and relates it to her background knowledge, (Block, 2003; Gass, 1997). At this level, activation of background knowledge for comprehension purposes is initiated. In the second category, interaction, the reader receives a new piece of information and conceptualizes and recycles the already acquired ones. The third part, output, feeds back into the two first parts: Input and interaction. It is at this level that proponents of DA acknowledge that any form of mediation should always be conducive to successful learning. To screen and overcome such deficiencies and reach cognitive modifiability, mediators should use very specific techniques to comprehend reading. Zone of Proximal Development (ZPD) and Mediate Learning Experience (MLE) are a case in point.

Vygotsky developed the idea of ZPD based on the SCT. It follows then that the ZPD basically stresses the importance of a socio-cultural environment that contributes to a higher performance level of the learners' language ability. Vygotsky notably draws in the lines of ZPD as the zone where learners can perform with the help of a more competent person and then at a later stage they can find routes to perform solely. This is the most important concept in DA according to Poehner and van Compernolle (2011). Poehner (2008) highlights the idea of ZPD as jointly constructed by learners and mediators where the latter can play a key role in maintaining appropriate scaffolding that could genuinely lead to learning independence. Sternberg and Grigorenko (2002) stress the effectiveness of mediation in the ZPD whether at the individual or group level. Poehner (2011) claims that this zone functions as a challenging teaching technique to potentially enable learners to go beyond their current thinking level. He states that highlighting the malleable abilities in the ZPD is produced not by individual "performance but through collaboration between teachers and learners (p. 247)" where the former mediates the latter in dealing with any learning input. The ZPD relates to the ongoing maturity process of cognitive and meta-cognitive strategies. Strategy awareness use should potentially depict the test-takers' actual level of a reading exam performance and they should be made aware of the fact that they have achieved good progress.

In defining the MLE, Feuerstein, Rand, and Rynder (1988, p. 58) define the MLE as the area where "the more a child is subjected to [MLE], the greater will be his capacity to benefit from direct exposure to learning," which is not the case in the absence of such MLE. Effective learning is at its best only when mediators interact with learners while accentuating their needs, pace and style. Feuerstein et al. (1988) highlight a basic premise in DA where learners are not directly influenced by the environment; rather they are exposed to the influence and mediation of other persons, who are supposed to be "an adult mediator" (p. 56). Thus, DA has been gaining territory to address the joint interactions of learners and mediators where they co-construct meaning in a shared knowledge activity. Although extant research has extensively investigated the relevance of dynamic reading in helping test-takers perform better, there are hardly any studies that have addressed the necessity of designing and validating a list of specs that would serve in the writing of fair dynamic reading comprehension test for learners of English in a related context.

2.1 Validation of the Reading Specs

Validation of DA reading tests has been overlooked in research. Cohen (2007) highlights the importance of specs validation by considering such variables as test-taking strategies, (reading mediation strategies are a case in point), scores, and other instruments, such as think-aloud protocols and interviews. For instance, using quantitative tools, such as analysis of scores (e.g., *FACETS*, (Bond & Fox, 2007)) is of great relevance in signposting inferences about item analysis and test specs. Such inferences can form a comprehensive view about test-takers' reading ability and, therefore, define the construct theoretically. It is a cyclical process that is intended to define and operationalize the construct, define specs, analyse scores and make inferences based on item analysis.

Validation lies at the core of any testing operation and it should establish sound theoretical foundations against which such specs can be properly operationalized into test items. In addition, score inferences should reflect the actual language ability of the learners (Kunnan, 1998a, b, 2000; McNamara, 2004, 2006). Bachman and Palmer's framework (1996, pp. 50–51) of test specs is a case in point. This framework includes setting, test rubrics, input, expected response, and the relationship between input and response. To comment on the framework, characteristics of the setting comprise three facets: Physical setting, participants, and time of the task. These aspects generally deal with the practicality of the test, such as place, noise, seating conditions, and lighting. As for characteristics of the test rubrics they deal with how test-takers work on test items and how raters grade such performances. Test rubrics give a detailed description of instructions, structure, time allotment, scoring method, criteria for correctness, explicitness of criteria and procedures, and procedures for scoring the response. The third part of the framework, characteristics of the input, describes content, format and language of input of the test. Part four, characteristics of expected responses, stresses the type of response that test-takers are supposed to produce, whether be it selected, limited or extended.

The last part, relationship between input and response, contains reactivity, scope, and directness of the relationship and constitutes the interconnection between the reading input and response. For instance, the reciprocal tasks embrace the interaction modes among test-takers, such as dynamic reading. Scope reflects the degree of response that candidates use to process dynamic reading questions. In the directness of relationship, three aspects constitute the input: Textual elements, context and background knowledge. Specs validation entails test interpretation and use of test scores (Messick, 1989).

Messick's framework of facets of validity (1989, p. 20), Table 1, has been very influential in language testing and validation. Messick stresses two facets of validity:

> One facet is the source of justification of the testing, being based on appraisal of either evidence or consequence. The other facet is the function or the outcome of the testing, being either interpretation or use. If the facet for source of justification (that is either an evidential basis or a sequential basis) is crossed with the function or outcome of the testing (that is, either test interpretation or test use), we obtain a four-fold classification.

Test purpose serves as a good rationale and scores might require appropriate inferences about the candidates' actual language reading ability. In addition, test use has a particular impact because of scores usefulness at the social level.

Table 1 Facets of validity (Messick, 1989, p. 20)

	Test interpretation	Test use
Evidential basis	Construct validity	Construct validity + relevance/utility
Consequential basis	Construct validity + value implications	Construct validity + relevance/utility + value implications + social consequences

The evidential basis for test interpretation depends on validation, test design and implementation. This contains test method, rating scale, and conditions. In test use, the evidential basis covers test-taking strategies, background knowledge and the candidates' profile. As for the consequential basis, test interpretation is related to the stakeholders' feedback, mainly test-takers and specialists, such as testing experts. The consequential evidence is about test use and it embraces test impact, such as washback effect, which can be both positive and negative (Alderson, 2004). Messick (1989, p. 13) defines validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions". Interpretations do not only depend on test scores, but also on the relevance and appropriateness of scores pertaining to the future of test-takers.

Addressing reading comprehension specs validation while accentuating these different facets of validity has become of great importance. Hence the necessity to use complex statistical software, such as IRT and specifically the one-parameter model, RASCH measurement. In this regards, Weiss (1980, p. 8) maintains that "IRT has the potential of offering solutions to the problem of measurement gains in achievement levels during the process of instruction." It follows then that the purpose of the study was to check whether the joint interactions in a dynamic reading test can be conducive to specs validation whose score can inform about the test-takers' performance in real-life contexts. Therefore, the following questions were considered in the present study:

(a) Can current dynamic reading interactions between mediator and test-takers contribute to test validation?

(b) What are the implications of specs validation to design dynamic tests for learners of English in a similar context?

3 Method

3.1 Participants and Setting

Participants of the study were a group of 25 students studying in a foundation program, a year before being specialised in one track, whether IT, business, or mathematics. The study took place in a college of applied sciences in Oman. Students' age range was 18–20 and they were speakers of Modern Standard Arabic. They studied 4 h of reading per week over a fourteen-week period for each term. Data collection was carried out during the second term of 2013. Students' level was labelled as Academic English Skills A-Level where they studied English for two terms before being specialised in one of the above-mentioned tracks. The textbook they used was *Headway*. To collect data, students were exposed to classroom mediated interactions between a mediator and two students on a reading input from *Headway* textbook. The other mediator was in charge of grading the students' performance. This reading constituted of interpreting graphs and answering questions. The second part of the study aimed at checking whether the input and interaction phases led to success on a reading exam that was made up of 20 test items and was graded by another teacher who was not involved in the input and interaction phases.

3.2 Data Analysis

DA was considered in the three phases and was aimed at helping students process the reading input. Data analysis implemented qualitative and quantitative approaches. The qualitative part consisted of analyzing the joint interactions undertaken by the mediator and learners during regular class hours. The purpose of the analysis was to attend to the different mediation forms, how students reacted to this mediation and whether they made any form of progress. Since the study consisted of three phases, input, interaction and output, students were administered a progress test whose scores were analyzed using the FACETS to measure the reading ability and whether the students' reading performance improved.

4 Results

4.1 Qualitative Analysis: An Example

Test interactions were meant to develop the learners' cognitive and metacognitive reading strategies to handle any text variety. The following Excerpts illustrate the test-takers' and mediator's interactions on a dynamic reading text about international tourism for Level-A Academic learners.

Phase one: Input

Excerpt 1

1. Mediator (M): do you happen to know very famous international touristic places?
2. Test-taker (TT 1): dubai
3. TT 2: oman
4. M: do you know other places in the world? europe? africa?

Thinking for a while

5. TT 4: england
6. TT 5: morocco (…) south africa.
7. M: ok. Good (…) what is the most touristic place in the world? can you guess?

No reply on the part of students

8. M: ok. can you choose from these three options: italy, china or turkey?
9. TT 10: turkey, i (…) think turkey
10. TT 2: Italy, i think (*in Arabic*)
11. M: why do you think it is turkey?

No reply on the part of students

12. M: it is italy. do you have an idea about the number of tourists who visit italy every year?
13. TT 12: two million
14. TT 16: No. i think (…) one million
15. M: ok. do you have an idea about the number of tourists that visit oman every year?
16. TT 9: many
17. TT 18: (…) four million

By activating their background knowledge to initiate interactions, the mediator in Excerpt 1, Turn 1, endeavoured to manoeuvre the test-takers into the scaffolding zone of negotiating meaning. In terms of content and cognitive load, the learners were not very knowledgeable in having some self-reflection to be able to retrieve the relevant piece of information on the famous international touristic places. At this level, the mediator's perception of the task, in large part, was shown to change the learners' reading strategies by having some more debate. Getting hold of the most touristic places is a case in point (e.g., Excerpt 1, Turns 1, 4 and 7). However, Turn 7 of Excerpt 1 posed some challenging load on the learners, as it asked about a content piece of information. In the absence of a more malleable interaction, the learners engaged in a silent pause wondering about the right answer on the most touristic places in the world. In order for the mediator to keep them in the ZPD, while preserving their modifiability, he leaned on providing the learners with a MC task containing options on the most touristic place (Excerpt 1, Turns 8) and the reasons for their selection (Excerpt 1, Turn 11). Gradually, the mediator tried to lead the test-takers to more challenging questions that called for some critical thinking, such as Turns 12 and 15, Excerpt 1. However, whenever the test-takers faced difficulties, the mediator straddled the task by modifying the strategies along with text context to make them think differently. Their silent pauses were very frequent, but the mediator reflectively endeavoured hard to push them to transcend their current level of thinking into a higher one (e.g., Excerpt 4, phase 2).

As the task became more and more complex, the mediator envisaged to change his mediation techniques by making the task more accessible to test-takers. This could be at the core of DA and it resulted in more talking time on the part of the mediator to explain the reading input. Interpreting data from charts or graphs akin to the text proved to be one of the most challenging tasks that learners were exposed to. Such difficulties impinged on their comprehension of the task. To counterbalance this handicap, a niche for interpreting data was implemented in a gradual and mediated way. To this end, noticing, as a mediation strategy, was offered to the learners (Excerpt 4, Turns 2, 5, 8 and 10). The gradual mediation that constituted focusing on details of the graph then on general items resulted in an appropriate relevant interpretation. Still, with the silent pauses on the part of the test-takers, the mediator again sought to probe for more details on the graph (Excerpt 4, Turn 11). Turn 13, Excerpt 3 was purposefully intended to reflect one of the major testing outcomes that the mediator tried to achieve: Using appropriate verbs to describe the graph about international tourists and touristic places in the world.

Phase two: Interaction

Excerpt 4

1. M: now, let's move to a very important task that you are supposed to use in your project. go to page 59. have a look at the first graph please. what do you see in the graph? try to describe it. ok. try to describe the graph.

No reply from students

2. M: what do you see on the horizontal and vertical axes?
3. TT 3: months and tourists
4. TT 9: months and number
5. M: is this graph about tourists in oman?
6. TT 11: yes
7. TT 4: no. everywhere (…) international
8. M: can you check the title of the graph please?
9. TT 20: international tourism
10. M: can you comment on the first graph? try to make sentences?

No reply on the part of students

11. M: ok. focus on january and august. do you think the number of tourists is the same during the two months?
12. TT 14: no. in august the number is (…)
13. M: what is the verb we should use here?

No reply on the part of students

14. M: Right. can you give me any verb that comes up to your mind? focus on the graph and try to suggest verbs, prepositions, nouns or any word that can be helpful in describing the graph.
15. TT 23: up (…) down (…) in (…) on (…)
16. M: try to use very relevant prepositions with verbs. let's take the verb "go" (…) which prepositions can we use here?
17. TT 12: go up (…) down (…) no, go down
18. TT 11: also go up
19. M: good. both prepositions can be used. what about the word "peak"? are you familiar with this word?

No reply on the part of students. The mediator showed them a visual of a peak.

20. M: do you know the meaning of the word "peak" now?
21. TT 10: yes. it high very much*
22. M: ok. how can we use the word "peak" to describe numbers? go back to the graph and check the month of august. any idea?
23. TT 2: number high peak, go peak (…)
24. TT 1: is peak, use peak
25. M: can you check the following verbs: reach, play, declare. which of these verbs can be used with the word "peak"?
26. TT 15: no. reach peak
27. M: great. we use it with the very reach", so, it's "reach a peak". now, how can you use it to describe the number of tourist? use this in relation to august. what do you see? is the number up or down?
28. TT 3: up (…)
29. M: so, the number of tourist in august (…) (*pause, expecting students to finish the sentence but no reply on part of students*)
30. M: reach a peak.
31. M: good, but which year? is it 2013?
32. TT 6: no, it is 2009. tourists reached a peak
33. M: good, the number of tourists reached a peak in august in 2009

Excerpt 9

1. M: ok. try to read this short text and underline or circle the words you think can be related to describe tourism

Students were engaged in a silent reading of the text.

Turn 14 Excerpt 4 is a case of using brainstorming to bridge the gap between the learners' background knowledge and content of the text. This had the purpose of manoeuvring the task to make it more accessible. This mediation strategy worked out for test-takers as they gradually started to respond to the mediator's instructions (Turns 15, 17, 18, Excerpt 4). Turns 16 and 19, Excerpt 4, were meant to potentially support the test-takers to attain better achievement and performance levels in their ZPD. They were also made aware that they made considerable progress in probing into the appropriate test items, such as Turn 19, Excerpt 4. As the task became more complex, especially in introducing novel vocabulary, the mediator employed a new mediation strategy which was the use of visuals. After some attempts, the learners remediated difficult vocabulary by getting hold of the meaning of the word "peak". Undoubtedly, the visual evidence helped them use the word "peak" in the tourism context to refer to a very high increase in the number of tourists (Excerpt 4, Turns 23, 24 and 26).

These mediation techniques were a case in point of how mediation changed the behaviour of test-takers and how it could help them use vocabulary in context.

Again, praise and encouragement on the part of the mediator (Excerpt 4, Turn 27) made the learners feel more motivated to process the task that was becoming more challenging. He made some analogy to assist the test-takers to use the word "peak" to describe the number of tourists (Excerpt 4, Turn 27). To this end, he tried to blur difficulty by using the right tense with the word "peak" (Excerpt 4, Turns 27, 29, 30 and 31). He kept repeating the correct sentences uttered by the learners to show them that he was adhering to their correct feedback. This was in fact a kind of encouragement, praise and support to endow them with the feeling that they had achieved good progress.

Phase three: Output

Excerpt 11

1. M: ok. can you mention the words related to international tourism to describe the graph?

A silent pause on the part of learners

2. TT 18: international tourists (…)
3. TT 17: January (…) december
4. M: ok. do you know the meaning of these words? (*the teacher drew a spider gram including international tourism, then inserted the following words*: approximately, rose, grew, stable, increase, very, increase, rise, reaching, a peak, rose, fall, dropped, remained, slight, dramatic, steadily, steady, suddenly (…)). can you mention the words you already know?
5. TT 3: rise (…) grow (…) fall (…)
6. TT 6: drop (…)
7. M: do you know the meaning of "approximately, remain, slight dramatic"?
8. TT 5: yes, but i not (…) know use with verbs*
9. M: you don't know how to combine them with these verbs?
10. TT 5: yes (…)
11. M: can you try to combine words to make useful expressions?
12. TT 8: rise a peak?
13. M: do you think so? can you think of another combination?

No reply from students

14. M: ok. try to put the following list of words under the appropriate heading
15. TT 10: in (…) go (…) up (…) we (…) rise (…) grow (…) decrease"
16. TT 13: no. increase (…) go down
17. M: ok. then?
18. TT 13: no. go down (…) decrease, drop, fall (…)
19. M: what about the third and fourth boxes?
20. TT 13: remain steady (…) stay the same
21. M: ok. good
22. TT 13: fluctuate, go up and down

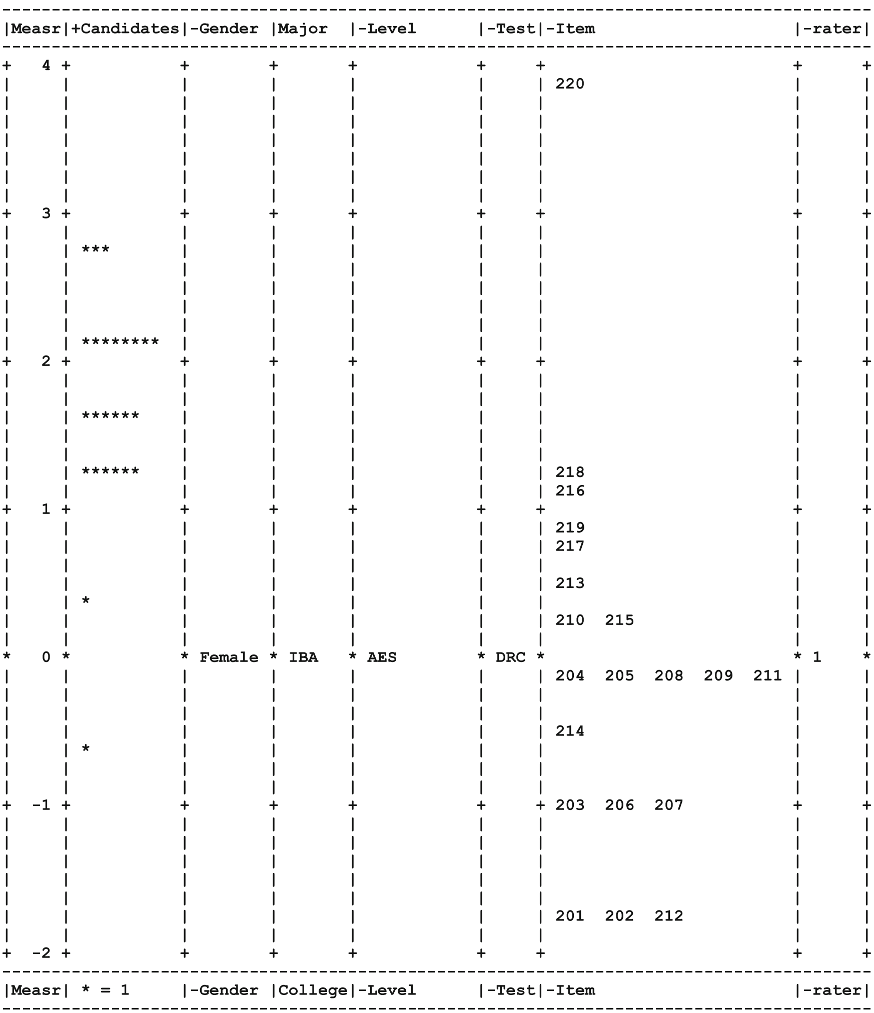
Again, the mediator employed different mediation techniques whenever the test-takers faced text difficulties, such as brainstorming tourism vocabulary to assume continual progress on the test items. The use of a spider gram to brainstorm on tourism had the aim of assisting the learners (Excerpt 11, Turns 4, 7 and 9). The mediator started with the smallest units of words, phrases and then progressively moved on to more complex sentences. This constituted a bottom-up approach to the teaching of dynamic reading for test-takers who faced tremendous reading problems. The use of brainstorming on the part of the mediator was done on purpose to make the learners value the newly acquired vocabulary against what they already knew (Excerpt 11, Turn 4).

The mediator targeted the learners' knowledge of these words both out of (Excerpt 11, Turn 7) and in context (Excerpt 11, Turn 14). The task became more complex for the learners when they tried to combine verbs with prepositions (Excerpt 11, Turns 15, 16 and 18), but they were functionally adapted to the task when they initiated categorization of the easier items under the appropriate heading (Excerpt 11, Turns 20 and 22). This could naturally be implemented by the mediator who continuously debated the right mediation strategy to handle difficult and novel vocabulary. After implementing the mediation strategies in the joint interactions in the input, interaction and output phases, it was felt more appropriate if the learners were administered a test to measure their success on the reading test items, based on the use of the different mediation techniques.

4.2 *Quantitative Analysis: An Example*

This section presents the *FACETS* analysis of the reading test. The purpose of this progress test was to evaluate the test-takers' adaptability to use the mediation techniques of the input, interaction and output phases and perform individually.

Figure 1 portrays all the test variables. The scale, column one, *measr*, of the students' ability ranges from −1 to +4. The second label "candidates," facet 2, represented the test-takers' ability. The asterisks in this column indicate the distribution of the candidates' ability estimates (n = 25 female test-takers). Facet 3 is the gender of the candidate. Facet 4 is major of the test-takers. Facet 5 is the level of the candidates, Academic English Skills (AES) A-Level; and facet 6 deals with the type of the test, dynamic reading comprehension. Facet 7 is the 20 test items. In the *FACETS* analysis, the items are referred to as items 201, 202, 203, etc. and facet 8 is the rater. The scale used is 0 = fully incorrect and 1 = fully correct. All the 20 test items were scored by another rater and were double-checked by the researcher. A few instances of scoring disagreements (n = 5) were identified and agreed on by the two raters. The candidates above the measure of scale 0 were said to be more able than those below this measure. What could be noticed was that 23 candidates out of 25 had an ability estimate that ranged from +1 to +2 and it was labeled as more than average, since the 0 represents the average difficulty of language ability. The tests-takers whose ability was above 0 were said to be more able, since they



```
-----------------------------------------------------------------------------------------
|Measr|+Candidates|-Gender |Major  |-Level      |-Test|-Item                     |-rater|
-----------------------------------------------------------------------------------------
+   4 +           +        +       +            +     +                          +      +
|     |           |        |       |            |     | 220                      |      |
|     |           |        |       |            |     |                          |      |
|     |           |        |       |            |     |                          |      |
|     |           |        |       |            |     |                          |      |
|     |           |        |       |            |     |                          |      |
|     |           |        |       |            |     |                          |      |
|     |           |        |       |            |     |                          |      |
+   3 +           +        +       +            +     +                          +      +
|     |           |        |       |            |     |                          |      |
|     | ***       |        |       |            |     |                          |      |
|     |           |        |       |            |     |                          |      |
|     |           |        |       |            |     |                          |      |
|     |           |        |       |            |     |                          |      |
|     |           |        |       |            |     |                          |      |
|     | ********  |        |       |            |     |                          |      |
+   2 +           +        +       +            +     +                          +      +
|     |           |        |       |            |     |                          |      |
|     |           |        |       |            |     |                          |      |
|     | ******    |        |       |            |     |                          |      |
|     |           |        |       |            |     |                          |      |
|     |           |        |       |            |     |                          |      |
|     | ******    |        |       |            |     | 218                      |      |
|     |           |        |       |            |     | 216                      |      |
+   1 +           +        +       +            +     +                          +      +
|     |           |        |       |            |     | 219                      |      |
|     |           |        |       |            |     | 217                      |      |
|     |           |        |       |            |     |                          |      |
|     |           |        |       |            |     | 213                      |      |
|     | *         |        |       |            |     |                          |      |
|     |           |        |       |            |     | 210  215                 |      |
|     |           |        |       |            |     |                          |      |
*   0 *           * Female * IBA   * AES        * DRC *                          * 1    *
|     |           |        |       |            |     | 204  205  208  209  211  |      |
|     |           |        |       |            |     |                          |      |
|     |           |        |       |            |     |                          |      |
|     |           |        |       |            |     | 214                      |      |
|     | *         |        |       |            |     |                          |      |
|     |           |        |       |            |     |                          |      |
|     |           |        |       |            |     |                          |      |
+  -1 +           +        +       +            +     + 203  206  207            +      +
|     |           |        |       |            |     |                          |      |
|     |           |        |       |            |     |                          |      |
|     |           |        |       |            |     |                          |      |
|     |           |        |       |            |     |                          |      |
|     |           |        |       |            |     | 201  202  212            |      |
|     |           |        |       |            |     |                          |      |
+  -2 +           +        +       +            +     +                          +      +
-----------------------------------------------------------------------------------------
|Measr| * = 1     |-Gender |College|-Level      |-Test|-Item                     |-rater|
-----------------------------------------------------------------------------------------
```

Fig. 1 Facets map of the reading progress test variables

had more chances to answer the items correctly than those who were below on the scale. Success on the test was probably due to the test-takers' familiarity with contents of the text that were jointly debated and mediated in the three phases. The candidates did not have a wide range of abilities as most of them clustered around +1 and +2. They should have performed much better than what was displayed on the measure scale. In order to probe the test-takers' ability, Table 2 describes the mean, SD and reliability of separation index to check the effective result of implementing the appropriate mediation strategies. The ability ranged between −0.57 and 2.76 logits, with a mean of 1.73 (SD = 0.71). The positive mean, 1.73, indicated that the test was accessible to the test-takers. The mean standard error

Table 2 Summary of test-taker facet statistics

M (model SE)	1.73 (0.67)
SD (model SE)	0.71 (0.10)
Min	−0.57
Max	2.76
Infit	
M	0.99
SD	0.28
Separation statistics	
Strata 0.68 Reliability of separation 0.09 Fixed chi-square statistic (d.f.) 34.6 (24), $p < 0.07$	

(SE) is 0.67 and it dealt with the precision or imprecision of the test-takers' ability estimates (McNamara, 1996). The fixed chi-square which hypotheses that all the test-takers had an equal ability estimates is of 34.6 (24) and it is statistically significant at $p < 0.07$. The reliability of the separation is 0.09 and it indicated that there was not much substantial variation in the test-takers' ability, i.e., they generally had the same ability in this test and that the test did not distinguish much between them in terms of the ability being measured.

4.3 Item Difficulty

Table 3 describes the twenty test items in terms of their difficulty. The purpose of analyzing item difficulty was to check which items had to remain in the test, which to be weeded out and which to be edited. This of course had direct implications for specs validation. The items were ordered from the least to the most difficult ones, with item 201 being the easiest with a measure of −1.73 and item 220 being the most difficult with a measure of 3.88. The reliability of the separation index is 0.71 which is not high. This means that the test items did not discriminate much in terms of difficulty. Also the chi-square of 66.9 with 19 d.f. is significant at $p \leq 0.00$, which means that the test items were not of equal difficulty. The mean of the model of SE is 0.65, ranging from 1.04 for item 201 to 0.63 for item 220. The mean of the infit MS (column 4) is close to 1, (0.99). This infit varies from 0.67 to 1.18. The data suggested that item 202 was said to be misfitting as it had an infit MS value of 0.67. According to McNamara (1996), the range of the infit MS could be set at the mean plus two SDs. So, the mean in this case is 0.99 and the SD is 0.15. Therefore, the range is 0.99 + 30 (0.15 × 2) = 1.30 in one direction and 0.99 − 30 (15 × 2) = 0.70 in the other direction. The range is then 0.70–1.30.

Table 3 Items facet summary of statistics (sorted by measure)

Tasks	Measure	SE model	Infit MS
201	−1.73	1.04	1.12
202	−1.73	1.04	0.67
212	−1.73	1.04	1.14
203	−0.94	0.77	1.00
206	−0.94	0.77	1.16
207	−0.94	0.77	1.00
214	−0.45	0.64	0.73
204	−0.08	0.57	0.94
205	−0.08	0.57	1.00
208	−0.08	0.57	1.03
209	−0.08	0.57	1.10
211	−0.08	0.57	0.76
210	0.22	0.53	0.87
215	0.22	0.53	1.04
213	0.48	0.49	1.00
217	0.71	0.47	1.18
219	0.92	0.45	1.10
216	1.12	0.44	0.81
218	1.31	0.43	1.13
220	3.88	0.63	0.99
M (n = 20)	0.00	0.65	0.99
SD	1.26	0.19	0.15

Reliability 0.71
Fixed (all same) chi-square: 66.9 d.f.: 19 significance (probability): 0.00

5 Discussion

The chapter explored how mediation strategy awareness could help test-takers to handle any text variety that is of relevance to their field of work. Assessing the test-takers in three phases was the goal of this study, as it tried to measure the students' gradual progress on the test, depending on its item difficulty. This was investigated in research (Jeltova et al., 2007, p. 276) who highlighted the fact that these phases address "the student's learning profile rather than on the learner's final score of the test." Also, specs validation will continue to be a major focus area in language testing and assessment. In this work, assessing dynamic reading was implemented and approached from a CBA perspective to evaluate the test-takers' progress on the instructional materials, such as reading. Advocates of DA (e.g., Lantolf & Poehner, 2011; Lidz, 2002) have praised its relevance and effectiveness to CBA and they even decried the limitations of mainstream assessment in unveiling the potential of cognitive and metacognitive reading strategies. Perhaps implementing both assessment modes would be the solution. This idea was discussed and maintained by Jeltova et al. (2007). Such an implementation can inform much about dynamic reading strategies and, consequently, activating such reading comprehension strategies should be the goal of teaching as well as testing. Integrating both assessment and classroom learning has been addressed in research (Black & William, 1998) where students are encouraged to adopt a self-assessment attitude or self-evaluation.

Writing dynamic reading test specs while leaning on test scores solely can by no means lead to validated test specs. Therefore, it was crucial to check and analyse the joint interactions undertaken by the mediator and test-takers. In other words, validation was carried out from two angles: Qualitative and quantitative. The mediated learning interactions between learners and mediator were done in a gradual way, where mediation was offered based on the outcomes of the interactions. This idea was further discussed in research and was referred to as "graduated prompt approach" (Campione & Brown, 1987).

In terms of mediation techniques, using the ZPD and MLE becomes paramount. Such zones should be established and developed both by mediators and test-takers where meaning is mutually constructed. Given its due importance, the ZPD yielded coherent interactions that served in the development of cognitive and meta-cognitive reading strategies, which in turn had direct implications for specs validation. All the mediation aspects were adjusted to the test-takers' needs in the teaching and testing cycle. Generally, test designers should have high esteem to the underlying abilities of the test-takers and they should endeavor hard to equip them with the most appropriate and most relevant reading skills and sub-skills to deal with text difficulty. In addition, all the reading materials should be adjusted to meet the test-takers' background knowledge and their needs. In other words, the taught materials should be partly new and partly known to the test-takers to process the test items. In this test, the items could be said to be accessible to the test-takers as the *FACETS* output indicated this "high" reading ability because of the familiarity with content and items. Assuredly, DA has been most of the time hailed for its effectiveness and practicality to unveil the cognitive and metacognitive reading strategies and develop them. Unfortunately, it is at this phase that traditional and mainstream assessments have failed to achieve; thus, resulting in no interaction between learners and mediators. Macrine and Sabbatino (2008) have stressed this idea. Also, the friendly behavior of the mediator, by using praise and encouragement, motivated the test-takers more and more.

Dynamic reading specs validation could reveal much about test-takers' reading ability. Even in terms of validity, the test-takers' interactions can also underscore a good understanding of the predictive validity about how these test-takers were supposed to behave in similar-related learning and testing contexts. Validation largely highlighted the fact that scores could also be a very informative source in validating dynamic reading specs. Construct validity, according to Messick (1989), plays an important role in test validation. Messick at the same time stresses the importance and relevance of the construct. Once the construct was defined theoretically, it was operationalized into test items that measured the right construct that it was intended to measure. This study stressed the necessity to consider reading from a sociocultural perspective where users of the act of reading meet together to construct meaning. This work meets the study on dynamic assessment and remediation undertaken by Macrine and Sabbatino (2008). This may sound challenging for the test writers, as they should be in control of these different strategies. Given the different reading problems, such as word recognition, mediators can employ the bottom-up approach to tackle these problems.

This study tackled the implementation of a reading test whose input was related to the curriculum. Such testing results probed the curriculum content. This idea was discussed in research (Pearson, Valencia & Wixson, 2014). The curriculum-based assessment should then target the fusion of instruction and assessment where students are expected to improve their language ability in the presence of mediation. In addition, the notion of more challenging thinking strategies can be akin to the strands of DA. This enhances testing reading items in context by making the task more accessible. Interactions on the reading input highlighted the idea that DA should be handled from a process, instead of a product perspective. This is what Cioffi and Carney (1997) called for.

6 Implications

The study had direct implications whose aim was to design a valid framework of test specs for designing dynamic reading tests in a similar-related context.

Dynamic reading test
Nature and purpose of the test: Progress test for General English Skills program for students majoring in IBA is intended to measure their progress in an EAP course over a fourteen-week term. Scores of this test are used to inform the test-takers about their reading ability.
Timing: One hour
Age: Eighteen to 19 years old level A-students who study 18 h of English as part of their Level-A Academic program.
Level: Upper intermediate
Language: English as a Foreign Language
Number of test sections: Three test sections: Input, interaction and output. In the input phase, the mediator(s) should introduce the task to the test-takers by activating their background knowledge and should set up rules for the turn-taking. In the interaction phase, the test-takers should be engaged in continuous scaffolding to process the input. In the output phase, the mediator(s) should reduce their support and help. The three sections should be carried out in one h, with 15 min for the input and interaction phases each and 30 min for the output phase.
Skills to be tested: Reading to extract general and specific information, locating details, synthesizing information, making inferences, guessing, and using words in context.

Target language situation: Reading in a problem-solving activity, reading any text, such as general, business, and other EAP varieties that meet the learners needs.
Task type: Close and open-ended questions and prompts.
Test method: All test items should be jointly answered by the test-takers and mediator(s). In the output phase, the test-takers should be given much more time to negotiate answers among themselves.
Text type: Any reading text variety that is of relevance to the test-takers' needs both in general and business contexts, such as business negotiations, ethics, and partnerships. It is preferable if this text variety has as many problem-solving and authentic activities as possible to engage the test-takers in meaning negotiations.
Text length: 350–400 words
Frequency of reading: Twice. Learners should be exposed to another reading whenever comprehension problems arise.
Number of items: Six test items in the input and interaction phases each and eight in the output phase.
Time: Ten minutes to read all the text then four minutes to think about the questions before answering.
Criteria for correctness: Test-takers' replies should be either fully correct or fully incorrect. Half correct answers should be avoided and test-takers should not be penalized for grammar, spelling, fluency or accuracy problems. All answers should be in English.

Certain rubrics should be highlighted and should function as guidelines for the reading test-designers. The following rubrics are meant to reflect the course objectives, the learning outcomes and the language ability of the learners:

- All the test procedures should be highlighted in a note as a cover page whose purpose is to guide the test-takers in how to deal with the test. The written instructions can be read by the mediators when they are engaged in joint interactions with the test-takers.
- All the test questions and prompts should be bolded and should be immediately followed by marking grades. They should also be worked on in pairs or in groups and they should reflect the same type of questions and prompts that the reading comprehension textbooks contain.
- The mediator(s) should select the two test-takers randomly. There should be no selection based on gender or ability level.
- Interactions are joint. The mediator(s) should know how and when to interfere to give the floor to all the test-takers to negotiate meaning.
- The mediator(s) should be able to score the relevant joint performance.
- In the post-testing phase, the mediator(s) support should be reduced to its lowest level.

Defining a clear list of dynamic reading comprehension test specs might not guarantee a successful testing operation. To remedy this, other parties are called upon to reconsider the assessment policy in similar contexts, such as policy-makers and textbook designers. Adhering to the use of DA might be a fruitful and very informative evaluation enterprise given the eventuality that learners of English have tremendous difficulties in processing the reading input, and therefore, in processing the other language skills and content courses. And it is at this level that the assessment policy should be revisited. At the same level, curriculum designers and program evaluators are invited to reconsider the nature and contents of tasks in the reading materials administered in class to cope with the socio-cognitive context where these learners are operating. Both parties should call for professional development of these teachers in how to implement DA in class.

7 Limitations and Directions for Future Research

This study raised a few limitations the first of which was the test-takers' inability to establish word recognition as they sometimes struggled hard to read sentences properly. This posed some comprehension problems for them, which distracted the mediator from focusing on the comprehension questions and contents of the text. The use of other research tools, such as interviews, could also lead to specs validation of the reading construct. These insightful comments from the test-takers or test interviewers would undoubtedly lead to these results, yet not that divergent and different from the ones of the current study. In addition, applying Gass's model (1997) to similar contexts seemed difficult to substantially lead to success on any verbal or non-verbal input. Sometimes, moving from input to interaction, for instance, led to comprehension breakdowns, especially if the mediators were not well equipped with the appropriate mediation strategies to implement dynamic reading or if the learners lacked motivation to be involved in the task. The difficulty of any input that led to a distortion in background knowledge was also conducive to comprehension failure. Some issues have been raised against DA, such as how to quantify measurement by establishing objective scoring to yield fair assessment of the language ability of the learners and subsequently many proponents of DA have lauded the relevance and effectiveness of this testing mode in mediating the learners to develop their reading ability.

Investigating the reading comprehension problems in two testing modes, such as static and dynamic, can inform much about the major differences between these two modes. This might serve as a basic indication for teachers on the most appropriate testing mode that can be conducive to learning autonomy. Researchers in similar contexts are encouraged to investigate DA in the other language skills as well as in content courses. In addition, mastery of the mediation strategies is a topic of further investigation given the eventuality that such adherence to given mediation strategies does in fact reflect the mediators' perceptions of language learning and attitudes to language teaching. Thus, addressing DA conceptions among teachers in similar contexts would be another further opportunity to tackle the teaching practices that, in fact, emanate from such conceptions (Hidri, 2015). Research should continue to investigate alternative forms of assessment where focus should be attended to underlying, latent variables that could genuinely be conducive to developing students' mental processing.

8 Conclusion

The purpose of this chapter was to address the validation of dynamic reading comprehension test specs for learners of English in an EAP program. Even though DA is still not recognized as a formal assessment policy to measure the language ability of test-takers in such a context, the different reading comprehension problems call for the adherence to this assessment mode. Specs validation is a cyclical process that should be intertwined with a myriad of facets, such as course objectives, language ability, learning outcomes and learners' needs. Item writers are called upon to adopt a challenging attitude to thoroughly address all these variables. Given the necessity to measure the test-takers' ability to know about and use language in context, test designers should move further ahead in measuring the ability of the test-takers to use language in context (Hidri, 2014). Many teaching methods have defined language ability differently, depending on the skill. Even though DA has become pervasive, many practitioners still have conflicting conceptions and practices about how to implement DA and how to score its performance. And may be the most effective way of assessing the test-takers' ability is to commingle dynamic with static assessments.

References

Ableeva, R. (2008). The effects of dynamic assessment on L2 listening comprehension. In J. P. Lantolf & M. E. Poehner (Eds.), *Sociocultural theory and the teaching of second languages* (pp. 57–86). London: Equinox.

Alderson, J. C. (2000). *Assessing reading*. Cambridge Language Assessment Series Cambridge: Cambridge University Press.

Alderson, J. C. (2004). Foreword. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods*. London: Lawrence Erlbaum.

Bachman, L. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

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

Black, P. J., & William, D. (1998). Assessment and classroom learning. *Assessment in Education: Principals, Policy and Practice, 5*(1), 57–74. doi:10.1080/0969595980050102

Block, D. (2003). *The social turn in second language acquisition*. Cambridge: Cambridge University Press.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum.

Campione, J. C., & Brown, A. L. (1987). Linking dynamic testing with school achievement. In C. S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 82–115). New York: Guilford Press.

Carrell, P. L. (1988). Introduction: Interactive approaches to second language reading. In P. L. Carrell, J. Devine, & D. E. Eskey (Eds.), *Interactive approaches to second language reading* (pp. 1–7). Cambridge: Cambridge University Press.

Cioffi, G., & Carney, J. J. (1997). Dynamic assessment of composing abilities in children with learning disabilities. *Educational Assessment, 4*(3), 175–202. doi:10.1207/s15326977ea0403_2

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

Cohen, A. D. (2007). The coming of age of research on test-taking strategies. In J. Fox, M. Weshe, D. Baylis, L. Cheng, C. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 89–111). Ottawa: Ottawa University Press.

de Beer, M. (2010). A modern assessment psychometric approach to dynamic assessment. *Journal of Psychology in Africa, 20*(2), 241–246. doi:10.1080/14330237.2010.10820372

Feuerstein, R., Rand, J., & Rynders, J. E. (1988). *Don't accept me as I am: Helping "retarded" people to excel*. New York: Plenum Press.

Gass, S. (1997). *Input, interaction, and the second language learner*. Mahwah, NJ: Lawrence Erlbaum.

Hamp-Lyons, L. (1997). Washback, impact and validity: Ethical concerns. *Language Testing, 14*, 295–303. doi:10.1177/026553229701400306

Haywood, H. C., & Lidz, C. S. (2007). *Dynamic assessment in practice. Clinical and educational applications*. Cambridge: Cambridge University Press.

Hidri, S. (2014). Developing and evaluating a dynamic assessment of listening comprehension in an EFL context. *Language Testing in Asia*, *4*(4). doi:10.1186/2229-0443-4-4

Hidri, S. (2015). Conceptions of assessment: Investigating what assessment means to secondary and university teachers. *Arab Journal of Applied Linguistics, 1*(1), 19–43.

Jeltova, I., Birney, D., Fredine, N., Jarvin, L., & Sternberg, R. E. L. (2007). Dynamic assessment as a process-oriented assessment in educational settings. *Advances in Speech-Language Pathology, 9*(4), 273–285. doi:10.1080/14417040701460390

Kozulin, A., & Gindis, B. (2007). Sociocultural theory and education of children with special needs: From defectology to remedial pedagogy. In H. Daniels, M. Cole, & J. V. Wertsch (Eds.), *The Cambridge companion to Vygotsky* (pp. 332–361). Cambridge: Cambridge University Press.

Kunnan, A. J. (1998a). Approaches to validation in language assessment. In A. J. Kunnan (Ed.), *Validation in language assessment: Selected papers from the 17th Language Testing Research Colloquium* (pp. 1–18). Mahwah, NJ: Lawrence Erlbaum.

Kunnan, A. J. (Ed.). (1998b). *Validation in language assessment: Selected papers from the 17th Language Testing Research Colloquium*, Long Beach. Mahwah, NJ: Lawrence Erlbaum.

Kunnan, A. J. (Ed.). (2000). Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida. *Studies in Language Testing*, (Vol. 9, pp. 1–14). Cambridge: UCLES, Cambridge University Press.

Lantolf, J. L., & Poehner, M. E. (2011). Dynamic assessment in the classroom: Vygotskian praxis for second language development. *Language Teaching Research, 15*(1), 1–23. doi:10.1177/1362168810383328

Lidz, C. S. (2002). Mediated learning experiences as a basis for an alternative approach to assessment. *School Psychology International, 23*(1), 68–84. doi:10.1177/0143034302023001731

Lidz, C. S., & Gindis, B. (2003). Dynamic assessment of the evolving cognitive functions in children. In A. Kozulin, V. S. Ageev, S. Miller, & B. Gindis (Eds.), *Vygotsky's educational theory in cultural context* (pp. 99–116). New York: Cambridge University Press.

Macrine, Sh L, & Sabbatino, E. D. (2008). Dynamic assessment and remediation approach: Using the DARA approach to assist struggling readers. *Reading and Writing Quarterly, 24*, 52–76. doi:10.1080/10573560701753112

McNamara, T. (1996). *Measuring second language performance*. New York: Longman.

McNamara, T. (2004). Language testing. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 763–83). Blackwell Publishing Ltd.

McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly, 3*(1), 31–51. doi:10.1207/s15434311laq0301_3

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

Pearson, P. D., Valencia, W. S., & Wixson, K. (2014). Complicating the world of reading assessment: Toward better assessments for better teaching. *Theory into Practice*, *53*, 236–246. doi:10.1080/00405841.2014.916958

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

Poehner, M. (2008). *Dynamic assessment: A Vygotskian approach to understanding and promoting second language development*. Berlin: Springer Publishing.

Poehner, M. (2011). Validity and interaction in the ZPD: Interpreting learner development through L2 Dynamic Assessment. *International Journal of Applied Linguistics, 21*(2), 244–263. doi:10.1111/j.1473-4192.2010.00277.x

Poehner, M., & van Compernolle, R. (2011). Frames of interaction in dynamic assessment: Developmental diagnoses of second language learning. *Assessment in Education: Principles, Policy and Practice, 18*(2), 183–198. doi:10.1080/0969594X.2011.567116

Reynold, M., Wheldall, K., & Madelaine, A. (2009). Building the WARL: The development of the Wheldall assessment of reading lists, a curriculum-based measure designed to identify young struggling readers and monitor their progress. *Australian Journal of Learning Difficulties, 14* (1), 89–111. doi:10.1080/19404150902783443

Sternberg, R. J., & Grigorenko, E. L. (2002). *Dynamic testing. The nature and measurement of learning potential*. Cambridge: Cambridge University Press.

Valsiner, J., & van der Veer, R. (1993). The encoding of distance: The concept of the zone of proximal development and its interpretations. In R. R. Cocking & K. A. Renninger (Eds.), *The development and meaning of psychological distance* (pp. 35–62). Hillsdale, NJ: Erlbaum.

Vygotsky, L. (1981). *Mind in society: The development of higher psychological process*. Cambridge, MA: Harvard University Press.

Vygotsky, L. (1986). *Thought and language*. Cambridge, MA: MIT Press.

Weiss, D. J. (1980). *Final report: Computerised adaptive performance evaluation*. Minneapolis, MN: University of Minnesota, Department of Psychology.