The influence of the CEFR on rating scale design
Bart Deygers & Koen Van Gorp

Abstract

The Certificate Dutch as a Foreign Language (CNaVT) offers Dutch task-based language exams for 6 different profiles which have been determined by an extensive needs analysis (Van Avermaet & Gysen, 2006). The task content is co-determined by a pool of subject specialists around the world who verify the authenticity and representativeness of each task and check the items for cultural bias.

For the past years the CNaVT's rating scale has been dichotomous and analytical. Even though this scale has a proven reliability and usability, it was decided to reshape it into a model that would better reconcile the CNaVT's philosophy with its stakeholders' needs: i.e. a clearer alignment with both the CEFR and domain experts' judgements of language performance (Jacoby & McNamara, 1999).

Redesigning the scale has proven to be an extensive undertaking which touches upon all aspects of language testing. Indeed, altering a dichotomous model into a polytomous band rating scale, which merges performance driven exemplifications (Weigle, 2007) with measurement driven descriptors is an operation so all-encompassing that it necessitates rethinking the entire testing process. Simultaneously, working closely with the CEFR has forced the rating scale developers to critically examine the level descriptors so as to operationalize them in a usable rating scale without neglecting known pitfalls such as validity reduction (Lumley, 2002) and a lack of concreteness (Fulcher, 2010).

This presentation focuses on the role of the CEFR in the rating scale redevelopment process, on its strengths, but also on its shortcomings which prevent it from being a readymade assessment tool. The presentation will include data resulting from the development and validation process. This includes focus groups with subject specialists, stimulated recall interviews with raters as well as qualitative test analyses (i.e. inter and intra rater reliability, correlation coefficients etc).

Rating scale typologies

Rating scales can be classified according to different parameters, such as the way in which the scoring criteria have been established or the way these criteria are presented to the rater. Naturally, these different types can be combined and modified to match the idiosyncrasies of each individual test.

Measurement driven rating scales have been drawn up by language experts and are typically not derived from real-life performances, which is the very basis of performance driven scales (Fulcher, Davidson and Kemp 2010, Weigle 2007). Since measurement driven scales are founded in theory and abstraction, their level descriptors may be too distant from reality. Conversely, given that performance driven scales are based on actual performances, their descriptions might be too detailed to allow for generalization (Fulcher et al. 2010).

Holistic rating scales compel raters to judge a performance as a whole, whereas their analytic counterparts take into account separate features of language, such as grammar, vocabulary and structure (Alderson, Clapham and Wall 1995). Previous studies have shown that the analytic scales are often more reliable than holistic ones, offer richer L2 diagnostic information and are better suited for novice raters (Barkaoui and Knouzi 2011, Barkaoui 2010, Knoch 2009, Weigle 2002). Holistic scales on the other hand, perform better than analytic ones in terms of authenticity and rating speed (Knoch 2009, Weigle 2002). A third possible way to categorize rating scales is according to the number of scoring categories they employ. "Items that are scored in two categories - right or wrong - are referred to as dichotomously scored items. Items that are scored in multiple-ordered categories are referred to as polytomously scored items" (Tang 1996: 2).

Whether or not a rating scale is performance driven or measurement driven, holistic or analytic, dichotomous or polytomous, it is always the rater and not the rating scale who decides on the score (Fulcher at al. 2010, Lumley 2002). Naturally, the quality of the descriptions, their level of complexity and abstraction will influence the consistency and accuracy of the rater (Alderson et al 1995, Fulcher et al 2010). Additionally rater training has proven to be of great value when streamlining the interpretations of rating criteria (Lumley 2002, Shohamy, Gordon and Kraemer 1992, Weigle 1994). Without rater training, it would be up to each individual rater to decide on the meaning of frequently used but unquantifiable terms such as "adequate", "good" and "sufficient". Even with such a training it is difficult to overcome the problems associated with vagueness in descriptors.

An Asymmetrical Framework

Upon its publication, the CEFR was to be a document that addresses concerns about multilingualism, stimulates the use of a common metalanguage, helps curriculum development and promotes professional mobility within Europe (Little 2007, Fucher 2004, Milanovic 2001). More than a decade later its actual use differs from these original intentions. As more and more schools, test developers and policy makers use the CEFR it is regarded as more than the theoretical model it actually is (Fulcher 2004) and has become a fixed standard in European language education and language testing. Still, the CEFR, being a measurement-driven language-independent model of L2 acquisition, lacks the empirical foundation and descriptional specificity to act as a real framework (Alderson 2007, Little 2007), let alone a scoring tool (Papageorgiou 2010, Weir 2005).

For one thing, the relative distance between the different CEFR levels is inconsistent (Fulcher 2004). This causes fundamental problems in a rating context since polytomous IRT analysis generally assumes that the distribution between different scoring levels is equal (Huyn 1994 & 1996, Tang 1996) or at least known (Roberts, Donoghue & Laughlin 2010).

Furthermore, the level descriptors often show overlaps and gaps (Alderson 2004), both of which may create the vagueness a rating scale constructor whishes to avoid.

> "When the scales, in particular, were examined closely, it became apparent that many terms lacked definitions, there were overlaps, ambiguities, and inconsistencies in the use of terminology, as well as important gaps in the CEFR scales."
> (Alderson 2007: 661)

Finally, the CEFR is asymmetrical in the attention it gives to receptive and productive skills. It focuses heavily on production while the receptive skills remain underdeveloped (Weir 2005, Alderson 2004, Staehr 2008, Milton 2010). The CEFR lacks usable specifications for quite a few skills that may be operationalized in receptive tasks, i.e. text complexity (Alderson 2004, Weir 2005, Alderson

2006, Davidson & Fulcher 2007), lexis (Alderson 2007, Milton 2010) and subject matter (Weir 2005, Fulcher 2004).

CNaVT Rating scale construction

The Certificate of Dutch as a Foreign Language (CNaVT) offers five functional task-based language tests (Van Gorp & Deygers 2013) that operate according to Bachman and Palmer's (2010) can-do typology. These tests correspond to five profiles and fall into three categories: societal, professional, and academic language use. The profiles have been determined by a needs analysis among end users (Van Avermaet & Gysen 2006), which continues to shape the exams to date. Currently, the CNaVT is a pass/fail exam: candidates either pass the examination in the domain of their choosing or they do not.

In 2009, the subject specialists of the two academic profiles suggested altering the binary approach of the existing dichotomous analytic rating scale so it would align more closely with their "indigenous criteria" (Jacoby and McNamara 1999). Around the same time quite a few stakeholders voiced their wish for the different tests to be more explicitly linked to the CEFR (a trend also observed by Fulcher 2004). More recently, the government organisation funding the exams has decreed that over the coming years the pass/fail approach should be abandoned in favour of a system in which each test contains two cut scores, each one linked to a CEFR level. These developments instigated both a revision of the testing process and a reconceptualization of the rating scale (see Deygers, Van Gorp, Luyten and Joos 2013 for a full discussion of the rating scale construction and validation process). The new rating scales were to reconcile the subject specialists' criteria with both the stakeholders' wish for a clearer CEFR alignment and with the test sponsor's demand for a double cut off score at two CEFR levels per test. Even though all rating scales are in the process of revision, this paper solely focuses on the scale of the new Dutch for academic purposes (DAP) test.

The composition of the DAP's team of raters may change from one year to the next. Since the judgment of novice raters is more reliable when using an analytic rather than a holistic scale (Barkaoui 2010), the new scales are analytic in nature. The criteria for these scales are derived from focus groups with subject specialists (N = 13), subject specialist questionnaires (N = 178) and literature reviews (Deygers et al. 2013). Each criterion can be scored on four levels, the third being up to the minimum standard, the fourth being above and the first and second below. Each scoring category corresponds to a CEFR level. In the case of the DAP test, three corresponds with the B2 level of the CEFR, four with C1.

After an iterative development process, the DAP rating scale was piloted using 4 trained raters who rated 250 tasks using both the original dichotomous scale and the newly developed polytomous scale. In order to avoid sequence or contamination effects, two raters first used the polytomous scale while the other two started with the dichotomous scale. Irrespective of the order in which the scales were used, the dichotomous scale consistently showed to be more reliable and to yield higher inter-rater agreement (Deygers et al. 2013).

Following the rating process, the four raters took part in a focus group. They preferred the dichotomous scale when judging written performances but the polytomous one for speaking tasks. All raters preferred the polytomous approach in theory because it allows for a more fine-grained judgment. In practice, they all reported confusion when using the CEFR-based level descriptors.

A second and third trial followed the initial pilot of the rating scale. Each new trial focused on rewriting the level descriptors so they would become more easily interpretable by novice raters. Before each trial, the raters received an intensive two-day rater training during which they reported vagueness in the level descriptors and suggested ways to reformulate the descriptions. Often these suggestions meant clarifying the difference between one level and the next, providing concrete examples and adding language-specific expectations. In the second trial, two trained raters judged 76 spoken performances and in the third trial two trained novice raters judged 27 written argumentative tasks and 28 presentation tasks.

After each trial the raters now reported to prefer the polytomous scale over the dichotomous one. They did not report feeling uncertain or confused when using the adapted level descriptors. Nonetheless, quantitative analysis of the rating process shows that the descriptors of grammar and vocabulary caused problems. For grammar, the distinction between level 2 (B1) and 3 (B2) was considered too harsh. For vocabulary, all descriptors remained too vague. Other CEFR tables such as "Orthographic control" and "Coherence and cohesion" also appeared quite challenging indeed to operationalize.

Discussion: The use of the CEFR for rating scale design

Even though the CEFR "was not designed specifically for test specifications and language testing contexts" (North 2004 in Papageorgiou 2010: 273), there is an apparent need within Europe among stakeholders to demand a clear link between a test score and a CEFR level.

> "For many producers of tests, one of the dangers lies in the desire to claim a link between scores on their tests and what those scores mean in terms of CEF levels, simply to get recognition within Europe. They do not have any choice in this, for if institutions begin to believe that the CEF is the truth against which all else must be measured, failure to claim a link to the CEF would equate to a commercial withdrawal from continental Europe." (Fulcher 2004: 260)

In the case of the CNaVT, the endeavor to link the test with the CEFR has surpassed the "intuitive guess" Fulcher (2004) observes. Each CNaVT test has undergone an extensive standard setting process and the rating scales combine input from subject specialists, language specialists, raters and the CEFR level descriptors. By working closely with the CEFR, the developers of the rating scale have critically examined the its level descriptors in order to operationalize them in a usable rating scale while avoiding validity reduction (Lumley, 2002) lack of concreteness (Fulcher, 2010) and other known pitfalls of rating scale construction.

The major shortcoming of the CEFR when used as a source for rating scale development appears to be its unsound theoretical foundation. It is partly based on empirical findings but at its core are the intuitions of language experts (Alderson 2004, Fulcher 2004, Hulstijn 2007, Little 2007, North 2007). This leads to inconsistency and vagueness on a meta and micro level. On a meta level, the unequal distance between levels causes problems for a polytomous IRT analysis. On a micro level, not all level descriptors can readily be operationalized in a rating scale.

One example of this is the CEFR's description of grammatical accuracy. The difference between "*relatively high degree of grammatical control [without] mistakes which lead to misunderstanding*" (lower end B2) and "*generally good control […] errors occur, but it is clear what he/she is trying to express*" (higher end B1) is too tentative to use in a rating context. Using either the lower B1 and the upper B2 or the upper B1 and upper B2 prove equally problematic. All raters involved in the pilot study claimed that the difference between 2 (B1) and 3 (B2) was either too vague or too harsh to be usable. Even though the criterion "grammar" caused some correlational problems among the raters, "vocabulary" yielded the lowest inter-rater agreement of all criteria. Indeed, the CEFR "provides little assistance in identifying the breadth and depth of productive or receptive lexis" (Weir 2005: 292).

Conclusion: a common basis?

The CEFR is a theory on second language acquisition, partly based on empirical data, partly on theoretical conceptions and partly on intuition (Hulstijn 2007, Little 2007, North 2007). It takes on a positive and descriptive approach to language learning by focusing on what learners can do at a given level. This has forced language teachers not to only think of their students in terms of deficit, but also in terms of accomplishment. Throughout Europe language practitioners and policy makers now not only know that the CEFR exists and use its terminology, they may also see what it entails and might even wish for classroom and testing practice to adhere to its logic. And that is where the problem begins.

For one thing, no theoretical model can strive towards universality without trading in specificity for generic applicability. Because of this, every CEFR descriptor used in the CNaVT rating scale development was too underdefined to be used without adaptation. For each criterion language-specific additions had to be made, differences between levels had to be clarified and examples had to be provided. Only then were raters able to maintain an acceptable level of consistency.

Furthermore, the CEFR occupies a somewhat dubious position in terms of malleability. In the minds of many stakeholders and policy makers the CEFR-levels appear etched in stone, B2 occupying an especially elevated position. At the same time however there is general agreement that the broadness of CEFR level descriptors allows for multiple interpretations, forcing users into interpretation and specification. And when generally accepted levels are universally interpreted differently, the CEFR can only provide "a common basis" (Milanovic, 2001: 1) on paper.

References

Alderson, C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2004). The Development of Specifications for Item Development and Classification within The Common European Framework of Reference for Languages: Learning, Teaching, Assessment Reading and Listening. Final report of the Dutch CEF construct project. Unpublished Document. Retrieved from http://eprints.lancs.ac.uk/44/1/final_report.pdf

Alderson, J. C. (2007). The CEFR and the Need for More Research. *The Modern Language Journal*, *91*(4), 659–663.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge University Press.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2006). Analysing Tests of Reading and Listening in Relation to the Common European Framework of Reference: The Experience of The Dutch CEFR Construct Project. *Language Assessment Quarterly: An International Journal*, *3*(1), 3–30.

Bachman, L., & Palmer, A. (2010). *Language Assessment in Practice*. Oxford University Press, USA.

Barkaoui, K, & Knouzi, I. (2011). Rating scales as frameworks for assessing L2 writing: examining their impact on rater performance. Presented at the ALTE 4th International Conference, Kraków, Poland.

Barkaoui, Khaled. (2010). Explaining ESL essay holistic scores: A multilevel modeling approach. *Language Testing*, *27*(4), 515–535.

Davidson, F., & Fulcher, G. (2007). The Common European Framework of Reference (CEFR) and the design of language tests: A matter of effect. *Language Teaching*, *40*(03), 231. doi:10.1017/S0261444807004351

Deygers, B., Van Gorp, K., Luyten, L., & Joos, S. (2013). Rating scale design: a comparative study of two analytic rating scales in a task-based test. In E. Galaczi & C. Weir (Eds.), *Exploring Language Frameworks. Proceedings from the ALTE Kraków Conference, July 2011.* (Vol. 36, pp. 273–289). Cambridge: Cambridge University Press.

Fulcher, G. (2004). Deluded by Artifices? The Common European Framework and Harmonization. *Language Assessment Quarterly*, *1*(4), 253–266.

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

Hulstijn, J. H. (2007). The Shaky Ground Beneath the CEFR: Quantitative and Qualitative Dimensions of Language Proficiency1. *The Modern Language Journal*, *91*(4), 663–667.

Huynh, H. (1994). On equivalence between a partial credit item and a set of independent Rasch binary items. *Psychometrika*, *59*(1), 111–119.

Huynh, H. (1996). Decomposition of a Rasch partial credit item into independent binary and indecomposable trinary items. *Psychometrika*, *61*(1), 31–39.

Jacoby, S., & McNamara, T. (1999). Locating Competence. *English for Specific Purposes*, *18*(3), 213–241.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, *26*(2), 275–304.

Little, D. (2007). The Common European Framework of Reference for Languages: Perspectives on the Making of Supranational Language Education Policy. *The Modern Language Journal*, *91*(4), 645–655.

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.

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing*.

Milanovic, M. (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Council of Europe.

Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. In G. Pallotti (Ed.), *communicative proficiency and linguistic development: intersections between SLA and language testing research*. Rome: Creative Commons.

North, B. (2007). The CEFR Illustrative Descriptor Scales. *The Modern Language Journal*, *91*(4), 656–659.

Papageorgiou, S. (2010). Investigating the decision-making process of standard setting participants. *Language Testing*, *27*(2), 261–282.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A General Item Response Theory Model for Unfolding Unidimensional Polytomous Responses. *Applied Psychological Measurement*, *24*(1), 3–32.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The Effect of Raters' Background and Training on the Reliability of Direct Writing Tests. *The Modern Language Journal*, *76*(1), 27–33. doi:10.2307/329895

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, *36*(2), 139–152.

Stæhr, L. S. (2009). Vocabulary Knowledge and Advanced Listening Comprehension in English as a Foreign Language. *Studies in Second Language Acquisition*, *31*(4), 577–607.

Tang, L. K. (1996). *Polytomous Item Response Theory (IRT) Models and their applications in large-scale testing programs: review of literature*. New Jersey: Educational Testing Service.

Van Avermaet, P., & Gysen, S. (2006). From needs to tasks: Language learning needs in a task-based approach. In K. van den Branden (Ed.), *Task-Based Language Education: From Theory to Practice*. Cambridge: Cambridge University Press.

Van Gorp, K., Deygers, B., & Kunnan, A. (2013). Task Based Language Assessment. In *The Companion to Language Assessment*. New Jersey: Wiley-Blackwell.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), 197–223.
Weigle, S. C. (2002). *Assessing Writing*. Cambridge University Press.

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, *22*(3), 281–300.

Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing*, *27*(1), 119–140.