



## Selected Bibliography for Conference in Salamanca, Spain, November 2018

### The Roles of Test Takers and Users in Striving for Fairness

#### List of Headings

Fairness in Test Creation (especially in the Spanish context) *Rosario Domínguez*

Engaging Test Takers and Test Users in Language Testing *Lia Plakans*

Online Proctoring: A Key Feature of Secure High-Stakes Tests in the Future *Roger Johnson*

Marking of Writing Tests from the Examiner's Point of View: Consequences for the Candidate  
*Juan Miguel Prieto*

Supervision and Security: Good Practices to Guarantee the Success of Exams *Marta García*

What is a Fair Test in a Multi-Cultural Society? *Gordon Stobart*

Immigration and Language Testing *Kamran Khan*

Language and Security *Kamran Khan*

## **Fairness in Test Creation (especially in the Spanish context)**

ALTE (1998). *The Multilingual Glossary of Language Testing Terms*, Cambridge, Cambridge University Press.

ALTE (1994). Code of Practice (*Código de buena práctica* de la Association of Language Testers in Europe (ALTE)), en [https://es.alte.org/resources/Documents/code\\_practice\\_es.pdf](https://es.alte.org/resources/Documents/code_practice_es.pdf).

ALTE (2011). *Manual for Language Test Development and Examining*.

ALTE (forthcoming). *Principles of Good Practice*.

Llorián, Susana (2007). *Entender y utilizar el Marco común europeo de referencia desde el punto de vista del profesor de lenguas*. Madrid: Santillana-Universidad de Salamanca.

Martínez, M<sup>a</sup> Rosario (1995). *Psicometría: teoría de los tests psicológicos y educativos*. Madrid: Síntesis.

Muñiz, José (2000). *Teoría Clásica de los Tests*. Madrid: Pirámide.

Muñiz, José (2010). "Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems". *Papeles del Psicólogo*, 31(1), 57-66.

Figueras, Neus (2004). "Estándares y calidad en la elaboración y administración de pruebas y exámenes. Criterios mínimos para el reconocimiento y la comparabilidad", en: *III Congreso de la Lengua Española (La certificación de la competencia lingüística en español como lengua extranjera. Hacia un enfoque hispánico del sistema)* celebrado en Rosario (Argentina) del 17 al 20 de noviembre de 2004. En: [http://www.congresosdelalengua.es/rosario/ponencias/internacional/figueras\\_n.htm](http://www.congresosdelalengua.es/rosario/ponencias/internacional/figueras_n.htm) (30 de noviembre de 2015).

Instituto Cervantes (2007). *Plan curricular del Instituto Cervantes. Niveles de referencia para el español*. Madrid: Instituto Cervantes - Biblioteca Nueva.

Instituto Cervantes. *Guía para los creadores de ítems niveles. Destrezas receptoras y productivas*.

## **Engaging Test Users and Test Takers in Language Testing**

Andrade, H., Du, Y., & Mycek, K. (2010). Rubric-referenced self-assessment in middle school student writing. *Assessment in Education: Principles, Policy, and Practice*, 17, 199-214.

Becker, A. (2016). Student-generated scoring rubrics: Examining their formative value for improving ESL students' writing performance. *Assessing Writing*, 29, 15-24.

Bradbury, H. & Reason, P. (2008). Issues and choice points for improving the quality of action research. In Minkler & Wallerstein (Eds.) *Community-based participatory research of health: From process to outcomes*. Jossey-Bass.

Brown, A. (1993). The role of test-taker feedback in the test development process: Test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing*, 10, 277-303.

Duran, B, Wallerstein, N., Alvial, M., Belone, L. Minker, M., & Foley, K. (2013). Developing and maintaining partnerships with communities. In Isreal, Eng, Shulz, & Parker (Eds.) *Method for community-based participatory research for health*. Jossey-Bass.

Gu, L., & So, Yo. (2015). Voices from stakeholders: What makes an academic English test 'international'? *Journal of English for Academic Purposes*, 18, 9-24.

Ockey, G., Koyama, D., & Setoguchi, E. (2013). Stakeholder input and test design: A case study on changing the interlocutor familiarity facet of the group oral discussion test. *Language Assessment Quarterly*, 10, 292-308.

Stiggins, R. J. (2004). *Student-involved assessment for learning*. Upper Saddle River, NJ: Prentice Hall.

So, Y. (2014). Are teacher perspectives useful? Incorporating EFL teacher feedback in the development of a large-scale international English test. *Language Assessment Quarterly*, 11, 283-299.

### **Online Proctoring: A Key Feature of Secure High-Stakes Tests in the Future?**

Berkey, Dennis & Halfond, Jay (2015). Cheating, Student Authentication and Proctoring in Online Programs. *New England Journal of Higher Education*, <http://www.nebhe.org/thejournal/cheating-student-authentication-and-proctoring-in-online-programs/>

D'Souza, K. A. and Siegfeldt, D. V. (2017). A Conceptual Framework for Detecting Cheating in Online and Take-Home Exams. *Decision Sciences Journal of Innovative Education*, 15: 370-391. doi:10.1111/dsji.12140

Karim, Michael & Kaminsky, Samuel & Behrend, Tara (2014). Cheating, Reactions, and Performance in Remotely Proctored Testing: An Exploratory Experimental Study. *Journal of Business and Psychology*, 29. 1-18. 10.1007/s10869-014-9343-z.

Ketab, Salam & Clarke, Nathan & Haskell-Dowland, Paul (2015). E-Invigilation of E-Assessments. *Conference: Proceedings of INTED2015 Conference 2nd-4th March 2015, Madrid, Spain*

Weiner, John A. & Hurtz, Gregory M (2017). A Comparative Study of Online Remote Proctored versus Onsite Proctored High-Stakes Exams. *Journal of Applied Testing Technology*, v18 n1 p13-20

### **Marking of Writing Tests from the Examiner's Point of View: Consequences for the Candidate**

Barret, Seven (2005). Raters and examinations. In Alagumalai, Sivakumar; Curtis, David C. and Hungi, Njora (Eds). *Applied Rasch Measurement: A Book of Exemplars – Papers in Honour of John P. Deeves*. Dordrecht: Springer, 159-177.

Casanova, Dominique and Demeuse, Marc (2011). Analyse des différentes facettes influant sur la fidélité d'une l'épreuve d'expression écrite d'un test de français langue étrangère. *Mesure et évaluation en éducation* 34 (1), 25-53.

Cooper, William H. (1981). Ubiquitous halo. *Psychological Bulletin*, 90, 218-244.

Cronbach, Lee J. (1990). *Essentials of Psychological Testing* (5<sup>th</sup> ed.). New York: Harper & Row.

Cuxart, Anna; Martí, Manuel and Ferrer, Ferran (1997). Algunos factores que inciden en el rendimiento y la evaluación en los alumnos de las pruebas de aptitud de acceso a la universidad. *Revista de Educación*, 314, 63-88. In <http://www.mecd.gob.es/dctm/revista-de-educacion/articulosre314/re3140400462.pdf?documentId=0901e72b81272c3d> (Accessed 7 November 2018).

DeCotiis, Thomas A. (1977). "An analysis of the external validity and applied relevance of three rating formats". *Organizational Behavior and Human Performance*, 19, 247-266.

Eckes, Thomas (2004). Beurteilerübereinstimmung und Beurteilerstrenge: Eine Multi-facetten-Rasch-analyse von Leistungsbeurteilungen im «Test Deutsch als Fremdsprache» (TestDaF) [Rater agreement and rater severity: A many-facet Rasch analysis of performance assessments in the «Test of German as a Foreign Language (TestDaF) »]. *Diagnostica*, 50, 65-77.

Eckes, Thomas (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, 2, 197-221.

Eckes, Thomas (2008). Assuring the quality of TestDaF examinations: A psychometric modeling approach. In Taylor, Lynda and Weir, Cyril J. (eds.), *Multilingualism and Assessment: Achieving Transparency, Assuring Quality, Sustaining Diversity – Proceedings of the ALTE Berlin Conference May 2005*. Cambridge: Cambridge University Press, 157-178

Eckes, Thomas (2009). Many-facet Rasch measurement. In Takala, Sauli (ed.), *Reference Supplement to the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (Section H). Strasbourg: Council of Europe.

Eckes, Thomas (2010). The TestDaF implementation of the SOPI: Design, analysis, and evaluation of a semi-direct speaking test. In Araujo, Luisa (ed.), *Computer-based assessment (CBA) of Foreign Language Speaking Skills*. Luxembourg: Office of Publications of the European Union, 63-83.

Eckes, Thomas (2011). *Introduction to Many-facet Rasch Measurement. Analyzing and Evaluating Rater-Mediated Assessments*. Frankfurt: Peter Lang.

Farrokhi, Farahman and Esfandiari, Rajab (2011). A Many-facet Rasch model to detect halo effect in three types of raters. *Theory and Practice in Language Studies*, 1 (11), 1531-1540.

Fisicaro, Sebastiano A. and Lance, Charles E. (1990). Implications of three causal models for the measurement of halo error. *Applied Psychological Measurement*, 14, 419-429.

Fisicaro, Sebastiano A. and Vance, Robert J. (1994). Comments on the Measurement of Halo. *Educational and Psychological Measurement*, 54 (2), 366-371.

Tejedor, F. J. and Montero, L. (1990). Indicadores de la calidad docente para la evaluación del profesorado universitario. *Revista española de Pedagogía*, 86, 259-279.

Grossman, Michele y Wood, Wendy (1993). Sex differences in emotional intensity: A social role explanation. *Journal of Personality and Social Psychology*, 65, 1010-1022. En <[http://dornsife.usc.edu/assets/sites/545/docs/Wendy\\_Wood\\_Research\\_Articles/Gender\\_Differences\\_in\\_Social\\_Behavior/Grossman.Wood.1993\\_Sex\\_differences\\_in\\_emotional\\_intensity.pdf](http://dornsife.usc.edu/assets/sites/545/docs/Wendy_Wood_Research_Articles/Gender_Differences_in_Social_Behavior/Grossman.Wood.1993_Sex_differences_in_emotional_intensity.pdf)> (Accessed 7 November 2018).

Guilford, Joy Paul (1954). *Psychometric Methods*. New York, NY: McGraw-Hill.

King, Larry M., Hunter, John E. and Schmidt, Frank L. (1980). "Halo in a multidimensional forced-choice performance evaluation scale". *Journal of Applied Psychology*, 65 (5), 507-516.

Knoch, Ute, Read, John, and von Randow, Janet. (2007). Re-training writing raters online: How does it compare with face-to-face training?. *Assessing Writing* 12.2, 26–43.

Kondo-Brown, Kimi (2002). An analysis of rater bias with FACETS in measuring Japanese L2 writing performance. *Language Testing*, 19, 1-29.

Landy, Frank J. and Farr, James L. (1980). Performance rating. *Psychological Bulletin*, 87 (1), 72-107.

Lane, Suzanne and Stone, Clement A. (2006). Performance assessment. In Brennan, Robert L. (ed.). *Educational Measurement*, (4th edition). Westport, CT: American Council on Education and Praeger, 387–431.

Linn, Robert L. and Gronlund, Norman E. (2000). *Measurement and assessment in teaching*. Columbus, OH: Merrill.

McManus, Ian Christopher, Thompson, Margaret and Mollon, Jennifer (2006). Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP (UK) clinical examination (PACES) using multi-facet Rasch modeling. *BMC Medical Education*, 6, 1272-1294. In <<http://link.springer.com/article/10.1186%2F1472-6920-6-42#page-2>> (Accessed 7 November 2018).

Morales, Pedro (1988). *Medición de actitudes en psicología y educación. Construcción de escalas y problemas metodológicos*, San Sebastián: Ttarttalo.

Myford, Carol M. and Wolfe, Edward W. (2004a). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. In Smith, Everett V. y Smith, Richard M. (eds.). *Introduction to Rasch measurement*. Maple Grove, MI: JAM Press, 460–517.

Myford, Carol M. y Wolfe, Edward W. (2004b). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. In Smith, Everett V. y Smith, Richard M. (eds.). *Introduction to Rasch measurement*. Maple Grove, MI: JAM Press, 518-574.

Myford, Carol M., Marr, Diana B., and Linacre, John Michael (1996). *Reader Calibration and its Potential Role in Equating for the Test of Written English* (TOEFL Research Report 52). Princeton, NJ: Educational Testing Service. In <<https://www.ets.org/Media/Research/pdf/RR-95-40.pdf>> (Accessed 7 November 2018).

Park, Taejoon (2004). An investigation of an ESL placement test of writing using Many-Facet Rasch measurement. *Papers in TESOL and Applied Linguistics*, 4, 1-21.

PISA (Programa para la Evaluación Internacional de Alumnos) en la encuesta que se utiliza en los cuestionarios del alumnado de la OCDE de 2012 <https://www.mecd.gob.es/inee/dam/jcr:ea7ff855-4a02-4cc8-880d-170f3e38e465/pisa-2012-cuestionario-del-alumno-b.pdf> y de 2015: <https://www.mecd.gob.es/dctm/inee/internacional/pisa-2015/informebienestar042517.pdf?documentId=0901e72b8249f043>.

Prieto, Juan Miguel (2016). Estudio del comportamiento de los examinadores de la prueba de expresión escrita mediante el modelo *Many-Facet Rasch Measurement* (MFRM) en el contexto de un examen de dominio: el diploma de español nivel A2. Tesis doctoral. En < <https://gredos.usal.es/jspui/handle/10366/128550> (Accessed 7 November 2018).

Prieto, Gerardo (2011). Evaluación de la ejecución mediante el modelo Many-Facet Rasch Measurement. *Psicothema*, 23: 2, 233-238.

Prieto, Gerardo y Nieto, Eloísa (2014). Analysis of rater severity on written expression exam using Many Faceted Rasch Measurement. *Psicológica*, 35, 363-375.

Robbins, Stephen P. (1989). *Organizational behavior*. Englewood Cliffs, NJ: Prentice-Hall.

Rodríguez, José Luis and Tejedor, Francisco Javier (1996). *Evaluación educativa: 1. Evaluación de los aprendizajes de los alumnos*. Salamanca: Ediciones Universidad de Salamanca.

Saal, Frank E.; Downey, Ronald G., and Lahey, Mary A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88 (2), 413-428.

Schriesheim, Chester A., Kinicki, Angelo J., y Schriesheim, Janet F. (1979). The effect of leniency on leader behavior descriptions. *Organizational Behavior and Human Performance*, 23, 1-29.

Stone, Gregory Ethan (2006). Whose criterion standard is it anyway?. *Journal of Applied Measurement*, 7 (2), 160-169.

Sudweeks, Richard R.; Reeve, Suzanne and Bradshaw, William S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, 9, 239-261.

Thorndike, Robert L. and Hagen, Elizabeth P. (1977). *Measurement and evaluation in psychology and education*, New York, NY: John Wiley and Sons.

Thurstone (1928). Attitudes can be measured. *American Journal of Sociology* 33, 529-554.

Tyndall, Belle and Kenyon, Dorry Mann (1996). Validation of a new holistic rating scale using Rasch multi-faceted analysis. En Cumming, Alister H. y Berwick, Richard (eds.), *Validation in language testing*. Clevedon: Multilingual Matters, 39-57.

Watts, Frances and García, Amparo (eds.) (2006). *La evaluación compartida: investigación multidisciplinar*. Valencia: Editorial de la UPV. In <<http://www.upv.es/gie/LinkedDocuments/descargar%20libro.pdf>> (Accessed 7 November 2018).

Wolfe, Edward W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, 46, 35-51.

Wolfe, Edward W. (2009). Item and rater analysis of constructed response items via the multi-faceted Rasch model. *Journal of Applied Measurement*, 10, 335-447.

Yorozuya, Ryuichi and Oller, John W., Jr. (1980). Oral proficiency scales: Construct validity and the halo effect. *Language Learning*, 30 (1), 135-153.

### **Supervision and Security: Good Practices to Guarantee the Success of Exams**

Alderson, J.C., Clapham, C. & Wall, D. (1995) *Language Test Construction and Validation*. Versión española: *Exámenes de idiomas*, Madrid (1998): Cambridge University Press.

ALTE (1993): *Content Analysis Checklists for Speaking and Writing*

ALTE (1994): *Code of Practice* <https://www.alte.org/Materials>

ALTE (1998): *Multilingual Glossary of Language Testing Terms, Studies in Language Testing 6*, UCLES/Cambridge University Press.

ALTE (2001): *Principles of Good Practice* <https://www.alte.org/Materials>

ALTE(2005): *Item-writer Guidelines*

[http://www.alte.org/attachments/files/item\\_writer\\_guidelines.pdf](http://www.alte.org/attachments/files/item_writer_guidelines.pdf)

ALTE Q-mark terms of use

[http://www.alte.org/setting\\_standards/the\\_alte\\_q\\_mark\\_questions\\_and\\_answers](http://www.alte.org/setting_standards/the_alte_q_mark_questions_and_answers)

ALTE (2013): *Quality Assurance Checklists* <http://www.alte.org/resources/filter>

Bachman, L. F. (1990): *Fundamental Considerations in Language Testing*, Oxford, Oxford University Press.

Bachman, L. F. (1990): *Fundamental Considerations in Language Testing*, Oxford, Oxford University Press.

Bachman, L. F. and Palmer, S. (1996): *Language Testing in Practice*, Oxford. Oxford University Press.

Baker, D. (1989): *Language Testing*, London, Edward Arnold.

Bordón, T. (2006): *La evaluación de la lengua en el marco de E/L2: Bases y procedimientos*. Madrid, Arco Libros.

Bordón, T. (2004): "Panorama histórico de algunas de las cuestiones fundamentales en la evaluación de segundas lenguas", *Carabela*, 55, Madrid, SGEL, págs. 5-30.

Council of Europe (2009): *Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment*



(CEFR). [http://www.coe.int/t/dg4/linguistic/Source/ManualRevision-proofread-FINAL\\_en.pdf](http://www.coe.int/t/dg4/linguistic/Source/ManualRevision-proofread-FINAL_en.pdf)

Council of Europe (2011): *Manual for Language Test Development and Examining*  
[http://www.coe.int/t/dg4/linguistic/ManualLanguageTest-Alte2011\\_EN.pdf](http://www.coe.int/t/dg4/linguistic/ManualLanguageTest-Alte2011_EN.pdf)

Dutch CEFR Construct Project (2004): *The CEFR Grid for the analysis of Reading and Listening*

Fulcher, G. (2010) *Practical Language Testing*. London: Hodder Education.

Gardner, J., Harlen, W., Hayward, L., and Stobart, G. (2008): *Changing Assessment Practice: Process, Principles and Standards*. Cambridge: University of Cambridge Faculty of Education.

Harris, M. and McCaan, P. (1996): *Assessment*. London: Heinemann.

Harrison, A. (1983): *A Language Testing Handbook*, London, MacMillan.

Kunnan, A. J. (2017): *Evaluating language assessments*. New York: Routledge.

Kunnan, A. J. (1995): *Test taker characteristics and test performance: a structural modelling approach*, Cambridge, University of Cambridge Local Examinations Syndicate and Cambridge University Press.

Lado, R. (1961): *Language Testing*, London, Longman.

Lee, Y. P. et alii. (eds.) (1985): *New Directions in Language Testing*, Oxford, Pergamon.

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

Messick, S. (1989): "Meaning and Values in test Validation: the Science and Ethics of Assessment. *Educational Researcher*, 18(2), 5-11.

Puig, F. (2008): *Evaluación*. Monográficos MarcoELE (7).

Weir, C. J. (1993): *Understanding and Developing Language Test*. New York: Prentice Hall.

### **What is a Fair Test in a Multi-Cultural Society?**

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014) *Standards for Educational and Psychological Testing*. Washington D.C.: American Educational Research Association.

Gee, J. P. (2003) Opportunity to learn: a language-based perspective on assessment, *Assessment in Education*, 10(1), 27-46.

Gipps, C. & Stobart, G. (2010) Fairness in B. McGraw, E. Baker, & P. Peterson (Eds) *International Encyclopedia of Education*, 3<sup>rd</sup> Edition, 56-60, Elsevier.

Secada, W. G. (1989) Educational equity versus equality in education: An alternative conception. In W. G. Secada (ed) *Equity and Education*, 68-88. New York: Falmer

Stobart, G. (2005) Fairness in multicultural assessment systems, *Assessment in Education*, 12, 3, 275-287.

### **Immigration and Language Testing**

ALTE (2016). *Language tests for access, integration and citizenship: an outline for policy makers*

Blackledge, A. (2005). *Discourse and Power in a Multilingual World*. Amsterdam: John Benjamin Publishing.

Extra, G, Spotti, M. and van Avermaet, P. (eds). (2009). *Migration, Citizenship and Language Testing*. London; New York: Continuum.

Khan, K., & McNamara, T. (2017). Immigration law, citizenship and testing. In S. Canagarajah (ed). *Language and Migration*. 451-467. Abingdon: Routledge.

Löwenheim, O., and Gazit, O. (2009). Power and Examination: A critique of citizenship tests. *Security Dialogue*, 40(2), 145-167.

McNamara, T. and Roever, C. (2006). *Language Testing: The Social Dimension*. Oxford: Blackwell Publishing.

McNamara, T. and Ryan, K. (2011). Fairness Versus Justice in Language Testing: The Place of English Literacy in the Australian Citizenship Test. *Language Assessment Quarterly*, 8 (2): 161-178.

Turner, J. (2014). Testing the liberal subject: (in)security, responsibility and 'self-improvement' in the UK citizenship test. *Citizenship Studies*, 18(3-4), 332-348.

### **Language and Security**

Khan, K. 2017. "Citizenship, securitization and 'suspicion' in UK ESOL policy." In Karel Arnaut, Martha Sif. Karrabæk & Massimiliano Spotti (eds). *Engaging with Superdiversity*. 303-320. Clevedon: Multilingual Matters.

Charalamabous, C., Charalambous, P., & Rampton, B. (2015). "De-securitizing Turkish: Teaching the Language of a Former Enemy, and Intercultural Language Education." *Applied Linguistics*: amv063.

Karrabæk, M., & Ghandchi, N. (2017). The Very Sensitive Question. *Pragmatics and Society*, 8: 1: 38-60.

Keefe, P.R. (2007). "The Challenge of Global Intelligence Listening." In L. K. Johnson. (ed). *Strategic Intelligence 2. The Intelligence Cycle: The Flow of Secret Information from Overseas to the Highest Councils of Government.* 23-39. London: Praeger Security International.

Pratt, M..L. (2004). Language and National Security: Making a New Public Commitment. *Modern Language Journal*, 88(2), 289-291.