Language Testing in Asia
a SpringerOpen Journal

RESEARCH Open Access

Organization of ideas in writing: what are raters sensitive to?

Rachael Ruegg[1*] and Yuko Sugiyama[2]

* Correspondence: rachaelruegg@gmail.com
[1]Akita International University, Akita, Japan
Full list of author information is available at the end of the article

Abstract

Whether foreign language writing is rated using analytic rating scales or holistically, the organization of ideas is invariably one of the aspects assessed. However, it is unclear what raters are sensitive to when rating writing for organization. Which do they value more highly, the physical aspects of organization, such as paragraphing and the existence of organization markers, or deeper textual aspects, such as the coherent flow of ideas? This study investigates whether raters of timed essays value paragraphing, cohesive devices or coherence more when assigning a score for organization. The current study used multiple regression to ascertain what raters of the writing section of an in-house proficiency test were sensitive to when rating writing for organization using an analytic rating scale. The number of paragraphs, number of cohesive devices and coherence within 116 timed essays were evaluated and it was found that raters seem to value the physical aspects of organization more than the deeper textual aspects. Specifically, the number of paragraphs and the number of cohesive devices predicted differences in scores assigned for organization. In addition, the scores assigned for content were significantly predictive of scores for organization.

Keywords: EFL writing, Assessment, Timed writing, Writing test, Organization, Essay structure, Coherence, Cohesive devices, Cohesion

Background

Whether holistic rating or analytic rating scales are used, in the assessment of writing the organization of ideas (sometimes called essay structure) is invariably taken into account. Likewise, in evaluation of writing by classroom teachers organization of ideas is often highly valued. However, the concept of organization is very broad (Cummings et al. 2001; Freedman, 1979). What constitutes effective organization seems to vary widely between teachers, institutions and geographical areas. There are many different rhetorical features that can be taken into account. Some in the field of language education may look for the presence of thesis statements and topic sentences (MacIntyre, 2007), while others may focus more on paragraphing, cohesive devices or coherence (Erdosy, 2004), for example.

The present study was carried out on the writing section of an in-house proficiency test with a timed essay task in a Japanese university which is rated by instructors from a wide range of cultural and linguistic backgrounds. Analytic rating scales are used to rate the essays and the organization scale specifies coherence and structure as its

defining features (See Additional file 1). Each essay is twice rated and Rasch modelling is used to scale the scores taking into account rater severity or leniency.

Coherence is a concept which may not be understood clearly and may be understood in different ways by different raters. Knoch (2007: 109) suggests that one reason for "vague descriptions of coherence might lie in the rather vague nature of coherence". Furthermore, the concept of structure is not defined within the analytic rating scales used in this study and raters from different backgrounds may have different conceptions of the ideal essay structure. In addition, anecdotal evidence from the test's administration suggests that the organization rating scale is one that raters apply less consistently.

Although the reliability of scores assigned to writing by raters can be increased through rater training, raters are humans and, as such, they will always vary in their interpretations. In fact, Knoch (2007) states that up to 35% of variance in writing scores may be attributable to a lack of interrater reliability. Likewise, Lumley (2002, 2005) asserts that it is the rater, rather than the rating scale that occupies the pivotal position in the centre of the rating process. Therefore, the purpose of this study is to ascertain what rhetorical features raters value when assessing writing for organization. This study considers whether essay length, content scores, paragraphing, cohesive devices and coherence, predict scores assigned for organization. The research question for the study is: Are raters more sensitive to paragraphing, the use of cohesive devices or coherence when rating timed writing for organization using an analytic rating scale?

Literature review

Defining organization

In both analytic and holistic scoring, the organization of academic writing is of importance to teachers and institutions. This is evidenced by the fact that in many analytic rating scales for the assessment of writing, organization is one of the analytic rating scales. The SAT test in the United States (SAT. n.d.), IELTS. (n.d.), the STEP test (Lumley, 2002) and the ESL Composition Profile (Jacobs et al. 1981) are a few examples of assessment instruments that include organization as one aspect to be considered.

Some studies have shown that raters highly value organization in essays. A study by Harris (1977) had 36 high school English teachers rate 12 papers, half of which were strong on content and organization and half of which were strong on mechanics and sentence structure. The papers were ranked based on the teachers' perception of the overall quality. There was a tendency for teachers to provide more feedback on content and organization than sentence structure and mechanics in the L1 setting. Freedman (1979) also conducted a study on raters' perceptions of a piece of academic writing and found that the rewriting of essays to be stronger in content and organization had a greater impact on scores in comparison to those that were rewritten to be stronger in sentence structure and mechanics. It can be concluded from these studies that raters consider organization to be of great importance in academic essays.

However, the raters in Freedman's (1979) study said that organization was the second most difficult construct to assess with the content construct being the most difficult. Although raters find organization to be important, assessing and measuring organization is a challenge. Often, raters complain "that the exact nature of the

construct they assess remains uncertain" (Cummings et al. 2001: 3). In the studies by Harris (1977) and Freedman (1979), organization was not clearly defined, which makes it difficult to identify which aspects of organization raters find challenging to assess.

In practice, the interpretation of organization may vary for each rater. As Erdosy (2004) states, "constructs such as 'content' and 'organization' have as many manifestations as there are raters" (p. 10). Furthermore, even if the constructs are defined in rubrics, each rater interprets the constructs differently (MacIntyre, 2007). Similarly, Vaughan (1991) states that raters have their own rating styles which they rely on during the rating process. Despite these differences, there are some common features that raters focus on when rating for organization.

MacIntyre (2007) conducted think aloud research on the KEPT and investigated whether the analytic rating scales used were effective in assessing writing. Six raters of the KEPT were asked to verbalize their thoughts while they were rating essays using the rating scales. In his study, the raters mentioned variables related to organization, such as transitions, the number of paragraphs and whether the essay has an introduction, body and conclusion or not, which indicate that these are some common features that raters tend to look for when rating writing for organization.

A similar study conducted by Erdosy (2004) examined think-aloud protocols of four experienced raters of the TOEFL TWE and compared the ways they evaluated a written text and how their backgrounds affected their ratings. In the protocols, teachers identified cohesive devices, length, and rigid paragraph structure to be important when judging writing proficiency. Raters in his study also took into consideration paragraph structure, organizational patterns, coherence and cohesion as part of organization.

To a certain degree, raters in the abovementioned studies focussed on similar concepts in organization. Both MacIntyre (2007) and Erdosy (2004) found that overall essay structure (i.e. introduction, body, conclusion), paragraph structure, use of cohesive devices and coherence are valued when assessing writing for organization. Similarly, according to Freedman (1979), strong organization included proper paragraphing, logical order of presentation, and appropriate cohesive devices. In a similar vein, this present study focused on the following characteristics: paragraphing, the use of cohesive devices and coherence.

Cohesion and coherence

An important concept related to organization is the concept of coherence (i.e. Erdosy, 2004; Kobayashi and Rinnert, 1992; MacIntyre, 2007). Coherence is an abstract concept (Connor, 1990) and it has not been clearly defined in the literature on the topic (Grabe & Kaplan, 1996; Johns, 1986; Lee, 2002). However, within discussions of coherence in the literature, two important distinctions have been made. The first is between coherence at the surface level (Lee, 2002), as achieved through the addition of cohesive devices, and coherence between propositions. The second is between coherence at the local level (between sentences) and at the global (discourse) level.

Early literature on coherence considered coherence at the surface level only and researchers often looked for cohesive ties. It was assumed that if a text was held together with cohesive ties then it was coherent. Witte and Faigley (1981) found that high-rated essays had more cohesive ties than low-rated essays when the essays were rated using a holistic scale. The researchers concluded that cohesion is an important property of writing.

More recently, the term 'metadiscourse' has been used. According to Hyland and Tse (2004), the word 'metadiscourse' refers to a range of different devices writers use to organize their text. It helps readers understand the connection between the ideas, thus making the text easy to follow.

Metadiscourse is important in all formal writing, but it is claimed to be especially crucial in argumentative writing, to facilitate persuasion (Crismore et al. 1993; Hyland, 2004; Hyland & Tse, 2004; Williams 1981). Vande Kopple (1985) states that without recourse to metadiscourse, we would be able to write lists of sentences, but not cohesive texts and we would only be able to express our ideas to a minimal extent.

The use of cohesive devices is often taught in academic writing courses to improve the flow of the written text (Folse et al. 2010; Hirose, 2005; Hyland, 2004; Raimes, 2004). Furthermore, the use of no, or minimal, cohesive devices would put onus on the reader to connect the ideas as opposed to the writer explicitly spelling out their meaning. Japanese is a reader responsible language (Carson, 1993; Hinds, 1987), whereas during English language instruction, Japanese students are often taught to state their ideas explicitly in English. Therefore, cohesive devices are particularly important for Japanese students in their attempt to make their text more writer responsible.

However, the definition of coherence as consisting of cohesive devices alone was criticized by later writers and the presence of cohesive devices came to be known as cohesion which was distinguished from coherence (Bamberg, 1984; Carrell, 1982; Johns, 1986). That a text is cohesive does not necessarily imply that the text is coherent (Carrell, 1982; Morgan & Sellner, 1980; Van Dijk, 1980; Witte & Faigley, 1981). Witte and Faigley (1981: 200) argue that "a cohesive text may be only minimally coherent." In addition to being cohesive, it is considered that a text must be coherent at the ideational level (Bamberg, 1984; Carrell 1982; Lee, 2002). Johns (1986), discusses the 'sticking to the point' feature of coherence. She states that some writers have investigated how the content of sentences combines to lead the reader through the text. Similarly, Carrell (1982) discusses 'content coherence.'

Other writers have noted that the connection between propositions creates local coherence, but it creates a globally coherent text if information is also connected at the discourse level (Bamberg, 1983 1984; Lee, 2002; Van Dijk, 1980). Global coherence is concerned with "what the essay is about" (Connor & Farmer, 1991: 128).

According to Bachman and Palmer (2010), organizational knowledge is one of the two areas of language knowledge and encompasses grammatical knowledge and textual knowledge. Textual knowledge in turn encompasses knowledge of cohesion and knowledge of rhetorical organization. More specifically, knowledge of cohesion is demonstrated by "producing. . .explicitly marked relationships among sentences in written texts" (2010: 45) using cohesive devices such as 'therefore', 'on the other hand', and 'however'. Knowledge of rhetorical organization is demonstrated by sequencing units of information appropriately within texts. In the present study, rhetorical organization is referred to as coherence. Based on the literature, it appears that both cohesion and coherence are essential in communicating meaning to raters; therefore, the researchers examined not only the use of cohesive ties, but also looked at whether there was any connection between sentences on the ideational level.

In this current study, the analytic rating scales used have an "organization" scale and "content" scale. The organization scale mentions coherence, whereas the content scale

is concerned with the logical connection between ideas, development and support. Since ideas such as topical development and logical connections between ideas are related to global coherence (i.e. Bamberg, 1983; Carrell, 1982; Connor 1990; Lee, 2002), this study assumes that global coherence should be encompassed by the content scale for these rating scales and therefore focuses on local coherence.

Methods

Although most studies (e.g. Erdosy, 2004; Lumley, 2005; MacIntyre, 2007) have focused on identifying variables that possibly lead to higher organization scores, the results of these in-depth and qualitative studies were based on a small number of participants. The present study was conducted with a more quantitative approach and with a larger number of participants in order to determine what elements of organization raters are more sensitive to when assigning analytic scores for organization. For this reason, rather than focusing on one particular aspect of organization, as many other studies have, this study attempts to take into account different aspects which may be considered by raters to constitute the construct 'organization'.

One hundred and sixteen essays were randomly selected from the March 2009 administration of the in-house test and used as samples for the current study. The writers were high school graduates who were entering a foreign language university in eastern Japan. The raters were 45 lecturers and learning advisors at the university. All the raters took part in a rater norming session within a week of the test administration.

Rater Norming

In the days prior to the rater norming session, raters were required to collect a rater packet which had been put together by members of the test committee. The committee met and discussed a number of sample essays from the previous administration at length until they came to an agreement on the ratings for those essays. Subsequently, rationale for the scores on three of those essays were discussed and a document was typed which contained both the agreed-upon scores and the rationale for those scores for three essays. The rationale for the organization scores mentioned: the overall essay structure, the presence of paragraph level and sentence level transitions and clarity. One example of a rationale in terms of clarity was "The reader is wondering, what exactly is the writer trying to say? While some amount of implication is justified in academic writing, the message does need to be clearly communicated."

An additional four essays were used during the rater norming session. In addition to this document, the packet contained the rating scales and photocopies of the three sample essays. Raters were instructed to at least read all the sample materials to refamiliarize themselves with the rating scales. Some of the raters preferred to rate the sample essays themselves before reading the scores and rationale prepared by the group.

The rating session lasted for two and a half hours. The raters sat in groups of six. Each group included one member of the test committee who had the role of facilitator. Raters were instructed to first refer to their group facilitator if they had any questions. If they needed further help the two norming session leaders could be asked.

Initially, the raters were reminded of the rating guidelines and how to document the grades. Following this, one essay was distributed. Each individual was required to rate the essay themselves before discussing the ratings with the other five members in their group. Through the group discussion, each group was required to reach a consensus about the ratings for the essay. Each group discussed the scores they had assigned the essay for each rating scale in turn, until they came to a score for each rating scale that everyone was satisfied with. The facilitator's role was to keep the discussion moving in a timely manner and ensure that raters were referring to the rating scales in their justification of the scores given. As long as the raters referred to the rating scales when justifying their scores, the facilitator would not comment on the actual numeric scores assigned. When a group consensus had been reached, the ratings were reported by each group and documented on the white-board. Once this process was complete, five of the six group members would move to a new group, using a pattern that ensured that each rater never worked with another rater more than once. This process was repeated three more times so that raters had rated and reached a consensus on four sample essays by the end of the norming session. At the end of the norming session, raters were asked not to discuss their ratings with anyone else during the rating so that they remained consistent in the way they rated.

Each essay was rated twice. The inter-rater reliability of the ratings given for organization for the 116 essays used in this study was 0.763. Inter-rater reliability rates of 0.7 or higher are acceptable (Zhang & Li, 2004). The ratings were then scaled using Rasch modelling.

Evaluation of texts

In order to investigate the research question, first of all it was necessary for the researchers to determine organizational qualities in each of the 116 sample essays. According to the literature reviewed, organization often refers to paragraphing, the use of cohesive devices and coherence. Therefore, the following qualities were evaluated in each of the essays: the number of paragraphs, the coherence of information within paragraphs and the number of cohesive devices used.

In this study, the analysis was done collaboratively rather than independently as it was considered important that the figures could be agreed upon by both researchers, therefore no inter-rater reliabilities were calculated (For a similar method see: Ferris, 2006).

The first evaluation was the number of paragraphs present in the essay. This may seem like an overly simple measure, both easy to evaluate and too simplistic to use to determine the organizational quality of an essay. However, this measure was far from simple. As most Japanese students have little, if any, experience of writing in English before entering university, they also mostly do not know the paragraphing conventions of English prose. Therefore, what seemed like a simple task; counting the number of paragraphs in each essay, turned out to be quite an endeavour. Many essays had paragraphs just slightly indented, making it unclear whether or not this was intentional. One possible reason for this practice is that in Japanese writing, a new paragraph is indented by just one character. Another paragraphing characteristic found in some essays was a change of margin which was the opposite of what one would expect, the first

line starting at the left margin and the remaining lines hanging. Ultimately, the judgment as to how many paragraphs an essay had was carried out on a case-by-case basis, with both researchers coming to agreement in every case as to how many paragraphs each essay had.

The reason for including this measure in the analysis is that, at first glance, the more paragraphs an essay has the more organized it appears to be. Furthermore, Charney (1984: 75) states that "in spite of training, readers' judgments are strongly influenced by salient, though superficial, characteristics of writing samples. . .one of these superficial characteristics is physical appearance." It was considered that with the time pressures involved in the rating process, raters may take this into account when assessing writing for organization. Indeed, the think-aloud research conducted by MacIntyre (2007) shows that some raters do take the number of paragraphs into consideration when rating writing for organization. Anecdotal evidence gained during the rater norming sessions in 2008 and 2009 also indicated that, although the number of paragraphs is not mentioned as a factor in the organization rating scale, some raters were taking this into consideration when rating writing for organization using the rating scales.

The number of cohesive devices used was another measure of organizational quality. Initially, a list of cohesive devices was built through a brain-storming session. However, ultimately writers used some cohesive devices that were not on the list and in the same way, many of the cohesive devices thought of by the researchers were not used by the writers. Therefore, each word was considered individually by the two researchers who subsequently met to discuss their findings. In this way, the final list of cohesive devices emerged from the data. The researchers agreed on what should and should not be counted as a cohesive device in every case. The cohesive devices were organized into the categories used by Hyland and Tse (2004).

Often cohesive devices were incorrectly used by the writers. It was considered by the researchers that, as misuse of cohesive devices would serve to confuse the message rather than making it clear, only correctly used cohesive devices would be included in the evaluation of cohesion. Additional file 1 shows all the cohesive devices that were used correctly by writers in the 116 essays.

In this study, local coherence was considered rather than global coherence, that is; coherence between sentences within paragraphs was evaluated rather than a coherent relationship between main ideas. It would be expected that topic shifts would be accompanied by a new paragraph, whereas all the information within one paragraph should be connected into a coherent whole. This concept is often referred to as 'paragraph unity' (Johns, 1986; Lee, 2002). To measure the coherence within an essay, first the number of sentences in the essay was counted. The number of paragraphs was deducted from the number of sentences as it was assumed that the first sentence in each new paragraph would not follow coherently from the last sentence of the preceding paragraph. The researchers then individually went through each essay determining which sentences followed coherently from the sentence previous. The overall coherence evaluation was a proportion; the number of sentences that followed coherently, divided by the number that should be expected to follow coherently from the previous. For example; if an essay had five paragraphs and in total 18 sentences. First of all, five was deducted from 18 resulting in 13 sentences that should expect to follow on coherently from each other. If in fact, 10 such coherent relationships were found, then the

proportion of coherence was evaluated as 10/13 (0.7692). The following sentences were judged to lack a coherent relationship: "The best important point is the feeling. It isn't easy to get married but they help each other and they can live together." The second sentence here does not appear to be about feeling; therefore, these sentences lack a coherent relationship. On the other hand, these sentences were judged to have a coherent relationship: "If people get married after the age of thirty, probably it will be hard for women. I have heard that it will be difficult that women who beyond the age of twenty-seven have babies." In the first sentence the writer states that it is difficult for women and in the second sentence the writer expands on this point by saying why it is difficult for women, therefore, these sentences were judged to have a coherent relationship.

The essay with the lowest overall scores and the essay with the highest overall scores are included in the appendices. (Additional file 1 respectively)

Analysis

Multiple regression was used to determine which of the organizational qualities predicted a higher score on the organization scale. The Rasch adjusted organization score for each essay was used as the dependent variable. The independent variables were: essay length, Rasch adjusted content score, number of paragraphs, number of cohesive devices and coherence.

Anecdotal evidence suggests that it may be difficult for raters to distinguish between the quality of the content and that of organization. For this reason, the content scores assigned by raters were included in the analysis to see what role they play in the prediction of organization scores.

Essay length may inflate writing test scores regardless the quality of the essay. Weigle (2002) states that "A number of L1 studies have demonstrated that length. . .is a significant predictor of holistic scores" (p. 69). On the other hand, in some studies (such as Whithaus et al. 2008) longer essays have been found to be penalized. However, in this 2008 study the essays ranged from 75 words up to over 600 words. It appears that extremely lengthy essays may be negatively affected, whereas at the shorter end of the spectrum, extremely short essays may be penalized. In the current study, the length of each essay was included in the analysis to ascertain the extent to which the quantity of writing predicts variation in organization scores. Length scores were determined by the number of words in the essay.

Results

The descriptive statistics for the dependent variable and the independent variables are shown in Table 1. The organization scores and the content scores both ranged from 0.2 to 4, on a scale of 0 to 4. The length of the essays ranged from 80 words to 271 words. The number of paragraphs in each essay ranged from one to five. The number of cohesive devices in each essay ranged from 0 to 9. The coherence evaluations ranged from 0.33 to 1.0. That is, one third to all of the sentences followed coherently from the previous one on the ideational level.

All skewness and kurtosis measures fell within the acceptable range (–2 to 2), indicating that the data are sufficiently close to standard distribution to validly conduct

multiple regression. The results of the multiple regression are shown in Table 2. Multiple regression shows to what extent changes in each independent variable are predictive of changes in the dependent variable (organization).

The R squared value demonstrates that the variables used account for a large portion of overall variance in organization scores. For multiple regression analysis, an R squared value of 0.8 or above can be considered a large effect (Cohen, 1992). The results show that high scores on the organization rating scale are significantly predicted by scores on the content scale (p = 0.000), the number of paragraphs an essay has (p = 0.000) and the number of cohesive devices used (p = 0.015). On the other hand, the negative results for the length and coherence variables show that longer essays and ideationally more coherent essays tended to predict lower scores on the organization rating scale, although these relationships are not statistically significant.

Discussion

Coherence is a variable that did not significantly predict organization scores. This finding was not surprising. Although coherence is listed as something to think about when rating essays for organization, the content scale specifies logical connection between ideas and the development of ideas as two important aspects. Through the literature (i.e. Bamberg, 1983 1984; Carrell, 1982; Connor, 1984; Knoch, 2007; Lee, 2002; Williams, 1981), a number of different definitions can be found, some of which relate to the ideational connection between sentences, which is referred to in the organization scale and some of which refer to topic development, which clearly pertains to the content scale in these rating scales. Indeed, coherence as a construct is not even well defined (Knoch, 2007; Lee, 2002). It appears that some consider local coherence, the way it was defined for this study. Others seem to consider global coherence, for example, topical development. In the analytic rating scales used in this study, coherence is mentioned quite clearly as a defining property of organization; however, coherence is not well defined in the rating scales. The possibility that raters all have different working definitions of coherence certainly poses an issue for the test statistics.

A significant finding of this study is the strong relationship between organization and content. The scores on the content scale predicted the scores on the organization scale to a greater extent than any of the organizational qualities analyzed. Currently, these scales are too strongly related to be able to say that they are assessing different constructs.

Table 1 Descriptive statistics

Variable	Mean	SD	N	Skewness	Kurtosis
Organization	2.1139	1.00985	115	0.259	−1.133
Content	2.1687	0.99617	115	−0.018	−1.287
Length	140.6140	50.67710	115	0.939	0.003
Paragraphs	2.1913	1.40735	115	0.652	−1.100
Cohesive devices	1.8696	1.93998	115	1.295	1.473
Coherence	0.8200	0.18284	115	−0.915	−0.020

Table 2 Multiple regression

Variable	B	Std. error	Beta	T	Sig.
Constant	.235	.207	-	1.134	.259
Content	.688	.065	.681	10.612	.000*
Length	-.001	.001	-.028	-.473	.637
Paragraphs	.177	.037	.249	4.769	.000*
Cohesive devices	.070	.028	.135	2.480	.015*
Coherence	-.060	.252	-.011	-.239	.811
$R^2 = .820$					

Dependent Variable: Organization.
* Significant at the 0.05 level.

Conclusions

The context of the present study poses a problem when it comes to interrater reliability for assessments of organization in writing. The 45 raters of the March 2009 administration of the in-house proficiency test came from 10 different countries on four different continents. Anecdotal evidence suggests that raters from different cultural and educational backgrounds hold different beliefs about writing. Furthermore, Charney (1984) states that for holistic ratings to be considered valid, the raters should come from similar backgrounds. Cooper (1977) goes further and says that raters should come from similar educational backgrounds so that they can draw on similar experience. When it comes to rating writing for organization, this appears also to be true of ratings assigned using analytic rating scales. One example of this can be seen in the think aloud data collected by MacIntyre (2007). One participant mentions the use of a question at the beginning of the essay as positive, calling it a 'hook', while another comments that it is inappropriate to use rhetorical questions in academic writing. This is just one of a myriad of writing conventions that are considered differently between even Anglophone cultures. With so many different raters from so many different cultural and educational backgrounds it is especially difficult to find agreement about issues such as organization. Even raters who are trained and use a common rating scale may rate idiosyncratically (Charney, 1984). In addition to this, the raters have hugely varying amounts of experience in teaching and rating writing. This may compound the problem even more.

Physical appearance is a characteristic that appears to be valued when rating. Charney (1984) states that "scores can only reflect agreement on salient but superficial features of the writing such as the quality of the handwriting or the presence of spelling errors" (p. 78). It seems that the three organizational qualities analyzed for the purpose of this study can be divided into two categories; the number of paragraphs and the number of cohesive devices both being superficial physical characteristics whereas coherence is a deeper textual characteristic. While a rater could evaluate the essay based on the number of paragraphs and cohesive devices present after merely scanning it, they would need to read more closely to evaluate the ideational coherence of the text. It seems that raters took the easy route in evaluating the relative organizational quality of the essays. In the context of this study all members of the English Language Institute rate essays as a contractual requirement. It is unclear whether the lack of deeper reading stems from the obligatory nature of the rating task or whether this also occurs in other contexts because of the time pressure involved in rating.

The relationship between organization and content

There are a number of possible reasons for the strong relationship between organization scores and content scores in this study. Several possible reasons will be outlined below.

The rating scales may be written in a way that makes it difficult for raters to clearly understand the two different skill sets they are supposed to be assessing. One indication that this may be the case is the existence of the word 'connection' in both scales. The word is intended to mean two different things; connection between sentences in the organization scale and 'logical connection between ideas' in the content scale, however, this is not clearly stated and therefore may create ambiguity for raters.

Some (such as Lumley, 2002) have stated that rater training is more important than the actual instrument used for rating. It is possible that the constructs of organization and content and their defining qualities is something that needs to be discussed and clarified to a greater extent during rater norming sessions.

It is possible that essays that are strong in terms of organization are also generally strong in terms of content. As it is a group of students who are culturally and linguistically very homogenous that wrote the essays analyzed for the present study it seems possible that those who had learnt English composition at high school were strong in both their content and their organization skills and those who had not were weak in both. Writing instruction may lead to the parallel development of both of these skill sets. Unfortunately, information about the backgrounds of the students involved in this study was not available and therefore this hypothesis could not be examined. Further investigation may be able to explore this possibility.

Suggestions for further research

The findings of this study suggest the need for more research into the different ways in which people understand the concept of coherence in academic writing. This could be done in relation to the rating of writing, through think aloud protocols, or simply through interviews and would represent a valuable contribution to the field of composition.

This study represented an observation of the way in which raters rated timed essays for organization using an analytic rating scale designed for the purpose. The results could be verified through a study with an experimental design involving the manipulation of the different organizational qualities of essays before rating.

Replicating this study with different rating scales, or without rating scales, would help to clarify whether the findings of this study result from the rating scales or whether raters in other contexts are sensitive to the same organizational qualities.

Additional file

Additional file 1: Appendix.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

RR and YS together designed the study, collected and analysed the data and wrote the manuscript. Both authors read and approved the final manuscript.

Authors' information
Rachael Ruegg holds an MA in Applied Linguistics and is currently a PhD candidate. She lectures at Akita International University in Akita, Japan. Her research interests include writing, assessment and vocabulary.
Yuko Sugiyama received her MA in TESOL in 2008. Yuko is currently a lecturer at Kanda University of International Studies in Chiba, Japan. Her research interests are assessment, writing and bilingualism.

Author details
[1]Akita International University, Akita, Japan. [2]Kanda University of International Studies, Chiba, Japan.

Received: 21 March 2013 Accepted: 21 March 2013
Published: 8 April 2013

References

Bachman, L, & Palmer, A. (2010). *Language assessment in practice*. Oxford: Oxford University Press.
Bamberg, B, (1983). What makes a text coherent? *Coll Compos Commun, 34*(4), 417–429.
Bamberg, B, (1984). Assessing coherence: A reanalysis of essays written for the National Assessment of Educational Progress, 1969–1979. *Res Teach Engl, 18*(3), 305–319.
Carrell, P, L, (1982). Cohesion is not coherence. *TESOL Q, 16*(4), 479–488.
Carson, J, (1993). Reading for writing: Cognitive perspectives. In J. Carson & I. Leki (Eds.), *Reading in the composition classroom: Second language perspectives* (pp. 85–104). Boston: Heinle & Heinle.
Charney, D, (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Res Teach Engl, 18*(1), 65–81.
Cohen, J, (1992). A power primer. *Psychol Bull, 112*(1), 155–159.
Connor, U, (1984). A study of cohesion and coherence in English as a second language students' writing. *Res Lang Soc Interact, 17*(3), 301–316.
Connor, U, (1990). Linguistic/rhetorical measures for international persuasive student writing. *Res Teach Engl, 24*, 67–87.
Connor, U, & Farmer, M, (1991). The teaching of topical structure analysis as a revision strategy for ESL writers. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom*. Cambridge: Cambridge University Press.
Cooper, C, (1977). Holistic evaluation in writing. In C. Cooper & L. Odell (Eds.), *Evaluating writing*. National Council of Teachers of English: Urbana, IL.
Crismore, A, Markannen, R, & Steffensen, M, (1993). Metadiscourse in persuasive writing: A study of texts written by American and Finnish university students. *Writ Commun, 10*(1), 39–71.
Cummings, A, Kantor, R, & Powers, D, (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making, and development of a preliminary analytic framework. TOEFL Monograph Series*. Princeton, NJ: Educational Testing Service.
Erdosy, M, U, (2004). *Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions. (TOEFL Research Report No. RR-70)*. Princeton, NJ: ETS.
Ferris, D, (2006). Does error feedback help student writers? In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (p. 89). New York: Cambridge University Press.
Folse, K, S, Solomon, E, V, & Clabeaux, D, (2010). *From great paragraphs to great essays*. Boston, MA: Heinle Cengage Learning.
Freedman, S, (1979). How characteristics of student essays influence teachers' evaluations. *J Educ Psychol, 71*(3), 328–338.
Grabe, W, & Kaplan, R, B, (1996). *Theory and practice of writing*. London: Longman.
Harris, W, (1977). Teacher response to student writing: A study of the response patterns of high school English teachers to determine the basis for teacher judgment of student writing. *Res Teach Engl, 11*, 175–185.
Hinds, J, (1987). Reader versus writer responsibility: A new typology. In U. Connor & R. B. Kaplan (Eds.), *Writing across languages: Analysis of L2 text* (pp. 141–152). Reading, MA: Addison-Wesley.
Hirose, K, (2005). *Product and process in the L1 and L2 writing of Japanese students of English*. Hiroshima: Keisuisha.
Hyland, K, (2004). Disciplinary interactions: Metadiscourse in L2 postgraduate writing. *J Second Lang Writ, 13*, 133–151.
Hyland, K, & Tse, P, (2004). Metadiscourse in academic writing. *Applied Linguistics, 25*(2), 156–177.
IELTS. (n.d.). *An overview of IELTS academic writing*. Retrieved from http://www.ielts.org/pdf/Writing%20Band%20descriptors%20Task%202.pdf.
Jacobs, H, L, Zinkgraf, S, A, Wormuth, D, R, Hartfiel, V, F, & Hughey, J, B, (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.
Johns, A, (1986). Coherence and academic writing: Some definitions and suggestions for teaching. *TESOL Quarterly, 20*(2), 247–265.
Lee, I, (2002). Teaching coherence to ESL students: A classroom inquiry. *J Second Lang Writ, 11*, 135–159.
Lumley, T, (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Lang Test, 19*(3), 246–276.
Lumley, T, (2005). *Assessing second language writing: A rater's perspective*. Frankfurt: Peter Lang Publishing.
Knoch, U, (2007). 'Little coherence, considerable strain for reader': A comparison between two rating scales for the assessment of coherence. *Assess Writ, 12*, 108–128.
Kobayashi, H, & Rinnert, C, (1992). Effects of first language on second language writing: Translation versus direct composition. *Lang Learn, 42*(2), 183–215.
MacIntyre, R, (2007). Revision of a criterion-referenced rating scale used to assess academic writing. *Studies in Linguistics and Language Teaching, 18*, 203–219.
Morgan, J, L, & Sellner, M, B, (1980). Discourse and linguistic theory. In R. J. Spiro, B. C. Bertram, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Raimes, A, (2004). *Grammar troublespots: A guide for writers*. New York: Cambridge University Press.

SAT. (n.d.). *Essay scoring: How it's scored and what the scores mean*. Retrieved from http://sat.collegeboard.org/scores/sat-essay-scoring-guide.
Vande Kopple, W, (1985). Some exploratory discourse on metadiscourse. *Coll Compos Commun, 36*(1), 82–93.
van Dijk, T, A, (1980). *Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognition*. Hillsdale, NJ: Erlbaum.
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* (pp. 111–125). Norwood, NJ: Ablex.
Weigle, S. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.
Whithaus, C, Harrison, S, & Midyette, J, (2008). Keyboarding compared to handwriting on a high stakes writing assessment: Student choice of composing medium, raters' perceptions and text quality. *Assess Writ, 13*(1), 4–25.
Williams, J, (1981). *Style: Ten lessons in clarity and grace*. Glenview: Scott Forseman.
Witte, S, P, & Faigley, L, (1981). Coherence, cohesion, and writing quality. *Coll Compos Commun, 32*(2), 189–204.
Zhang, P, & Li, N, (2004). An assessment of human-computer interaction research in management information systems: Topics and methods. *Computers in Human Behaviour, 20*, 125–147.

doi:10.1186/2229-0443-3-8
Cite this article as: Ruegg and Sugiyama: **Organization of ideas in writing: what are raters sensitive to?.** *Language Testing in Asia* 2013 **3**:8.