A basic analysis of reliability and validity of A Descriptors-based Rating Scales

Yi-qi Wang

APD 1292: Instrument Design and Analysis

Final Assignment

Ontario Institute for Studies in Education,University of Toronto

A basic analysis of reliability and validity of A Descriptors-based Rating Scales

Introduction

"Various quality indicators, assessment tools, and measurement methods are widely used in education and social sciences. They are crucial for informing theories and advancing practice." (Jang, 2015) Through taking the course of instrument design and analysis, the author has learned core concepts and principles required for high-quality instrument design and analysis and have gained hands-on experiences with instrument design, analysis, and interpretations.

This paper is the final project for this course. Empirical data from a writing test that used a descriptor-based rating scale was supplied so that the author has the opportunity to report some findings of interest.

I Descriptor-based rating scales

A rating scale is a set of categories designed to elicit information about a quantitative or a qualitative attribute. In the language assessment, a rating scale is a method that requires the rater to assign a value, sometimes numeric, to the rated object, as a measure of some rated attribute. A rating scale can be "used to model examples of good work as a success criteria, to distinguish good from poor quality work, to provide formative feedback for learning by highlighting what a competent learner can do, and to facilitate self regulation". (Jang, 2015)

Several classifications of rating scales have been proposed in the literature. The most commonly cited categorization is that of holistic and analytic scales (Hamp-Lyons, 1991; Weigle, 2002). Weigle summarizes the differences between these two scales in terms of

six qualities of test usefulness (p. 121), showing that analytic scales are generally accepted to result in higher reliability, have higher construct validity for second language writers. Because analytic scales measure writing on several different aspects, better diagnostic information can be expected.

Another possible classification of rating scales represents the way the scales are constructed. Fulcher (2003) distinguishes between two main approaches to scale development: intuitive methods or empirical methods. Intuitively developed scales are developed based on existing scales or what scale developers think might be common features at various levels of proficiency. Typical examples of these scales are the FSI family of scales. In recent years, a number of researchers have proposed that scales should be developed based on empirical methods. Examples of such scales are those produced by North and Schneider (1998) who proposed the method of scaling descriptors, Fulcher's data-based scale (1996) as well as Upshur and Turner (1999) and Turner and Upshur's (2002) empirically derived, binary-choice, boundary definition (EBB) scales.

Rating scales commonly used in the assessment of writing have been criticized for a number of reasons. The first criticism is that they are usually intuitively designed and therefore often do not closely enough represent the features of candidate discourse. Furthermore, Brindley (1998) and others have pointed out that the criteria often use impressionistic terminology which is open to subjective interpretations (Upshur & Turner, 1995; Watson Todd et al., 2004). The band levels have moreover been criticized for often using relativistic wording to differentiate between levels (Mickan, 2003), rather than offering precise and detailed descriptions of the nature of performance at each level.

The problems with intuitively developed rating scales described above might affect the raters' ability to make fine-grained distinctions between different traits on a rating scale. This might result in important diagnostic information being lost. Similarly, if raters resort to letting an overall, global impression guide their ratings, even when using an analytic rating scale, the resulting scoring profile would be less useful to candidates. It is therefore doubtful whether intuitively developed rating scales are suitable in a diagnostic context.

The descriptor-based rating scale analyzed in this study is an analytic one developed by empirical method. It was used in a diagnostic assessment research to G5 and G6 students in Ontario, Canada. According to the traits of itself and the targeted research, this rating scale can be considered a good choice.

II Inter-Item Correlation (Chronbach's Alpha)

Chronbach's Alpha is more commonly used in psychometric and risk assessment tool research and is seen by many to be the most important index of test reliability (Kline, 2000). Essentially the Cronbach's Alpha is a measure of the correlation of each item in a test with each and every other item in a test. Calculating a Cronbach's Alpha is usually done with the help of statistical packages, such as SPSS.

Test developers often start with a large number of items that they pilot on a population. They calculate the Chronbach's Alpha and then they delete items (those that correlate least with the other items in the test) until they obtain an Alpha of an acceptable standard (usually 0.7 or above). This method is more common in psychometric tests that assess a specific psychological construct such as depression, self-esteem etc. A good tool (where all factors are linked to recidivism) should have a reasonable level of internal consistency.

III Internal validity

Internal validity refers to how well an construct of instrument is designed, especially whether it avoids confounding (more than one possible independent variable [cause] acting at the same time). The less chance for confounding in a construct, the higher its internal validity is. Therefore, internal validity refers to how well the construct of a instrument allows you to choose among alternate explanations of something. A construct with high internal validity lets you choose one explanation over another with a lot of confidence, because it avoids (many possible) confounds.

IV The current study

This study sought to analyze item and scale statistics, including reliability and internal validity for a descriptors-based rating scale for writing assessment used in Jang et al.s' 2015 research, to establish whether this rating scale has acceptable inter-item reliability and internal validity.

The study was conducted in two main phases. During the first phase, the reliability analysis phase, 18 items with 123 variables each were analyzed using Chranbach's alpha . The Chranbach's alpha measure was selected because it is a function of the number of items in a test, the average covariance between item-pairs, and the variance of the total score. (Wikipedia) Based on the findings in Phase 1, whether the inter-item reliability is acceptable can be observed.

During the second phase of this study, the internal validity analysis phase, principal axis factor analysis was applied with the same data to demonstrate whether the 3 factors construct in the rating scale is valid.

This paper reports on the findings from the both phases. The overarching research question for the whole study is as follows:

Does this empirically developed rating scale of writing with level descriptors based on discourse analytic measures have acceptable inter-item reliability and internal validity so as to be used in the diagnostic writing assessment?

V Method

5.1 Data collection and Participants

Data for this paper was supplied by Professor Eunice Eunhee Jang. The sample consisted of 44 students from two Grades 5 and 6 classrooms, taught by the same female literacy teacher, in a private school located in southern Ontario, Canada.

3 writing assessment tasks was used to estimate students' writing skills mastery levels. In the rating scales, the 20 writing items were divided into 3 factors as organization, convention and content. These items assess students' ability to organize main ideas and supporting details using correct spelling, grammar and punctuation in variety of written formats, as described in the Ontario curriculum. Unfortunately, the data of item 7 and 18, and the data of 9 writing tasks out of 132(44x3) are all missing. Then after the cleaning up, the data of 18 items with 123(44x3-9) variables are valid.

5.2 Measures

5.2.1 Inter item reliability

To assess the inter item reliability, the Chranbach's alpha as described earlier was used (see Appendix for sample items).

In analyzing the data, firstly the author intended to ensure that these items (1 through 18) all reliably measure the same latent variable (i.e., writing competence). To

test the internal consistency, the Chronbach's alpha test was run using the reliability analysis command in SPSS.

Chronbach's alpha is a measure of internal consistency, that is, how closely related a set of items are as a group. It is considered to be a measure of scale reliability. (Technically speaking, Chronbach's alpha is not a statistical test - it is a coefficient of reliability or consistency).

5.2.2 Internal validity

To assess the inter-item reliability, the principal axis factor analysis as described earlier was used (see Appendix for sample items).

In addition to computing the alpha coefficient of reliability, the author also want to investigate the dimensionality of the scale, although the original rating scale has already been categorized into 3 factors as organization, convention and content. The author used the factor Analysis command in SPSS to do this.

Factor analysis is a method of data reduction. It does this by seeking underlying unobservable (latent) variables that are reflected in the observed variables (manifest variables, here namely the items).

For this scale, the author did a rather "plain vanilla" factor analysis. The author used iterated principal axis factor with eigenvalues greater than one as the method of extraction and a varimax rotation. The determination of the number of factors to extract should be guided by theory, but also informed by running the analysis extracting different numbers of factors and seeing which number of factors yields the most interpretable results.

VI. Results and discussion

6.1 Inter-item reliability

Table 1

Case Processing Summary

		N	%
Cases	Valid	123	99.2
	Excluded[a]	1	.8
	Total	123	100.0

a. Listwise deletion based on all variables in the procedure.

Table 2

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.868	.863	18

Table 3

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
O1	35.85	50.160	.259	.172	.868
O2	35.80	49.619	.329	.380	.866
O3	34.53	46.383	.549	.563	.858
O4	34.83	44.540	.613	.698	.855
O5	35.13	43.536	.696	.674	.850
O6	35.34	43.798	.736	.717	.849
O8	34.63	46.499	.551	.585	.858
C9	34.35	47.519	.538	.664	.859
C10	34.43	47.008	.592	.552	.857
C11	35.61	46.670	.512	.445	.860
C12	35.22	48.951	.282	.316	.869
C13	34.25	51.117	.171	.144	.870
C14	34.45	46.993	.578	.639	.857
C15	35.18	47.537	.346	.355	.868
C16	34.69	46.629	.448	.599	.863
CO17	34.30	50.114	.257	.484	.868
CO19	35.85	48.804	.461	.412	.862
CO20	35.38	44.832	.646	.584	.853

Table 1-3 show item-total correlations, and reliabilities of the writing scales. The findings show that scale items had acceptable part-whole corrected item-total correlations for all scales, with none of the correlations falling short of .10 and only 4 of the correlations falling short of the .30 threshold. To interpret the total internal reliability, the author followed the rule of George and Mallery (2003):

> .9 (Excellent), > .8 (Good), > .7 (Acceptable), > .6 (Questionable), > .5

The alpha coefficient for the 18 items is .868, suggesting that the items have relatively high internal consistency. (a reliability coefficient of .70 or higher is considered "acceptable" in most social science research situations.)

Of all the items, only C12, C13, if deleted, would slightly increase the value of Chranbach's alpha. However, the little increments caused separately by these two items are acceptable. So all the items in the rating scale are considered valid. In sum, these findings indicate that the rating scale show that reliabilities range from acceptable to good.

6.2 Internal validity **Table 4**

Total Variance Explained

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.857	32.541	32.541	5.475	30.419	30.419	2.814	15.635	15.635
2	1.869	10.383	42.923	1.441	8.004	38.423	2.804	15.579	31.215
3	1.708	9.487	52.410	1.300	7.222	45.644	1.601	8.893	40.108
4	1.303	7.239	59.649	.863	4.795	50.439	1.495	8.305	48.413
5	1.015	5.638	65.287	.550	3.057	53.496	.915	5.083	53.496
6	.983	5.458	70.745						
7	.913	5.069	75.815						
8	.780	4.335	80.150						
9	.600	3.331	83.481						
10	.572	3.181	86.662						
11	.493	2.737	89.398						
12	.398	2.213	91.611						
13	.348	1.931	93.542						
14	.327	1.815	95.357						
15	.254	1.411	96.769						
16	.224	1.246	98.014						
17	.183	1.016	99.031						
18	.174	.969	100.000						

Extraction Method: Principal Axis Factoring.

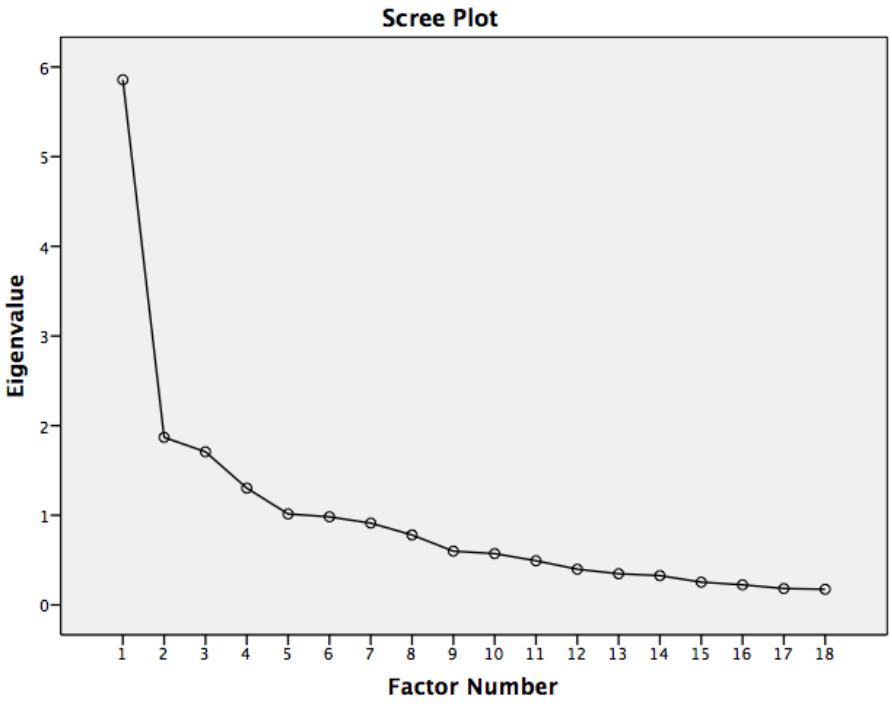

Figure 1

Table 5
Rotated Factor Matrix[a]

	Factor				
	1	2	3	4	5
O6	.708	.264	.253	.228	.110
CO20	.700	.230	.113	.062	.323
C11	.642	.163	.076	.067	.072
O2	.582	.080	-.004	.093	-.242
CO19	.550	.056	.119	.015	.266
O5	.544	.425	.255	.243	-.052
C14	.180	.819	.128	.005	.128
C9	.119	.799	.077	.163	.056
O3	.294	.647	.058	-.011	.191
C10	.151	.622	.135	.385	.091
C16	.181	.063	.833	.114	.094
O4	.226	.252	.783	.280	.078
CO17	.080	-.016	.111	.745	-.074
O8	.152	.373	.144	.709	.075
O1	.201	.027	.159	.202	.014
C15	.402	-.061	.094	.042	.525
C12	.052	.335	-.032	-.015	.497
C13	-.026	.161	.140	-.021	.219

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 9 iterations.

The author was curious about whether groups of traits the data was measuring was 3 or not. Then a principal axis factor analysis (or principal factor analysis – PFA) was performed on the rating data. PFA reduces the data in hand into a number of components, each with an eigenvalue representing the amount of variance of the components. Components with low eigenvalues (below 1.0) are discarded from the analysis, as they are not seen to be contributing enough to the overall variance. Table 4 above shows the results from the principal factor analysis. Both the scree plots (Figure 1) and the tables displaying the results from the PFA show that when the existing rating scale was analyzed, five major components but three was found. These components had eigenvalues above 1.0 and cumulatively accounted for about 54% of the entire variance. All other eigenvalues were clearly below 1 (following Kaiser, 1960) and there was no further leveling off point on the scree plot. The next step in the PFA was to identify which variables load onto which component. For this, a rotation of the data was necessary. A varimax rotation was chosen to facilitate the interpretation of the factors of the scale. A trait was considered to be loading on a factor if the loading was higher than .2 (as indicated in bold font). (see table 5)The five factors loadings for the scale can be seen in Table 5. The largest factor, accounting for 16% of the variance, was made up of item 6, 20,11, 2, 19 and 5. This factor can be described as a factor of construct, coherence and cohesion. The second factor, which accounted for a further 16% of the variance, was made up of item 14, 9, 3 and 10. This factor can be described as lexical and syntactical ability. The third factor, which accounted for 9% of the variance, consisted of item 16 and 4, the section in which writers are required to use appropriate genre and tone to finish the tasks. The fourth factor, which accounted for 8% of the variance, comprises item 17, 8 and 1. This factor assesses the competency of effectively addressing the task and topic.

Item 12, 13 and 15 were the measures that loaded on the fifth factor, which accounted for another 5% of the variance. This factor is also about the lexical ability but in a more detailed level compared to factor two. The five factors together accounted for 55% of the entire variance of the score.

Then another PFA and varimax rotation were performed on these rating data with fixed number of factors as 3. The three factors loadings for the scale can be seen in Table 6. These three factors don't have the same categories of the items as the three factors in the original rating scale and the three factors together only accounted for 44% of the entire variance of the score.

Table 6

Rotated Factor Matrix[a]

	Factor			
	1	2	3	4
CO20	.747	.291	.065	.123
O6	.707	.249	.324	.204
C11	.639	.161	.140	.038
CO19	.595	.116	-.008	.141
O5	.492	.363	.399	.178
O2	.477	.012	.238	-.076
C15	.474	.078	-.082	.172
C14	.166	.815	.121	.106
C9	.083	.776	.278	.042
O3	.300	.673	.061	.047
C10	.124	.596	.456	.111
C12	.148	.424	-.116	.058
C13	.019	.213	-.067	.182
CO17	.048	-.076	.686	.090
O8	.132	.326	.684	.134
O1	.192	.009	.230	.140
C16	.205	.066	.181	.798
O4	.226	.229	.372	.761

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

It can therefore be argued that the five factors construct generated by PFA and varimax rotation not only accounted for more aspects of writing ability, but it also accounted for a larger amount of variation of the scores. In other words, there was less unaccounted variance when the 3 factor construct was used than the new 5 construct.

VII Conclusion

The findings of this little study have two of implications. The first refers to the reliability of rating scale. This descriptor based rating scale has acceptable inter-item reliability. Another implication relates to internal validity. The three factors construct (organization, convention and content) of the rating scale accounts for less aspects of writing ability and less variance. A five factors construct (construct, coherence and cohesion; lexical and syntactical ability; ability to use appropriate genre and tone; competency of effectively addressing the task and topic; specific lexical ability), which could account for more aspects of writing ability and more variance, is suggested to the instructors.

VIII Reference

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

ETS. (2004). TOEFL iBT writing rubrics. Retrieved April 10, 2014, from

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: Longman/ Pearson Education.

Jang, E, E., Dunlop, M., Wagner, M., Kim,Y & Gu, Z. (2013). Elementary School ELLs' Reading Skill Profiles Using Cognitive Diagnosis Modeling: Roles of Length of Residence and Home Language Environment. Language Learning, 63(3), 400–436,

Jang, E, E., Dunlop, M., Park, G & Boom, E. (2015). How do young students with different profiles of reading skill mastery, perceived ability, and goal orientation respond to holistic diagnostic feedback?. Language Testing 0265532215570924

Jang, E, E. (2015). Instrument design [PowerPoint Slides]. Retrieved from lecture notes online: https://portal.utoronto.ca/webapps/portal/frameset.jsp?tab_tab_group_id=_2_1&url=%2Fwebapps%2Fblackboard%2Fexecute%2Flauncher%3Ftype%3DCourse%26id%3D_755335_1%26url%3D

Jang, E, E. (2015). Design rating scale [PowerPoint Slides]. Retrieved from lecture notes online: https://portal.utoronto.ca/webapps/portal/frameset.jsp?tab_tab_group_id=_2_1&url=%2Fwebapps%2Fblackboard%2Fexecute%2Flauncher%3Ftype%3DCourse%26id%3D_755335_1%26url%3D

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

Mickan, P. (2003). 'What's your score?' An investigation into language descriptors for rating written performance. Canberra: IELTS Australia.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. Language Testing, 15(2), 217–263.

Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary Educational Psychology*, *36*(1), 36-48.

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. A., & 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.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L.Hamp-Lyons (Ed.), Assessing second language writing in academic contexts. Norwood, NJ: Ablex.

Weigle, S. C. (2002). Assessing writing . Cambridge: Cambridge University Press.

Appendix

Figure 2 Traditional rating scale for writing assessment

Level 1	**Level 2**	**Level 3**	**Level 4**
include **a few** simple ideas with **minimal** development structure writing through simple sequencing or listing, but ideas may be repeated or confusing use **some** simple sentences that may or may not include basic punctuation	provide **few** details to support and develop ideas use simple logical structures (e.g., simple sequence, introduction/ conclusion) but may include details that are confusing or sound like a simple list use **some** common transition words (e.g., first, next, secondly) to link ideas make simple sentences with accurate punctuation spell familiar grade-level words correctly or phonetically	clearly express ideas with relevant supporting details, but **some** details may be vague or limited organize ideas into paragraphs use dialogue, quotations, word choice, etc., to help the flow of ideas use conventional spelling, punctuation and grammar	develop ideas with details that make their main idea clear and consistent select words and phrases that make their meaning clear organize ideas logically into well-developed paragraphs with effective transition words use a variety of organizational patterns to structure their writing combine sentences in different ways using a variety of connecting words

Figure 3 Descriptor-based rating scale for writing assessment

	Student can:	Mastery (60%-100%)	Transition (40%-60%)	In need of support (0%-40%)
Organization (8 descriptors) Generate, gather, and organize ideas and information to write for an intended purpose and audience	1. Organize ideas into paragraphs			
	2. Use some common transition words (e.g., first, second, next) to link ideas			
	3. Organize ideas logically			
	4. Use genre-appropriate organizational patterns to structure writing			
	5. Clearly express main ideas with relevant supporting details			
	6. Organize writing using a clear introduction, body, and conclusion			
	7. Present ideas by considering different perspectives and make connections between ideas			
	8. Generate ideas relevant to audience and appropriate for the purpose			

Category	**Student can:**	Mastery (60%-100%)	Transition (40%-60%)	In need of support (60%-100%)
Convention: Use knowledge of language conventions to edit errors, refine expressions, and present writing effectively	9. Choose appropriate words for conveying the intended meaning			
	10. Use parts of speech correctly to communicate meaning			
	11. Combine sentences in a variety of ways using various connecting words			
	12. Use conventional spelling, punctuation and grammar			
	13. Spell familiar words correctly			
	14. Select words and phrases that make meanings clear			
	15. Word choices are appropriate for the purpose			
	16. Use a tone appropriate for the purpose			

Content (4 Descriptors) write a topic relevant to the task and develops ideas with details that make the main idea clear and strong	17. Write a topic relevant to the writing task			
	18. Use relevant information from other resources (reading selection)			
	19. Use vivid (figurative) language and innovative expressions to enhance interest			
	20. Use relevant details, personal thoughts and effective word choices to make writing interesting and engaging			

Category	Descriptors: Student can:	Mastery	Transition	Need help
Organization: Generate, gather, and organize ideas and information to write for an intended purpose and audience	1. Organize ideas into paragraphs	3	2	1
	2. Use some common transition words (e.g., first, second, next) to link ideas			
	3. Organize ideas logically			
	4. Use genre-appropriate organizational patterns to structure writing			
	5. Clearly express main ideas with relevant supporting details			
	6. Organize writing using a clear introduction, body, and conclusion			
	7. Present ideas by considering different perspectives and make connections between ideas			
	8. Generate ideas relevant to audience and appropriate for the purpose			
Convention: Use knowledge of language conventions to edit errors, refine expressions, and present writing effectively	9. Choose appropriate words for conveying the intended meaning			
	10. Use parts of speech correctly to communicate meaning			
	11. Combine sentences in a variety of ways using various connecting words			
	12. Use conventional spelling, punctuation and grammar			
	13. Spell familiar words correctly			
	14. Select words and phrases that make meanings clear			
	15. Can make word choices are appropriate for the purpose			
	16. Use a tone appropriate for the purpose			
Content: write a topic relevant to the task and develops ideas with details that make the main idea clear and strong	17. Write a topic relevant to the writing task			
	18. Use relevant information from other resources (reading selection)			
	19. Use vivid (figurative) language and innovative expressions to enhance interest			
	20. Use relevant details, personal thoughts and effective word choices to make writing interesting and engaging			