IAEA 2010 Conference
General Speaking Ability: An Assessable Construct?

CHEUNG Kwai Mun Amy PhD
Hong Kong Examinations and Assessment Authority
Macquarie University, Australia
cheung_km@yahoo.com

Abstract

In a study on large-scale oral assessment in Hong Kong, high correlations emerged between ratings on the various rating criteria and verifiable quantitative measures (VQM) for these criteria. The initial conclusion was that the raters were paying attention to the textual features measured by the VQM. However an alternative hypothesis (AltH) was suggested i.e. 'raters were simply assessing 'general speaking ability' which would naturally correlate with VQM. To investigate AltH, Rasch Fair Averages for rated criteria were correlated against each other. Ratings on all criteria correlated with each other, 0.963 to 0.979, initially indicating that AltH was correct. However, upon investigating 'raw score' correlations between criteria for our best rater (clearly acceptable Rasch fit values and good correlations with both expert panel and VQM), it was found that correlations between the various criteria were only 0.787 to 0.871, indicating that there was a 'smoothing function' in the Rasch fair average which exaggerated correlations between different criteria. Without this function, correlations dropped, indicating that various rating criteria may well be separate entities, reflecting attention to separate axiomatically-related textual phenomena rather than showing AltH was correct. Further challenge to AltH emerged with correlations of VQM against each other. When no axiomatic relationship between VQM was apparent, correlations ranged from 0.597 to 0.789. When an axiomatic relationship was obvious, correlations ranged from 0.959 to 0.996.

Key words: oral testing, neurolinguistics, testing constructs

1. Background

In the literature on language assessment, a number of studies including the author's own study[1] (Cheung, 2010) found strong correlations between verifiable quantitative measures (VQM) of assessment criteria in student language and subjective ratings of equivalent criteria (Banerjee, et al., 2007; Iwashita 2006; Ortega 1999, 2003; Wolfe-Quintero, et al., 1998). The strength of such correlations in the author's own study caused one commentator to suggest that the VQM and the ratings were all being driven by a hidden variable, 'general speaking ability' (GSA)[2] rather than a group of discrete language criteria and that correlations between VQM and ratings were due to this driving variable rather than the fact that raters were estimating the criteria measured by VQM. This suggestion caused the author to reflect on the nature of 'general speaking ability'. Two possibilities occurred in this process. Firstly, that general speaking ability was some kind of real-world entity which could 'drive' other measures of language which purported to measure distinct (though related) aspects of speaking ability: i.e. CAUSE. Secondly, GSA might exist only in the minds of the rater as informal averaging of proficiency on a number of discrete functions which have to be performed at once in order for speaking to take place i.e. EFFECT. Assuming the second possibility, GSA would not be able to 'drive' VQM since these are trans-subjective count-based indices performed by non-raters on transcriptions of student language. These two models are represented in Figures 1 and 2 as follows.

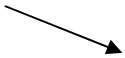
	SPEECH	**RESULTS**
GENERAL SPEAKING ABILITY	- stress - intonation - pronunciation - syntactic complexity - grammatical accuracy - T-units - ideas and organization	Verifiable Quantitative Measures (VQM) - stress VQM - intonation VQM - pronunciation VQM - syntactic complexity VQM - grammatical accuracy VQM - T-units VQM - meaningful clauses VQM
RATERS' RATING - ideas and organization - pronunciation and delivery - vocabulary and language patterns		

Figure 1: General Speaking Ability as CAUSE

[1] The author's study involves a stratified sample of 150 Secondary 3 (Grade 9) student oral pretest performances from the Hong Kong SAR.

[2] The author is grateful to Dr Glenn Fulcher for suggesting that she consider this alternative interpretation.

	SPEECH	Verifiable Quantitative Measures (VQM)
RATER ESTIMATES	- stress - intonation - pronunciation - syntactic complexity - grammatical accuracy - T-units - ideas and organization	- stress VQM - intonation VQM - pronunciation VQM - syntactic complexity VQM - grammatical accuracy VQM - T-units VQM - Meaningful clauses VQM



RATERS' WEIGHTING PROCESS
- rating GSA

Figure 2: General Speaking Ability as EFFECT

2. Neurolinguistic evidence as to GSA

Neurolinguistic evidence supports the view that general speaking ability (GSA) is an EFFECT in that such evidence indicates that speech results from separate functions exercised by separate parts of the brain. As early as the turn of the 20th century, studies of aphasic showed that one part of the brain, Broca's area (inferior frontal gyrus of dominant hemisphere) was required for grammatical speech and another part of the brain Wernicke's area (posterior superior temporal gyrus of dominant hemisphere) was required to make speech meaningful. Patients with damage to Broca's area were able to produce speech which was meaningful but lacking in grammar. Conversely, patients with damage to Wernicke's area were able to produce speech which was grammatical but lacking in meaning. Moreover, Hartsuiker et. al., (1999) all found that Broca's area damage was not a comprehension deficit, i.e. it did not interfere with comprehension even when it was interfering with grammar processing. The subjects in their study were able to comprehend a sentence spoken to them but when they tried to repeat the same sentence they could not produce its grammatical features (e.g. tense and subject-verb agreement). Similar findings were made by: Hagiwara (1995) with Japanese, Beretta et. al., (1999) with regard to Spanish, Penke (2000) with regard to German, Italian, French and Dutch, and Friedmann (2001) with regard to Hebrew and Arabic. However, as Berreta (2006) points out there are some differences in findings depending on the language of subjects in these studies (and the degree of damage to Broca's area) and there are two main hypotheses as to how damage to Broca's area causes agrammaticism; the Trace Deletion Hypothesis (Grodzhisky, 1995) and the Double Dependency Hypothesis (Mauner, et. al., 1993). Broca's area also seemed to be required for comprehension of sentences which were grammatical but atypical in word order (such as sentences in passive voice).

More recently left-side ideomotor apraxia has been found to be caused by a lesion of the anterior corpus callosum. Damage to this area interferes with the physical production of the syllables in what they patient is trying to say (phonetic encoding). This indicates yet another separate skill required for speaking.

Furthermore, Pell (1999) found that prosodic features of language such as stress and intonation were controlled by sections of the right brain hemisphere. Patients with damage to the right hemisphere were not able to produce these functions although they were able to produce other aspects of language such as meaning and syntax. Moreover, Shipley-Brown, et. al., (1988) found non brain damage that subjects were better able to interpret prosodic features if they went into the left ear (meaning that the stimulus went straight to the right hemisphere before it went to the left hemisphere) rather than into the right ear. They concluded that processing of intonation was occurring at least primarily in the right hemisphere. Reicker et. al., (2008) used nuclear magnetic resonance (NMR) imaging to discover that speech motor coordination (actually producing the syllables) relied on sites in the inferior frontal gyrus of the dominant hemisphere (usually left for right handers) with a supplementary area in the anterior cingualte gyrus.

3. What are the Components of Speech?

Neurolinguistic research tells us that there are at least four components to speech all sited in different areas of the brain: 1) Meaning: based in Werenicke's area (left hemisphere for right handers); 2) Syntax: based in Broca's area (left hemisphere for right handers); 3) Prosodic features (stress intonation) based in the right hemisphere (of right handers); 4) Speech motor coordination: based in the inferior frontal gyrus of the dominant hemisphere (usually right for right handers) with a supplementary area in the anterior cingualte gyrus).

All of these sites have an analogue of themselves in the opposite hemisphere to which they are located. These analogous sites show some excitation when the main brain sites, (e.g. in a right handed person Broca's area is in the left hemisphere) are active. However, the analogous sites have much lower levels of activity than the primary sites which they mirror.

In summary, psycholinguistic evidence indicates that that there are at least four different processes going on in the brain to produce speech while there is still some debate over the precise localization of syntactic functions. For example, Grodzisky (2000) holds that not all syntax processes occur in Broca's area even he maintains that 'syntactic abilities are nonetheless distinct from other cognitive skills and are represented entirely and exclusively in the left cerebral hemisphere' (for right handers) and that 'language is a distinct modularly organized neurological entity' (p. 1).

4. On Fluency

Fluency is usually measured as number of utterances over time. Logically, **fluency** must be an empirical measure of the speed and accuracy with which all of the 'modules' involved in language controlling 'meaning', 'grammar', 'prosodic features' and 'speech motor coordination' can be integrated. It is what we perceive as fluency, i.e. number of correct utterances over time. A glitch in any one of the speech production 'modules' can create hesitations and false starts which reduce fluency as measured by verifiable quantitative measures (VQM) and as perceived by raters/listeners. Since hesitations and false starts occur in almost all unrehearsed speech, it is reasonable to suppose that they are not always a result of neurological damage. Since and false starts occur about 16 time more often in L2 speakers than in an educated native speakers (Cheung, 2010, p. 107), it seems reasonable to infer that they can also be caused by any number of combinations of L1 interference, improper learning of syntax, speech motor functions and lexis. Therefore, fluency is an effect generated by the functioning of all the language 'modules' in the brain. However, fluency is not any sort of measure of 'general speaking ability'. Cheung (2010) found that numerous examples of student speech which was fluent but not accurate in terms of syntax or pronunciation and many more who were not accurate in stress and intonation.

5. A Brief Digression on the Neurolinguistics of Raters

As well as telling us about the language processing of speakers, neurolinguistic research also gives insight into the language processing of raters. Phillips et. al., (2001) show that the brain contains abstract representations of phonemes and compares sounds heard with these phonemes. Kutas and Hillyard (1980) identified a specific response to semantic anomalies, the P400. Osterhout and Holcomb (1992) shows that there was a specific brain response to syntactic errors, the P600 response. For example Hahne and Friederici (2002) show that when subjects were instructed to judge the 'acceptability' of sentences they did not show an N400 brain response (a response commonly associated with semantic processing), but when instructed to ignore grammatical acceptability and only judge whether or not the sentences made sense, the subjects did show the N400 brain response (Friederici, 2002). This is a highly specialized area and there is not enough space to cover it all in this paper but essentially neurolinguistics is indicating that human brains have the hardware and software for rating syntax and semantics and phonetics. It is not too much of a jump to suggest that raters are estimating levels of accuracy in various aspects of language. This would explain correlations between subjective ratings of performances and trans-subjective counts of particular aspects of a performance (e.g. pronunciation accuracy and grammatical accuracy) (Cheung, 2010, pp. 114 & 116). This explanation is empirically testable given co-operations between language assessment people and neurolinguists and does not require the invention of a vague entity such as GSA.

6. What is GSA and How if at all does it Exist?

As we have seen in sections 2.0 to 4.0 there are at least four distinct neurologically-based aspects to speech, 'meaning', 'syntax' and 'prosodic features' (stress/intonation) and 'speech motor coordination' (pronunciation). The fifth aspect of speech is fluency which logically must result from the integrated functioning of speech modules in the brain operating on learned syntax and lexis. However, one can see from the foregoing discussion, fluency stops way short of being a measure of general speaking ability (GSA). Therefore, the only way GSA can exist is as a perception in the mind of the listener as a kind of non-mathematical 'average' of how well a speaker is performing separate and distinct language functions, i.e. GSA exists only as 'effect'. Such an effect would explain why ratings on supposedly separate criteria and (even VQM for such criteria) correlate to an excessive degree as found in Cheung (2010, pp. 296-298) but

there are two other explanations for this phenomenon: 1) the smoothing effect of Rasch statistics; 2) axiomatic relationships between criteria.

7. The Smoothing Effect of Rasch Statistics

Some statistical means for representing scores and ratings on a uniform scale such as Rasch fair average can have a smoothing effect on ratings.

8. Axiomatic Relationships between Criteria and the research question

Axiomatic relationships between measures occur when results on one measure logically determine results in another measure. For example: getting a score for syntactic complexity (SC) requires that the structures being counted toward the SC score be CORRECT, although more points are given for more advanced structures (advanced being defined as less common in the sample). Therefore a student's grammatical accuracy (GA) is a necessary but not sufficient condition for getting a high score on SC. Imagine three students: Student A uses only simple common structures but gets them all correct. Student B uses simple common structures but gets them all wrong. Student C uses advanced (less common) structures but gets them all wrong. Student D uses advanced (less common) structures but gets them all right.

Grammar Accuracy	Rank Order	Syntactic Complexity	Rank Order
Student A	5	Student D	5
Student D	5	Student A	4
Student B	1	Student B	1
Student C	1	Student C	1

The above example shows how an axiomatic relationship can exist between grammatical accuracy (GA) and syntactic complexity (SC). GA contributes substantially to scores for SC although GA and SC are not the same thing. SC = accuracy + complexity so a high SC score is not possible without a high degree of GA. Moreover, SC and GA are both largely controlled by Broca's area. Conversely, SC and GA (controlled by Broca's area) should not have a high degree of axiomatic relationship with stress and intonation because the latter are prosodic features which we know are controlled largely by sites in the right hemisphere quite some distance away from Broca's area. However, we would expect a high direct axiomatic relationship between pronunciation accuracy (PA) and GA and a strong indirect axiomatic relationship PA→GA→SC because most grammatical features rely on word endings for their realization and hence require PA. The notable exception is canonical word order which is not so PA dependent. Stress (SA) and intonation (IA) are another story. They both affect intelligibility (Hahn, 2004, p. 201); therefore, they can ultimately affect IO ratings but axiomatically related to each other although they are both right hemisphere controlled. Proficiency in stress does not automatically mean proficiency in intonation, particularly for an L2 learner whose L1 is syllable timed and whose L2 is stress timed.

In a syllable timed language (e.g. Cantonese, Mandarin or Thai), intonation operates within syllable boundaries to differentiate between homophones and all syllables get equal stress. In a stress timed language particularly (e.g. English, Arabic, French), intonation contours run over strings of syllables to subtly modify clause or sentence meaning and stress sits on key syllables of key words. A learner coming from a syllable timed language background and trying to learn a stress timed language is bound to encounter more problems with intonation than with stress because: 1) mastering L2 stress only requires this learner to identify key syllables and make them louder and longer than unstressed syllables; 2) mastering L2 intonation requires the learner to suppress their own syllable bounded L1 intonation curves and run an intonation curve over a whole string of syllables (after first identifying the strings to be intoned). Given the foregoing, it seems reasonable that students such as those in Cheung (2010) would acquire SA later than pronunciation grammar and lexis and would acquire IA later than they acquired SA. Moreover, the distribution of these abilities found in Cheung (2010) bears this out. However, since SA and IA seem late acquired in developmental stage, they may correlate with other variables due to an intervening variable.

There should also be an axiomatic relationship between the rated criteria 'ideas and organization' (IO) and GA, SC and number of meaningful clauses (NMC). This relationship would pertain because SC and NMC all relate to the rating of IO because syntax is the organizing principle in language; its purpose is to provide a framework for semantic interpretation (Powers, 2004). However, there should also be an indirect axiomatic relationship between PA and IO i.e. PA→GA→SC→IO, since PA is a basic requirement for the realization of

many grammatical features. There would also be weaker relationships between SA and IO and IA and IO since both SA and IA contribute to intelligibility.

Our hypothesis was that: If GSA was only the effect on the rater caused by a set of separate but related linguistic functions, then correlations between ratings and VQM measures of assessment criteria should vary according to the degree of axiomatic relationship between these criteria and the degree of statistically smoothing function in the representation of criteria in results.

9. Methodology

In order to test the hypothesis that correlations between ratings on different criteria and between ratings and VQM of these criteria were affected by the degree of axiomatic relationship between them, we correlated all possible pairs of rating criteria and VQMs and analyzed these correlations in terms of number and degree of axiomatic relationships involved. A table of possible axiomatic relationships between criteria was drawn up and this was annotated with the relevant correlation figures so the effect of axiomatic relationships would be apparent. There were so many possible axiomatic relationships, they could not all be presented within the length constraints of this paper. Therefore, only results for axiomatic relationships for pronunciation accuracy (PA) of fluency (F), stress accuracy (SA) and intonation accuracy (IA) and assessment criteria 'ideas and organization' (IO), 'vocabulary and language patterns' (VL) and 'pronunciation and delivery' (PD) are shown. See Table 1 in Section 10.

10. Results and Discussion

Table 1.0 shows that ratings are ultimately being driven by actual measurable aspects of the text via axiomatic relationships between textual features. These can be direct, e.g. VQM of pronunciation accuracy (PA) and rating of 'pronunciation and delivery' (PD) or they can be indirect: PA→ grammatical accuracy (GA)→ syntactic complexity (SC) and rating ideas and organization (IO). If we look at relationships between verifiable quantitative measures (VQM) and the rated criteria IO, VL and PD (as well as IO, VL and PD against each other), we can see that there are two correlation figures in each box. The top figure is the figure derived from Rasch fair average of all raters. The (bottom) figure in brackets is derived from the raw rating of the best rater (in terms of Rasch statistics, correlation with expert panel and correlations against VQM). The difference between the figures shows the strength of the smoothing function of the Rasch fair average and how it can exaggerate correlations between ratings on different criteria and between ratings and VQM. While it is likely that some raters get an impression of GSA which drives their rating of ALL criteria, there is nevertheless evidence that these rating are driven by some trans-subjectively measurable features of the text the raters are examining. It can be seen that the raters' perception of IO, VL and PD correlates highly with the trans-subjective measures of the texts by VQM such as number of PA and fluency (F) and to a lesser extent stress accuracy (SA) and intonation accuracy (IA).

It seems SA and IA have a weaker relationship to other VQM than the other VQM have to each other, IA in particular. This is the same pattern is apparent between PA, SA and IA and rated criteria IO, VL and PD. The most likely explanation for this is that the stress and intonation are right brain functions whereas all other VQM represent left brain functions (see Section 2.0). We also have to remember the special problems that Cantonese background speakers (or anybody from a syllable timed language background) have with stress and intonation. Furthermore, Hong Kong students do not get much instruction in stress and intonation because large class sizes make this difficult in school and most Hong Kong students have little opportunity to mingle with native speakers and acquire stress and intonation by osmosis. Interestingly, SA did correlate relatively highly (0.810) with number of meaningful clauses (NMC). This is to be expected because correct stress helps to make a clause more meaningful to a listener by helping him/her identify key words. SA can even help intelligibility by making multi-syllabic words more comprehensible than they would be if uttered without correct stress, e.g. fifty, fifteen. Correlation between ratings and SA may also be due to developmental factors in that students have difficulty processing stress and intonation (especially intonation) until they are at ease with the more fundamental aspects of language. Developmental stage may well be a factor in the relationships between PA, SA and IA and between Fluency (F), SA and IA, since students need to acquire PA (which assists fluency) before they can acquire SA and IA.

Another point which becomes apparent from Table 1 is correlations between ratings and VQM are most likely driven by strong axiomatic relationships between VQM themselves. For example, PA correlates highly with IO rating at 0.896. This is most likely because PA is essential for grammatical accuracy (GA) which in turn is essential for syntactic complexity (SC). In other words, good pronunciation facilitates GA (because it allows

the realization of word endings). This in turn facilitates SC because only grammatically correct items are counted toward the SC score. In its turn SC facilitates the organization of ideas. This impresses raters to give a higher rating for IO. Direct axiomatic relationships between VQM are usually stronger than indirect ones. For example the axiomatic relationship is weaker between PA and SC (0.959) than between PA and GA (0.996) because some syntactic features, e.g. canonical word order, contribute heavily to the SC index but do not require fine distinctions in pronunciation. However, since PA facilitates GA which facilitates greater SC which impresses the rater in terms of the VL rating they give the student so we find that PA correlates highly 0.853 with VL. Since better pronunciation results in greater GA which in turn results in higher SC a student can with high PA can produce lots of meaningful clauses (NMC) which impress the raters in terms of the IO mark they give the student. Hence, PA drives IO ratings because of its effect on NMC via GA→SC. Finally, PA has a direct effect in the obvious place the PD mark 0.852. However, the PA drives IO more than the seemingly obvious PD. This is because PA only affects PD via one mechanism yet it affects IO through two mechanisms, PA→GA→SC and PA→GA→NMC.

Sometimes a textual feature can drive others by multiple mechanisms. For example, PA drives fluency in two ways. Firstly, PA is needed for grammatical accuracy (GA) which in turn needed for fluency PA→GA→F. Therefore, PA correlates with GA at 0.996 and GA correlates with fluency at 0.995. Secondly PA in itself enables fluency (PA→F) by preventing false starts. Hence the correlation between PA and fluency is high at 0.998. However, it is obvious that fluency cannot consist purely of grammar and pronunciation and in fact there is a high correlation between the token index (TI) (which measures the student's variety of lexis) and fluency, 0.961. However TI also seems to be driven by PA with a correlation of 0.958.

Table 1. Correlations between Criteria and Postulated Axiomatic Relationships between Criteria

Criteria	Postulated Axiomatic Relationships between Criteria	Corr
PA/F	Facility with pronunciation drives both their pronunciation accuracy (PA) score and their fluency (F) score although fluency is more concerned with speed.	0.998
PA/SA	PA and stress accuracy (SA) are related only in the sense of developmental level. They are functions of different areas of the brain.	0.791
PA/IA	PA and intonation accuracy (IA) are related only in the sense of developmental level. They are functions of different areas of the brain. Intonation facility is rarer in the sample than stress facility.	0.597
PA/IO	PA is a driver of SC (PA-SC) and token index (TI) (which measures lexical ability) both of which are essential for the organization of ideas.	0.896 (0.834)
PA/VL	PA drives GA and SC (PA-GA-SC). Since the range of lexis in the texts was limited GA and SC were by default the major determinants of PD rating since raters had little else to go on.	0.853 (0.743)
PA/PD	PA was the major determinant of PD scores since very few of the students in the sample showed any facility with stress or intonation.	0.852 (0.782)
F/GA	A thorough knowledge of grammar underlies both F and GA although GA is not the only component of fluency. Facility for recall of lexis as shown by TI is also a component of fluency as well as facility with Stress and Intonation (SA and IA respectively).	(0.995)
F/SA	SA drives F and both SA and F are related to a student's developmental stage.	(0.789)
F/IA	IA is probably a partial driver of F; both IA and F are related to a student's developmental stage.	(0.586)
F/IO	Fluency scores are driven by the same factors which contribute to IO ratings, P, GA, IA and SA. Fluency also makes the discourse more listenable and thus adds to intelligibility. In turn, intelligibility enhances the listener's impression of the organization of ideas (IO).	0.891 (0.826)
F/VL	Fluency scores are driven by the same factors which contribute to VL ratings, (P, GA, IA, TI and SA); however, fluency also requires these language processes to be done quickly.	0.844 (0.732)
F/PD	Fluency scores are driven by some of the same factors which contribute to PD scores, P, SA, and IA.	0.843 (0.781)
SA/IO	SA drives IO in that accurate stress placement contributes to intelligibility. SA and IO are probably both related to the student's developmental stage, very few students in the sample showed high SA levels and many showed SA levels close to zero. The students with high SA levels also got high IO ratings as well as high GA, SC, PA scores and high VL and PD ratings.	0.774 (0.673)
SA/PD	Stress accuracy is a partial driver for PD, the others being PA (major) and IA (minor). SA and PD are both related to developmental stage.	0.748 (0.662)
SA/VL	SA and VL are probably both related to the student's developmental stage, very few students in the sample showed high SA levels and many showed SA levels close to zero. The students with high SA levels also got high VL ratings as well as high GA, SC, PA scores	0.738 (0.610)
IA/IO	IA drives IO in that accurate intonation contributes to intelligibility. IA and IO are probably both related to the student's developmental stage, very few students in the sample showed high IA levels and many showed IA levels close to zero (IA was overall less coming than PA or SA). The students with high SA levels also	0.690 (0.516)

	got high IO ratings as well as high GA, SC, PA scores and high VL and PD ratings	
IA/VL	IA is a minor driver of VL in that it contributes slightly to intelligibility (less than SA which is in turn less than IA). IA and VL are probably both related to the students developmental stage, very few students in the sample showed high IA levels and many showed IA levels close to zero (IA was overall less common than PA or SA). The students with high SA levels also got high IO ratings as well as high GA, SC, PA scores and high VL and PD ratings.	0.722 (0.535)
IA/PD	IA a minor driver of PD Less than SA which is in turn less than PA. IA and PD are probably both related to the students developmental stage, very few students in the sample showed high IA levels and many showed IA levels close to zero (IA was overall less common than PA or SA). The students with high SA levels also got high PD ratings as well as high GA, SC, PA scores and high IO and VL ratings.	0.720 (0.537)

Remark: Correlation is significant at the 0.01 level (2-tailed).

11. Conclusion

It is possible to argue that correlations between verifiable quantitative measures (VQM) and ratings do not indicate causality i.e. that raters are 'paying attention' to aspects of the text measured by VQM. Some might argue that 'if each VQM is highly related to the others and are predictive of general speaking ability (GSA), then they will correlate highly with any reasonable set of speaking scores'. However, the problem with this hypothesis is that it assumes an unproven, complicating factor (see Occam's Razor[3] (Ockham, 1495)). It assumes the existence of something called 'general speaking ability' (GSA) to explain correlations without any proof that this entity exists. In fact as we have seen from the neurolinguistics evidence presented earlier in Section 3 'language is a distinct modularly organized neurological entity' (Grodzisky, 2000, p. 1).

On the other hand, VQM do exist independent of ratings and they are counted from observable aspects of the student performances. Ratings also exist and this study observes raters making them. Thus, the hypothesis that raters are paying attention to textual features which are measured by VQM is the simplest theory which fits the facts (i.e. correlations between VQM and ratings). Also, the fact that ratings for different criteria correlate well with each other does not mean they are all measures of GSA. Correlations between ratings and on different criteria e.g. vocabulary and language patterns (VL) and pronunciation and delivery (PD) can be explained by axiomatic relationships between distinct aspects of language, e.g. pronunciation accuracy (PA) which affects PD rating directly and affects VL indirectly via its effect on grammar PA→GA→SC→VL or PA→GA→VL and its effect on lexis PA→token index (TI)→VL. Moreover, correlations between VQM show it is possible for a student to give good performance on VQM like syntactic complexity (SC) and grammatical accuracy (GA) (and poor performance on others like intonation and stress). We have also seen that high correlations between ratings for different criteria seem to result in part from the smoothing effect of the Rasch fair average. With 'raw' scores these correlations drop to a level explicable by axiomatic relationships between textual features measured by VQM. Finally, the relationship between stress accuracy (SA) and intonation accuracy (IA) and the other VQM aspects of the text is an area which needs some attention. Stress and intonation are right brain functions which indirectly affect the realization of left brain functions such as semantics, syntax, phonology and lexis. SA and IA do this by increasing intelligibility, thus in turn, increasing rater perceptions of PD and VL. Stress and intonation also assist in textual organization thus increasing IO ratings.

Although statistically significant at a high level, SA and IA do not seem to correlate as well with other VQM or with ratings as do other features of the text. There may be a developmental pattern peculiar to Hong Kong students. This may result from the way English is taught in Hong Kong. It may also result from the fact that the first language of most Hong Kong is Cantonese: a syllable timed language with intonation curves locked to syllables. The answer to this question may lie in studies of language learners who come from stress timed languages more akin to English. Hebrew and Arabic are obvious candidatures since they are stress timed with intonation curves running across syllables (like English) but (like Cantonese) they have few items of lexis in common with English. A study where students produced oral presentations on a topic say 'Christmas in Oman' (similar to the authors topic 'Christmas in Hong Kong') could yield recordings which could be transcribed for VQM calculation and all VQM could be correlated against each other. This would enable us to see if stress and intonation still had the same relativity to other VQM as that found in the author's study. This in turn would enable us to learn more about L1/L2 relativity in the ontogenesis of stress and intonation.

[3] The principle is called Occam's Razor although its namesake William of Ockham spelt his surname differently: it states that 'entities should not be multiplied without necessity' in the construction of scientific theory. In other words, one should prefer the simplest explanation which fits the facts.

Another vital area of future research is the neurolinguistics of raters. Studies in this vein could look for distinct patterns of brain excitation raters looking criteria such as IO, VL and PD. Such studies may identify the formation of an impression of GSA by the raters and examine how this relates to specific excitation patterns formed when raters rate specific aspects of language. If there is a GSA, it will be found not in the speech of students but in the brains of raters. Even if GSA exists it is something we need to work around rather than towards. If assessment is to be valuable for teaching and learning, it must chart specific strengths and weaknesses for capitalization and remediation respectively. GSA is of no use for this purpose and may even defeat it because GSA perceptions may colour ratings of one criterion in the light of other unrelated criteria which for some reason, have an unduly powerful effect on particular raters.

References

Banerjee, J., Franceschina, F., & Smith, A. M. (2007). *IELTS Research Reports Volume 7*. IELTS Australia and British Council.

Beretta, A., Pinango, M., Patterson, J., & Hartford, C. (1999). Recruiting comparative crosslinguistic evidence to address competing accounts of agrammatic aphasia. *Brain and Language 67*, 149-168

Beretta, A. (2006). Agrammatism II: Linguistic models. *Encyclopedia of Linguistics and Languages*, pp.126-131. Oxford: Elsevier..

Cheung, K. M. A. (2010). Reliability and validity in practice: Hong Kong's Key Stage 3 oral assessment. Macquarie University, Australia, Unpublished PhD thesis.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *TRENDS in Cognitive Sciences 6* (2): 78-84.

Friedmann, N. (2001). Agrammatism and the psychological reality of syntactic tree. *Journal of Psycholinguistic Research, 30*(1), 71-90.

Grodzhisky, Y. (1995). A restrictive theory of trace deletion agrammatism. *Brain and Language 51*, 26-51.

Grodzhisky, Y. (2000). The neurology of syntax: Language use without Broca's area. *Behavior and Brain Sciences 23*, 1-71.

Hagiwara, H. (1995). The breakdown of functional categories and economy of derivation. *Brain and Language 50*, 92-116.

Hahne, A., & Friederici, A. D. (2002). Differential task effects on semantic and syntactic processes as revealed by ERPs. *Cognitive Brain Research 13*, 339-356.

Hahn, L. D. (2004). Primary Stress and Intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly 38* (2), 201-223.

Hartsuiker, J., Kolk, H., & Huinck, W. (1999). Agrammatic production of subject-verb agreement: the effect of conceptual number. *Brain and Language 69*, 119-160.

Iwashita, N. (2006). Syntactic complexity measures and their relation to oral proficiency in Japanese as a foreign language. *Language Assessment Quarterly, 3*, 151-169. Lawrence Erlbaum Associates, Inc.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science* 207 (4427): 203–205.

Mauner, G., Fromkin, V. A., & Cornell, T. L. (1993). Comprehension and acceptability judgments in agrammatism: disruptions in the syntax of referential dependency. *Brain and Language, 45*(3):340-70.

Ockham, W. (1495). *Sentences of Peter Lombard* (Quaestiones et decisiones in quattuor libros Sententiarum Petri Lombardi (ed. Lugd., 1495), i, dist. 27, qu. 2, K). Summa Totius Logicae, i. 12.

Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second_Language Acquisition, 21*, 109-48.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24(4)*, 492-518.

Osterhout, L., & Holcomb, P. J. (1992). Event-related potentials elicited by grammatical anomalies. *Psychophysiological Brain Research*: 299–302.

Pell, M.D. (1999). Fundamental frequency encoding of linguistic and emotional prosody by right hemisphere-damaged speakers. *Brain and Language, 69* (2), 161-192.

Penke, M. (2000): Unpruned trees in German Broca's aphasia. *Behavioral and Brain Sciences, 23*(1), 46-47.

Phillips, C., Pellathy, T., Marantz, A., Yellin, E., Wexler, K., McGinnis, Poeppel, M.D., & Roberts, T. (2001). Auditory cortex accesses phonological category: an MEG mismatch study. *Journal of Cognitive Neuroscience 12* (6): 1038–1055.

Powers, D. M. W. (2004). Robot babies: What can they teach us about language acquisition? In J. Leather, & J. Van Dam, (Ed.), *The ecology of language acquisition* (pp. 160-182), New York: Kluwer Academic.

Riecker, A., Brendel, B., Ziegler, W., Erb, M., & Ackermann, H. (2008). The influence of syllable onset complexity and syllable frequency on speech motor control. *Brain and Language, 107*(2), 102-113.

Shipley-Brown, F., Dingwall, W. O., Berlin, C. I., Yeni-Komshian, G., & Gordon-Salant, S. (1988). Hemispheric Processing of Affective and Linguistic Intonation Contours in Normal Subjects. *Brain and Language 33*, 16-26.

Wolfe-Quintero, K., Inagaki, S., & Kim, H-Y. (1998). Second language development in writing: Measures of fluency, accuracy and complexity. *Technical Report 17*. Honolulu: University of Hawaii at Manoa, Second Language Teaching and Curriculum Centre.