Development of a Spanish generic writing skills scale for the Colombian Graduate Skills Assessment (Saber Pro)

Ana Maria Ducasse
RMIT University, La Trobe University
Kathryn Hill
La Trobe University

While many higher education institutions list the generic skills their graduates are intended to acquire during a course of study (Barrie, 2006), the relevant skills are rarely directly assessed at graduation. In Colombia, exit assessment is compulsory for all post-secondary training and education. To this end, a Spanish-language version of the Australian Graduate Skills Assessment (GSA) was developed for the Colombian context. However, problems were identified with the reliability of the Spanish version of the GSA writing scale.

This paper describes the process of replacing the original version of the Spanish-language version of the GSA scale (an intuitively based writing scale) with an empirically based scale developed using a question tree method. Forty raters constructed two holistic (combined trait) and three analytic (individual trait) writing scales using benchmarked scripts from a previous test administration. The five scales were then trialled. Comparison of the scales showed the eight-level holistic scale provided the widest distribution of scores.

This research provides insights into generic writing skill testing for higher education graduates in Colombia. In addition, the study uniquely provides a detailed description of the development of empirically-based analytic and holistic scales for assessing the writing of Spanish-L1 speaking graduates in Colombia.

Key words: writing assessment, writing scales, performance decision trees, generic skills, Saber Pro

Ana Maria Ducasse; Email: anamaria.ducasse@RMIT.edu.au.

Introduction

While many higher education institutions list the generic skills their graduates are intended to acquire during a course of study (Barrie, 2009), the relevant skills are rarely directly assessed at graduation. In Colombia, however, national exit assessment is mandated for all post-secondary training and education and is an emerging area of research (Delgado-Ramírez, 2012; Garizabalo Dávila, 2012; Gil et al., 2013). This national testing forms part of a quality assurance system that underpins the assessment of educational outcomes (OECD, 2013). The Graduate Skills Assessment (GSA) (ACER, 2001), an English-language test originally designed to test Australian university graduate skills, was modified and translated into Spanish in co-operation with the Australian test developers at the Australian Council for Educational Research (ACER) and the Colombian test adapters at the Colombian Institute for the Evaluation of Education (ICFES 2011a,). The new Colombian test is known as Saber Pro. This paper reports on a further revision of the Saber Pro writing scales commissioned after previous revisions and adaptations of the writing task and assessment had not produced the expected mark distribution.

The GSA was developed based on interviews with stakeholders in academia and business regarding the skills required to enter the workforce or to continue with further studies. It consists of four sections designed to measure a combination of generic skills that higher education students are expected to develop: problem solving, critical thinking, interpersonal skills and written communication. The written communication section of the GSA comprises two tasks, an expository and an argumentative text on general topics, and is rated using a 5-level, 3-trait analytic scale (Appendix 1). These tasks were selected because they elicit generic skills such as obtaining, analysing, organizing and communicating ideas and information. Students have 70 minutes to complete the GSA, consisting of 10 minutes for reading the stimuli and 30 for producing each text.

In the original Spanish-language version of Saber Pro, writing performance was rated on a six-level scale comprising structure, language and communicative intent. However, a study carried out during an early administration of the test found raters had difficulty differentiating between the middle levels of the scale (Ducasse, 2011). ICFES commissioned a project to replace the original Spanish-language version of the Saber Pro writing scale with a more reliable scale. The following section is a summary of the rating scale construction methodology used to generate the revised scales.

Rating scale construction

Alternative methods of scale development include 'expert judgement' (or 'intuitive'), 'measurement-driven' and 'performance-'or 'data-driven' approaches. Scales based on expert judgement have been criticised for lacking empirical underpinnings (e.g. Fulcher, 2003) in that they have not evolved from, and are not connected to, the language sample elicited in the test. Yet, developing scales by consensus with a committee of experts persists as a common way of developing intuitive scales. Scales of this type, by definition, are not based on candidate performance; scale developers make assumptions about the candidates' performance before the test is taken. However, experience of using the scale over time can result in scales being revised and adapted.

In contrast, measurement-based scales are empirical in nature. Those developed with Rasch modelling, for example, use scaling of descriptors as in the Common European Framework of Reference (CEFR) where scales are set to a pre-determined number of levels (see North 1996, 2000). These 'empirical' scales are not, however, based on candidate performance but on an analysis of existing scales. For this reason Fulcher, Davidson and Kemp (2011) argue that "measurement-driven approaches generate impoverished descriptions of communication" (p.1). In contrast, performance- or data-driven approaches are entirely based on written or spoken candidate discourse, "have the potential to provide richer descriptions that offer sounder inferences from score meaning to performance in specified domains" (Fulcher, et al, p.1). Discourse is used for ranking scripts to levels that emerge before deciding on level cut-off points. Clearly defined levels with shared interpretation by the raters (Brindley, 1998) help examiners arrive at qualitative writing decisions (Shaw & Weir, 2007). According to North (2003), the advantage of a data-driven scale is that 'the development of the categories tends to involve detailed investigation and discussion of the performances involved and the categories selected as a result are embedded in the context concerned' (p. 2). Focus on actual performance is a strength of scales developed in this way.

Clearly defined levels of proficiency in a scale facilitate rater training and rater application of scales, and are an important goal for test developers during scale development. Use of concrete divisions using criterial questions (e.g., *"does the text use paragraphs?"*) is one way to achieve this. The main argument for using concrete divisions between bands for each level is that the scales are formulated in concrete terms based on language use. The concrete formulations can be qualitative, such as those used in the empirically derived, binary-choice, boundary-definition (EBB) scales developed by Upshur and Turner (1995;

1999), Turner (2000) and Turner and Upshur 1996, 2002) for writing and speaking; or quantitative, such as the discriminant analysis used by Fulcher (1993; 1996; 2003) to develop accuracy/fluency scales.

Another consideration for scale design is whether to develop holistic and/or analytic scales for the rating context. Both are well described and distinguished in the literature on second language (L2) writing assessment (e.g. Weigle, 2002). For holistic scoring the rater uses a single scale to assign an impressionistic mark based on their perception of the overall performance. For analytic scoring, separate scores are awarded for multiple traits on which the raters focus during the performance. Moving from L2 to first language (L1) writing research, Hamp-Lyons (1995) contends that a holistic scale is more appropriate for scoring L1 essays; when "a student's essay is internally congruent, and the qualities of the writing … may be adequately represented by a single score" (p. 760). This position is presented in contrast to an L2 diagnostic context, where students' performance on individual traits provides more information for score interpretation. Schoonen (2005) examined analytic and holistic scoring for language and organisation in L1 essays; using Generalizability Theory (e.g. Chiu, 2001), he found that tasks contributed more to variance than raters, and the effect sizes (for task) were greater for analytic (cf. holistic) ratings.

In their article describing a large study in which they revised speaking scales in the Cambridge suite of tests, Taylor and Galaczi (2011) explained the importance of combining intuitive and data-driven scale construction methodologies and presented a scale construction model for adaptation in different contexts. In addition to being open to a combination of methodologies to achieve scoring validity, they emphasised the value of experience and expertise that experts can bring to the revision or development process. However, despite the amount that could be learned from experts' reflections at different stages of the revision process, information on how individuals or committees come to agreement on descriptors for rating scales is rarely published, nor is detailed information regarding revision stages during scale development (Barkhoui, 2010; Fulcher, 2003; Knoch, 2011). The intention of this paper is to address this lack of information on test revision by detailing the most recent stage of scale development for the Saber Pro.

Rating scale development for this study

Table 1 provides an overview of the three-phase Saber Pro development process, based on Galaczi, Ffrench, Hubbard, and Green (2011), detailing the chronological and methodological dimensions that informed their process of test revision: aims, procedures, outcomes, data, data analysis, results and key decisions. A summary of the first two phases of the three-phase Saber Pro development is provided for context in columns two and three of Table 1. The study described in this paper (Phase 3) is summarised in the final column of Table 1. The particular focus of this paper is the pilot stage of Phase 3, conducted in the first half of November, 2011, which was to trial the EBB methodology using local raters and scripts. Once this pilot phase was completed, an expert group of raters was tasked with using the methodology to independently develop five evidence-based scales, which were then trialled in order to select the best performing scale. Note that for Phase 3, the decision was made to reduce the two writing tasks to a single, expository task (see Data). The prompt comprises a context and a question. Candidates write a one to three page response over thirty minutes.

Table 1. An overview of development of the Saber Pro writing scale.

Saber Pro	Phase 1	Phase 2	Phase 3
Aims	Trial the Spanish version of the GSA scale (5-level analytic scale with 3 traits)	Revise Phase 1 scale to improve discrimination	Develop 5 evidence-based scales Selection of the best performing scale
Procedures	Rater training & moderation	Local experts use test scripts to revise the scale. Rater training & moderation	Raters training (EBB question tree scales) Raters develop 5 EBB scales
Outcome	-	6-level holistic scale	3 x single-trait analytic scales 1 x 7-level holistic scale 1 x 8-level holistic scale
Data	Writing scripts (2 tasks)	Writing scripts (2 tasks)	Writing scripts (1 task)
Data analysis	Rasch FACETS Candidate scores Rater performance	Rasch FACETS Candidate scores Rater performance Task performance	Rasch FACETS Candidate scores Rater performance Task performance Scale comparison
Results	Unsatisfactory discrimination	Unsatisfactory discrimination	8-level holistic scale provided best discrimination
Key Decisions	Convert the trait-based grid into a holistic 6 band scale	Develop 5 new EBB scales (3 single trait & 2 holistic)	Trial the selected scale at next test administration

As a result of ongoing issues with discrimination in the scales developed in Phases 1 and 2, the ICFES psychometricians recommended the development of a set of five scales (three analytic and two holistic) based entirely on the candidate scripts with the aim of selecting the scale which produced the widest distribution of scores (Phase 3). The analytic scales were designed to investigate each of the three traits in the Phase 1 scale separately and the holistic scales (with 7 and 8 bands respectively) were designed to investigate whether lengthening the scale would improve discrimination.

The rating scale development procedure used was derived and adapted from the method known as EBB (Upshur & Turner, 1995), described in the scale construction section above. This methodology uses samples of candidate scripts to elicit judgements of the differences between levels. The levels are defined in terms of yes/no questions, known as 'criterial questions' because the criteria used to separate levels are embedded within them. Therefore, using this methodology it is not possible to specify the number of levels in advance. This EBB procedure reflected the features of written Spanish attended to by trained Saber Pro raters in their analysis of scripts of writing performance on tasks from three earlier national test administrations from 2010 and 2011.

Question tree development workshop

A group of 40 of the 78 ICFES raters who met the conditions of intra and inter-rater reliability during Phase 2 (Table 1) were selected, inter alia, on the basis of experience and performance as raters in Phases 1 and 2 for involvement in a pilot study. The purpose of the pilot study was to investigate whether they could be trained as scale developers using an application of Turner and Upshur's (2002) EBB method, called the binary question tree technique (hereafter 'question tree').

The aim of the methodology was for raters to verbalise the most salient difference between levels of writing by using a set of criterial questions to split sets of scripts into two, i.e., 'Does the writing response have paragraphs?' Next, the 'no paragraphs' pile might be divided using the question 'Does it have run-on sentences?' and so on until no more divisions can be made. After applying these questions, raters can quickly determine a rating and assign scripts to levels.

A set of 24 Phase 2 benchmark scripts (selected to represent each of the six band levels of the Phase 2 scale) were selected for the workshop. After the methodology was explained, the raters were divided into five groups of eight, each responsible for developing a different scale, analytic (single trait) or holistic (Table 2) with researcher 1 in the role of facilitator.

Table 2. Scale development groups

	Group 1 *(n=8)*	Group 2 *(n=8)*	Group 3 *(n=8)*	Group 4 *(n=8)*	Group 5 *(n=8)*
Scale type	Analytic (*Structure*)	Analytic (*Language*)	Analytic (*Communicative intention*)	Holistic	Holistic

Each group was divided into a further two subgroups. Each subgroup of four was then assigned 12 of the 24 benchmark scripts (randomly selected) with each subgroup potentially receiving a different mix of scripts. Each subgroup then worked independently to rank the 12 assigned scripts. This step was introduced to ensure that the participants were very familiar with at least half of the benchmark scripts. The two subgroups then reconvened to discuss and jointly rank the full set of 24 scripts and divide them into two sets of 12, representing levels of performance above and below the midpoint (nominally bands 3 and 4 on the Phase 2 scale, with band 4 representing a 'pass'). In deciding on the ranking, the groups responsible for each of the three analytic scales were directed to focus exclusively on the target trait (e.g., *structure*) while the two holistic scale groups worked without any predetermined criteria (i.e., drawing on their previous experience as ICFES raters). As each group based their judgements on different criteria, there was not necessarily any cross-group agreement about how the 24 benchmark scripts were ranked.

At the end of the ranking process the groups were asked to discuss their rankings with the aim of making explicit the features which distinguished the scripts falling above and below the midpoint, consistent with their assigned scale type. For example, the group using the analytic scale for 'structure' focussed exclusively on 'structure' in their discussions. The next task for each group was to use the results of their discussions to develop a set of concrete criterial questions (e.g., "Do they use paragraphs?"), which could be used to assign a given script to one level or the other. These questions were then placed on a question tree. At the end of this process, each group then checked their question tree against the 24 benchmark scripts and revised and rephrased, as necessary.

Finally, the whole group of 40 participants came together to discuss and to reach consensus on a final version of each of the five scales, with facilitation by researcher 1. By this stage all the participants were very familiar with the features of the 24 benchmark scripts. Each scale was accompanied by the criterial yes/no questions, band level descriptors and the script identification numbers corresponding to each band. The whole process took approximately four hours to complete. This process is summarised in Figure 1.

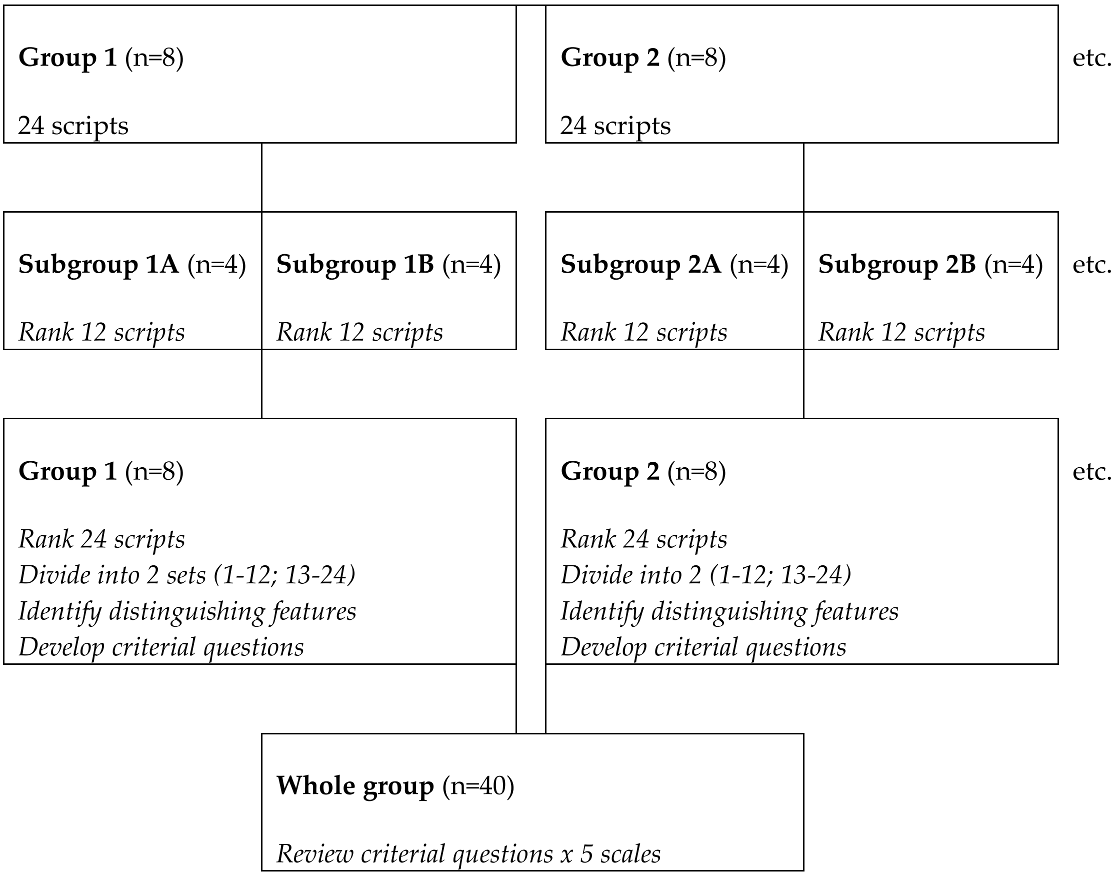


Figure 1. Question-tree development process

Figure 2 shows a question tree for an eight-level holistic scale that was developed in the pilot workshop where the numbers of the matching benchmark scripts are matched to the relevant levels. If the answer to the question for Band 5 (*Does the text address the question in a coherent manner?*) is 'yes', the rater goes 'up' the question tree to the question for Band 6. If the answer to this question is 'no', Band 6 is awarded, and so on.

<table>
<tr><td colspan="4">¿El texto plantea una idea central de manera estructurada?
Is there a structured main idea?</td></tr>
<tr><td colspan="2">Yes 5</td><td colspan="2">No 4</td></tr>
<tr><td colspan="2">¿Hay progresión temática y una estrategia de apoyo? (Is there evidence of topic development and support for the ideas?)</td><td colspan="2">¿El texto mantiene un tópico pero su estructura es incompleta o fragmentada? (Does the script respond to the task but with a faulty structure?)</td></tr>
<tr><td>No 5</td><td>Yes 6</td><td>Yes 3</td><td>No 4</td></tr>
<tr><td colspan="2">¿El texto desarrolla un plan argumental por medio de diversos recursos semánticos y sintácticos (razonamientos ejemplos citas etc.) (Does the text use different semantic and syntactic resources to develop the topic?)</td><td colspan="2">¿Las ideas u opiniones se pierden por problemas serios del uso del lenguaje? (Are the main ideas lost due to problems with expression?)</td></tr>
</table>

No 6	Yes 7	Yes 2	No 3
¿El texto tiene unidad global a través de la relación adecuada entre sentido y forma? (asociaciones, inferencia reflexiones, fluidez estilo propio etc.) (Is the text unified by meaning and form through inferences, reflections, flow and personal style?)		¿ El texto responde al planteamiento propuesto y o se limita a la copia literal del estímulo? (Is the response limited to copying the stimulus?)	
No 7	Yes 8	Yes 1	No 2

Figure 2. Eight-level holistic scale question tree developed in the pilot workshop

Formal feedback from raters about their experience of this process was collected via a questionnaire and will be reported on in another paper. However, as workshop facilitator, researcher 1 noted the participants working confidently with the benchmark scripts as they became increasingly familiar with the discourse with each stage of the process. As participants justified their rankings or defended their question trees and scale descriptors they were also able to refer to scripts that the whole group were familiar with from previous rating and training sessions (during Phases 1 and 2). This indicates that experienced raters were able to build on their previous experiences of rating when identifying levels, asking criterial questions to discriminate between levels and matching benchmark texts to the newly defined levels in Phase 3.

The pilot workshop demonstrated to the ICFES test developers that the Colombian raters could produce scales using the question tree method. This led to a second, more intense 2-day scale development workshop to develop the final set of scales for trialling and comparison. This involved 30 raters (selected from the group of 40 who participated in the pilot phase) working under the supervision of a senior rater and an ICFES test developer. Details of this process are reported elsewhere (ICFES, 2011b). (An example of the single trait analytic scale for 'communicative intent' is provided in the Appendix 2).

Results from the methodology trial, and the scale development workshop that followed, showed that a yes/no question tree scale could be successfully created by expert raters given appropriate training.

Over twelve days fifty-four raters were trained to apply each new scale then rated 135 benchmark scripts thus using each of the five scales resulting from the intense scale development workshop.. Data analysis and scale comparison (which also included the 54 raters using the phase 2 descriptive holistic scale) were conducted by ICFES and so the details are not reported here. (A full report of the analyses can be found in ICFES 2011c and Cuchimaque, Ordóñez,

& Pardo, 2012). The results of the comparison showed that the eight level holistic scale yielded the best discrimination and rater reliability.

The chosen scale is described in Table 3. In contrast to the approach described in Turner and Upshur (2002) and used in the pilot study, where raters start from the middle of the question-tree, for final scale (Table 3) raters ask the questions in sequence starting from the first question at the top of the page (i.e., *Is the task attempted but the problems…?*) and move downwards.

Table 3. The eight-level holistic scale (with English translations)

8 LEVEL HOLISTIC DATA DRIVEN SCALE	Score
¿Se aborda la tarea pero los problemas en el manejo de la convención escrita o el desarrollo insuficiente impiden la comprensión y/o la valoración del escrito? Is **the task attempted but the problems with use of writing conventions or insufficient topic development impede comprehension and or appraisal of the text?**	No (0) Si (1)
¿Se expresan ideas aunque sean desarticuladas, incongruentes y/o con un mínimo desarrollo? **Are ideas expressed even though unlinked, incongruent and or minimally developed?**	No (1) Si (2)
¿Se evidencia una intención comunicativa pese a tener problema de organización (fragmentación y/o repetición) **Is there evidence of a communicative intention despite problems in organization (fragmentation and or repetition?**	No (2) Si (3)
¿Elabora un texto estructurado (elemental) con un planteamiento básico haciendo uso aceptable del lenguaje? **Is there a (basic) structured text taking a position that makes use of appropriate language?**	No (3) Si (4)
¿Hay una unidad y progresión temática aunque pueda presentar fallos en alguna parte de la estructura? **Is there a single idea and a thematic progression despite there being faults in a part of the structure?**	No (4) Si (5)
¿Hay conexiones lógico semánticas consistentes entre las proposiciones del texto? **Are there consistent logical semantic connections between the propositions in the text?**	No (5) Si (6)
¿Hay una estrategia comunicativa satisfactoria que se evidencia en un plan textual? **Is there a satisfactory communication strategy that can be seen through planning in the text?**	No (6) Si (7)
¿Hace un uso contundente de uno o varios recursos estilísticos semánticos pragmáticos y del manejo del lenguaje? **Is there a solid use of one or two pragmatic and semantic stylistic devices and use of expression?**	No (7) Si (8)

In summary, descriptors for scales based on actual performance and question tree methodology were found to be a valid means of achieving rater reliability and widening the distribution of scores across all levels. The eight-level holistic scale has been used for test administrations in Colombia since November 2011, and further government reports on this national testing program are pending (http://www2.icfes.gov.co/resultados/Saber Pro). Unfortunately, until now there has been little published research in the area of comparisons of analytic versus holistic scales in L1 generic skills essay scoring. However, Hamp-Lyons' (1995) contention that scoring L1 essays holistically is more appropriate for L1 English College essays is consistent with evidence for Spanish L1 essays in this study.

Conclusion

This paper describes the lengthy and rigorous process of revision and testing of scales being used for assessing Spanish L1 exit level writing skills in Colombian post-secondary education and training. Following ongoing issues with scale discrimination, expert local raters used benchmark student writing samples and a question tree methodology to successfully develop five different scales for trialling and comparison, resulting in the selection of an 8-level holistic scale.

Some limitations of the question-tree scale development method have already been pointed out by its creators (Upshur & Turner, 1999; Turner & Upshur, 2002). They remind us that the success of the data-driven scale development procedure depends on the range of tasks (and discourses) available as input, as well as the skill and experience of the scale makers. Hence, as the scale was developed using scripts from a specific set of tasks, it cannot be generalised to any other test or task. Nevertheless, in this instance, the advantages of improving score distribution, and potentially easier training and faster marking, outweigh the fact that the scales are not necessarily transferable. Furthermore, as the participants in the scale development were expert raters with experience in marking L1 Spanish-language writing tests a strong commitment to group moderation processes the success demonstrated in this project may not necessarily be reproduced with a more inexperienced group.

Finally, the process described in this paper demonstrates the importance of ensuring externally-derived instruments, such as the GSA, are appropriate for local settings. It has particular relevance for tests that are translated into other languages without further modifications. That is, simple translation, without reference to examples of actual performance in the new context, may not meet the requirements of the test adopters. In contrast, the scale described in this study is grounded in locally produced scripts written in the local language and the tacit knowledge of local expert markers and, it is argued, provides a better representation of the construct (graduate-level writing in Colombia) than the imported version.

The final ICFES public report on the psychometric properties of the writing task and rating scales is still in preparation. Ongoing validation will be conducted as part of routine test administration procedures at ICFES. As previously mentioned a further paper is planned reporting on the results of the rater feedback survey for Phase 3. In addition, following calls for more research into how agreement on rating scale descriptors is achieved (Barkhoui, 2010; Fulcher, 2003; Knoch, 2011) a further paper analysing rater discussions recorded during the whole group training and moderation process is planned.

Acknowledgements

The first author would like to thank ICFES, ACER and La Trobe University for their institutional support during the preparation for Phase 3 of the revision process and the two anonymous reviewers who provided invaluable feedback on earlier drafts of this article.

References

Australian Council for Educational Research ACER. (2001). *Graduate Skills Assessment Summary Report 01/E,* Occasional Paper Series Higher Education Division. Melbourne: Department of Education, Training and Youth Affairs. Retrieved from http://www.acer.edu.au/gsa/test-reports

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

Barrie, S. C. (2006). Understanding what we mean by the generic attributes of graduates. *Higher education, 51*(2), 215-241.

Brindley, G. (1998). Describing language development? Rating scales and second language acquisition. In Bachman, L. F. And Cohen, A. D. (Eds.), Interfaces between second language acquisition and language testing research (pp. 112–140). Cambridge: Cambridge University Press.

Chiu, C.W.C. (2001). *Scoring performance assessments based on judgements: generalizability theory*. New York: Kluwer.

Cuchimaque, E., Ordóñez, C., & Pardo, C. (2012). *Evaluación de la escritura en el examen de estado de evaluación de la calidad de la educación superior en Colombia: La experiencia de ICFES.* Bogotá: ICFES

Delgado-Ramírez M. B. (2012) Examen de Estado de la Calidad de la Educación Superior -Saber Pro ¿Qué indican sus resultados? *Revista Colombiana de Anestesiología*. http://dx.doi.org/10.1016/j.rca.2013.06.005

Ducasse, A. M. (2011). Spanish writing: What do raters look for? In I. Candel Torres, L. Gómez Chova, A. López Martínez (Eds.) (pp. 4528-4538). ICERI 2011 Proceedings CD. www.iated.org Madrid: IATED.

Fulcher, G. (1993). *The construction and validation of rating scales for oral tests in English as a foreign languag*e (Unpublished PhD Dissertation). University of Lancaster, UK.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*(2), 208–238.

Fulcher, G. (2003). *Testing second language speaking*. London: Pearson Longman.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5–29. doi:10.1177/ 0265532209359514

Galaczi, E. D., Ffrench, A. Hubbard, C., & Green, G. (2011). Developing assessment scales for large-scale speaking tests: a multiple-method approach, *Assessment in Education: Principles, Policy & Practice, 18*(3), 217–237, doi:10.1080/0969594X.2011.574605

Garizabalo Dávila, C. M. (2012). Estilos de aprendizaje en estudiantes de enfermería y su relación con el desempeño en las pruebas SABER Pro. *Journal of Learning Styles,* 5(9) 97-110.

Gil, F. A., Rodríguez, V.A., Sepúlveda, L. A., Rondón, M.A, & Gómez-Restrepo, C. (2013). Impacto de las facultades de medicina y de los estudiantes sobre los resultados en la prueba nacional de calidad de la educación superior (SABER PRO)*Revista Colombiana de Anestesiología*, *41*(3) 196-204.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241–276). Norwood, NJ: Ablex.

Hamp-Lyons, L. (1995). *Uncovering possibilities for a constructivist paradigm for writing assessment,* College Composition and Communication, *46*(3), 446–455.

ICFES. (2011a). *Prueba de Habilidades Genéricas GSA Colombia: Resultados del pilotaje* Bogotá: ICFES.

ICFES. (2011b). *Proceso de construcción y aplicación de propuestas de rejillas y/o escalas de evaluación para la prueba 'Saber Pro,* Bogotá: ICFES.

ICFES. (2011c). *Diseño Experimental para la Evaluación de la Calificación de la Prueba Escrita en la Prueba GSA,* Bogotá: ICFES.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behaviour—a longitudinal study. *Language Testing*, *28*(2), 179–200. doi:10.1177/0265532210384252

North, B, (2000). *The development of a common framework scale of language proficiency.* New York: Peter Lang.

North, B. (1996). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. In A. Huhta, V. Kohonen, L. Kurki-Suonio, and S. Luoma (Eds.), *Current developments and alternatives in language assessment. Proceedings of the LTRC 1996* (pp. 423–447). Jyvaskyla: University of Jyvaskyla Press.

North, B. (2003). *Scales for rating language performance: Descriptive models, formulation styles, and presentation formats*. TOEFL Monograph Series, *24*. Princeton, NJ: Educational Testing Service.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales, *Language Testing, 15*(2), 217–262. doi: 10.1177/026553229801500204

OECD. (2013). Quality and relevance of tertiary education in Colombia. In OECD/IBRD/ The World Bank, *Reviews of National Policies for Education: Tertiary Education in Colombia 2012,* OECD Publishing. DOI: 10.1787/9789264180697-7-en

Schoonen, R. (2005). Generalizability of writing scores: an application of structural equation modelling. *Language Testing 22*, 1–30.

Shaw, S. D., & Weir, C. J. (2007). *Examining Writing*. Cambridge: Cambridge University Press.

Taylor, L., & Galaczi, E. D. (2011). Scoring validity. In L. Taylor (Ed.), *Examining Speaking* (pp. 171-233). Cambridge, UK: Cambridge University Press.

Turner, C. E. (2000). Listening to the voices of rating scale developers: Identifying salient features for second language performance assessment. *Canadian Modern Language Review*, *56*(4), 555–584.

Turner, C. E., & Upshur, J. A. (1996). Developing rating scales for the assessment of second language performance. In G. Wigglesworth & C. Elder (Eds.), *The language testing cycle: From inception to washback* (pp. 55–79). Melbourne, Australia: Applied Linguistics Association of Australia.

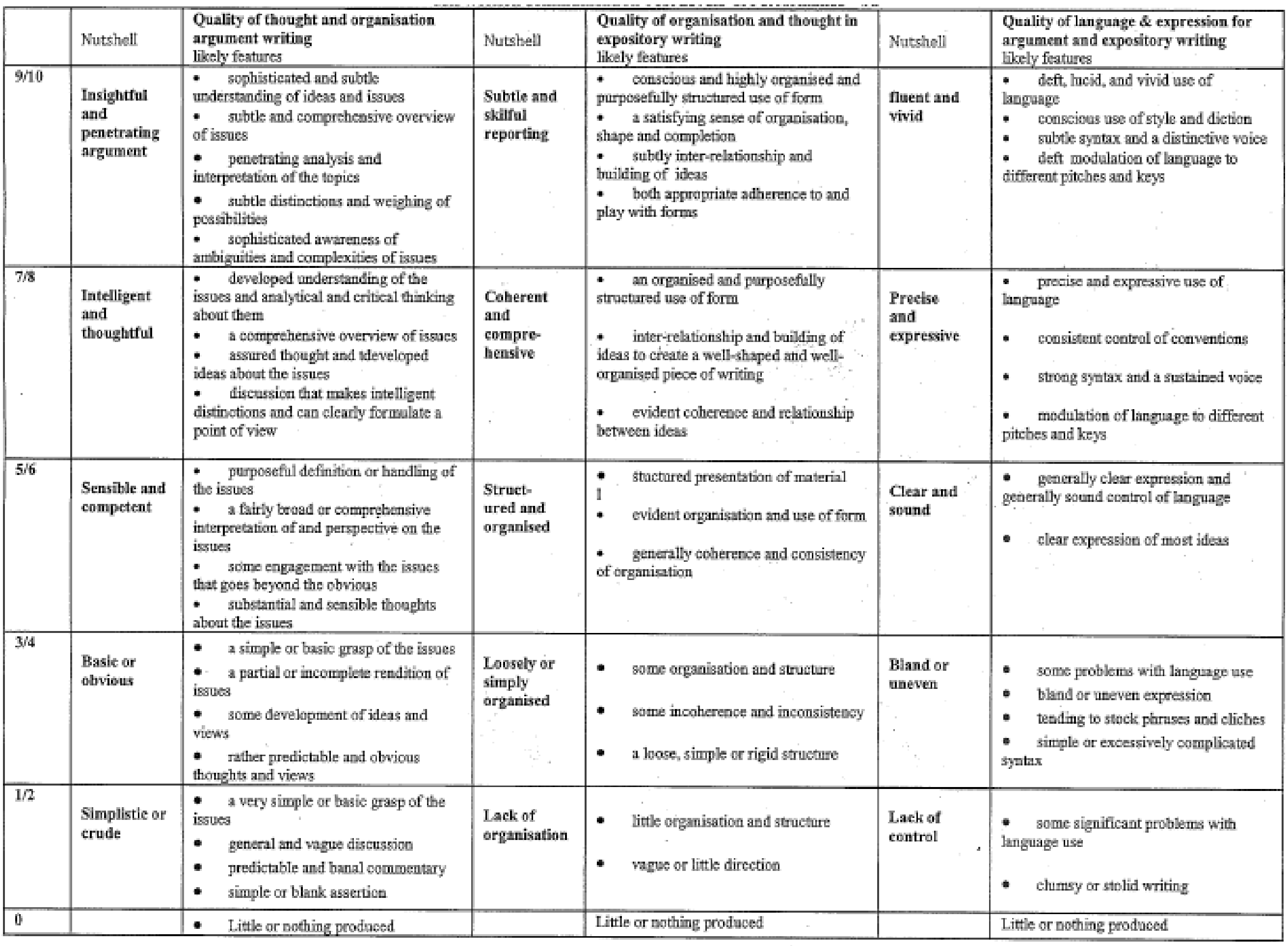
Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly*, *36*(1), 49–70.

Upshur, J., & Turner C. E. (1995). Constructing rating scales for second language tests. *English Language Teaching Journal*, *49*(1), 3–12.

Upshur, J., & Turner C. E. (1999). Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing*, *16*(1), 82–111.

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

Appendix 1 The GSA (English) scale



	Nutshell	**Quality of thought and organisation argument writing** likely features	Nutshell	**Quality of organisation and thought in expository writing** likely features	Nutshell	**Quality of language & expression for argument and expository writing** likely features
9/10	**Insightful and penetrating argument**	• sophisticated and subtle understanding of ideas and issues • subtle and comprehensive overview of issues • penetrating analysis and interpretation of the topics • subtle distinctions and weighing of possibilities • sophisticated awareness of ambiguities and complexities of issues	**Subtle and skilful reporting**	• conscious and highly organised and purposefully structured use of form • a satisfying sense of organisation, shape and completion • subtly inter-relationship and building of ideas • both appropriate adherence to and play with forms	**fluent and vivid**	• deft, lucid, and vivid use of language • conscious use of style and diction • subtle syntax and a distinctive voice • deft modulation of language to different pitches and keys
7/8	**Intelligent and thoughtful**	• developed understanding of the issues and analytical and critical thinking about them • a comprehensive overview of issues • assured thought and ideveloped ideas about the issues • discussion that makes intelligent distinctions and can clearly formulate a point of view	**Coherent and compre-hensive**	• an organised and purposefully structured use of form • inter-relationship and building of ideas to create a well-shaped and well-organised piece of writing • evident coherence and relationship between ideas	**Precise and expressive**	• precise and expressive use of language • consistent control of conventions • strong syntax and a sustained voice • modulation of language to different pitches and keys
5/6	**Sensible and competent**	• purposeful definition or handling of the issues • a fairly broad or comprehensive interpretation of and perspective on the issues • some engagement with the issues that goes beyond the obvious • substantial and sensible thoughts about the issues	**Struct-ured and organised**	• stuctured presentation of material • evident organisation and use of form • generally coherence and consistency of organisation	**Clear and sound**	• generally clear expression and generally sound control of language • clear expression of most ideas
3/4	**Basic or obvious**	• a simple or basic grasp of the issues • a partial or incomplete rendition of issues • some development of ideas and views • rather predictable and obvious thoughts and views	**Loosely or simply organised**	• some organisation and structure • some incoherence and inconsistency • a loose, simple or rigid structure	**Bland or uneven**	• some problems with language use • bland or uneven expression • tending to stock phrases and cliches • simple or excessively complicated syntax
1/2	**Simplistic or crude**	• a very simple or basic grasp of the issues • general and vague discussion • predictable and banal commentary • simple or blank assertion	**Lack of organisation**	• little organisation and structure • vague or little direction	**Lack of control**	• some significant problems with language use • clumsy or stolid writing
0		• Little or nothing produced		Little or nothing produced		Little or nothing produced

Appendix 2 Analytic single trait scale (*communicative intent*)

¿El texto aborda el estímulo de alguna manera? **Does the text respond to the stimulus in any way?**	No (award level 0) Sí (go to level1)
¿El texto presenta fallas en la configuración del mensaje (legibilidad, ruido, fallas en el uso de la convención) que dificultan identificar una intención comunicativa o el escaso desarrollo impide una valoración sin embargo se toca el estímulo? **Does the text present errors in conveying the message (legibility, noise, errors in convention) that make it difficult to identify a main idea or does the lack of development impede judgement, however it responds to the stimulus**?	No (award level 1) Sí (go to level 2)
¿A lo largo del texto se aprecia un desarrollo incipiente de la intención comunicativa? **Can evidence of the beginnings of message development be identified throughout the text?**	No (award level 2) Sí (go to level 3)
¿Los enunciados dan lugar a ambigüedades en relación a la intención comunicativa? **Do the statements give rise to ambiguities in relation to the main idea?**	No (award level 3) Si (go to level 4)
¿Son pertinentes los recursos (estilo, silueta textual, manejo de voces, marco de referencia o fuente de experiencia) que utiliza para hacer eficaz su intención? **Are the resources relevant (style, text outline, register and tone management reference to authorities or sources of experience) that are used to efficiently communicate the message?**	No (award level 4) Sí (go to level 5)
¿Hay un texto contundente que evidencia una problematización sobre el estímulo múltiples intenciones comunicativas interrelacionadas)? **Is there a strong text that is evidence of problematizing the stimulus with many inter-related ideas communicated?**	No (award level 5) Sí (award level 6)