Towards an Examiners Training Model for Standardized Oral Assessment Qualities in Vietnam

ANH TUAN NGUYEN
University of Languages and International Studies
Vietnam National University, Hanoi

ABSTRACT

There are many variables that may affect the reliability of speaking test results, one of which is rater reliability. The lessons learnt from world leading English testing organizations such as International English Testing System (IELTS) and Cambridge English Language Assessment show that oral examiner training plays a fundamental role in sustaining the highest consistency among test results. This paper presents a model of oral examiner training presently at its early stage in standardizing the English speaking test in Vietnam, as part of the country's National Foreign Languages Project 2020. Training sessions have been conducted using localized training materials, aiming to guarantee the professionalism of English teachers as oral examiners to enable maximum reliability when marking speaking tests using a standardized procedure. A post-training study on EFL university teachers shows that the model has generally met the teachers' professional development demands in oral assessment, though more research into real speaking test scores is needed to be able to guarantee that the goals of the model could be accomplished.

KEYWORDS: Oral, examiner, training, oral assessment

Introduction

Vietnam's National Foreign Languages Project, known as Project 2020, is coming to its critical stage of implementation. One of its most important targets is to upgrade Vietnamese EFL teachers' English language proficiency to required CEFR (Common European Framework of Reference) levels corresponding to B1 for Elementary School, B2 for Secondary and C1 for High School. In order to achieve this target, there have been courses and proficiency tests to upgrade proficiency of unqualified teachers whose English proficiency is lower than the required levels (B1, B2 and C1). These courses and tests have been administered by 10 universities and education centres specializing in foreign languages from the North, South and Central Vietnam.

Although there is a good rationale for such an extensive upgrading campaign, some critical questions have been raised regarding the reliability of such tests of highly subjective nature as speaking and writing. Concerns have been raised about whether the speaking test results provided by, for example, University of Languages and International Studies (ULIS) are the same as those by Hanoi University (HANU) in terms of reliability. It is clear that a good English teacher, who may not necessarily be a good examiner, requires professional training. How many university teachers of English among those employed as oral examiners in the speaking tests over the past five years of Project 2020 have been trained professionally?

The following data were collected from six universities in September 2014, which prove how urgent it is to seriously consider examiner training. It can be clearly seen that there is a great difference between the numbers of trained and untrained oral examiners, who are actually EFL university teachers, for both international English tests in Vietnam and Project 2020 proficiency tests.

Table 1. Oral Examiner Training at six universities specializing in foreign languages in Vietnam

University	**Total of English teachers**	**Total of English teachers trained as professional oral examiners in international English tests**	**Total of English teachers trained as oral examiners in Project 2020**
1	150	13	120
2	40	1	3
3	70	unknown	4
4	80	5	30
5	64	10	45
6	55	0	55
Total	459	>29	257

Luoma (2004) believes that rater training is the most common procedure used for reliability assurance for formal examinations as training helps change the individual's perception of the world to ensure reliability. Weigle (1994), investigating verbal protocols of four inexperienced raters of ESL placement compositions scoring the same essays, points out that rater training helps clarify the intended scoring criteria for raters, modify their expectations of examinees' performances and provide a reference group of other raters with which raters could compare themselves.

Further investigation by Weigle (1998) on sixteen raters (eight experienced and eight inexperienced) shows that rater training helps increase intra-rater reliability as "after training, the differences between the two groups of raters were less pronounced" (Weigle, 1998, p. 263). Eckes (2008) even provides evidence for a proposed rater type hypothesis which is that each rater type has his or her own characteristics on a distinct scoring profile due to rater background variables and suggests that training can redirect attention of different rater types and thus reduce imbalances.

In terms of oral language assessment, different factors that are not part of the scoring rubric have been spotted to influence raters' validation of scores, which confirms the important role of oral examiner training. Eckes (2005, p. 216), in examining rater effects in TestDaF states that "raters differed strongly in the severity with which they rated examinees … and were substantially less consistent in relation to rating criteria (or speaking tasks, respectively) than in relation to examinees." Most recently, Winke, Gass and Myford (2011) report that "rater and test taker background characteristics may exert an influence on some raters' ratings… when there is a match between the test taker's L1 and the rater's L2, some raters may be more lenient toward the test taker and award the test taker a higher rating than expected" (p. 50).

In order to increase rater reliability, besides improving oral test methods and scoring rubrics, Barnwell (1989, cited in Douglas, 1997, p.24) suggests that "further training, consultation, and feedback could be expected to improve reliability radically". This suggestion comes from Barnwell's study of native speakers of Spanish who used guidelines in the form of the American Council on the Teaching of Foreign Language (ACTFL) oral proficiency scales, but had no training in their use. There was evidence of patterning in the ratings although inter-rater reliability was not high for such untrained raters. In addition, for successful oral examiner training, "if raters are given simple roles or guidelines (such as may be found in many existing rubrics for rating spoken performances), they can use 'negative evidence' provided by feedback and consultation with expert trainers to calibrate their ratings to a standard" (Douglas, 1997, p.24).

In an interesting report by Xi and Mollaun (2009), the vital role and effectiveness of a special training package for bilingual or multilingual speakers of English and one or more Indian languages was investigated. It was found that with training similar to that which the operational U.S.-based raters receive, the raters from India performed as well as the operational raters in scoring both Indian and non-Indian examinees. Training also helped the raters score Indian examinees more consistently, leading to increased score reliability estimates, and boosted raters' levels of confidence in scoring Indian examinees.

Similarly, Karavas and Delieza (2009) reported a standardized model of oral examiner training in Greek which includes two main components of training seminars and on-site observation. The first component aims to train 3000 examiners who are fully and systematically trained in assessing candidate's oral performance at A1/A2, B1, B2, C1 levels. The second component makes an attempt to identify whether and to what extent examiners adhere to exam guidelines and the suggested oral exam procedure It also seeks to gain information about the efficiency of the oral exam administration and the efficiency of oral examiner conduct, of the applicability of the oral assessment criteria and of inter-rater reliability. The observation phase is considered a crucial follow-up activity in pointing out the factors which threaten the validity and reliability of the oral test and the ways in which the oral test can be improved.

Despite oral examiner training being of such great importance to assessment practices, in Vietnam’s context, as shown in Table 1, not all EFL university teachers employed in national English speaking tests have been trained. It should be emphasized that if Vietnam’s education policy makers have an ambition to develop Vietnam’s own speaking test, EFL teachers in Vietnam must be trained under a *national standardized oral examiner training procedure* so as to make sure that speaking test results are reliable across the country. In other words, there exists an urgent need for a standardized model of oral examiner training for Vietnamese EFL teachers, and building oral assessment capacity for Vietnamese teachers of English must be considered as top priority for the purpose of maximizing the reliability of speaking scores.

Oral Examiner Training Model

December 2013 could be considered a historic turning point in Vietnam’s EFL oral assessment when key oral examiner trainers from 10 universities and education centres specializing in foreign languages from the North, South and Central Vietnam gathered in Hanoi for a pioneering national workshop on oral examiner training. The primary aim of the four-day workshop was to provide the representatives with a chance to reach an agreement on how to operate an English speaking test systematically on a national scale. After the workshop, these key trainers would return to their school and conduct similar oral examiner training workshops to other speaking examiners. The model might look as follows:

Figure 1. Illustration of a general training model

(Tech.digesttouch, 2010)

What made the workshop successful was the agreement among 42 key trainers on fundamental issues in assessing speaking abilities, which can be summarized as follows:

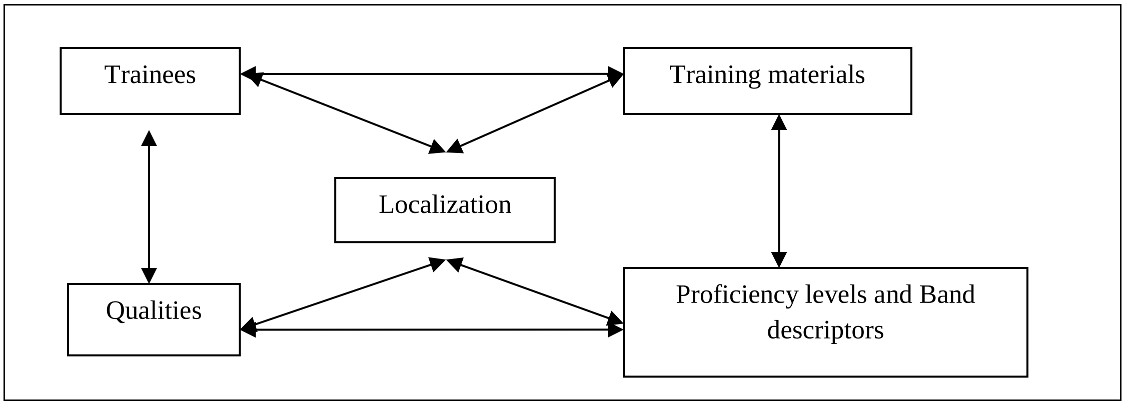
- Examiners must adhere to interlocutor frame during the course of the test
- Examiners assess students analytically instead of holistically (Key trainers agreed on how key terms in assessment scales should be understood across four criteria including grammar range, fluency and cohesion, lexical resources and pronunciation).
- A friendly interviewer style is preferred.
- Examiners must assess candidates based on their present performances instead of examiners' knowledge of candidates' background.

In fact, such a training model is common in many other fields and industries as it helps to cascade skills from top to bottom efficiently. It is also similar to the way world leading English testing organizations such as International English Testing System (IELTS) and Cambridge English Language Assessment (CELA) train their oral examiners. For example, CELA speaking tests are conducted by trained Speaking Examiners (SEs) whose quality assurance is managed by Team Leaders (TLs) who report to a Professional Support Leader (PSL), who is the professional representative of University of Cambridge ELA for the Speaking tests in a given country or region. However, the Hanoi workshop has a number of distinctive features which pave the way towards creating a national standardized oral examiner training model, including:

- ❖ An agreement on **localized** CEFR levels and speaking band descriptors
- ❖ Use of **authentic** training video clips in which participants are local students and teachers
- ❖ An agreement on certain qualities of a Vietnamese professional speaking examiner in terms of rating process, interviewer style, and use of test scripts.

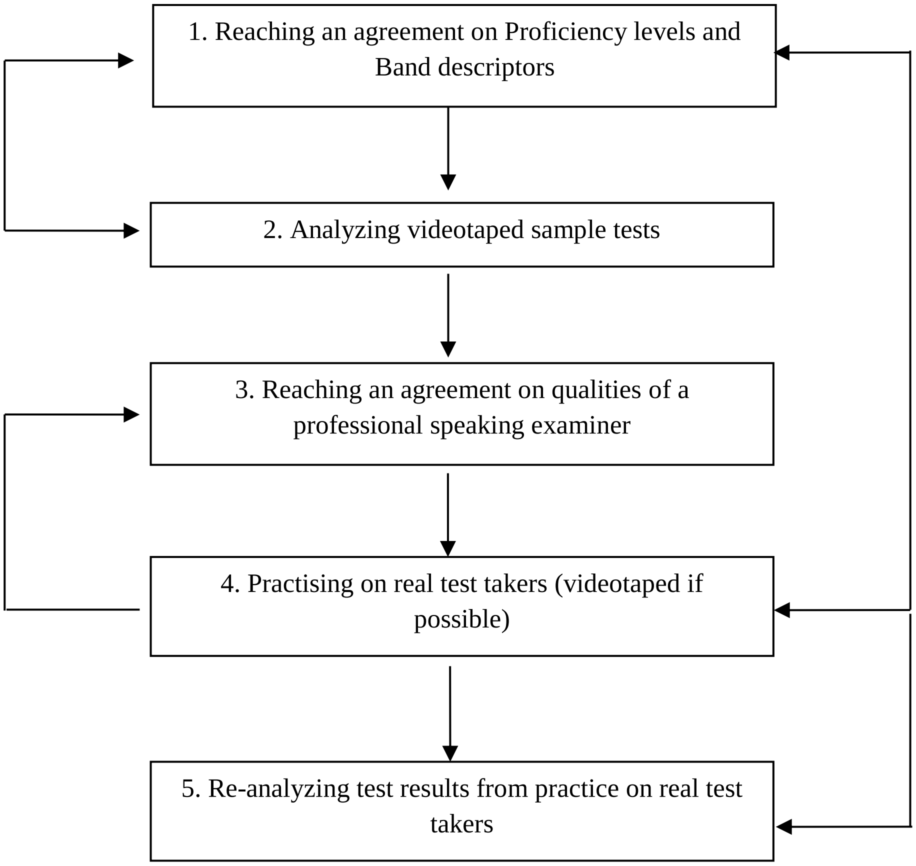
It is understandable that the term "localization" is the core of this workshop as it reflects the true nature of the training where the primary goal is to train local professional examiners believed by Xi and Mollaun (2009) as the best choices. A model built on this term is shown in Figure 2:

Figure 2. Localization Model



Based on the Localization Model, a step-by-step procedure can illustrate how a speaking examiner training workshop works, which is shown in Figure 3:

Figure 3. Oral examiner training procedure


1. Reaching an agreement on Proficiency levels and Band descriptors
2. Analyzing videotaped sample tests
3. Reaching an agreement on qualities of a professional speaking examiner
4. Practising on real test takers (videotaped if possible)
5. Re-analyzing test results from practice on real test takers

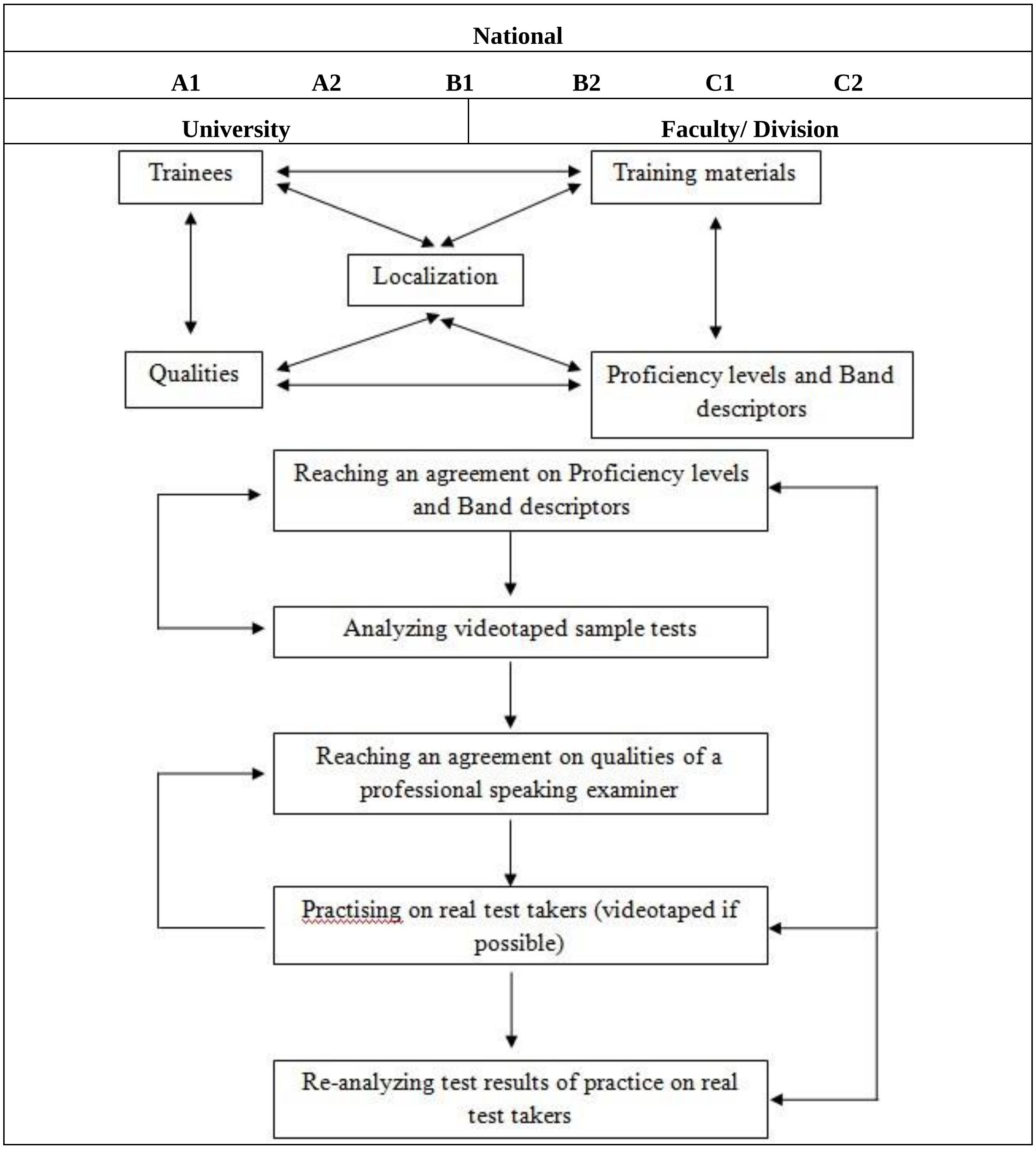
Steps 1 and 2 focus on helping examiners analyse and then understand how the test and marking criteria work through the use of standard videotaped sample performances. Each performance has been processed in advance so that it perfectly corresponds to a certain level. Step 3 deals with how an oral examiner operates an interview, uses test scripts and proceeds with marking scales. The last two steps aim to get examiners ready by providing them an opportunity to apply what they have learnt so far on simulated oral tests. Re-analysis of simulated test results allows collection of examiner trainees' peer-feedback on each other's given scores, allowing them to adjust their decisions to standard scores.

Multi-layered oral examiner training model

Upgrading English teachers' proficiency levels has been just part of Vietnam's ambitious Project 2020; in other words, the above training model is reflected in the progression of only one layer where university teachers as speaking examiners in upgrading courses are the target trainees. If CEFR levels in Vietnam are applied throughout the country, it is worth questioning whether these levels of specifications will be well understood by those teachers who are not used as oral examiners in upgrading courses but are still working in undergraduate programs. As required, undergraduates must achieve B1 or B2 for non-English major and C1 for English major, which means undergraduate teachers must be trained for the assurance of speaking test quality.

A multi-layered oral examiner training model (Figure 3), therefore, is expected to be able to help solve the problem. Multi-layered can be understood as either layers of administration including National, University, and Faculty or different levels of proficiency ranging from A1 to C2. In other words, the training model can be applied on any scale of test administration for either proficiency-based or level-based English tests.

Figure 4. Multi-layered oral examiner training model


National
A1
A2
B1
B2
C1
C2
University
Faculty/ Division
Trainees
Training materials
Localization
Qualities
Proficiency levels and Band descriptors
Reaching an agreement on Proficiency levels and Band descriptors
Analyzing videotaped sample tests
Reaching an agreement on qualities of a professional speaking examiner
Practising on real test takers (videotaped if possible)
Re-analyzing test results of practice on real test takers

There are several factors that can be inferred from this multi-layered model. First, the national layer is responsible for developing a comprehensive set of speaking assessment criteria across six CEFR levels. This set is the basis for any other proceeding action plans. Second, universities and faculties/divisions must provide training for their teachers at each CEFR level, using Localization Model and a step-by-step procedure, so that the national standardization of criteria can be maintained.

Preliminary evaluation of model

The University of Languages and International Studies (ULIS – Vietnam National University, Hanoi) has been one of the institutions authorised by the Project 2020 Administration Board to manage upgrading courses and proficiency tests. The school’s EFL teachers have taken part in oral examiner training workshops for the past two years, where this model has been used for training purposes. For a preliminary evaluation of the model, 25 EFL teacher trainees from the Faculty of English Language Teacher Education (ULIS-VNU) were chosen to complete a questionnaire of five questions centring on their feedback on training practices. Table 2 shows the results obtained:

Table 2. Feedback on oral examiner training model

Questions’ content	**Results**			Other comments
Q1: General satisfaction	Very satisfied	Satisfied	Dissatisfied	No
	15%	**85%**	**0%**	
Q2: Effectiveness in Vietnam’s context	Very effective	Effective	Ineffective	No
	12%	**88%**	**0%**	
Q3: Use of localized videos	Totally Agree	Agree	Disagree	Both Vietnamese and non-Vietnamese candidates
	44%	**48%**	**4%**	
Q4: Practice on real test takers	Very effective	Effective	Ineffective	Careful selection is suggested.
	36%	**60%**	**0%**	
Q5: Things improved	Better use of band descriptors	Marking sample performances	Strictly follow rules for examiner	No
	92%	**88%**	**68%**	

In terms of general satisfaction and effectiveness in Vietnam’s context, it can be seen that the majority of the teachers gave positive feedback on the model (85% satisfied and 88% effective respectively). The use of sample videos for marking practice, in which candidates are Vietnamese, was favoured by nearly all the teachers, except for one disagreement and one suggestion that there should be videos of both Vietnamese and non-Vietnamese candidates. Practice on real test takers was also highly appreciated (96% very effective and effective). In other words, the training model appears to have satisfied participants, which means the model can be applied successfully with EFL university teachers.

The teachers’ satisfaction was reflected through their improvement in assessing speaking after the training session. Around 90% said that they used band descriptors better in real tests and marking videotaped sample performances. This is a very important outcome of the training activity as the primary goal of training is to enhance teachers’ scoring reliability. However, a lower proportion of the teachers reported strictly following rules for examiners. It can be construed that EFL teachers used to grant themselves the right to do what they wanted, for example, with the test script instead of adhering to it to assure fairness among candidates. Further training and closer monitoring are believed to better foster examiner professionalism.

Conclusion

This paper presents a multi-layered model of oral examiner training presently at its early stage in standardizing the English speaking test in Vietnam as part of the country's National Foreign Languages Project 2020. The aim of the model is to guarantee the professionalism of English teachers as oral examiners by helping them have a complete understanding of speaking assessment criteria at certain proficiency levels, appropriate manners of a professional examiner, and better awareness of what they must do to minimize subjectivity. A survey of EFL university teachers as oral examiner trainees shows that the model has generally met the teachers' professional development demands in oral assessment, and, as a result, a new generation of oral examiners who can give the most reliable speaking test marks on a standardized procedure can be created.

It should be noted that training sessions are just an initial step of the whole training project. Their effectiveness in improving Vietnam's EFL teachers' ability in assessing students' speaking ability in the long-term still requires more quantitative research evidence.

Acknowledgements

An earlier version of this paper was presented at the 12th International Asia TEFL Conference and 23rd MELTA International Conference, held in Kuching, Sarawak, Malaysia in August 2014, and published in the conference proceedings.

References

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. (2000). TOEFL 2000 Speaking Framework: a working paper. *TOEFL Monograph Series*, MS-20 June. New Jersey: Princeton.

Douglas, D. (1997). Testing speaking ability in academic contexts: Theoretical considerations. *TOEFL Monograph Series*, MS-8 April. New Jersey: Princeton.

Douglas, D.,& Smith, J. (1997). Theoretical underpinnings of the Test of Spoken English Revision Project *TOEFL Monograph Series*, MS-9 May. New Jersey: Princeton.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: a many-facet Rasch Analysis. *Language Assessment Quarterly*, *2*(3), 197-221.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155-185.

Erlam, R., Randow, J. V., & Read, J. (2013). Investigating an online rater training program: product and process. *Papers in Language Testing and Assessment*, *2*(1), 1-29.

Karavas, E.,& Delieza, X. (2009). On site observation of KPG oral examiners: Implications for oral examiner training and evaluation. *Apples – Journal of Applied Language Studies*, *3*(1), 51-77.

Luoma, S. (2004).*Assessing speaking*. New York: Cambridge University Press.

Pizarro, M. A. (2004). Rater discrepancy in the Spanish university entrance examination. *Journal of English Studies*, *4*, 23-36.

Tannenbatum, R.,& Wylie, E. C. (2008). Linking English-language test scores onto the Common European Framework of Reference: an application of standard-setting methodology. *TOEFL iBT Research Report*, July 2008. ETS.

Weigle, S.C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), 197-223.

Weigle, S.C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263-87.

Weir, C. J. (2005).*Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.

Winke, P., Gass, S., & Myford, C. (2011). The relationship between raters' prior language study and the evaluation of foreign language speek samples. *TOEFL iBT Research Report*, July 2011. ETS.

Xi, X., & Mollaun, P. (2009). How do raters from India perform in scoring the TOEFL iBT Speaking section and what kind of training helps? *TOEFL iBT Research Report*, August 2009. ETS.

Tech.digesttouch (2010). Retrieved from http://tech.digesttouch.com/tapping-asian-e-commerce-mitochondria-multiplication-and-real-world-e-commerce/