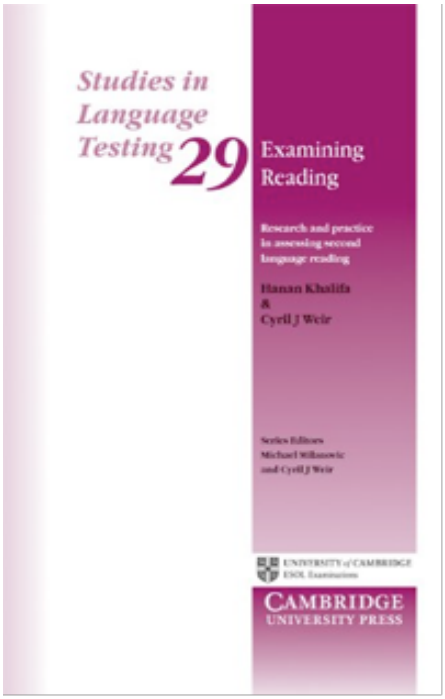
Reading in a Foreign Language
ISSN 1539-0578

October 2011, Volume 23, No. 2
pp. 225–230



Reviewed work:

Studies in Language Testing 29: Examining Reading: Research and Practice in Assessing Second Language Reading. (2009). Hanan Khalifa & Cyril J. Weir. Cambridge & New York: Cambridge University Press. Pp. 342 +xiv. ISBN 9780521736718. (paperback) $53.00

Reviewed by
Zahir Mumin
University at Albany, SUNY
United States

http://www.cambridge.org/

Khalifa and Weir present *Examining Reading* in order to address controversial issues concerning the structure and content of Cambridge English for Speakers of Other Languages (ESOL) exams. In this 29th volume of the *Studies in Language Testing* (SiLT) series, the authors focus on the dynamic relationships between students' cognitive skills, the test's scoring criteria, and its assessment tasks that exhibit a wide variety of different contexts. Khalifa and Weir fundamentally contend that test developers must clearly explicate and attentively examine how reading comprehension expectations align with students' learning needs in order to create reliable scoring assessments.

The book is organized into eight chapters that emphasize the importance of exam validity: Chapter 1 'Introduction; ' Chapter 2 'Test-taker characteristics; ' Chapter 3 'Cognitive validity;' Chapter 4 'Context validity;' Chapter 5 'Scoring validity;' Chapter 6 'Consequential validity;' Chapter 7 'Criterion-related validity;' and Chapter 8 'Conclusions and recommendations.' The authors also provide six appendices (Appendix A-F), which include reading comprehension tasks that substantiate their main argument in favor of meticulous learner outcome expectations.

Chapter 1 analyzes Weir's (2005) theoretical "framework for conceptualizing reading test validity" (p. 5, Figure 1.1) to demonstrate how this framework effectively employs a socio-

cognitive approach by associating validity components with students' mental processes as they complete reading comprehension tasks. In this chapter, Khalifa and Weir contextualize the content of Chapters 2–7, which further develop and apply Weir's theoretical framework to Cambridge ESOL exam constructs. Chapter 8 summarizes Chapters 1–7 in chronological order.

In Chapter 1, Khalifa and Weir foreground the content of Chapters 2–7 by defining the six validity components named in the chapter titles of the book. Test-taker characteristics consider how students' learning needs and sociocultural background affect their ability to complete different tasks. Context validity encompasses the sociocultural and sociolinguistic contexts in which tasks are performed. Cognitive validity examines the extent to which reading tasks are relevant to authentic language use. Scoring validity determines if the evaluation procedures of assessments are quantitatively consistent and reliable. Consequential validity explores how bias, washback, and social values affect testing procedures. Criterion-related validity analyzes the integration of test scores and external criteria to infer reliability of assessment procedures. The authors assert that the intense interaction amongst these components significantly impacts students' reading proficiency levels.

Chapter 2 outlines O'Sullivan's (2000) "categories of test-taker characteristics" (p. 18, Table 2.1) to investigate the physical/physiological, psychological, and experiential factors that may influence ESOL students' performance on reading proficiency exams. Khalifa and Weir contend that individual personal characteristics of test takers such as learning disabilities, motivation, and exam preparedness greatly impact their performance on reading tasks. The authors initiate the defense of this argument through analysis of international and the U.S. test accommodation policies which permit various types of exam modifications for students with learning disabilities (physical/physiological factors), such as extended time, Braille, scribers, readers, and different environmental settings without distractions. Khalifa and Weir focus on extended time and Braille as two of the most important accommodations that relate to students' ability to complete reading tasks.

Khalifa and Weir note that Sireci, Scarpati, and Li (2005) spark insightful controversy in the field of second language reading with their review of empirical studies which show that extended time enhances the academic disposition of students with disabilities giving an advantage over those without disabilities. However, Khalifa and Weir also point to Pring's (1994) findings concerning Braille readers which demonstrate that they read at a much slower pace than print readers and therefore, should be permitted extended time. The authors provide this example of extended time use to allay the controversy regarding the possible unfair advantage for students with learning disabilities. Khalifa and Weir also address issues dealing with fairness in their examination of psychological (personality) and experiential (educational background) factors.

They argue that Cambridge ESOL test development procedures effectively gather personal and demographic information from exam candidates using Candidate Information Sheets (CIS) and Differential Item Functioning (DIF) in an attempt to prevent bias and ameliorate the content of reading activities. They supplement this argument with further claims that the pretest conducted by Cambridge ESOL administrators focuses on eliciting information from students about their previous knowledge of exam formats and popular topics of personal interest. This information, in turn, makes exams contextually more relevant to students' personal lives and prior exam

experiences. This chapter successfully furnishes solid background knowledge on the interrelatedness of test policies, test procedures, and students' personal characteristics to set up the discussion in Chapter 3 concerning cognitive validity.

Chapter 3 investigates the importance of cognitive validity in ESOL exams. Khalifa and Weir initiate the chapter by defining and comparing exam validation approaches such as the factorial approach, the reading subskills approach, and the cognitive processing approach. Factorial approaches focus on quantitatively describing specific abilities that are generally necessary for reading comprehension success: recognizing key words, understanding the main ideas of paragraphs, and identifying similar patterns of sentence structure. Reading subskills approaches examine the effectiveness of different skills (automatized abilities/subconscious acquisition) and strategies (learned abilities/conscious learning) used to comprehend texts. Khalifa and Weir demonstrate the impact of cognitive processing approaches on exam development by creating a functional reading model (Figure 3.1, p. 43), describing its components (types of reading and cognitive processes), and applying these components to exam constructs of Cambridge ESOL Main Suite Reading papers from the Key English Test (KET), Preliminary English Test (PET), First Certificate in English (FCE), Certificate in Advanced English (CAE), and Certificate of Proficiency in English (CPE).

The authors first analyze types of reading and confirm Ashton's (2003, p. 128) findings which show that the gapped-text tasks on the CAE and CPE exams oblige students to employ careful global reading skills (profound higher-order thinking skills) in order to successfully comprehend texts. In contrast, Khalifa and Weir's examination of cognitive processes shows that the majority of the content of the Cambridge ESOL Main Suite Reading papers' tasks does not require students to use their most complex cognitive processing skills ("creating a text level structure [and] creating an organized representation of several texts" p. 70) to integrate the meaning of sentences in different paragraphs and/or explicate the intertextual meaning throughout multiple texts. This chapter could be enhanced with linguistic analyses of cognitive processes in order to determine, for example, if the development of ESOL students' comprehension skills of inference are primarily due to syntactic parsing, lexical access, or word recognition.

Chapter 4 explores polemical issues related to the development of contextually valid ESOL exams. Khalifa and Weir argue that tasks must exhibit real-life contexts associated with students' personal lives in order for exam administrators to reliably determine the extent to which students acquire certain levels of English reading proficiency. The authors substantiate this argument by applying the fifteen components of Weir's (2005) context validity model (p. 82, Figure 4.1) to different exam formats (e.g., multiple choice and true/false) and specific content (e.g., grammar and lexicon) included in the Cambridge Main Suite Reading papers exams. Of these fifteen components, Khalifa and Weir highlight the significance of the following two components: "order items" (p. 82) from the task setting category and "content knowledge" (p. 82) from the linguistic demands category. For order items, the authors contend that the order in which students are required to provide answers to reading comprehension questions should match the order in which the relevant reading material is presented to them. Hughes (1989) and Weir (1993) support this argument by concluding that students normally process reading material in a chronological order. Therefore, when these students are exposed to exercises using a random order of responses, this order may impede comprehension and degrade the reliability of test

performance results. Khalifa and Weir further solidify the aforementioned argument by illuminating the inconsistency between the chronological order (careful reading activities) and random order (expeditious reading activities) of responses required on Cambridge ESOL exams. The authors conclude the chapter by arguing in favor of matching general test topics (content knowledge) of the Main Suite Reading papers to students' background knowledge in order to stimulate their reading comprehension and enhance the authenticity of context validity.

Chapter 5 focuses on the employment of appropriate statistical analyses of exam tasks, results, and scoring procedures, which help determine the reliability of Cambridge ESOL reading comprehension exams. Khalifa and Weir postulate that scoring validity is one of the most imperative components because the lack of scoring validity constitutes grave deficiencies in cognitive validity and context validity. The authors sustain this postulate by applying the six components of Weir's (2005) scoring validity model (p. 144, Figure 5.1) to Cambridge ESOL scoring practices. The six components are item difficulty, item discrimination, internal consistency, error of measurement, marker reliability, and grading and awarding.

Item difficulty and item discrimination measure the relationship between task performance, task facility, and number of examinees. Internal consistency determines the extent to which students achieve similar scores when identical skills are assessed. Error of measurement assures that students with scores close to the borderline of pass/fail are not adversely affected by human errors.

Marker reliability develops detailed procedures for maintaining scoring consistency in manually graded exams. Grading and awarding establishes grading norms for quantifying cut-off passing/failing grades and develops criteria for written results notifications sent to students. Khalifa and Weir demonstrate, through their analyses of various statistical procedures such as Rasch-based statistics, the Cronbach Alpha method, and the Standard Error of Measurement, that Cambridge ESOL exams possess high levels of reliability. The authors' summary of Chapter 5's in-depth analysis of scoring validity clashes with Chapter 6's analysis of consequential validity because the former is an internal validation process whereas the latter is an external validation process.

Chapter 6 delves into prior and current research concerning the influence of consequential validity on test development procedures. Khalifa and Weir claim that *impact,* the effect of tests on society; *washback*, the effect of tests on teaching and learning; and potential *test bias* are three key factors of exam validation which represent a tug-of-war relationship between stakeholders who have major interests in exam criteria and formats and teachers and learners who are often concerned with all exam conditions. The authors successfully defend this claim through analysis of Taylor's (2000) "Stakeholders in the Test Community" model (p. 177, Figure 6.3) within the context of Cambridge ESOL test development procedures. Khalifa and Weir also review two recent studies, one on the washback effect of CPE textbooks and the other on the impact of International English Language Testing System (IELTS) preparation, to show that test administrators are attempting to create exams that are more amenable to test takers' real-life reading comprehension situations. Although these attempts reflect major improvements in the creation of reading comprehension tasks, the authors do not propose a clear resolution for the aforementioned tug-of-war which could be resolved through direct communication and feedback

exchanges between test takers and stakeholders. I argue that the integration of DIF analysis and test taker/stakeholder interaction mollifies and, in some cases, eliminates test bias, negative impact, and negative washback. The need for this integration is clearly demonstrated through prior research such as Geranpayeh and Kunnan's (2007) study which examines test-taker characteristics using DIF analysis without addressing stakerholders' superordinate logistical exam power.

Chapter 7 features a detailed examination of criterion-related validity which encompasses contentious issues regarding cross-test comparability, test equivalence, and external standards. Khalifa and Weir insist that the development of test comparability frameworks should be focused on aligning with the English language proficiency levels of the Common European Framework of Reference for Languages (CEFR): A1 Breakthrough, A2 Waystage, B1 Threshold, B2 Vantage, C1 Effective Operational Proficiency, and C2 Mastery. The authors substantiate this insistence with a multifaceted language proficiency model (p. 192, Figure 7.2) which demonstrates how other English proficiency exams such as IELTS, Business English Certificate (BEC), and International Certificate in Financial English (ICFE) are aligned with CEFR proficiency standards.

Khalifa and Weir also elucidate the effectiveness of manual alignment procedures (familiarization, specification, standardization, and empirical validation) that Cambridge ESOL test administrators use during their test development processes to link Cambridge ESOL proficiency standards to those of the CEFR. However, at the same time, the authors avouch the need for further research that empirically validates the tendency to compare English proficiency levels of other exams to those of the CEFR. Khalifa and Weir close off the chapter with Taylor's (2004) opposing arguments addressing this unresolved CEFR alignment issue by identifying benefits concerning the facilitated interpretation of English proficiency guidelines and risks dealing with the oversimplification of these guidelines.

Chapter 8 summarizes the previous seven chapters in chronological order, which metaphorically supports Khalifa and Weir's arguments dealing with Chapter 4's "order items" concept. The authors maintain that the reliability of construct validity—the combination of cognitive validity, context validity, and scoring validity—must be empirically examined and the differences between English proficiency levels must be clearly operationalized in order to provide solid evidence that accounts for test takers' real-life reading comprehension experiences. With regard to consequential and criterion-related validity, Khalifa and Weir reinforce the importance of constantly cross-evaluating exam content and procedures so as to develop and/or maintain high levels of reliability in Cambridge ESOL exams. Offering suggestions for empirical research studies at the end of this chapter would illuminate the need for quantitative and qualitative analyses of ESOL test takers, teachers, and stakeholders as research participants.

This book bedazzles educators, exam administrators, and language acquisition professionals by applying a socio-cognitive theoretical framework of exam validity components to Cambridge ESOL exams with the intention of exhibiting reliability levels of different exam constructs. Within the context of the book, however, there is not a clear definition for test takers, (who are not stockholders) and stockholders, (who are not test takers). It is clear that the term "test takers" refers to mid/low-level stockholders (students) and the term "stockholders" constitutes high-level

stockholders (governmental agencies). These high and mid/low-level stockholders provide feedback to test administrators, but not to each other, to try and enhance reliability and validity of ESOL exams.

References

Ashton, M. (2003). The change process at the paper level. Paper 1 Reading. In. C. J. Weir & M. Milanovic (Eds.), *Continuity and innovation: Revising the Cambridge proficiency in English examination 1913-2002*, *Studies in language testing 15* (pp. 121–164). Cambridge: UCLES/Cambridge University Press.

Geranpayeh, A., & Kunnan, A. J. (2007). Differential item functioning in terms of age in the Certificate in Advanced English examination. *Language Assessment Quarterly*, *4*(2), 190–222.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

O'Sullivan, B. (2000). *Towards a model of performance in oral language testing*. Unpublished PhD thesis. Berkshire, ENG: University of Reading.

Pring, L. (1994). Touch and go: Learning to read braille. *Reading Research Quarterly*, *29*(1), 67–74.

Sireci, S. G., Scarpati, S., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research*, *75*(4), 457–490.

Taylor, L. (2000). Stakeholders in language testing. *Research Notes*, *2*, 2–4.

Taylor, L. (2004). Issues of test comparability. *Research Notes*, *15*, 2–5.

Weir, C. J. (1993). *Understanding and developing language tests*. New York, NY: Prentice Hall.

Weir, C. J. (2005*). Language testing and validation: An evidenced-based approach*. Basingstoke, ENG: Palgrave Macmillan.

About the Reviewer

Zahir Mumin teaches Spanish courses at the University at Albany, SUNY, USA and conducts research in the field of linguistics. His primary research interests include L1 and L2 reading comprehension, sociolinguistics, phonology, morphosyntax, language contact, language change, L1 and L2 acquisition, semantics, pragmatics, multilingualism, and dialect variation. Email: zm227418@albany.edu