Title of Project:
Investigating the Combined Effects of Rater Expertise, Working Memory Capacity, and Cognitive Functionality on the Scoring of Second Language Speaking Performance

Researcher:
Qie Han
Teachers College, Columbia University
qh2139@tc.columbia.edu

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


Qie Han

TIRF Research Topic Investigated:
Language Assessment

Final Report

Motivation for the Research

In second language (L2) speaking assessment, raters can significantly affect test validity due to rater variability, conceptualized as a source of construct-irrelevant variance in scores caused by individual differences in raters' characteristics rather than examinees' ability. To improve the validity of interpretations and decisions based on scores, we must investigate what rater characteristics are likely to contribute to rater variability. One of the most frequently examined rater characteristics is rater expertise. Rater expertise refers to raters' relevant knowledge and experience in L2 performance assessment, and it has been found to associate with different levels of scoring performance. Despite its effect on scoring performance, rater expertise is likely *not* the only rater characteristic that contributes to rater variability. The possible effects of two other rater characteristics (i.e., raters' working memory capacity [WMC] and cognitive functionality), also need to be investigated. The reason is that, on the one hand, WMC has been found to impact, either independently or in combination with expertise, a variety of complex cognitive performances, such as language comprehension and acquisition. On the other hand, the cognitive functionality of raters, defined in terms of the strategies related to information processing, underlies the different stages of raters' scoring process and the mental processes associated with those stages. Through an exploration of the combined effects of these rater characteristics on scoring performance, we can strengthen our assumptions about what rater-associated factors lead to rater variability, thereby shedding light on rater selection, training, and scoring practices. With improved scoring performance, test scores can capture a more accurate portrayal of L2 learners' speaking ability.

Research Questions

Recently, an increasing amount of research has been conducted to investigate the effect of rater expertise on scoring performance in L2 performance assessment. However, few studies have

investigated the joint influence of rater expertise and raters' cognitive characteristics on scoring performance. The current study was thus designed to address this gap by exploring the following three main research questions:

1. What were the joint influences of rater expertise and WMC on raters' scoring performance?
    - 1-1. What were the relative contributions of rater expertise and WMC to scoring performance?
    - 1-2. What possible interaction was there between rater expertise and WMC in their joint influence on scoring performance?
2. What strategies did raters use during their scoring process?
3. How did the expert and novice raters differ in the patterns of strategy use?

Research Methodology

To address these questions, the current study employed a mixed-methods research design to examine the combined effects of three rater characteristics, i.e., rater expertise, WMC, and cognitive functionality, on the scoring of L2 speaking performance. Ninety raters were recruited on a voluntary basis from major universities in the United States and one large testing organization in the United Kingdom to participate in a multi-stage research project, where the raters filled out a rater background survey designed to measure their rater expertise, scored 27 test takers' speech samples from the Aptis speaking test, re-scored 10 of those samples at least a week later, and completed a cognitive task (i.e., a listening span) that measured their verbal WMC. The quantitative data collected from these sessions included the 90 raters' holistic ratings, which were analyzed using many-facet Rasch measurement analysis to calculate three scoring performance indices that respectively reflected the raters' scoring accuracy, severity, and consistency. In addition, the raters' responses to the rater background survey were used to calculate a composite rater-expertise score, and the raters' scores for the listening span were used as measures for their verbal WMC. Then, a hierarchical regression analysis was performed to explore the joint influences of rater expertise and WMC on the three aforementioned aspects of scoring performance. Subsequently, six raters (three experts and three novices) were randomly selected from the 90 to participate in a cognitive lab session, where the raters verbally reported their thoughts and mental processes while scoring three speech samples. The raters' verbal reports were transcribed and analyzed in terms of the strategies that they had used while scoring and how the expert and novice raters differed in strategy use. The results of the quantitative and qualitative data analyses were integrated to reveal the combined effects of the three rater characteristics (i.e., rater expertise, WMC, and cognitive functionality) on scoring performance.

Summary of Findings

The results from the quantitative data analyses (mainly hierarchical regression analysis) have demonstrated that rater expertise significantly predicted scoring accuracy. This result seems to align with previous research indicating that more proficient raters are more accurate and appropriate in their application of the rating scale than less proficient raters (Davis, 2012; H. J. Kim, 2011), especially if the scale is holistic (Barkaoui, 2010). This finding also seems to corroborate previous research which has found significant contributions of expertise to complex cognitive performance in various domains (e.g., Alptekin & Erçetin, 2011; Hambrick & Engle, 2002; Joh & Plakans, 2017; Payne et al., 2009, to name a few), and may be explained by theories

about expertise and expert cognition, such as experts' deeper, better-organized and more functional knowledge representation (Chase & Simon, 1973a, 1973b; Chiesi et al., 1979; de Groot, 1978; Simon & Chase, 1973; Spilich et al., 1979) and superior skills of using pre-established knowledge structures during task performance (Endsley, 2018; Ericsson & Kintsch, 1995; Feltovich et al., 1984). Despite the significant effect of rater expertise on scoring accuracy, the effect size of rater expertise found was a bit small, probably due to the fact that rater expertise was measured in the current study as raters' *general*, assessment-related experiences rather than *test-specific* knowledge and skills.

What was a bit surprising in my findings was that WMC was not found to significantly predict scoring performance. This result may be explained by the fact that the current study was designed to reduce raters' cognitive load during the scoring process, which might have mitigated the contribution of WMC to scoring performance. Specifically, the current study attempted to make raters' scoring process as authentic as possible to real-life scoring practices. As a result, not only were the raters allowed to replay spoken responses during the scoring process, but were also given constant access to the scoring rubric. Replays during the scoring process could have reduced the essentialness of a larger WMC. In the same way, the raters' WMC could have contributed more to scoring performance if the raters were not given constant access to the scoring rubric, which could have forced them to rely more on WM to process and integrate the rubric criteria with their pre-established conceptualization of L2 speaking ability. As a result of these factors, the effect of WMC on scoring performance may have been moderated. As for the interaction between rater expertise and WMC, the non-significant interaction effects found in the current study seem to support the independent influences model, which hypothesizes no interaction between expertise and WMC in their joint influences on cognitive performance (Alptekin & Erçetin, 2011; Hambrick & Oswald, 2005; Payne et al., 2009).

The results from the qualitative analyses identified fourteen major (meta)cognitive strategies that the raters had reported using during the scoring process. Amongst these strategies, three were most frequently reported by the raters: applying criteria from the rubric, evaluating, and retrieving relevant response information from WM. These three strategies were the building blocks of a rater's scoring process and were thus most commonly used by the raters. Moreover, the raters were frequently found to have used these strategies as a strategy cluster, which seems to align with the information processing model and the different stages of scoring proposed by Purpura (2012) and Bejar (2012).

Based on the strategies identified from the raters' verbal reports, differences in the quantity and quality of strategy use between the expert and novice raters were explored. The results have revealed that the two groups of raters used similar ranges of strategies. However, the expert raters used eight strategies considerably more frequently than the novice raters, which include applying criteria from the rubric, comprehending, evaluating, linking to prior knowledge or experience, noticing, reasoning, reflecting (especially on one's own scoring performance), and retrieving relevant response information from WM. These differences, as discussed in detail in my dissertation, can mostly be attributed to the experts' higher level of expertise in L2 speaking assessment. Not only were differences found in the quantity of strategy use between the expert and novice raters, but the two groups of raters were also found to differ in the *quality* of strategy use. The results have shown that the expert raters had demonstrated more success and competence in using strategies and strategy clusters than the novice raters. These differences may be explained in light of the characteristics of expertise, such as experts' better recognition and recall of new materials involving domain-relevant information and better integration of new

information within a coherent and interconnected framework of existing knowledge and information to make useful inferences (Chiesi et al., 1979; Spilich et al., 1979). Relevant details can be found in the results and discussion sections of my dissertation.

Implications

The present study has a number of implications for the assessment of L2 speaking performance. First, the significant effect of rater expertise on scoring accuracy seems to support our assumption about the contribution of raters' accumulated experience in L2 assessment to scoring performance. However, the low effect size of rater expertise has raised questions about measuring rater expertise as raters' *general*, assessment-related experiences rather than test-specific knowledge and skills. This finding may warrant a reconsideration of our commonly-used criteria (e.g., *general* rating and teaching experiences) for rater selection and categorization in existing L2 research and rating practices, especially when it comes to predicting raters' scoring performance for a *specific* test. Second, although WMC is a basic ability required for performing a variety of complex cognitive tasks, the relative contribution of WMC to scoring performance, compared to that of rater expertise, was found much less appreciable in the current study. This finding seems to corroborate existing findings about the more predominant role of expertise in cognitive performance (e.g., Hambrick & Engle, 2002; Joh & Plakans, 2017; Payne et al., 2009). From another perspective, it also seems to support the effectiveness of our commonly-used measures (e.g., giving raters opportunities to re-listen to responses and constant access to the scoring rubric) to reduce raters' cognitive load in real-life rating practices. Lastly, differences in the expert and novice raters' strategy use, both in terms of the frequency of using certain strategies and the efficiency of using strategies in general, were discovered. These differences can not only help explain scoring performance on a deeper, cognitive-processing level (Purpura, 2014) but also indicate the potential of training novice raters on strategy use similar to an expert rater's during the scoring process to improve scoring performance.

References

Abernethy, B., Neal, R. J., & Koning, P. (1994). Visual–perceptual and cognitive differences between expert, intermediate, and novice snooker players. *Applied Cognitive Psychology*, *8*(3), 185–211. https://doi.org/10.1002/acp.2350080302

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General*, *117*(3), 288–318. https://doi.org/10.1037/0096-3445.117.3.288

Ackerman, P. L., Beier, M. E., & Boyle, M. D. (2002). Individual differences in working memory within a nomological network of cognitive and perceptual speed abilities. *Journal of Experimental Psychology: General*, *131*(4), 567–589. https://psycnet.apa.org/doi/10.1037/0096-3445.131.4.567

Adams, J. W., & Hitch, G. J. (1997). Working memory and children's mental addition. *Journal of Experimental Child Psychology*, *67*(1), 21–38. https://doi.org/10.1006/jecp.1997.2397

Alamargot, D., Plane, S., Lambert, E., & Chesnet, D. (2010). Using eye and pen movements to trace the development of writing expertise: case studies of a 7th, 9th and 12th grader, graduate student, and professional writer. *Reading and Writing*, *23*(7), 853–888. https://doi.org/10.1007/s11145-009-9191-9

Allard, F., & Starkes, J. L. (1991). Motor-skill experts in sports, dance, and other domains. In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 126–152). Cambridge, UK: Cambridge University Press. https://doi.org/10.7551/mitpress/3080.003.0034

Alloway, T. P., Gathercole, S. E., & Pickering, S. J. (2006). Verbal and visuospatial short-term and working memory in children: Are they separable?. *Child Development*, *77*(6), 1698–1716. https://doi.org/10.1111/j.1467-8624.2006.00968.x

Alptekin, C., & Erçetin, G. (2011). Effects of working memory capacity and content familiarity on literal and inferential comprehension in L2 reading. *TESOL Quarterly*, *45*(2), 235–266. https://doi.org/10.5054/tq.2011.247705

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*, *89*(4), 369– 406. https://doi.org/10.1037/0033-295x.89.4.369

Anderson, N. J. (1989). *Reading comprehension tests versus academic reading: What are second language readers doing?* [Unpublished doctoral dissertation]. The University of Texas, Austin.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. https://doi.org/10.1007/bf02293814

Andringa, S., Olsthoorn, N., van Beuningen, C., Schoonen, R., & Hulstijn, J. (2012). Determinants of success in native and non-native listening comprehension: An individual differences approach. *Language Learning*, *62*, 49–78. https://doi.org/10.1111/j.1467-9922.2012.00706.x

Ang-Aw, H. T., & Goh, C. C. M. (2011). Understanding discrepancies in rater judgement on national-level oral examination tasks. *RELC Journal*, *42*(1), 31–51. https://doi.org/10.1177/0033688210390226

Anzai, Y., & Simon, H. A. (1979). The theory of learning by doing. *Psychological Review*, *86*(2), 124–140. https://doi.org/10.1037/0033-295x.86.2.124

Azevedo, R., Faremo, S., & Lajoie, S. P. (2007). Expert-novice differences in mammogram interpretation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *29*(29), 65–70. https://escholarship.org/uc/item/9vs3q436

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

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, *12*(2), 238–257. https://doi.org/10.1177/026553229501200206

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

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

Baddeley, A. (1986). *Working memory*. Oxford, UK: Clarendon Press.

Baddeley, A. (2000). The episodic buffer: A new component of working memory?. *Trends in Cognitive Sciences*, *4*(11), 417–423. https://doi.org/10.1016/s1364-6613(00)01538-2

Baddeley, A. (2007). *Oxford psychology series: Vol. 45. Working memory, thought, and action.* Oxford, UK: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198528012.001.0001

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. A. Bower (Ed.), *The psychology of learning and motivation*, (pp. 47–89). New York, NY: Academic Press. https://doi.org/10.1016/s0079-7421(08)60452-1

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, *12*(2), 86–107. https://doi.org/10.1016/j.asw.2007.07.001

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*(1), 54–74. https://doi.org/10.1080/15434300903464418

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice, 18*(3), 279–293. https://doi.org/10.1080/0969594X.2010.526585

Batalova, J., & Fix, M. (2010). A profile of limited English proficient adult immigrants. *Peabody Journal of Education*, *85*(4), 511–534. https://doi.org/10.1080/0161956x.2010.518050

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice*, *31*(3), 2–9. https://doi.org/10.1111/j.1745-3992.2012.00238.x

Benton, S. L., Kraft, R. G., Glover, J. A., & Plake, B. S. (1984). Cognitive capacity differences among writers. *Journal of Educational Psychology*, *76*(5), 820–834. https://doi.org/10.1037/0022-0663.76.5.820

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

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*(1), 89–110. https://doi.org/10.1191/0265532203lt245oa

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *26*(2), 211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Brisbois, J. E. (1995). Connections between first-and second-language reading. *Journal of Reading Behavior*, *27*(4), 565–584. https://doi.org/10.1080/10862969509547899

British Council. (2016, April). *Aptis candidate guide*. https://www.academia.edu/31198744/Aptis_Candidate_Guide

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, *12*(1), 1–15. https://doi.org/10.1177/026553229501200101

Brown, A. (2000). An investigation of the rating process in the IELTS oral interview. In R. Tulloh (Ed.), *IELTS research reports* (Vol. 3, pp. 49–84). Canberra, Australia: IELTS Australia. https://www.ielts.org/-/media/research-reports/ielts_rr_volume03_report3.ashx

Brown, A. (2005). *Interviewer variability in oral proficiency interviews*. Frankfurt, Germany: Peter Lang.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test taker performance on English-for-academic-purposes speaking tasks*. (ETS Research Report No. RR-05-05). Princeton, NJ: ETS. http://dx.doi.org/10.1002/j.2333-8504.2005.tb01982.x

Bugg, J. M., Zook, N. A., DeLosh, E. L., Davalos, D. B., & Davis, H. P. (2006). Age differences in fluid intelligence: contributions of general slowing and frontal decline. *Brain and Cognition*, *62*(1), 9–16. https://doi.org/10.1016/j.bandc.2006.02.006

Butterworth, B. (2018). Mathematical expertise. In Ericsson, K. A., Hoffman, R. R., Kozbelt, A., & Williams, A. M. (Eds.), *The Cambridge handbook of expertise and expert performance* (2nd ed., pp. 616–633). New York, NY: Cambridge University Press. https://doi.org/10.1017/9781316480748.032

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming.* New York, NY: Routledge. https://doi-org.ezproxy.cul.columbia.edu/10.4324/9780203807644

Cai, R., Dong, Y., Zhao, N., & Lin, J. (2015). Factors contributing to individual differences in the development of consecutive interpreting competence for beginner student interpreters. *The Interpreter and Translator Trainer*, *9*(1), 104–120. https://doi.org/10.1080/1750399x.2015.1016279

Canale, M. (1983). From communicative competence to language pedagogy. In J. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 2–28). London, UK: Longman.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1–47. https://doi.org/10.1093/applin/1.1.1

Cañas, A. J., Coffey, J. W., Carnot, M. J., Feltovich, P., Hoffman, R. R., Feltovich, J., & Novak, J. D. (2003). *A summary of literature pertaining to the use of concept mapping techniques and technologies for education and performance support*. https://eventos.unipampa.edu.br/seminariodocente/files/2011/03/Oficina-9-A-Summary-of-Literature-Pertaining-to-the-Use-of-Concept.pdf

Cantor, J., & Engle, R. W. (1993). Working-memory capacity as long-term memory activation: An individual-differences approach. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(5), 1101–1114. https://doi.org/10.1037/0278-7393.19.5.1101

Carroll, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. In H. B. Allen & R. N. Campbell (Eds.), *Teaching English as a second language: A book of readings* (2nd ed.) (pp. 313–321). New York, NY: McGraw-Hill.

Carson, J., Carrell, P., Silberstein, S., Kroll, B., & Kuehn, P. (1990). Reading-writing relationships in first and second language. *TESOL Quarterly*, *24*(2), 245–266. https://doi.org/10.2307/3586901

Case, R., Kurland, D. M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology*, *33*(3), 386–404. https://doi.org/10.1016/0022-0965(82)90054-6

Cavalluzzo, L. C. (2004). *Is national board certification an effective signal of teacher quality?*. (Technical Report No. 11204). Alexandria, VA: CNA Corp.. https://files.eric.ed.gov/fulltext/ED485515.pdf

Charness, N. (1976). Memory for chess positions: Resistance to interference. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(6), 641–653. https://doi.org/10.1037/0278-7393.2.6.641

Chase, W. G., & Simon, H. A. (1973a). The mind's eye in chess. In W.G. Chase (Ed.), *Visual information processing* (pp. 215–281). New York, NY: Academic Press. https://doi.org/10.1016/b978-0-12-170150-5.50011-1

Chase, W. G., & Simon, H. A. (1973b). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81. https://doi.org/10.1016/0010-0285(73)90004-2

Chenoweth, N. A., & Hayes, J. R. (2003). The inner voice in writing. *Written Communication*, *20*(1), 99–118. https://doi.org/10.1177/0741088303253572

Chi, M. T. H. (2006). Laboratory methods for assessing experts' and novices' knowledge. In Ericsson, K. A., Charness, N., Feltovich, P. J., & Hoffman, R. R. (Eds.), *The Cambridge handbook of expertise and expert performance* (1st ed., pp. 167–184). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/cbo9780511816796.010

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*(2), 121–152. https://doi.org/10.1207/s15516709cog0502_2

Chi, M. T. H., Glaser, R., & Farr, M. (1988). *The nature of expertise*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Chiesi, H. L., Spilich, G. J., & Voss, J. F. (1979). Acquisition of domain-related information in relation to high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior*, *18*(3), 257–274. https://doi.org/10.1016/s0022-5371(79)90146-4

Chiswick, B. R., & Miller, P. W. (1995). The endogeneity between language and earnings: International analyses. *Journal of Labor Economics*, *13*(2), 246–288. https://doi.org/10.1086/298374

Clark, J. L. D. (1975). Theoretical and technical considerations in oral proficiency testing. In R. L. Jones & B. Spolsky (Eds.), *Testing language proficiency* (pp. 10–28). Arlington, VA: Center for Applied Linguistics.

Cohen, A. D. (2011). *Strategies in learning and using a second language* (2nd ed.). New York, NY: Pearson. https://doi.org/10.4324/9781315833200

Cohen, A. D., & Aphek, E. (1981). Easifying second language learning. *Studies in Second Language Acquisition*, *3*(2), 221–236. https://doi.org/10.1017/s0272263100004198

Cohen, A. D., & Upton, T. (2007). I want to go back to the text: Response strategies on the reading subtest of the new TOEFL. *Language Testing*, *24*(2): 209–250. https://doi.org/10.1177/0265532207076364

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, P., Cohen, J., West, S. G., & Aiken, L. S. (2002). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Conway, A. R. A., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, *30*(2), 163–183. https://doi.org/10.1016/s0160-2896(01)00096-4

Conway, A. R.A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin and Review*, *12*(5), 769–786. https://doi.org/10.3758/bf03196772

Council of Europe. (2001). *Common European framework of reference for languages: learning, teaching, assessment*. https://rm.coe.int/1680459f97

Cowan, N. (1993). Activation, attention, and short-term memory. *Memory & Cognition*, *21*(2), 162–167. https://doi.org/10.3758/bf03202728

Cowan, N. (1998). *Attention and memory: An integrated framework*. Oxford, UK: Oxford University Press.

Cowan, N. (1999). An embedded-process model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory* (pp. 62–101). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/cbo9781139174909.006

Cowan, N. (2005). *Working memory capacity*. New York, NY: Psychology Press.

Cowan, N. (2008). What are the differences between long-term, and short-term, and working memory? In W. S. Sossin, J.-C. Lacaille, V. F. Castellucci, & S. Belleville (Eds.), *Progress in Brain Research: Essence of memory* (Vol. 169, pp. 323–338). Amsterdam: Elsevier. https://doi.org/10.1016/s0079-6123(07)00020-9

Cowan, N. (2014). Working memory underpins cognitive development, learning, and education. *Educational Psychology Review*, *26*(2), 197–223. https://doi.org/10.1007/s10648-013-9246-y

Cowan, N., Elliott, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive psychology*, *51*(1), 42–100. https://doi.org/10.1016/j.cogpsych.2004.12.001

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage publications.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, *7*(1), 31–51. https://doi.org/10.1177/026553229000700104

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, *86*(1), 67–96. https://doi.org/10.1111/1540-4781.00137

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466. https://doi.org/10.1016/s0022-5371(80)90312-6

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, *3*(4), 422–433. https://doi.org/10.3758/bf03214546

Dansereau, D. F., & Gregg, L. W. (1966). An information processing analysis of mental multiplication. *Psychonomic Science*, *6*(2), 71–72. https://doi.org/10.3758/bf03327962

Davies, A. (1968). Introduction. In A. Davies (Ed.), *Language testing symposium: A psycholinguistic approach* (pp. 1–18). London, UK: Oxford University Press.

Davies, A. (1977). The construction of language tests. In J. P. B. Allen & A. Davies (Eds.), *The Edinburgh Course in Applied Linguistics: Testing and experimental methods* (Vol. 4, pp. 38–104). Oxford, UK: Oxford University Press.

Davis, L. E. (2012). *Rater expertise in a second language speaking assessment: The influence of training and experience* [Unpublished doctoral dissertation]. The University of Hawai‘i, Manoa. https://scholarspace.manoa.hawaii.edu/bitstream/10125/100897/1/Davis_Lawrence_r.pdf

Davis, L. E. (2016). The influence of training and experience on scoring performance in scoring spoken language. *Language Testing, 33*(1), 117–135. https://doi.org/10.1177/0265532215582282

Dawson, R. (2011). How significant is a boxplot outlier?. *Journal of Statistics Education, 19*(2), 1–13. https://doi.org/10.1080/10691898.2011.11889610

Dechert, H. (1987). Analysing language processing through verbal protocols. In C. Faerch & G. Kasper (Eds.), *Introspection in second language research* (pp. 96-112). Clevedon, UK: Multilingual Matters.

de Groot, A. D. (1978). *Thought and choice in chess* (2nd ed.). Berlin, Germany: Walter de Gruyter. https://doi.org/10.1515/9783110800647

Dehn, M. J. (2008). *Working memory and academic learning*. Hoboken, NJ: John Wiley & Sons.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation, 14*(20), 1–11. https://doi.org/10.7275/da8t-4g52

Dunn, K. (2019). *Aptis scoring system*. (Technical Report No. 2019/001). London, UK: British Council. https://www.britishcouncil.org/sites/default/files/aptis_scoring_system_layout_final.pdf

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly: An International Journal, 2*(3), 197–221. https://doi.org/10.1207/s15434311laq0203_2

Eckes, T. (2011). *Introduction to many-facet Rasch measurement.* Frankfurt, Germany: Peter Lang. https://doi.org/10.3726/978-3-653-04844-5

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly, 9*(3), 270–292. https://doi.org/10.1080/15434303.2011.649381

Endsley, M. R. (2018). Expertise and situation awareness. In Ericsson, K. A., Hoffman, R. R., Kozbelt, A., & Williams, A. M. (Eds.), *The Cambridge handbook of expertise and expert performance* (2nd ed., pp. 714–744). New York, NY: Cambridge University Press. https://doi.org/10.1017/9781316480748.037

Engle, R. W. (1996). Working memory and retrieval: An inhibition-resource approach. In J. T. E. Richardson, R. W. Engle, L. Hasher, R. H. Logie, E. R. Stoltzfus, & R. T. Zachs (Eds.), *Working memory and human cognition* (pp. 89–119). New York, NY: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195100990.003.0004

Engle, R. W. (2001). What is working memory capacity?. In H. L. Roediger, J. S. Nairne, I. Neath, & A. M. Suprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 297–314). Washington, D.C.: American Psychological Association Press. https://psycnet.apa.org/doi/10.1037/10394-016

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, *11*(1), 19–23. https://doi.org/10.1111/1467-8721.00160

Engle, R. W., & Bukstel, L. (1978). Memory processes among bridge players of differing expertise. *American Journal of Psychology*, *91*(4), 673–689. https://doi.org/10.2307/1421515

Engle, R. W., Carullo, J. J., & Collins, K. W. (1991). Individual differences in working memory for comprehension and following directions. *Journal of Educational Research, 84*(5), 253–262. https://doi.org/10.1080/00220671.1991.10886025

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, *128*(3), 309–331. https://doi.org/10.1037/0096-3445.128.3.309

Erdosy, M. U. (2004). *Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions.* (ETS Technical Report No. 70). Princeton, NJ: ETS. https://doi.org/10.1002/j.2333-8504.2003.tb01909.x

Ericsson, K. A. (1996). The acquisition of expert performance: An introduction to some of the issues. In K. A. Ericsson (Ed.), *The road to excellence: The acquisition of expert performance in the arts and sciences, sports and games* (pp. 1–50). Mahwah, NJ: Lawrence Erlbaum Associates.

Ericsson, K. A. (2002). Attaining excellence through deliberate practice: Insights form the study of expert performance. In M. Ferrari (Ed.), *The pursuit of excellence in education* (pp. 21–55). Hillsdale, NJ: Lawrence Erlbaum Associates. https://doi.org/10.1002/9780470690048.ch1

Ericsson, K. A. (2015). Acquisition and maintenance of medical expertise: A perspective from the expert-performance approach with deliberate practice. *Academic Medicine, 90*(11), 1471–1486. https://doi.org/10.1097/acm.0000000000000939

Ericsson, K. A. (2018a). An introduction to the second edition of the Cambridge handbook of expertise and expert performance: Its development, organization, and content. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), *The Cambridge handbook of expertise and expert performance* (2nd ed., pp. 3–20). New York, NY: Cambridge University Press. https://doi.org/10.1017/9781316480748.001

Ericsson, K. A. (2018b). The differential influence of experience, practice, and deliberate practice on the development of superior individual performance of experts. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), *The Cambridge handbook of expertise and expert performance* (2nd ed., pp. 745–769). New York, NY: Cambridge University Press. https://doi.org/10.1017/9781316480748.038

Ericsson, K. A. (2018c). Superior working memory in experts. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), *The Cambridge handbook of expertise and expert performance* (2nd ed., pp. 696–713). New York, NY: Cambridge University Press. https://doi.org/10.1017/9781316480748.036

Ericsson, K. A. (2018d). Capturing expert thought with protocol analysis: Concurrent verbalizations of thinking during experts' performance on representative tasks. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), *The Cambridge handbook of expertise and expert performance* (2nd ed., pp. 191–212). New York, NY: Cambridge University Press. https://doi.org/10.1017/9781316480748.012

Ericsson, K. A., Chase, W. G., & Faloon, S. (1980). Acquisition of a memory skill. *Science*, *208*(4448), 1181–1182. https://doi.org/10.1126/science.7375930

Ericsson, K. A., Delaney, P. F., Weaver, G., & Mahadevan, R. (2004). Uncovering the structure of a memorist's superior "basic" memory capacity. *Cognitive Psychology*, *49*(3), 191–237. https://doi.org/10.1016/j.cogpsych.2004.02.001

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*(2), 211–245. https://psycnet.apa.org/doi/10.1037/0033-295X.102.2.211

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363–406. https://doi.org/10.1037/0033-295x.100.3.363

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology*, *47*(1), 273–305. https://doi.org/10.1146/annurev.psych.47.1.273

Ericsson, K.A., & Polson, P. G. (1988). A cognitive analysis of exceptional memory for restaurant orders. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. 23–70). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis*. Cambridge, MA: MIT Press.

Ericsson, K. A., & Smith, J. (1991). Prospects and limits of the empirical study of expertise: An introduction. In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 1–38). New York, NY: Cambridge University Press.

Ericsson, K. A., & Ward, P. (2007). Capturing the naturally occurring superior performance of experts in the laboratory: Toward a science of expert and exceptional performance. *Current Directions in Psychological Science, 16*(6), 346–350. https://doi.org/10.1111/j.1467 8721.2007.00533.x

Fairbairn, J., & Dunlea, J. (2017). *Speaking and writing rating scales revision*. (Technical Report No. 2017/001). London, UK: British Council. https://www.britishcouncil.org/sites/default/files/aptis_scale_revision_layout.pdf

Feigenbaum, E. A. (1989). What hath Simon wrought? In D. Klahr & K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert A. Simon* (pp. 165–180). Hillsdale, NJ: Lawrence Erlbaum Associates.

Feltovich, P. J., & Barrows, H. S. (1984). Issues of generality in medical problem solving. In H. Cl. Devolder & M. C. Schmidt (Eds.), *Tutorials in problem-based learning* (pp. 128–142). Assen, Holland: Van Gorcum.

Feltovich, P. J., Johnson, P. E., Moller, J. H., & Swanson, D. B. (1984). LCS: The role and development of medical knowledge in diagnostic expertise. In W. Clancey & E. Shortliffe (Eds.), *Readings in medical artificial intelligence: The first decade* (pp. 275–319). Reading, MA: Addison-Wesley.

Feltovich, P. J., Prietula, M. J., & Ericsson, K. A. (2018). Studies of expertise from psychological perspectives: Historical foundations and recurrent themes. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), *The Cambridge handbook of expertise and expert performance* (2nd ed., pp. 59–83). New York, NY: Cambridge University Press. https://doi.org/10.1017/9781316480748.006

Fincher-Kiefer, R., Post, T. A., Greene, T. R., & Voss, J. F. (1988). On the role of prior knowledge and task demands in the processing of text. *Journal of Memory and Language, 27*(4), 416–428. https://doi.org/10.1016/0749-596x(88)90065-4

Fitts, P. M., & Posner, M. I. (1967). *Human performance*. Belmont, CA: Brooks/Cole.

Fleishman, E. A., & Hempel, W. E., Jr. (1954). Changes in the factor structure of a complex psychomotor test as a function of practice. *Psychometrika, 19*, 239–252. https://doi.org/10.1007/bf02289188

Freedman, S.W., & Calfee, R. C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P. Mosenthal, L. Tamor, & S. A. Walmsley (Eds.), *Research on writing: Principles and methods* (pp. 75–98). New York, NY: Longman.

Friedman, N. P., & Miyake, A. (2005). Comparison of four scoring methods for the reading span test. *Behavior Research Methods, 37*(4), 581–590. https://doi.org/10.3758/bf03192728

Friedman, N. P., Miyake, A., Young, S. E., DeFries, J. C., Corley, R. P., & Hewitt, J. K. (2008). Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental Psychology: General, 137*(2), 201–225. https://doi.org/10.1037/0096-3445.137.2.201

Fulcher, G. (2003). *Testing second language speaking*. Harlow, UK: Longman. https://doi.org/10.4324/9781315837376

Gagné, E. D., Yekovich, C. W., & Yekovich, F. R. (1993). *The cognitive psychology of school learning*. New York, NY: Harper Collins College Publishers.

Gardner, R. C. (2004). *Attitude/motivation test battery: International AMTB research project*. http://hyxy.nankai.edu.cn/jingpinke/buchongyuedu/Motivation%20measurement-AMTB.pdf

Gass, S. M., & Mackey, A. (2017). *Stimulated recall methodology in applied linguistics and L2 research* (2nd ed.). New York, NY: Routledge. https://doi.org/10.4324/9781315813349

George, D., & Mallery, M. (2010). *SPSS for windows step by step: A simple guide and reference* (10th ed.). Boston, MA: Pearson.

Gilhooly, K. J., McGeorge, P., Hunter, J., Rawles, J. M., Kirby, I. K., Green, C., & Wynn, V. (1997). Biomedical knowledge in diagnostic thinking: the case of electrocardiogram (ECG) interpretation. *European Journal of Cognitive Psychology, 9*(2), 199–223. https://doi.org/10.1080/713752555

Glaser, R. & Chi, M. T. H. (1988). Overview. In Glaser, R., Chi, M. T., & Farr, M. J. (Eds.), *The nature of expertise* (pp. xv–xxviii). Hillsdale, NJ: Lawrence Erlbaum Associates.

Glaser, B., & Strauss, A. L. (2017). *The discovery of grounded theory: Strategies for qualitative research*. New York, NY: Routledge. https://doi.org/10.4324/9780203793206-1

Goodwin, S. (2016). A Many-Facet Rasch analysis comparing essay rater behavior on an academic English reading/writing test used for two purposes. *Assessing Writing, 30*, 21–31. https://doi.org/10.1016/j.asw.2016.07.004

Grabowski, K. C. (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.

Groebel, L. (1980). A comparison of students' reading comprehension in their native language with their reading comprehension in the target language. *English Language Teaching Journal, 35*, 54–59. https://doi.org/10.1093/elt/xxxv.1.54

Groeger, J. A., & Brady, S. J. (2004). *Differential effects of formal and informal driver training*. (Road Safety Research Report No. 42). London, UK: Department of Transport. https://www.researchgate.net/profile/John_Groeger/publication/255601556_Differential_effects_of_formal_and_informal_driver_training/links/53fde98b0cf2364ccc092b4b/Differential-effects-of-formal-and-informal-driver-training.pdf

Groen, G. J., & Patel, V. L. (1988). The relationship between comprehension and reasoning in medical expertise. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. 287–310). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gruson, L. M. (2001). Rehearsal skill and musical competence: Does practice make perfect. In J. A. Sloboda (Ed.), *Generative processes in music: The psychology of performance, improvisation, and composition* (pp. 91–112). Oxford, UK: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198508465.003.0005

Gui, M. (2012). Exploring differences between Chinese and American EFL teachers' evaluations of speech performance. *Language Assessment Quarterly, 9*(2), 186–203. https://doi.org/10.1080/15434303.2011.614030

Hair, F. J. Jr., Sarstedt, M., Hopkins, L., & Kuppelwieser, V. G. (2014). Partial least squares structural equation modeling (PLS-SEM): An emerging tool in business research. *European Business Review, 26*(2), 106–121. https://doi.org/10.1108/ebr-10-2013-0128

Hambrick, D. Z., & Engle, R. W. (2002). Effects of domain knowledge, working memory capacity, and age on cognitive performance: An investigation of the knowledge-is-power hypothesis. *Cognitive Psychology*, *44*(4), 339–387. https://doi.org/10.1006/cogp.2001.0769

Hambrick, D. Z., & Oswald, F. L. (2005). Does domain knowledge moderate involvement of working memory capacity in higher-level cognition? A test of three models. *Journal of Memory and Language*, *52*(3), 377–397. https://doi.org/10.1016/j.jml.2005.01.004

Henning, G. (1975). Measuring foreign language reading comprehension. *Language Learning*, *25*(1), 109–114. https://doi.org/10.1111/j.1467-1770.1975.tb00111.x

Henning, G. (1992). *Scalar analysis of the Test of Written English.* (TOEFL Research Report No. RR-92-30). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1992.tb01461.x

Hinsley, D. A., Hayes, J. R., & Simon, H. A. (1977). From words to equations: Meaning and representation in algebra word problems. In M. Just & P. Carpenter (Eds.), *Cognitive processes in comprehension* (pp. 89–106). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hogan, T., & Rabinowitz, M. (2009). Teacher expertise and the development of a problem representation. *Educational Psychology*, *29*(2), 153–169. https://doi.org/10.1080/01443410802613301

Howell, D. C. (2010). *Statistical methods for psychology* (7th ed.). Belmont, CA: Wadsworth.

Huber, P. (1981), *Robust Statistics*. New York, NY: John Wiley. https://doi.org/10.1002/0471725250

Hummel, K. M. (2009). Aptitude, phonological memory, and second language proficiency in nonnovice adult learners. *Applied Psycholinguistics*, *30*(2), 225–249. https://doi.org/10.1017/s0142716409090109

Isaacs, T., & Trofimovich, P. (2011). Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgements of second language speech. *Applied Psycholinguistics*, *32*(1), 113–140. https://doi.org/10.1017/s0142716410000317

Jeffries, R., Turner, A., Polson, P., & Atwood, M. (1981). Processes involved in designing software. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 255–283). Hillsdale, NJ: Lawrence Erlbaum Associates.

Jeon, E. H. (2015). Multiple regression. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 131–158). New York, NY: Routledge. https://doi.org/10.4324/9781315870908-7

Joh, J., & Plakans, L. (2017). Working memory in L2 reading comprehension: The influence of prior knowledge. *System*, *70*, 107–120. https://doi.org/10.1016/j.system.2017.07.007

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, *26*(4), 485–505. https://doi.org/10.1177/0265532209340186

Jones, R. L. (1985). Second language performance testing: An overview. In P. C. Hauptman, R. LeBlanc, & M. B. Wesche (Eds.), *Second language performance testing* (pp. 15–24). Ottawa, Canada: University of Ottawa Press.

Juffs, A., & Harrington, M. (2011). Aspects of working memory in L2 learning. *Language Teaching*, *44*(2), 137–166. https://doi.org/10.1017/s0261444810000509

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*, 122–149. https://doi.org/10.1037/0033-295x.99.1.122

Kane, M. J., Conway, A. R., Bleckley, M. K., & Engle, R. W. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology: General*, *130*(2), 169–183. https://doi.org/10.1037/0096-3445.130.2.169

Kane, M. J., & Engle, R. W. (2000). Working-memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(2), 336–358. https://doi.org/10.1037/0278-7393.26.2.336

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*(2), 189–217. https://doi.org/10.1037/0096-3445.133.2.189

Kane, M. T. (2006). Validation. *Educational measurement*, *4*(2), 17–64.

Kang, O. (2008). Ratings of L2 oral performance in English: Relative impact of rater characteristics and acoustic measures of accentedness. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *6*, 181–205.

Kellman, P. J., & Massey, C. M. (2013). Perceptual learning, cognition, and expertise. *Psychology of Learning and Motivation*, *58*, 117–165. https://doi.org/10.1016/b978-0-12-407237-4.00004-9

Kim, B. (2016, May 20). *Hierarchical linear regression*. https://data.library.virginia.edu/hierarchical-linear-regression/

Kim, H. J. (2011). *Investigating raters' development of rating ability on a second language speaking assessment* [Unpublished doctoral dissertation]. Teachers College, Columbia University, New York. https://search.proquest.com/openview/37eef0ed7846f5c3ee5d60219693abc0/1?pqorigsite=gscholar&cbl=18750&diss=y

Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly*, *12*(3), 239–261. https://doi.org/10.1080/15434303.2015.1049353

Klein, G. (1998). *Sources of power: How people make decisions*. Cambridge, MA: MIT Press.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford.

Koda, K. (1989). The effects of transferred vocabulary knowledge on the development of L2 reading proficiency. *Foreign Language Annals*, *22*(6), 529–540. https://doi.org/10.1111/j.1944-9720.1989.tb02780.x

Koschmann, T., LeBaron, C., Goodwin, C., & Feltovich, P. (2001). Dissecting common ground: Examining an instance of reference repair. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *23*(23). https://escholarship.org/uc/item/976933jc

Kossoudji, S. A. (1988). English language ability and the labor market opportunities of Hispanic and East Asian immigrant men. *Journal of Labor Economics*, *6*(2), 205–228. https://doi.org/10.1086/298181

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?!. *Intelligence, 14*(4), 389–433. https://doi.org/10.1016/s0160-2896(05)80012-1

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London, UK: Longman.

Larkin, J., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science*, *208*(4450), 1335–1342. https://doi.org/10.1126/science.208.4450.1335

Lawson, M. J., & Hogben, D. (1996). The vocabulary-learning strategies of foreign-language students. *Language Learning*, *46*(1), 101–135. https://doi.org/10.1111/j.1467-1770.1996.tb00642.x

Lehmann, A. C., & Ericsson, K. A. (1993). Sight-reading ability of expert pianists in the context of piano accompanying. *Psychomusicology: A Journal of Research in Music Cognition, 12*(2), 182–195. https://doi.org/10.1037/h0094108

Lim, G. S. (2009). *Prompt and rater effects in second language writing performance assessment* [Unpublished PhD dissertation]. The University of Michigan, Ann Arbor.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*(4), 543–560. https://doi.org/10.1177/0265532211406422

Linacre, J. M. (1992). *Many-facet Rasch measurement*. Chicago, Illinois: MESA Press.

Linacre, J. M. (2002). What do infit, outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2004). *Facets Rasch measurement computer program*. Chicago, IL: Winsteps.com.

Linacre, J. M. (2014). *Facets Rasch measurement computer program* (version 3.71.4). Beaverton, Oregon: Winsteps.com.

Linck, J. A., Hughes, M. M., Campbell, S. G., Silbert, N. H., Tare, M., Jackson, S. R., Smith, B. K., Bunting, M. F., & Doughty, C. J. (2013). Hi-LAB: A new measure of aptitude for high-level language proficiency. *Language Learning*, *63*(3), 530–566. https://doi.org/10.1111/lang.12011

Linck, J. A., & Weiss, D. J. (2011). Working memory predicts the acquisition of explicit L2 knowledge. In C. Sanz & R. P. Leow (Eds.), *Implicit and explicit language learning: Conditions, processes, and knowledge in SLA and bilingualism* (pp. 101–113). Washington, DC: Georgetown University Press.

Lindhardsen, V. (2009). *From independent ratings to communal ratings: A study of CWA raters' decision-making behaviors* [Unpublished doctoral dissertation]. Copen Copenhagen Business School, Frederiksberg, Denmark.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters?. *Language Testing*, *19*(3), 246–276. https://doi.org/10.1191/0265532202lt230oa

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt, Germany: Peter Lang.

Lumley, T., & Brown, A. (2005). Research methods in language testing. In E. Hinkle (Ed.), *Handbook of research in second language teaching and learning* (pp. 833–856). Mahwah, NJ: Lawrence Erlbaum Associates.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54–71. https://doi.org/10.1177/026553229501200104

Luoma, S. (2004). *Assessing speaking*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/cbo9780511733017

Mackey, A., Philp, J., Egi, T., Fujii, A., & Tatsumi, T. (2002). Individual differences in working memory, noticing of interactional feedback and L2 development. In P. Robinson (Ed.), *Individual differences and instructed language learning* (Vol. 2, pp. 181–209). Amsterdam, Netherlands: John Benjamins. https://doi.org/10.1075/lllt.2.12mac

Matsumoto, K. (1993). Verbal-report data and introspective methods in second language research. *RELC Journal*, *24*, 32-60.

May, L. A. (2006). An examination of rater orientations on a paired candidate discussion task through stimulated verbal recall. *Melbourne Papers in Language Testing (MPLT)*, *11*(1), 29–51. https://eprints.qut.edu.au/15747/1/15747.pdf?ev=pub_ext_prw_xdl

May, L. A. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly*, *8*(2), 127–145. https://doi.org/10.1080/15434303.2011.565845

McCutchen, D. (2000). Knowledge, processing, and working memory: Implications for a theory of writing. *Educational Psychologist*, *35*(1), 13–23. https://doi.org/10.1207/s15326985ep3501_3

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

Mecartty, F. H. (2000). Lexical and grammatical knowledge in reading and listening comprehension by foreign language learners of Spanish. *Applied Language Learning*, *11*(2), 323–348.

Meinz, E. J., & Salthouse, T. A. (1998). The effects of age and experience on memory for visually presented music. *Journal of Gerontology Series B: Psychological Sciences and Social Sciences, 53B*(1), 60–69. https://doi.org/10.1093/geronb/53b.1.p60

Meiron, B. E. (1998). *Rating oral proficiency tests: A triangulated study of rater thought processes* [Unpublished master's thesis]. California State University, Los Angeles.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York, NY: American Council on Education and Macmillan.

Milanovic, M., Saville, N., & Shen, S. (1996). A study of the decision-making behavior of composition markers. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium (LTRC), Cambridge and Arnhem* (pp. 92–114). New York, NY: Cambridge University Press.

Minsky, M., & Papert, S. (1974). *Artificial intelligence*. [Condensed lectures]. Oregon State System of Higher Education.

Miyake, A., & Friedman, N. P. (1998). Individual differences in second language proficiency: Working memory as language aptitude. In A. F. Healy & L. E. Bourne Jr. (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention* (pp. 339–364). Mahwah, NJ: Lawrence Erlbaum Associates.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Morrow, K. (1979). Communicative language testing: Revolution or evolution?. In C. J. Brumfit & K. Johnson (Eds.), *The communicative approach to language teaching* (pp. 143–157). Oxford, UK: Oxford University Press.

Myford, C. M., & Wolfe, E. W. (2000). *Monitoring sources of variability within the Test of Spoken English assessment system* (TOEFL Research Report No. 65). Princeton, NJ: Educational Testing Service. https://www.ets.org/Media/Research/pdf/RR-00-06-Myford.pdf

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386–422.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring scoring performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement, 46*(4), 371–389. https://doi.org/10.1111/j.1745-3984.2009.00088.x

Norman, D. A., & Shallice, T. (1986). Attention to action. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and self-regulation: Advances in research and theory* (Vol. 4, pp. 1–18). New York, NY: Plenum.

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments*. Honolulu, HI: University of Hawai'i, National Foreign Language Resource Center.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects?. *Educational Evaluation and Policy Analysis*, *26*(3), 237–257. https://doi.org/10.3102/01623737026003237

Oller, Jr. J. W. (1979). *Language tests at school*. London, UK: Longman.

Oller, Jr. J. W. (1983). A consensus for the eighties? In J. W. Oiler, Jr. (Ed.), *Issues in language testing research* (pp. 351–356). Rowley, MA: Newbury House.

O'Malley, J. M., & Chamot, A. U. (1990). *Learning strategies in second language acquisition*. Cambridge, UK: Cambridge University Press.

Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. *System*, *30*(2), 143–154. https://doi.org/10.1016/s0346-251x(02)00002-7

O'Sullivan, B. (2008). *Notes on assessing speaking*. Cornell University Language Resource Center. http://www.lrc.cornell.edu/events/past/20082009/papers08/osull1.pdf.

O'Sullivan, B. (2015a). *Aptis test development approach*. (Technical Report No. 2015/001). London, UK: British Council. https://www.britishcouncil.org/sites/default/files/tech_001_barry_osullivan_aptis_test_-_v5_0.pdf

O'Sullivan, B. (2015b). *Linking the Aptis reporting scales to the CEFR*. (Technical Report No. 2015/003). London, UK: British Council. https://www.britishcouncil.org/sites/default/files/tech_003_barry_osullivan_linking_aptis_v4_single_pages_0.pdf

O'Sullivan, B., & Dunlea, J. (2015). *Aptis technical manual Version 1.0*. (Technical Report No. 2015/005). London, UK: British Council. https://www.britishcouncil.org/aptis-general-technical-manual-version-10

Park, J. H. (1999). The earnings of immigrants in the United States: The effect of English-speaking ability. *American Journal of Economics and Sociology*, *58*(1), 43–56. https://doi.org/10.1111/j.1536-7150.1999.tb03282.x

Patel, V. L., & Groen, G. J. (1991). The general and specific nature of medical expertise: A critical look. In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 93–125). Cambridge, UK: Cambridge University Press.

Payne, T. W., Kalibatseva, Z., & Jungers, M. K. (2009). Does domain experience compensate for working memory capacity in second language reading comprehension?. *Learning and Individual Differences*, *19*(1), 119–123. https://doi.org/10.1016/j.lindif.2008.05.003

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press. https://doi.org/10.17226/10019

Penrose, A. M. (1993). Writing and learning: Exploring the consequences of task interpretation. In Penrose, A. M. and B. M. Sitko (Eds.), *Hearing ourselves think: Cognitive research in the college classroom*, (pp. 52–69). New York, NY: Oxford University Press.

Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language Testing*, *20*(1), 26–56. https://doi.org/10.1191/0265532203lt243oa

Phakiti, A. (2007). *Strategic competence and EFL reading test performance*. New York, NY: Peter Lang.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium (LTRC), Cambridge and Arnhem* (pp. 74–91). Cambridge, UK: Cambridge University Press.

Poulisse, N., Bongaerts, T., & Kellerman, E. (1987). The use of retrospective verbal reports in the analysis of compensatory strategies. In C. Faerch & G. Kaasper (Eds.), *Introspection in second language research* (pp. 213–229). Clevedon, UK: Multilingual Matters.

Prystowsky, J. B. (2005). Are young surgeons competent to perform alimentary tract surgery?. *Archives of Surgery*, *140*(5), 495–502. https://doi.org/10.1001/archsurg.140.5.495

Purpura, J. E. (1997). An analysis of the relationships between test-takers' cognitive and metacognitive strategy use and second language test performance. *Language Learning*, *47*(2), 289–325. https://doi.org/10.1111/0023-8333.91997009

Purpura, J. (1999). *Strategy use and second language test performance: A structural equation modeling approach*. Cambridge, UK: Cambridge University Press.

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

Purpura, J. E. (2012). *What is the role of strategic competence in a processing account of L2 learning or use?*. Paper presented at the American Association for Applied Linguistics Conference, Boston, MA.

Purpura, J. E. (2014). Cognition and language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1–23). Hoboken, NJ: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118411360.wbcla150

Purpura, J. E. (2017). *Assessing meaning*. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language testing and assessment* (3rd ed., pp. 33–61). New York, NY: Springer. https://doi.org/10.1007/978-3-319-02261-1_1

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Denmarks Paedagogiske Institute.

Richards, J. C. (2008). *Teaching listening and speaking: From theory to practice*. New York: Cambridge University Press.

Ricks, T. R., & Wiley, J. (2009). The influence of domain knowledge on the functional capacity of working memory. *Journal of Memory and Language, 61*(4), 519–537. https://doi.org/10.1016/j.jml.2009.07.007

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 129–152). Cambridge, UK: Cambridge University Press.

Sakyi, A. A. (2003). *A study of the holistic scoring behaviors of experienced and novice ESL instructors* [Unpublished doctoral dissertation]. Toronto, Canada: University of Toronto.

Savignon, S. J. (1972). *Communicative competence: An experiment in foreign language teaching.* Philadelphia: The Center for Curriculum Development.

Schmidt, R. (1994). Deconstructing consciousness in search of useful definitions for applied linguistics. *AILA Review, 11,* 11-26.

Seedhouse, P., Harris, A., Naeb, R. & Üstünel, E. (2014) The relationship between speaking features and band descriptors: A mixed methods study. *IELTS Research Reports Online Series, 2,* 1–30. IELTS Partners: British Council, Cambridge English Language Assessment and IDP: IELTS Australia.

Shah, P., & Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing: An individual differences approach. *Journal of Experimental Psychology: General, 125*(1), 4–27. https://doi.org/10.1037/0096-3445.125.1.4

Simon, H. A., & Chase, W. G. (1973). Skill in chess. *American Scientist*, *61*(4), 394–403.

Smith, D. (2000). Rater judgements in the direct assessment of competency-based second language writing ability. In G. Brindley (Ed.), *Studies in immigrant English language assessment* (pp. 159–190). Sydney, Australia: National Centre for English Language Teaching and Research, Macquarie University.

Spilich, G. J., Vesonder, G. T., Chiesi, H. L., & Voss, J. F. (1979). Text processing of domain-related information for individuals with high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior*, *18*(3), 275–290. https://doi.org/10.1016/s0022-5371(79)90155-5

Spolsky, B. (1973). What does it mean to know a language; or how do you get someone to perform his competence? In J. W. Oiler & J. C. Richards (Eds.), *Focus on the learner: Pragmatic perspectives for the language teacher* (pp. 164–176). Rowley, MA: Newbury House Publishers.

Stemmer, B. (1991). *What's on a C-examinee's mind? Mental processes in C-test taking*. Bochum, Germany: Universitätsverlag Dr. N. Brockmeyer.

Swanson, H. L., & Berninger, V. W. (1996). Individual differences in children's working memory and writing skill. *Journal of Experimental Child Psychology*, *63*(2), 358–385. https://doi.org/10.1006/jecp.1996.0054

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson Education.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent?. *Journal of Memory and Language*, *28*(2), 127–154. https://doi.org/10.1016/0749-596x(89)90040-5

Unsworth, N., & Engle, R. W. (2007). On the division of short-term and working memory: An examination of simple and complex span and their relation to higher order abilities. *Psychological Bulletin*, *133*(6), 1038–1066.https://doi.org/10.1037/0033-2909.133.6.1038

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111–125). Norwood, NJ: Ablex.

Vickers, A. J., Bianco, F. J., Gonen, M., Cronin, A. M., Eastham, J. A., Schrag, D., ... & Scardino, P. T. (2008). Effects of pathologic stage on the learning curve for radical prostatectomy: Evidence that recurrence in organ-confined cancer is largely related to inadequate surgical technique. *European Urology*, *53*(5), 960–966. https://doi.org/10.1016/j.eururo.2008.01.005

Vickers, A. J., Bianco, F. J., Serio, A. M., Eastham, J. A., Schrag, D., Klein, E. A., ... & Scardino, P. T. (2007). The surgical learning curve for prostate cancer control after radical prostatectomy. *Journal of the National Cancer Institute*, *99*(15), 1171–1177. https://doi.org/10.1093/jnci/djm060

Vickers, A. J., Savage, C. J., Hruza, M., Tuerk, I., Koenig, P., Martínez-Piñeiro, L., ... & Guillonneau, B. (2009). The surgical learning curve for laparoscopic radical prostatectomy: A retrospective cohort study. *The Lancet Oncology*, *10*(5), 475–480. https://doi.org/10.1016/s1470-2045(09)70079-8

Voss, J. F., Greene, T. R., Post, T. A., & Penner, B. C. (1983). Problem-solving skill in the social sciences. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 17, pp. 165–213). New York, NY: Academic Press. https://doi.org/10.1016/S0079-7421(08)60099-7

Wallace, M. J. (1991). *Training foreign language teachers: A reflective approach*. Cambridge, UK: Cambridge University Press.

Wei, J., & Llosa, L. (2015). Investigating differences between American and Indian raters in assessing TOEFL iBT speaking tasks. *Language Assessment Quarterly*, *12*(3), 283–304. https://doi.org/10.1080/15434303.2015.1037446

Weigle, S. C. (1994a). Effects of training on raters of English as a second language compositions: Quantitative and qualitative approaches [Unpublished PhD dissertation]. University of California, Los Angeles.

Weigle, S. C. (1994b). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), 197–223. https://doi.org/10.1177/026553229401100206

Weigle, S. C. (2002). *Assessing writing*. Cambridge, UK: Cambridge University Press.

Whitehead, A. E., Taylor, J. A., & Polman, R. C. (2015). Examination of the suitability of collecting in event cognitive processes using Think Aloud protocol in golf. *Frontiers in psychology*, *6*, 1083. https://doi.org/10.3389/fpsyg.2015.01083

Wind, S. A., Stager, C., & Patil, Y. J. (2017). Exploring the relationship between textual characteristics and rating quality in rater-mediated writing assessments: An illustration with L1 and L2 writing assessments. *Assessing Writing*, *34*, 1–15. https://doi.org/10.1016/j.asw.2017.08.003

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*(2), 231–252. https://doi.org/10.1177/0265532212456968

Wiseman, C. S. (2008). Investigating selected facets in measuring second language writing ability using holistic and analytic scoring methods [Unpublished doctoral dissertation]. Teachers College, Columbia University, New York.

Wolfe, E. W. (1995). A study of expertise in essay scoring [Unpublished doctoral dissertation]. University of California, Berkeley.

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing*, *4*(1), 83–106. https://doi.org/10.1016/s1075-2935(97)80006-2

Wolfe, E. W., & Feltovich, B. (1994). *Learning to rate essays: A study of scorer cognition*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Wolfe, E. W., & Kao, C. W. (1996). *Expert/novice differences in the focus and procedures used by essay scorers*. Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Wolfe, E. W., Kao, C. W., & Ranney, M. (1998). Cognitive differences in proficient and non-proficient essay scorers. *Written Communication*, *15*(4), 465–492. https://doi.org/10.1177/0741088398015004002

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

Xi, X., & Mollaun, P. (2011). Using raters from India to score a large-scale speaking test. *Language Learning*, *61*(4), 1222–1255. https://doi.org/10.1111/j.1467-9922.2011.00667.x

Zhang, J. (2016). Same text different processing? Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring. *Assessing Writing*, *27*, 37–53. https://doi.org/10.1016/j.asw.2015.11.001

Zhang, Y., & Elder, C. (2011). Judgements of oral proficiency by non-native and native English-speaking teacher raters: Competing or complementary constructs?. *Language Testing*, *28*(1), 31–50. https://doi.org/10.1177/0265532209360671

Zucker, S., Sassman, C., & Case, B. J. (2004). *Cognitive labs*. (Technical Report). San Antonio, TX: Harcourt Assessment Inc.. http://images.pearsonclinical.com/images/PDF/CognitiveLabs_Final.pdf