Title of Project:
Investigating the Construct of Topical Knowledge in a Scenario-Based Assessment Designed to Simulate Real-Life Second Language Use

Researcher:
Heidi Liu Banerjee
Teachers College, Columbia University
heidi.liu@tc.columbia.edu

Research Supervisor:
Dr. James E. Purpura
Teachers College, Columbia University

Heidi Liu Banerjee

Final Report

Motivation for the Research

The vast development of digital technology and the widespread use of social network platforms have reshaped how we live in the world. For second language (L2) learners to maximally utilize their language proficiency to function effectively as members of modern society, they need not only the necessary L2 knowledge, skills, and abilities (KSAs) but also essential topical knowledge. While many researchers believe that topical knowledge should be viewed as an integral component of L2 communicative competence, the role of topical knowledge has not always been accounted for in an assessment context due to the difficulty of operationalizing the construct.

Scenario-based assessment, an innovative, technology-based assessment approach, allows great affordances for expanding the measured constructs of an assessment. It is designed expressly for learners to demonstrate their KSAs in a context that simulates real-life language use. Through the utilization of a sequence of thematically-related tasks, along with simulated character interaction, scenario-based assessment offers opportunities to examine L2 learners' communicative competence in a purposeful, interactive, and contextually meaningful manner.

Acknowledging the importance of including topical knowledge as part of the broadened construct of L2 proficiency, as well as the potential of utilizing scenario-based assessment to simulate the complexity of real-life language use, the primary purpose of this study was to investigate the relationships between topical knowledge and L2 KSAs in a scenario-based language assessment (SBLA). A set of language tasks surrounding a particular theme was designed to elicit L2 learners' reading, listening, and writing abilities. Additionally, a topical knowledge task related to the same theme was designed as a performance moderator (O'Reilly & Sabatini, 2013) to activate prior knowledge, as well as to track topical learning. L2 learners' various language and topical background characteristics were also explored to determine whether and to what extent they played a role in L2 learners' demonstration of topical learning and overall L2 performance as measured by the SBLA.

Research Questions

The following research questions were addressed in the current study:

1. What is the relationship between L2 learners' pre-scenario topical knowledge, post-scenario topical knowledge, and their L2 performance in an SBLA?
2. How well can the items in the topical knowledge task provide evidence for construct validity?
    a. To what extent do the topical knowledge items display adequate psychometric properties for their function as a measure of L2 learners' topical knowledge?
    b. To what extent is there evidence in support of the use of the same set of topical knowledge items to track L2 learners' topical learning?
3. To what extent do L2 learners with different language and topical background characteristics, specifically, L2 proficiency levels (intermediate, high-intermediate, advanced), fields of expertise (English, science, other), and degrees of prior personal experience with the topic (low, medium, high), vary in their topical learning and L2 performance?
4. In what ways are L2 learners' abilities to achieve the scenario goal dependent on their topical knowledge and L2 KSAs? In other words, what inferences can be made about L2 learners who display certain levels of topical knowledge, listening and reading abilities, and their eventual success in achieving the scenario goal via summary writing?

Research Methodology

Context of the Study

The context of the current study, specifically for the assessment development, was set in the Community English Program (CEP) at a major research university in the United States. In order to place students entering the Program into appropriate course levels, a requisite placement exam is administered to all new CEP students at the beginning of every semester. Based on the results, incoming students are assigned to one of the 19 beginning, intermediate, or advanced levels the CEP offers. As an ongoing project, the CEP has been revamping its placement exam to better reflect the evolving construct of communicative language ability. The assessment instrument (i.e., the SBLA) developed in this study was designed for the high-intermediate level scenario component of the new CEP placement exam

Participants

The participants in this study included 118 adult, Taiwanese English as a foreign language (EFL) learners, who served as test-takers, and two experienced English as a second language (ESL) teachers, who served as raters. In order to ensure that the characteristics of the participants were relevant to the research context, purposeful sampling (Wiersma & Jurs, 2009) was used. The EFL learners did not have direct connection to the CEP at the time of data collection; however, their proficiency levels were deemed appropriate to represent prospective test-takers who would be assigned to take the high-intermediate scenario module in the new placement exam. Both of the raters have had extensive ESL teaching and testing experience in the United States.

Instruments

The instruments used in the current study included an SBLA titled "*Nutrition Ambassador*," the SBLA experience survey, and the analytic scoring rubric developed to score test-takers' responses to the writing task in the SBLA.

Data Collection Procedures

The test data were collected from EFL learners in Taiwan who voluntarily signed up to take the SBLA at a university computer lab. The SBLA was administered a total of six times over the span of three weekends. Two university students were hired as research assistants to proctor and monitor the test-taking process on-site, and the Researcher monitored the entire process remotely. The test-takers had a maximum of 110 minutes to finish the entire test. All of the test-takers finished the SBLA between 60 and 110 minutes. The assessment platform, Qualtrics, recorded and saved the test-takers' responses automatically. All of the test-takers completed the test, and their responses were recorded successfully.

Data Analysis Procedures

The first research question, which investigated the relationships between topical knowledge and L2 performance within the SBLA, was examined by way of correlation and path analyses. Three hypothesized path models were analyzed to determine the effects between learners' proficiency levels, topical knowledge, and L2 KSAs, as measured in this study.

The second research question, which investigated the extent to which the topical knowledge task in the SBLA could properly measure topical knowledge and track topical learning, was examined through a four-step Rasch analysis. The Rasch results were cross-compared to ensure that there was sufficient validity evidence in support of the intended use of the topical knowledge task.

The third research question, which investigated the roles L2 learners' language and topical background characteristics (i.e., proficiency levels, fields of expertise, and degrees of prior personal experience with food additives) played in their topical learning and L2 performance, was examined through a series of one-way ANOVAs.

Finally, the fourth research question, which investigated the ways in which L2 learners' ability to achieve the scenario goal (i.e., sharing knowledge of unsafe food additives) was dependent on their topical knowledge and L2 KSAs as measured in the preceding tasks within the SBLA, was examined using Bayesian networks.

Summary of Findings

Through correlation and path analyses, L2 learners' pre-scenario topical knowledge and post-scenario topical knowledge were found to have different relationships with their L2 performance in the SBLA. The two aspects of topical knowledge as operationalized in this study, content knowledge (i.e., knowledge of factual information) and lexical knowledge (i.e., knowledge of lexical meanings), showed effects on L2 abilities differently; however, L2 learners' proficiency level was found to account for their L2 performance the most.

Following that, validity evidence of the construct of topical knowledge as measured in the SBLA was provided using a four-step Rasch analysis. The results showed that the topical knowledge items were generally shown to fit the Rasch models well, demonstrating adequate psychometric properties for their functions as a measure of L2 learners' topical knowledge. Because the same topical knowledge task was administered both in the beginning and at the end of the SBLA, the test-takers' topical learning was tracked through the changes in item difficulty parameters. The estimated item difficulty change showed that, through contextualizing the knowledge building and sharing process, L2 learners were able to demonstrate substantial content learning. However, L2 learners did not seem to have learned the lexical items as much as they had the factual information of food additives. This is likely related to the ultimate goal of the SBLA: the test-takers were asked to share the information they had learned about food additives with their community. Therefore, the test-takers had to rely heavily on the content of the article in the reading task. It is important to note however, they were not asked to use the lexical items in any part of the assessment, nor was there explicit instruction of the lexical items during the knowledge building and sharing process. Such an assessment design may have contributed to the type of topical learning L2 learners demonstrated.

Then, taking a closer look at the roles L2 learners' language and topical background characteristics played in their topical learning and L2 performance, a series of one-way ANOVAs was employed. The results revealed that L2 learners of different proficiency levels (intermediate, high-intermediate, and advanced) and fields of expertise (English, science, and other) varied in their content learning, lexical learning, and L2 performance. However, their prior personal experience with food additives did not appear to play a role here, suggesting that L2 learners' self-identified life experience with a particular topic, which may be related to their episodic memory (Tulving, 1972), does not necessarily transform to readily accessible knowledge that can be utilized while the learners perform L2 tasks.

Finally, considering that the tasks in the SBLA were all thematically-related, a Bayesian network was constructed to examine the ways in which L2 learners' ability to achieve the scenario goal depended on their topical knowledge and L2 KSAs. While the Bayesian network constructed in this study was exploratory in nature, and the sample size was not sufficient to yield robust generalizability of the results, it provided a holistic understanding of how L2 learners' ability to achieve a communicative goal depended on their topical knowledge and L2 KSAs in the context of an SBLA. The results from Bayesian network also revealed that L2 learners' ability to gain topical knowledge while completing a sequence of thematically-related L2 tasks appeared to be an essential part of their L2 communicative competence, and therefore, should be considered as a component of their "L2 proficiency score." Lastly, with the increasing interest in adopting game-, scenario-, and simulation-based assessments to measure complex constructs of learners' KSAs, this study demonstrated how Bayesian networks may be utilized to interpret the relationships between the measured constructs, so that results from these complex assessments can yield meaningful interpretations.

Implications

The current study carries a number of possible theoretical, methodological, and pedagogical implications for the field of applied linguistics, particularly in language assessment.

Theoretically, this study contributes to the understanding of the role of L2 learners' topical knowledge and its relation to L2 KSAs in a language assessment. In order to properly capture the nature of topical knowledge, this study operationalized topical knowledge as both knowledge of topical content and knowledge of lexical meanings, and a set of items related to the theme of the scenario-based language assessment were designed to measure L2 learners' topical knowledge. By administering the same set of items to the test-takers both before and after the scenario, it can be observed how much topical knowledge the test-takers already had (i.e., prior topical knowledge), and how much they learn about the topic in the process of achieving the scenario goal (i.e., topical learning). Through the test design, this study may provide insights into how L2 learners utilize their prior topical knowledge to complete the language tasks, and how a highly-contextualized scenario-based assessment may facilitate topical learning.

Methodologically, the current study informs the use of scenario-based assessment in L2 assessment contexts as well as the test design and the statistical procedure for an assessment with complex constructs. Because scenario-based assessment aims to simulate real-life language use, it is crucial for the storyline embedded in the scenarios to be coherent so that test-takers can perform in a way that is natural to their cognitive functioning (O'Reilly et al., 2015). With a coherent structure, the assessment results subsequently may allow test users to make meaningful interpretations of the evidence collected from test-takers' performance. In order to do so, the scenario-based language assessment in this study adopts the key literacy practice of building and sharing knowledge as its theoretical framework, and an ECD framework as its design principle. The coherent test design also allows for the development of a Bayesian network to model the dependencies among L2 learners' topical knowledge and their L2 KSAs, a measurement method rarely used in the context of L2 assessment. The scarcity of its use is primarily due to the fact that, until fairly recently, the technical constraints have made it difficult to measure complex constructs or simulate real-life language use coherently and systematically within an assessment.

Pedagogically, the current study attempts to address the facilitation of learning through meaningfully contextualizing an assessment, where test-takers can apply their L2 KSAs in an authentic manner (Hidalgo, Sata, & Suzuki, 2015). By examining the extent to which test-takers gained topical knowledge while completing the language tasks to fulfill the scenario goal, this study demonstrates how a purpose-driven, high-contextualized scenario-based language assessment could both be used to gauge L2 learners' language proficiency and serve as a learning medium.

References

Agostinho, S., Meek, J., & Herrington, J. (2005). Design methodology for the implementation and evaluation of a scenario-based online learning environment. *Journal of Interactive Learning Research, 16*, 229–242.

Alderson, C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: UK: Continuum.

Alderson, J. C., Haapakangas, E. L., Huhta, A., Nieminen, L. & Ullakonoja, R. (2014). *The diagnosis of reading in a second or foreign language*. New York, NY: Routledge.

Alderson, J. C., & Urquhart, A. H. (1985a). The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing, 2*, 192–204. doi:10.1177/026553228500200207

Alderson, J. C., & Urquhart, A. H. (1985b). This test is unfair: I'm not an economist. In P. C. Hauptman, R. Leblanc, & M. B. Wesche (Eds.), *Second-language performance testing* (pp. 25–43). Ottawa, Canada: University of Ottawa Press.

Alexander, P. A., Schallert, D. L., & Hare, V. C. (1991). Coming to terms: How researchers in learning and literacy talk about knowledge. *Review of Educational Research, 61*, 315–343. doi:10.3102/00346543061003315

Almond, R. G., Mislevy, R. J., Steinberg, L., Yan, D., & Williamson, D. (2015). *Bayesian networks in educational assessment*. New York, NY: Springer.

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment, 1*(5). Retrieved from http://www.bc.edu/research/intasc/jtla/journal/v1n5.shtml

Anderson, J. R. (1990). *Cognitive psychology and its implications* (3rd ed.). New York, NY: W.H. Freeman.

Anderson, R. C., Reynolds, R. E., Schallert, D. L., & Goetz, E. T. (1977). Frameworks for comprehending discourse. *American Educational Research Journal, 14*, 367–381. doi:10.3102/00028312014004367

Anderson, R. D., & Vastag, G. (2004). Causal modeling alternative in operations research: Overview and application. *European Journal of Operational Research, 156*(1), 92–109. doi:10.1016/s0377-2217(02)00904-9

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

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 41–71). Ottawa, Canada: University of Ottawa Press.

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*, 449–465. doi:10.2307/3586464

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

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

Baddeley, A. D. (2015). What is memory? In A. Baddeley, M. W. Eysenck, & M. C. Anderson (Eds.), *Memory* (2nd ed.) (pp. 18–50). Hove, East Sussex: Psychology Press.

Baecher, L., Farnsworth, T., & Ediger, A. (2014). The challenges of planning language objectives in content-based ESL instruction. *Language Teaching Research, 18*(1), 118–136. https://doi.org/10.1177/1362168813505381

Banerjee, H. L. (2017, October). *Investigating the construct of topical knowledge in second language assessment: A scenario-based assessment approach*. Best student paper award presented at 2017 annual MwALT conference, Dayton, OH.

Banerjee, H. L. (2018, March). *Measuring topical knowledge and tracking topical learning through a scenario-based language assessment.* Paper presented at 2018 AAAL annual conference colloquium "The affordances of scenario-based assessment for broadening measurement opportunities in second or foreign language assessment", Chicago, IL.

Banerjee, H. L. (forthcoming). Investigating the construct of topical knowledge in second language assessment: A scenario-based assessment approach. *Language Assessment Quarterly*.

Bao, L. (2006). Theoretical comparisons of average normalized gain calculations. *American Association of Physics Teachers*, *74*, 917-922. doi: 10.1119/ 1.2213632

Basturkmen, H. (2006). *Ideas and options in English for specific purposes*. Mahwah, NJ: Erlbaum.

Beaujean, A. A. (2014). *Latent variable modeling using R: A step-by-step guide*. New York, NY: Routledge.

Belcher, D. (2009). What ESP is and can be: An introduction. In D. Belcher (Ed.), *English for specific purposes in theory and practice* (pp. 1–20). Ann Arbor, MI: University of Michigan.

Bennett, R. E. (2010). Cognitively Based Assessment of, for, and as Learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives, 8*, 70–91. doi:10.1080/15366367.2010.508686

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education, 39*(1), 370–407. doi:10.3102/0091732X14554179

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K-12 assessment: Integrating accountability testing, formative assessment and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Assessment issues of the 21st century* (pp. 43–61). New York, NY: Springer.

Bereiter, C. & Scardamalia, M. (1987). *The psychology of written composition.* Hillsdale, NJ: Lawrence Erlbaum.

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Defining twenty-first century skills. In P. Griffin, B McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17–66). Springer Science+Business Media B.V. doi: 10.1007/ 978-94-007-2324-5_2.

Black, H. (1986). Assessment for learning. In D. L. Nuttall (Ed.), *Assessing educational achievement* (pp. 7–18). London: Falmer Press.

Black, P. J., & Wiliam, D. (2004). The formative purpose: assessment must first promote learning. In Wilson, M. (Ed.). *Towards coherence between classroom assessment and accountability: 103rd Yearbook of the National Society for the Study of Education (part 2)* (vol. Part II, pp. 20–50). Chicago, IL: University of Chicago Press.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE–Life Sciences Education, 15*(4), 1–7. doi:10.1187/cbe.16-04-0148

Brinton, D. M. (2013). Content-based instruction in English for specific purposes. In C. A. Chapelle (Ed.), The encyclopedia of applied linguistics. Blackwell Publishing Ltd. doi: 10.1002/9781405198431.wbeal0191

Brinton, D. M., & Holten, C. (2001). Does the emperor have no clothes? A re-examination of grammar in content-based instruction. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 239–251). Cambridge, UK: Cambridge University Press.

Brinton, D. M., & Snow, M. S. (2017). The evolving architecture of content-based instruction. In M. A. Snow & D. M. Brinton (Eds.), *The content-based classroom: New perspectives on integrating language and content* (pp. 2–20), Ann Arbor, MI: Michigan University Press.

Brinton, D. M., Snow, M. S., & Wesche, M. B. (1989). *Content-based second language instruction.* New York, NY: Newbury House Publishers.

Broukal, M. (2016). *Weaving it together 4*. San Francisco, CA: Cengage Learning.

Brown, J. D. (1984). A norm-referenced engineering reading test. In A. K. Pugh & J. M. Ulijn (Eds.), *Reading for professional purposes* (pp. 213–222). London, UK: Heinemann Educational Books.

Canale, M. (1983). On some dimensions of language proficiency. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 333–342). Rowley, MA: Newbury House.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1–47.

Carrell, P. L. (1983). Three components of background knowledge in reading comprehension. *Language and Learning, 33*, 183–207. https://doi.org/ 10.1111/j.1467-1770.1983.tb00534.x

Carrell, P. L. (1991). Second language reading: Reading ability or language proficiency? *Applied Linguistics, 12*, 159–179. https://doi.org/10.1093/ applin/12.2.159

Carroll, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. In *Testing the English Proficiency of Foreign Students* (pp. 30–40). Washington, DC: Center for Applied Linguistics.

Carroll, B., Liu, H., & Oh, S. (2015, April). *Design considerations in the assessment of L2 integrated skills through scenarios*. The First TC/ETS Forum on Teaching, Learning, and Assessment of English Language Learners. New York, NY.

Chalhoube-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20*, 369–383. https://doi.org/10.1191/ 0265532203lt264oa

Chang, W.-C., & Chan, C. (1995). Rasch analysis for outcomes measures: some methodological considerations. *Archives of Physical Medicine and Rehabilitation, 76*, 934–939. https://doi.org/10.1016/ S0003-9993(95)80070-0

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). New York, NY: Cambridge University Press.

Chiang, S. C., & Dunkel, P. (1992). The effect of speech modification, prior knowledge, and listening proficiency on EFL lecture learning. *TESOL Quarterly, 26,* 345–374. doi:10.2307/3587009

Chung, T., & Berry, V. (2000). The influence of subject knowledge and second language proficiency on the reading comprehension of scientific and technical discourse. *Hong Kong Journal of Applied Linguistics*, *5*(1), 187–222.

Clapham, C. (1996). *The development of IELTS: A study of the effect of background knowledge on reading comprehension*. Cambridge, UK: Cambridge University Press.

Clark, J. (1972). *Foreign language testing: Theory and practice.* Philadelphia, PA: Center for Curriculum Development.

Clarke, M. A. (1980). The short circuit hypothesis of ESL reading: When language competence interferes with reading performance. *Modern Language Journal*, *64,* 203–209. doi:10.2307/325304

Cohen, G. (1983). *The psychology of cognition*. New York, NY: Academic Press.

Cohen, J. (1988) *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J., Cohen P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Collier, L. (2016). *Do scenario-based assessments hold new promise for uncovering reading ability?* (ETS Open Notes). Princeton, NJ: Educational Testing Service. Retrieved from http://news.ets.org/stories/scenario-based- assessments-hold-new-promise-uncovering-reading-ability/

Council of Chief State School Officers (2012). *Framework for English language proficiency development standards corresponding to the Common Core State Standards and the Next Generation Science Standards.* Washington, DC: Author.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334. https://doi.org/10.1007/BF02310555

Culbertson, M. J. (2016). Bayesian networks in educational assessment: The state of the field. *Applied Psychological Measurement, 40,* 3–21. https://doi.org/ 10.1177/0146621615590401

Cumming, A. (2014). Assessing integrated skills. In A. J. Kunnan (Ed.), The companion to language assessment. John Wiley & Sons, Inc. doi:10.1002/ 9781118411360.wbcla131

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper* (Research Report No. RM-00-05). Princeton, NJ: Educational Testing Service.

Dag, O., Dolgun, A., & Konar, N. M. (2018). onewaytests: An R package for one-way tests in independent groups designs. *The R Journal, 10*(1), 175–199.

Dalton-Puffer, C., Nikula, T., & Smit, U. (Eds.). (2010). *Language use and language learning in CLIL classrooms*. Amsterdam, the Netherlands: John Benjamins Publishing.

Davies, M. (2001). Explicit and implicit knowledge: Philosophical aspects. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (pp. 8126–8132). Amsterdam, the Netherlands: Elsevier Science.

Davison, C. (2005). Learning your lines: negotiating language and content in subject English. *Linguistics and Education, 16* , 219–237. https://doi.org/10.1016/ j.linged.2006.01.005

de Klerk, S., Veldkamp, B. P., & Eggen, T. J. H. M. (2015). Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a Bayesian network example. *Computer & Education*, *85*, 23–34. https://doi.org/10.1016/j.compedu.2014.12.020

de Zarobe, Y. R., & Catalán, R. M. J. (2009). *Content and language integrated learning: Evidence from research in Europe*. Buffalo, NY: Multilingual Matters.

Deane, P. (2011). *Writing assessment and cognition* (Research Report No. RR-11-14). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02250.x

Deane, P., Sabatini, J., Feng, G., Sparks, J., Song, Y., Fowles, M., . . . Foley, C. (2015), *Key practices in the English Language Arts (ELA): Linking learning theory, assessment, and instruction.* (Research Report No. RR-15-17). Princeton, NJ: Educational Testing Service. doi:10.1002/ets2.12063

Deane, P., Sabatini, J., & O'Reilly, T. (2012). *English language arts literacy framework.* Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/s/research/pdf/ela_literacy_framework.pdf

DeGroff, L. J. C. (1987). The influence of prior knowledge on writing, conferencing, and revising. *Elementary School Journal. 88*, 105–118. https://doi.org/ 10.1086/461527

Delacruz, G. C. (2011). *Games as formative assessment environments: Examining the impact of explanations of scoring and incentives on math learning, game performance, and help seeking* (CRESST Report 796). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Dinsmore, D. L., & Parkinson, M. M. (2013). What are confidence judgments made up? Students' explanations for their confidence ratings and what that means for calibration. *Learning and Instruction, 24*, 4–14.

Dochy, F. J. R. C., & Alexander, P. A. (1995). Mapping prior knowledge: A framework for discussion among researchers. European *Journal of Psychology of Education, 10*, 225–242. doi:10.1007/BF03172918

Dochy, F. J. R. C., Moerkerke, G., & Martens, R. (1996). Integrating assessment, learning, and instruction: Assessment of domain-specific and domain-transcending prior knowledge and progress. *Studies in Educational Evaluation*, *22,* 309–339. https://doi.org/10.1016/0191-491X(96)00018-1

Dochy, F. J. R. C., Segers, M., & Buehl, M. M. (1999). The relation between assessment practices and outcomes of studies: The case of research on prior knowledge. *Review of Educational Research, 69*, 145–186. doi:10.3102/00346543069002145

Dörnyei, Z. (1998). Motivation in second and foreign language learning. *Language Teaching, 31*, 117–135. doi:10.1017/S026144480001315X

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

Douglas, D. (2005). Testing languages for specific purposes. In E. Hinkel (Ed.), *The handbook of research in second language teaching and learning* (pp. 857–868). Mahwah, NJ: Erlbaum.

Douglas, D. (2013). ESP and assessment. In B. Paltridge & S. Starfield (Eds.), *The handbook of English for specific purposes* (pp. 367–384). West Sussex, UK: John Wiley & Sons.

Douglas, D., & Selinker, L. (1992). Analyzing oral proficiency test performance in general and specific purpose contexts. *System*, *20*, 317–328. http://dx.doi.org/ 10.1016/0346-251X(92)90043-3

Dudley-Evans, T., & St. John, M. J. (1998). *Developments in English for specific purposes: A multidisciplinary approach*. Cambridge, UK: Cambridge University Press.

Dueñas, M. (2004). The whats, whys, hows, and whos of content-based instruction in second/foreign language education. *International Journal of English Studies*, *4*(1), 73–96. Retrieved from https://revistas.um.es/ijes/article/view/48061

Dunlea, J. (2014). *Investigating the relationship between empirical task difficulty, textual features, and CEFR levels.* Paper presented at the 2014 EALTA Conference, Coventry, UK.

Educational Testing Service (2005). *TOEFL iBT writing sample responses*. Princeton, NJ: Author. Retrieved from http://www.ets.org/Media/Tests/TOEFL/pdf/ibt_writing_sample_responses.pdf

Educational Testing Service (2014). *TextEvaluator*®. Princeton, NJ: Author. Available from https://textevaluator.ets.org/TextEvaluator/

Educational Testing Service (2015). *Mapping the TOEIC® and TOEIC Bridge™ tests on the Common European Framework of Reference for Languages*. Princeton, NJ: Author.

Educational Testing Service (2017). *About the TextEvaluator® Technology*. Princeton, NJ: Author.

Emery, H. J. (2014). Developments in LSP testing 30 years on? The case of aviation English. *Language Assessment Quarterly, 11*, 198–215. https://doi.org/10.1080/15434303.2014.894516

Engestrom, Y., Miettinen, R., & Punamaki, R. (1999). *Perspectives on activity theory*. Cambridge, UK: Cambridge University Press.

Eskey, D. E. (1997). Syllabus design in content-based instruction. In M.A. Snow & D. M. Brinton (Eds.), *The content-based classroom: Perspectives on integrating language and content* (pp. 132–141). White Plains, NY: Longman.

Farhady, H. (1983). On the plausibility of the unitary language proficiency factor. In J. Oller (Ed.), *Issues in language testing research* (pp. 11–28). Rowley, MA: Newbury House.

Ferris, D. R. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly, 28*, 414-420. doi:10.2307/3587446

Finn, B. (2017). A framework of episodic updating: An account of memory updating after retrieval. In B. H. Ross (Ed.), *Psychology of learning and motivation (Vol. 67)* (pp. 173–211). Academic Press. https://doi.org/10.1016/ bs.plm.2017.03.006.

Flowerdew, L. (2013). English for academic purposes. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Blackwell Publishing Ltd. doi:10.1002/9781405198431.wbeal0376

Flowerdew, J., & Peacock, M. (2001). Issues in EAP: A preliminary perspective. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 8–24). Cambridge, UK: Cambridge University Press.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*, 193–202. http://dx.doi.org/ 10.1037/0003-066X.39.3.193

Genesee, F., & Lindholm-Leary, K. (2013). Two case studies of content-based language education. *Journal of Immersion and Content-Based Language Education, 1*(1), 3–33. https://doi.org/10.1075/jicb.1.1.02gen

Goh, C. C., & Aryadoust, V. (2014). Examining the notion of listening sub-skill divisibility and its implication for second language listening. *International Journal of Listening*, *29*, 109–133. doi:10.1080/10904018.2014.936119

González-Brenes, J. P., Behrens, J. T., Mislevy, R. J., Levy, R., & Dicerbo, K. E. (2016). Bayesian Networks. In A. A. Rupp, & J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 328-353). West Sussex, UK: John Wiley & Sons, Ltd.

Gordon Commission on the Future of Assessment in Education. (2013). *A public policy statement*. Princeton, NJ: Author. Retrieved from http://www.gordoncommission.org/rsc/pdfs/gordon_commission_public_policy_report.pdf/

Grabe, W., & Jiang, X. (2014). Assessing reading. In A. J. Kunnan (Ed.), *The companion to language assessment*. John Wiley & Sons, Inc. doi:10.1002/9781118411360.wbcla060

Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: An applied linguistic perspective*. New York: Longman.

Grabowski, K. (2009). *Investigating the construct validity of a test designed to measure grammatical and pragmatic knowledge in the context of speaking*. (Unpublished doctoral dissertation). Teachers College, Columbia University, New York.

Graesser, C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004) Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*. *36*, 193–202. https://doi.org/10.3758/ BF03195564

Granger, C. (2008). Rasch analysis is important to understand and use for measurement. *Rasch Measurement Transactions*, *21*, 1122–1123.

Gustilo, L., & Magno, C. (2015). Explaining L2 writing performance through a chain of predictors: A SEM approach. *The Southeast Asian Journal of English Language Studies, 21*, 115–130. doi:10.17576/3L-2015-2102-09

Hailikari, T., Nevgi, A., & Lindblom-Ylänne, S. (2007). Exploring alternative ways of assessing prior knowledge, its components, and their relation to student achievement: A mathematics based case study. *Studies in Educational Evaluation*, *33*, 320–337. doi:10.1016/j.stueduc.2007.07.007

Hailikari, T., & Nevgi, A. (2010). How to diagnose at-risk students in chemistry: The case of prior knowledge assessment: International Journal of Science Education, 32, 2079–2095. doi:10.1080/09500690903369654

Hake, R. (1998). Interactive-Engagement Versus Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses. *American Journal of Physics*, *66*(1), 64–74. http://dx.doi.org/ 10.1119/1.18809

He, A. W., & Young, R. (1998). Language proficiency interviews: A discourse approach. In R. Young & A. W. He (Eds.), *Talking and testing* (pp. 1–24). Philadelphia, PA: John Benjamins.

He, L., & Shi, L. (2012). Topical knowledge and ESL writing. *Language Testing*, *29*, 443–464. doi: 10.1177/0265532212436659

Herrington, J., & Oliver, R. (2000). An instructional design framework for authentic learning environments. *Educational Technology Research and Development, 48*(3), 23–48. https://doi.org/10.1007/BF02319856

Hill, Y. Z., & Liu, O. L. (2012). *Is there any interaction between background knowledge and language proficiency that affects TOEFL iBT® reading performance?* (Research Report No. RR-12-22)Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2012.tb02304.x

Hoekje, B. (2016). “Language,” “communication,” and the longing for the authentic in LSP testing. *Language Testing*, *33*, 289–299. doi:10.1177/ 0265532215607921/

Hönig, I. (2009). *Assessment in CLIL: A case study*. (Unpublished masters thesis). University of Vienna, Vienna, Austria.

Hothorn, T., Hornik, K., van de Wiel, M. A., & Zeileis, A. (2008). Implementing a class of permutation tests: The coin package. *Journal of Statistical Software 28*(8), 1–23. http://www.jstatsoft.org/v28/i08/.

Hoz, R., Bowman, D., & Kozminsky, E. (2001). The differential effects of prior knowledge on learning: A study of two consecutive courses in earth sciences. Instructional Science, *29*, 187–211. https://doi.org/10.1023/ A:1017528513130

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55. https://doi.org/10.1080/ 10705519909540118

Huang, H.-T. D. (2010). *Modeling the relationships among topical knowledge, anxiety, and integrated speaking test performance: A structural equation modeling approach.* (Unpublished doctoral dissertation). The University of Texas at Austin, Austin, TX.

Huang, H.-T. D., Hung, S.-T. A., & Plakans, L. (2018). Topical knowledge in L2 speaking assessment: Comparing independent and integrated speaking test tasks. *Language Testing, 35*(1), 27–49. https://doi.org/10.1177/ 0265532216677106

Hyland, K. (2002). Specificity revisited: How far should we go now? *English for Specific Purposes, 21*, 385–395. https://doi.org/10.1016/ S0889-4906 (01)00028-X

IELTS (2018). *Common European Framework for IELTS.* Retrieved from https://www.ielts.org/en-us/ielts-for-organisations/ common-european-framework

Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes, 19*, 213–241. https://doi.org/10.1016/S0889-4906(97)00053-7

Jang, E. E., Wagner, M., & Dunlop, M. (2016, June). *Construct validation of multimodal scenario-based language assessment (SBLA) tasks for diagnostic placement purposes.* Plenary presented at the 38th LTRC conference, Palermo, Italy.

Jensen, C., & Hansen, C. (1995). The effect of prior knowledge on EAP listening-test performance. *Language Testing, 12*, 99–119. https://doi.org/10.1177/ 026553229501200106

Jones, R. (1985). Second language performance testing: An overview. In P. C. Haupman, R. LeBlanc, & M. B. Wesche (Eds.), *Second language performance testing* (pp. 104–115). Ottawa, Canada: University of Ottawa Press.

Karpicke, J. D., (2016). A powerful way to improve learning and memory: Practicing retrieval enhances long-term, meaningful learning. *American Psychological Association Psychological Science Agenda.* Retrieved from http://www.apa.org/science/about/psa/2016/06/learning-memory.aspx.

Kenny, D. A. (2015). *Measuring model fit.* Retrieved from http://davidakenny.net/cm/ fit.htm

Khabbazbashi, N. (2017). Topic and background knowledge effects on performance in speaking assessment. *Language Testing, 34*, 23–48. https://doi.org/ 10.1177/0265532215595666

Kim, A.-Y. A. (2009). Investigating second language reading components: Reading for different types of meaning. Teachers College, Columbia University *Working Papers in TESOL & Applied Linguistics, 9*(2), 1–28. doi:10.7916/ D85M6J9Q

Kim, A. Y. (2011). *Examining second language reading components in relation to reading test performance for diagnostic purposes: A fusion model approach.* (Unpublished doctoral dissertation). Teachers College, Columbia University, New York.

Kim, Y. J., Almond, R. G., & Shute, V. J. (2016). Applying Evidence-Centered Design for the development of game-based assessments in Physics Playground. International Journal of Testing, *16*, 142–163. http://dx.doi.org/10.1080/ 15305058.2015.1108322

Kline, R. B. (2016). *Methodology in the social sciences. Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford Press.

Knoch, U., & McNamara, T. (2015). Rasch analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 275–304). New York, NY: Routledge.

Koester, A. (2013). English for occupational purposes. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Blackwell Publishing Ltd. https://doi.org/10.1002/9781405198431.wbeal0381

Korb, K. B., & Nicholson, A. E. (2011). *Bayesian artificial intelligence* (2nd ed.). Boca Raton, FL: CRC Press.

Krekeler, C. (2006). Language for special academic purposes (LSAP) testing: The effect of background knowledge revisited. *Language Testing*, *23*(1), 99–130. https://doi.org/10.1191/0265532206lt323oa

Lado, R. (1961). *Language testing*. New York, NY: McGraw-Hill.

Lambert, W. E., & Tucker, G. R. (1972). *The bilingual education of children: The St. Lambert Experiment.* Rowley, MA: Newbury House.

The Language Training and Testing Center (2016). *GEPT-CEFR alignment.* Retrieved from https://www.lttc.ntu.edu.tw/e_lttc/E_GEPT/alignment.htm

Lantolf, J. P., & Poehner, M. E. (2004). Dynamic assessment: Bridging the past into the future. *Journal of Applied Linguistics*, *1*, 49–74.

Lantolf, J. P., & Poehner, M. E. (2011). Dynamic assessment in the classroom: Vygotskyan praxis for second language development. *Language Teaching Research*, *15*(1), 11–33. https://doi.org/10.1177/1362168810383328

Lathrop, R. H. (2012). *Bayesian networks*. Retrieved from https://www.ics.uci.edu/~rickl/courses/cs-171/2012-wq-cs171/2012-wq-cs171-lecture-slides/2012wq171-17-BayesianNetworks.pdf

Lee, H. K., & Anderson, C. (2007). Validity and topic generality of a writing performance test. *Language Testing*, *24*, 307–330. https://doi.org/10.1177/ 0265532207077200

Lee, H., & Pang, N. (2018). Understanding the effects of task and topical knowledge in the evaluation of websites as information path. *Journal of Documentation*, *74*(1), 162–186. https://doi.org/10.1108/JD-04-2017-0050

Lee, J.-W., & Schallert, D. L. (1997). The relative contribution of L2 language proficiency and L1 reading ability to L2 reading performance: A test of the threshold hypothesis in an EFL context. *TESOL Quarterly*, *31*, 713–739. doi:10.2307/3587757

Linacre, J. M. (2018) *Facets computer program for many-facet Rasch measurement*, (version 3.80.4). Beaverton, Oregon: Winsteps.com

Lipson, M. (1982). Learning information from text: The role of prior knowledge and reading ability. *Journal of Reading Behavior, 14*, 243–261. doi:10.1080/10862968209547453

Liu, O. L., Schedl, M., Malloy, J., & Kong, N. (2009). *Does content knowledge affect TOEFL iBT reading performance? A confirmatory approach to differential item functioning* (Research Report No. RR-09-29). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02186.x

Llosa, L. (2017) Assessing Students' Content Knowledge and Language Proficiency. In: E. Shohamy, I. G. Or., & S. May (Eds.), *Encyclopedia of language and education (Vol. 7): Language testing and assessment* (pp. 3–14). New York, NY: Springer Cham. https://doi.org/10.1007/978-3-319-02261-1_33

Long, D. R. (1990). What you don't know can't help you: An exploratory study of background knowledge and second language listening comprehension. *Studies in Second Language Acquisition, 12*, 65–80. https://doi.org/10.1017/ S0272263100008743

Lüdecke D (2018). *sjstats: Statistical functions for regression models* (R package version 0.17.2). Zenodo. http://doi.org/10.5281/zenodo.1489175

Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience, and topic on task performance in tape-mediated assessment of speaking. *Language Testing, 22*, 415–437. https://doi.org/10.1191/0265532205lt303oa

Lund, A., & Lund, M. (2018). *One-way ANOVA*. Retrieved from https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide.php

Lyster, R. (2007). *Learning and teaching languages through content: A counterbalanced approach*. Amsterdam, the Netherlands: John Benjamins Publishing.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130-149. http://dx.doi.org/10.1037/1082-989X.1.2.130

MacIntyre, P. D., & Gardner, R. C. (1991). Language anxiety: Its relationship to other anxieties and to processing in native and second languages. *Language Learning, 41*, 513–534. https://doi.org/10.1111/j.1467-1770.1991.tb00691.x

Mangiafico, S. S. (2016). *Summary and analysis of extension program evaluation in R* (version 1.15.0). Retrieved from rcompanion.org/documents/ RHandbookProgramEvaluation.pdf.

Mangiafico, S. S. (2018). *rcompanion: Functions to support extension education program evaluation* (R package version 2.0.0). https://CRAN.R-project.org/ package=rcompanion

Marais, I. (2009). Response dependence and the measurement of change. *Journal of Applied Measurement*, *10*, 17–29.

Markham, P., & Latham, M. (1987). The influence of religion-specific background knowledge on the listening comprehension of adults second-language students. *Language Learning*, *37*, 157–170. doi:10.1111/j.1467-1770.1987.tb00563.x

Marsh, D., & Frigols Martín, M. J. (2013). Content and language integrated learning. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics.* Blackwell Publishing Ltd. https://doi.org/10.1002/9781405198431.wbeal0190

Marx, J. D., & Cummings, K. (2007). Normalized change. *American Association of Physics Teachers*, *75*(1), 87-91. doi: 10.1119/1.2372468

Marzano, R. J. (2004). *Building background knowledge for academic achievement: Research on what works in school*s. Alexandria, VA: The Association for Supervision and Curriculum Development (ASCD).

Mayer, R. E. (2002). Multimedia learning. *Psychology of Learning and Motivation*, *41*, 85–139. doi:10.1016/s0079-7421(02)80005-6

McCarthy, K. S., Geurrero, T. A., Kent, K. M., Allen, L. K., McNamara, D. S., Chao, S.-F., . . . Sabatini, J. (2018). Comprehension in a scenario-based assessment: Domain and topic-specific background knowledge. *Discourse Processes*, *55*, 510-524. doi:10.1080/0163853X.2018.1460159

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. New York, NY: Cambridge University Press.

McNamara, T. (1996). *Measuring second language performance*. London, UK: Longman.

Messick, S. J. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13–23. https://doi.org/10.3102/0013189X023002013

Met, M. (1991). Learning language through content: Learning content through language. *Foreign Language Annals*, *24*, 281–295. doi:10.1111/ j.1944-9720.1991.tb00472.x

Mislevy, R. J. (1995). Test theory and language-learning assessment. *Language Testing*, *12*, 341–369. https://doi.org/10.1177/026553229501200305

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (Research Report No. RR-03-16). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ j.2333-8504.2003.tb01908.x

Mislevy, R. J., Haertel, G., Riconscente, M., Rutstein, D. W., & Ziker, C. (2017). *Assessing model-based reasoning using evidence-centered design: A suite of research-based design patterns*. SpringerBriefs in Statistics. doi:10.1007/978-3-319-52246-3_1

Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A., Hao, J., Corrigan, S., . . . John, M. (2014). *Psychometric considerations in game-based assessment*. Retrieved from http://www.instituteofplay.org/ wp-content/ uploads/2014/02/GlassLab_GBA1_WhitePaperFull.pdf (PDF)

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). *Evidence-centered assessment design*. Princeton, NJ: Educational Testing Service.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62. http://dx.doi.org/10.1207/S15366359MEA0101_02

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (2002). Making sense of data from complex assessments. *Applied Measurement in Education, 15*, 363–389. https://doi.org/10.1207/S15324818AME1504_03

Mitchell, R. (1992). *Testing for learning*. New York, NY: Free Press, Macmillan.

Morgan, C. (2007). Appropriate language assessment in content and language integrated learning. *Language Learning Journal, 33,* 59–67. doi:10.1080 /09571730685200121

National Academies of Sciences, Engineering, and Medicine. (2018). *How People Learn II: Learners, Contexts, and Cultures.* Washington, DC: The National Academies Press. https://doi.org/10.17226/24783.

National Center for Education Statistics. (2017). *The condition of education 2016* (NCES 2017-144). Washington, DC: U.S. Department of Education.

National Council of Teachers of English. (2013). The *NCTE definition of 21st century literacies*. Retrieved from http://www.ncte.org/positions/statements/ 21stcentdefinition

Neuman, S. B., Kaefer, T., & Pinkham, A. (2014). Building background knowledge. *The Reading Teacher, 68*, 145–148. doi: 10.1002/trtr.1314

The New Media Consortium. (2005). *A global imperative: The report of the 21st century literacy summit*. Austin, TX: Author.

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessment*. Honolulu, HI: University of Hawai'i Press.

Norsys, Inc. (2018). *Netica* [Computer software]. Vancouver, Canada: Author. Available from http://www.norsys.com/download.html

O’Brien, H. L., & Toms, E. G. (2010). The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology, 61*, 50–69. https://doi.org/10.1002/asi.21229

O’Reilly, T., Deane, P. & Sabatini, J. (2015). *Building and sharing knowledge key practice: What do you know, what don't you know, what did you learn?* (Research Report No. RR-15-24). Princeton, NJ: Educational Testing Service. doi:10.1002/ets2.12074

O’Reilly, T., & Sabatini, J. (2013). *Reading for understanding: How performance moderators and scenarios impact assessment design* (Research Report No. RR-13-31). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02338.x

O’Reilly, T., Weeks, J., Sabatini, J., Halderman, L., & Steinberg, J. (2014). Designing reading comprehension assessments for reading interventions: How a theoretically motivated assessment can serve as an outcome measure. *Educational Psychology Review, 26*, 403–424. https://doi.org/10.1007/ s10648-014-9269-z

Oh, S. (2018). *Investigating test-takers’ use of linguistic tools in second language academic writing assessment* (Unpublished doctoral dissertation). Teachers College, Columbia University, New York.

Oller, J. (1979). *Language tests at schools: A pragmatic approach.* London, UK: Longman.

Ortega, L. (2003). Syntactic Complexity Measures and their Relationship to L2 Proficiency: A Research Synthesis of College-level L2 Writing. *Applied Linguistics, 24*, 492–518. doi:10.1093/applin/24.4.492

Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). *The association between TOEFL iBT® test scores and the Common European Framework of Reference (CEFR) levels* (Research Report No. RM-15-06). Princeton, NJ: Educational Testing Service.

Papajohn, D. (1999). The effect of topic variation in performance testing: The case of the chemistry TEACH test for international teaching assistants. *Language Testing, 16*(1), 52–81. https://doi.org/10.1177/026553229901600104

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA: Morgan Kaufmann.

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect.* New York, NY: Basic Books.

Pearl, J., & Russell, S. (2003). Bayesian networks. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (2nd ed) (pp. 157–160). Boston, MA: The MIT Press.

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

Peterson, B. G., & Carl, P. (2018). PerformanceAnalytics: Econometric tools for performance and risk analysis (R package version 1.5.2). https://CRAN.R-project.org/ package=PerformanceAnalytics

Petty, G. (2009). *Evidence-based teaching: A practical approach*. Cheltenham, UK: Nelson Thornes.

Prawat, R. S., (1989). Promoting access to knowledge, strategy, and disposition in students: A research synthesis. *Review of Educational Research, 59*, 1–41. https://doi.org/10.3102/00346543059001001

Purpura, J. E. (2004). *Assessing grammar*. Cambridge, UK: Cambridge University Press.

Purpura, J. E. (2008). Assessing communicative language ability: Models and their components. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education (Vol. 7): Language testing and assessment* (pp. 53–68). New York, NY: Springer.

Purpura, J. E. (2010, March). *The design and development of the web-based Online Oxford Placement Exam*. The TESOL-ILTA joint session presented at the 2010 annual TESOL conference, Boston, MA. Retrieved from https://www.slideshare.net/JEPurpura/tesol-2010-boston

Purpura, J. E. (2013). Assessment of grammar. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Blackwell Publishing Ltd. https://doi.org/10.1002/9781405198431.wbeal0045

Purpura, J. E. (2014). Assessing grammar. In A. J. Kunnan (Ed.), *The companion to language assessment*. John Wiley & Sons, Inc. doi:10.1002/ 9781118411360.wbcla147

Purpura, J. E. (2015). B*roadening the construct of second and foreign language proficiency through scenario-based language assessment.* Paper presented at the First TC/ETS Forum on Teaching, Learning, and Assessment of English Language Learners. New York, NY.

Purpura, J. E. (2016). Second and foreign language assessment. *The Modern Language Journal, 100* (S1), 190–208. doi:10.1111/modl.12308

Purpura, J. E. (2017). Assessing meaning. In E. Shohamy, I. G. Or., & S. May (Eds.), *Encyclopedia of language and education (Vol. 7): Language testing and assessment* (pp. 33–62). New York, NY: Springer Cham. https://doi.org/ 10.1007/978-3-319-02261-1_1

Purpura, J. E. (2018, March). *An introduction to scenario-based assessment: Opportunities for construct expansion*. Paper presented at 2018 AAAL annual conference colloquium “The affordances of scenario-based assessment for broadening measurement opportunities in second or foreign language assessment", Chicago, IL

Purpura, J. E., & Banerjee, H. L. (2018, March). *The affordances of scenario-based assessment for broadening measurement opportunities in second or foreign language assessment.* Colloquium presented at the 2018 annual AAAL conference, Chicago, IL.

Purpura, J. E., & Turner, C. E. (2014, October). *A learning-oriented assessment approach to understanding the complexities of classroom-based language assessment*. Paper presented at the Roundtable on Learning-Oriented Assessment in Language Classrooms and Large-Scale Contexts, Teachers College, Columbia University, New York.

Purpura, J. E., & Turner, C. E. (forthcoming). *Learning-oriented assessment in language classrooms: Using assessment to gauge and promote language learning*. New York, NY: Routledge.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Nielsen & Lydiche.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. In M. Blegvad (Ed.), *The Danish yearbook of philosophy* (pp. 58–94). Copenhagen, Denmark: Munksgaard.

R Core Team (2018). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. https://www.R-project.org/

Revelle, W. (2018) *psych: Procedures for personality and psychological research* (R package version 1.8.4). Northwestern University, Evanston, IL. https://CRAN.R-project.org/package=psych

Rivers, W. M. (1968). *Teaching foreign language skills*. Chicago, IL: University of Chicago Press.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. http://www.jstatsoft.org/v48/i02/

Rosson, M. B., & Carroll, J. M. (2002). Scenario-based design. In J. Jacko, & A. Sears (Eds.), *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications* (pp. 1032–1050). Mahwah, NJ: Lawrence Erlbaum Associates.

Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *The Journal of Technology, Learning, and Assessment, 8*(4), Retrieved from http://www.jtla.org.

Sabatini, J., & O’Reilly, T. (2013). Rationale for a new generation of reading comprehension assessments. In Miller, B., Cutting, L., & P. McCardle (Eds.), *Unraveling reading comprehension: Behavioral, neurobiological, and genetic components* (pp. 100–111). Baltimore, MD: Brookes Publishing.

Sabatini, J., O’Reilly, T., Halderman, L., & Bruce, K. (2014). Integrating scenario-based and component reading skill measures to understand the reading behavior of struggling readers. *Learning Disabilities Research & Practice*, *29*(1), 36–43. doi:10.1111/ldrp.12028

Sasaki, M. (1996). *Second language proficiency, foreign language aptitude, and intelligence.* New York, NY: Peter Lang.

Sawilowsky, S (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods. 8*, 467–474. doi:10.22237/jmasm/1257035100

Schmidt-Rinehart, B. C. (1994). The effects of topic familiarity on second language listening comprehension. *The Modern Language Journal, 78*, 179–189. doi:10.2307/329008

Schumacker, R. E., & Lomax, R. G. (2016). *A beginner’s guide to structural equation modeling* (4th ed.). New York, NY: Routledge.

Seong, Y. (2018, March). *Examining the cognitive dimension of academic speaking ability through a scenario-based speaking test.* Paper presented at 2018 AAAL annual conference colloquium "The affordances of scenario-based assessment for broadening measurement opportunities in second or foreign language assessment", Chicago, IL

Shapiro, A. M. (2004). How including prior knowledge as a subject variable may change outcomes of learning research. *American Educational Research Journal, 41*(1), 159–189. doi:10.3102/00028312041001159

Sharma, S. (1996). *Applied multivariate techniques.* New York, NY: John Wiley & Sons.

Sheehan, K. M. (2015). *Aligning TextEvaluator scores with the accelerated text complexity guidelines specified in the Common Core State Standards*. (ETS Research Report No. RR-15-21). Princeton, NJ: Educational Testing Service.

Sheehan, K., & O’Reilly, T. (2012). The case for scenario-based assessments of reading competency. In J. Sabatini, T.O’Reilly, & L. Albro (Eds.), *Reaching an understanding: Innovations in how we view reading assessment* (pp. 19–33). Lanham, MD: Rowman & Littlefield.

Shore, J. R., Wolf, M. K., & Blood, I. (2013). *English learner formative assessment (ELFA): ELFA teacher's guide.* Princeton, NJ: Educational Testing Service.

Shore, J. R., Wolf, M. K., O'Reilly, T., & Sabatini, J. P. (2017). Measuring 21st-centruy reading comprehension through scenario-based assessment. In M. K. Wolf, & Butler, Y. G. (Eds.), *English language proficiency assessment for young learners* (pp. 234–252) New York, NY: Routledge.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503–523). Charlotte, NC: Information Age Publishing.

Shute, V. J., Masduki, I., Donmez, O., Dennen, V. P., Kim, Y.-J., Jeong, A. C., & Wang, C.-Y. (2010). Modeling, assessing, and supporting key competencies within game environments. In D. Ifenthaler, P. Pirnay-Dummer, & N. M. Seel (Eds.), *Computer-based diagnostics and systematic analysis of knowledge* (pp. 281–310). New York. NY: Springer Science+Business Media, LLC.

Shute, V. J. & Sun, C. (in press). Games for assessment. In J. Plass, B. Homer, & R. Mayer (Eds.), *Handbook of game-based learning*. Cambridge, MA: MIT Press.

Shute, V. J. & Ventura, M. (2013). *Measuring and supporting learning in games: Stealth assessmen*t. Cambridge, MA: The MIT Press.

Shute, V. J., Ventura, M., & Ke, F. (2015). The power of play: The effects of Portal 2 and Lumosity on cognitive and noncognitive skills. *Computers & Education, 80*, 58-67. https://doi.org/10.1016/j.compedu.2014.08.013

Shute, V. J., & Wang, L. (2017). Assessing and supporting hard-to-measure constructs in video games. In A. A. Rupp., & J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 535–562). West Sussex, UK: John Wiley & Sons.

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior, 63*, 106–117. http://dx.doi.org/10.1016/j.chb.2016.05.047

Sick, J. (2008). Rasch measurement in language education: Part 1. S*hiken: JALT Testing & Evaluation SIG Newsletter, 12*(1), 1–6. Retrieved from http://jalt.org/test/PDF/Sick1.pdf

Signorell, A., Aho, K., Alfons, A., Anderegg, N., Aragon, T., Arppe, A., . . . Zeileis, A. (2018). *DescTools: Tools for descriptive statistics* (R package version 0.99.25). https://CRAN.R-project.org/package=DescTools

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford, UK: Oxford University Press.

Smith, J. (1989). Topic and variation in ITA oral proficiency: SPEAK and field-specific tests. *English for Specific Purposes, 8*, 155–167. https://doi.org/10.1016/0889-4906(89)90027-6

Snow, A., & Katz, A. (2014). Assessing language and content. In A. J. Kunnan (Ed.), *The companion to language assessment*. John Wiley & Sons. doi:10.1002/ 9781118411360.wbcla122

Song, Y., Deane, P., Graf, E. A., & van Rijn, P. (2013). *Using argumentation learning progressions to support teaching and assessments of English language arts.* (R&D Connections No. 22). Princeton, NJ: Educational Testing Service.

Spada, N. (2016). Focusing on language in meaning-based and content-based instruction. *JACET International Convention Selected Papers, 4*, 3–30.

Strevens, P. (1977). Special-purpose language learning. *Language teaching and Linguistics Abstracts, 10*, 145–163. https://doi.org/10.1017/ S0261444800003402

Sukin, T. (2010). Analysis of gain scores. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 519-523). Thousand Oaks, CA: Sage Publications.

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the p value is not enough. *Journal of Graduate Medical Education, 4*, 279–282. doi:10.4300/ JGME-D-12-00156.1

Sutton, R. (1995). *Assessment for learning*. Salford, UK: RS Publications.

Tan, M. (2011). Mathematics and science teachers' beliefs and practices regarding the teaching of language in content learning. *Language Teaching Research, 15*, 325–342. https://doi.org/10.1177/1362168811401153

Tedick, D. J. (1990). ESL writing assessment: Subject-matter knowledge and its impact on performance. *English for Specific Purposes, 9*, 123–143. doi:10.1016/0889-4906(90)90003-U

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381–403). New York: Academic Press.

Turner, C. E., & Purpura, J. E. (2016). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari, & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 255–273). Berlin, Germany: De Gruyter.

Upton, T. A., & Connor, U. (2013). Language for specific purposes: Overview. In C. A.Chapelle (Ed.), *The encyclopedia of applied linguistics*. Blackwell Publishing. doi: 10.1002/9781405198431.wbeal0891

Usó-Juan, E. (2006). The compensatory nature of discipline-related knowledge and English language proficiency in reading English for academic purposes. *The Modern Language Journal*, *90*, 210–227. https://www.jstor.org/stable/ 3876871

Valencia, S. W., Stallman, A. C., Commeyras, M., Pearson, P. D., & Hartman, D. (1991). Four measures of topical knowledge: A study of construct validity. *Reading Research Quarterly*, *26*(3), 204–233. doi:10.2307/747761

Vandergrift, L. (2006). Second language listening: Listening ability or language proficiency? *The Modern Language Journal*, *91*(1), 6–18. https://www.jstor.org/stable/3588811

van Loon, M., de Bruin, A. B. H., van Gog , T., & van Merriënboer, J. J. G. (2013). Activation of inaccurate prior knowledge affects primary-school students' metacognitive judgments and calibration. *Learning and Instruction, 24*, 15–25. https://doi.org/10.1016/j.learninstruc.2012.08.005

Vollmer, H. & Sang, F. (1983). Competing hypotheses about second language ability: A plea for caution. In J. Oller (Ed.), *Issues in language testing research* (pp. 29–79). Rowley, MA: Newbury House.

Wagner, M. J. (2006). *Utilizing the visual channel: An investigation of the use of video texts on tests of second language listening ability*. (Unpublished doctoral dissertation). Teachers College, Columbia University, New York.

Wang, L., Shute, V. J., & Moore, G. R. (2015). Lesson learned and best practices of stealth assessment. *International Journal of Gaming and Computer-Mediated Simulations, 7*(4), 66–87. doi:10.4018/IJGCMS.2015100104

WIDA Consortium. (2012). *Amplification of the English language development standards*. Madison, WI: Board of Regents of the University of Wisconsin System. Retrieved from http://www.wida.us/standards/eld.aspx

Wiersma, W., & Jurs, S. G. (2009). *Research methods in education: An introduction* (9th ed.). Boston, MA: Pearson.

Wiliam, B. (2011). What is assessment for learning? *Studies in Educational Evaluation, 37*, 3-14. doi:10.1016/j.stueduc.2011.03.001

Wolf, M. K., Shore, J. R., & Blood, I. (2014). *English learner formative assessment (ELFA): A design framework*. Princeton, NJ: Educational Testing Service.

Wolfe, E. W., & Chiu, C. W. (1999). Measuring change across multiple occasions using the Rasch rating scale model. *Journal of Outcome Measurement, 3,* 360-381.

Worthington, D. L., & Bodie, G. D. (2018). *The sourcebook of listening research: Methodology and measures* (1st ed.). Hoboken, NJ: John Wiley & Sons.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*, 370.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural research, 20*, 557–585.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics, 5*, 161–215.

Zareva, A. (2005). Models of lexical knowledge assessment of second language learners of English at higher levels of language proficiency. *System, 33*, 547–562. doi:10.1016/j.system.2005.03.005