Research Forum

Examining Personality Bias in Peer Assessment of EFL Oral Presentations: A Preliminary Study

Mitsuko Tanaka
Ritsumeikan University

In this study, I explored a potential personality bias in peer assessment of EFL oral presentations. First-year Japanese university students enrolled in an oral presentation class (N = 21) made presentations and evaluated their classmates' presentations over two semesters. Rater severity was estimated using the many-facet Rasch measurement model. Raters' personality traits were assessed based on their responses to a questionnaire containing 4 variables: dogmatism, individuality, evaluation apprehension, and dependency on others. The results of 2 multiple stepwise regression analyses showed that whereas personalities were not associated with rater severity in the beginning, dependency on others and evaluation apprehension significantly predicted rating severity as time went by. Whereas those with high dependency on others (who valued harmony with others) became more lenient, those with high evaluation apprehension became more severe in their assessment of their classmates' presentations. These findings indicate a potential personality bias in peer assessment of EFL oral presentations.

本研究は英語で行うプレゼンテーションに対する学生間相互評価において評価者の性格によるバイアスの有無を検証することを目的とする。一クラス21名の大学一年生が二学期にわたりプレゼンテーションおよび相互評価を行った。評価者の厳しさは多相ラッシュモデルで分析した。評価者の性格は、「独断性」、「個の認識・主張」、「評価懸念」、「他者への親和・順応」を含む質問紙への回答によって測定した。重回帰分析の結果、初めは評価者の性格と評価の厳しさに関連はないものの、時間の経過とともに「評価懸念」と

「他者への親和・順応」が評価の厳しさに影響を及ぼすことが分かった。具体的には、他者との関係を重視する学生の評価は甘くなり、他者からの評価を気にする学生の評価は厳しくなることが明らかになった。以上のことからオーラルプレゼンテーションの学生間相互評価においては学生評価者の性格に起因するバイアスが生じえることが示された。

Oral presentation is one of the tasks that are often used in EFL speaking classes in Japanese tertiary education. Peer assessment is incorporated into class activities in some EFL oral presentation courses. In general, peer assessment benefits learners as it tends to improve the quality and effectiveness of learning (Topping, 2009). Researchers in the EFL setting have also pointed out the numerous positive effects of peer assessment on learning (e.g., Azarnoosh, 2013; Cheng & Warren, 2005; Otoshi & Heffenen, 2007). For example, through peer assessment students can recognize assessment as a shared responsibility and thus can be involved in learning more autonomously. Additionally, they can understand the assessment criteria more clearly and reflect on their performance and learn more deeply by observing their peers' performance critically. Despite the acknowledged educational benefits of peer assessment, many teachers might feel hesitant about incorporating it into a formal grading system because its reliability has not been empirically established.

In general, rater variability, which has been characterized as "variability of scores awarded to examinees that is associated with characteristics of the raters and not with the performance of examinees" (Eckes, 2015, p. 39), exists in performance assessments regardless of rater types (e.g., teachers and students). One such rater variability is rater severity. Examinees of the same performance ability may pass or fail depending on the severity of raters. Raters differ in the severity or leniency with which they rate (Eckes, 2005). Student raters are also assumed to display such variance in rater severity in peer assessment.

Although many factors may affect rater severity—such as personality traits, rating experience, rating purposes, workload, and demographic characteristics (Eckes, 2015)—the present study focused on personality traits. When rapport is built among students in class, some students, such as those who value harmony with others, may give more supportive ratings to their peers' performances than other students do. Thus, personality traits may be a source of systematic variance affecting rater severity. The aim of the present study was to examine a potential rater bias derived from personality traits in peer assessment in an EFL oral presentation classroom.

Literature Review

There has been very little research on the roles of personality traits on peer assessment in EFL settings. To my knowledge, AlFallay (2004) is the only researcher to carry out a study that incorporated personality factors to examine rater effects in peer assessments. AlFallay investigated the effects of psychological and personality traits (i.e., self-esteem, anxiety, and motivation) on the accuracy of peer- and self-assessments in EFL oral presentations in Saudi Arabia. The results of correlational analysis showed that peer assessments were more highly associated with teacher-assessment when students had high anxiety, high integrative orientation, and low motivational intensity compared to students with low anxiety, high instrumental orientation, and high motivational intensity. Although the study did not address the issue of rater severity, it clearly demonstrated that individual difference variables, including personality traits, were associated with rating behaviors in peer assessment.

Currently, the Big Five model is the dominant model for investigating personality (Dörnyei & Ryan, 2015). The present study, however, employs variables for self-construal, or "how individuals see the self in relation to others" (Cross, Hardin, & Gercek-Swing, 2011, p. 143), to measure personality traits. I adapted Takata's (2000) questionnaire instrument to measure self-construal (see Appendix). The questionnaire consisted of 20 items used to measure four variables: dogmatism, individuality, dependency on others, and evaluation apprehension. Dogmatism represents assertive attitudes and behaviors people display based on their own beliefs. Those with higher dogmatism express their opinions assertively and clearly. Individuality refers to a type of personality that values its own beliefs and decisions. Those with higher individuality do not care even when their opinions and behaviors are different from others and they think that their own decision is the best decision. Dependency on others revolves around relatedness and harmony with others. Those with higher dependency on others think that maintaining harmony with others is important and tend to give others' opinions more weight than their own opinions when opinions conflict. Evaluation apprehension refers to a type of personality that cares about being evaluated by others. Those with higher evaluation apprehension care about what others think of them.

When students enjoy rapport with their classmates, those with higher dependency on others might give more lenient ratings to their peers' performances due to the value they place on relatedness with their peers. On the other hand, even when students build a strong bond with their peers, those

with higher dogmatism and individuality might maintain their severity level, as their decisions are usually not affected by their relationships with their peers. As no research has been conducted to investigate the impact of these personality traits on rater severity in peer assessment, the present study examined a potential rater bias derived from the personality traits in peer assessment in an EFL oral presentation classroom. The following research question was posited in this study:

RQ To what extent do personality traits influence rater severity as student raters become familiar with their classmates and with peer assessment?

Method

Participants

The participants were Japanese university students majoring in sports and health science at a private university in Japan. They were all members of the author's class. The students in this department take four oral presentation courses that are conducted once a week over 2 years (one course extending over four semesters) as a requirement. The present study focused on 1st-year students in one class during the 2014 academic year. The students were placed in the class in association with an introductory academic seminar course regardless of their English proficiency levels. The students were engaged in many academic and social activities in the main academic seminar class. The author observed that through these activities they had built good rapport with their classmates by the second semester. Although the class comprised 27 students, the data for only 21 student raters were used for the main analysis as data on personality traits, peer assessment, or both were missing for the remaining students.

Oral Presentations

Each student made two presentations (mid-term and final presentations) in each semester. This study focuses on the mid-term presentations they made in the first semester (Weeks 8 and 9; hereinafter, Time 1) and the second semester (Weeks 21 and 22; hereinafter, Time 2). The duration of the presentations was 3 minutes for Time 1 and 4 minutes for Time 2. Students made presentations on topics of their own choice both times. At Time 1, they made a presentation based on information from books and articles. Example presentation topics were *How to get better sleep* and *The effects of music*.

At Time 2, they conducted a survey and made a presentation based on the results. Example presentation topics were *Experiences of flow in sports* and *Burnout syndromes*.

Peer Assessment

Each student rater evaluated his or her classmates' presentations both times with a peer assessment form used in the English program of the department. The assessment form contained four categories (English language use, content and organization, preparation and nonverbal delivery, and question and answer session) to rate each presenter using a 5-point Likert scale (from 1 = *very poor* to 5 = *very good*) and space to write a short comment on each presentation. The present study focused on the first three categories.

The student raters were informed of the three criteria through the teacher's explanations in advance. As peer assessment was part of the course assignments for which their final course grade was calculated, students were generally seriously engaged in peer assessment and wrote a comment for each presentation (see the section on rater severity for more detailed discussion). The peer assessment was not disclosed to the presenters. No feedback was given for the peer assessments at either Time 1 or Time 2.

Personality Traits

Takata's (2000) questionnaire on self-construal was administered around Time 1 to measure the students' personality traits (see Appendix for the English translation of the questionnaire items). As illustrated in the literature review, the questionnaire contained items to measure four variables: dogmatism (four items), individuality (six items), dependency on others (six items), and evaluation apprehension (four items). The questionnaire was answered on a 6-point Likert scale (1 = *strongly disagree* to 6 = *strongly agree*) and was administered to 219 students, including the participants of this study (n = 21). The reliability analysis was conducted based on the responses from the 219 students using Winsteps 3.80.0 (Linacre, 2013b) and SPAA 24.0. Table 1 shows the summary of the reliabilities and unidimensionality of the four questionnaire constructs. Each construct is acceptably unidimensional as the Rasch model accounted for more than or approximately half of the total variance and the eigenvalue of the first residuals was less than 2.0, which is the variance of two items and the minimum value for construing a secondary dimension (Linacre, 2012). Concerning construct reliability, whereas the three constructs besides dependency on others dis-

played acceptable Cronbach's alpha coefficients (min. = .71) and Rasch person reliabilities (min. = .68), dependency on others showed a low reliability estimate (Cronbach's α = .57, Rasch person reliability = .53). Despite its low reliability, dependency on others was retained for further analysis due to its importance in the present study. Thus, the results must be interpreted with caution, especially as the sample is stratified into only one or two levels with a person reliability estimate of .50 (Linacre, 2012), which may suppress the effect of dependency on others in the main analysis.

Table 1. The Summary of the Reliability Analysis for the Questionnaire Constructs (*N* = 219)

	DOG	IND	DEP	EVA
Variance explained by measures	47.30	56.70	48.60	63.20
The first residuals	14.10	18.60	12.90	17.90
(eigenvalue)	1.60	1.70	1.50	1.90
Item separation	5.81	6.98	10.18	8.72
Item reliability	.97	.98	.99	.99
Person separation	1.68	1.47	1.05	1.78
Person reliability	.74	.68	.53	.76
Cronbach's α	.74	.71	.57	.76

Note. DOG = dogmatism (4 items); IND = individualism (6 items); DEP = dependency on others (6 items); EVA = evaluation apprehension (4 items).

Results and Discussion

Descriptive Statistics

Rater Severity

Rater severity of each student rater was estimated for both Times 1 and 2 using the many-facet Rasch measurement model with Facets 3.71.2 (Linacre, 2013a). Although the class comprised 27 students, data on 26 presenters and 25 raters were submitted to the Rasch analysis as the remaining data were unavailable. The data were specified to have four facets: the ability of student presenters, the severity of student raters, the difficulty of two sessions (Times 1 and 2), and the difficulty of three assessment categories (English language use, content and organization, and preparation and non-

verbal delivery). Figure 1 presents the Wright map plotting measures for these four facets with the logit scale in Column 1 on the left and the scale used in the assessment in the last column.

Column 2 shows presenter abilities. Higher ability presenters were mapped at the top of the vertical ruler and lower ability presenters at the bottom. The presenters are largely spread out along this measure, revealing a large variance in the presentation abilities of the participants of this study as perceived by their peers.

Column 3 shows rater severity. More severe raters are located at the top and more lenient raters at the bottom. As only 10 out of 25 student raters were located below 0.00 logits, the majority of the student raters scored their peers' presentations critically. The data from the calibration report for the student raters revealed that rater severity varied considerably, ranging between -1.82 and 1.16 logits (M = 0.50, SD = 0.10), with a rater separation reliability (rater separation index) of .97 (5.28). The significant fixed (all-same) chi-square, $\chi^2(24) = 620.7$, $p < .001$, also confirmed the significant variations in the level of severity among the student raters.

Column 4 shows the session difficulty for Times 1 and 2. Although the difficulty span between the two sessions was small (0.28 logits), the presentations at Time 2 (M = 0.14) were more severely scored than at Time 1 (M = -0.14). The separation reliability (separation index) of .96 (5.10) and the significant chi-square, $\chi^2(1) = 27.0$, $p < .001$, also confirmed the significant difference between the two sessions.

Column 5 shows the category difficulty. Although all three categories were clustered around the center, preparation and nonverbal delivery was scored the most severely, followed by English language use and content and organization, respectively.

Concerning consistency of the student raters' ratings, two of the 25 student raters (Raters A and B) were identified as misfitting based on the criteria of the infit and outfit mean square (MNSQ) statistics between 0.50 and 1.50 (Linacre, 2013c). Rater A (rater severity = 0.34 logits, infit MNSQ statistics = 1.76, outfit MNSQ statistics = 1.77) and Rater B (rater severity = 0.86 logits, infit MNSQ statistics = 1.91, outfit MNSQ statistics = 1.89) underfit the model. Although use of fit MNSQ statistics above 2.0 "distorts or degrades the measurement system," MNSQ statistics between 1.5 and 2.0 are indicated as "unproductive for construction of measurement, but not degrading" (Linacre, 2013c, p. 266). Accordingly, the two misfitting raters with fit MNSQ statistics below 2.0 were retained for the main analysis. The fit MNSQ statistics of 25 student raters ranged between 0.67 and 1.91 (M =

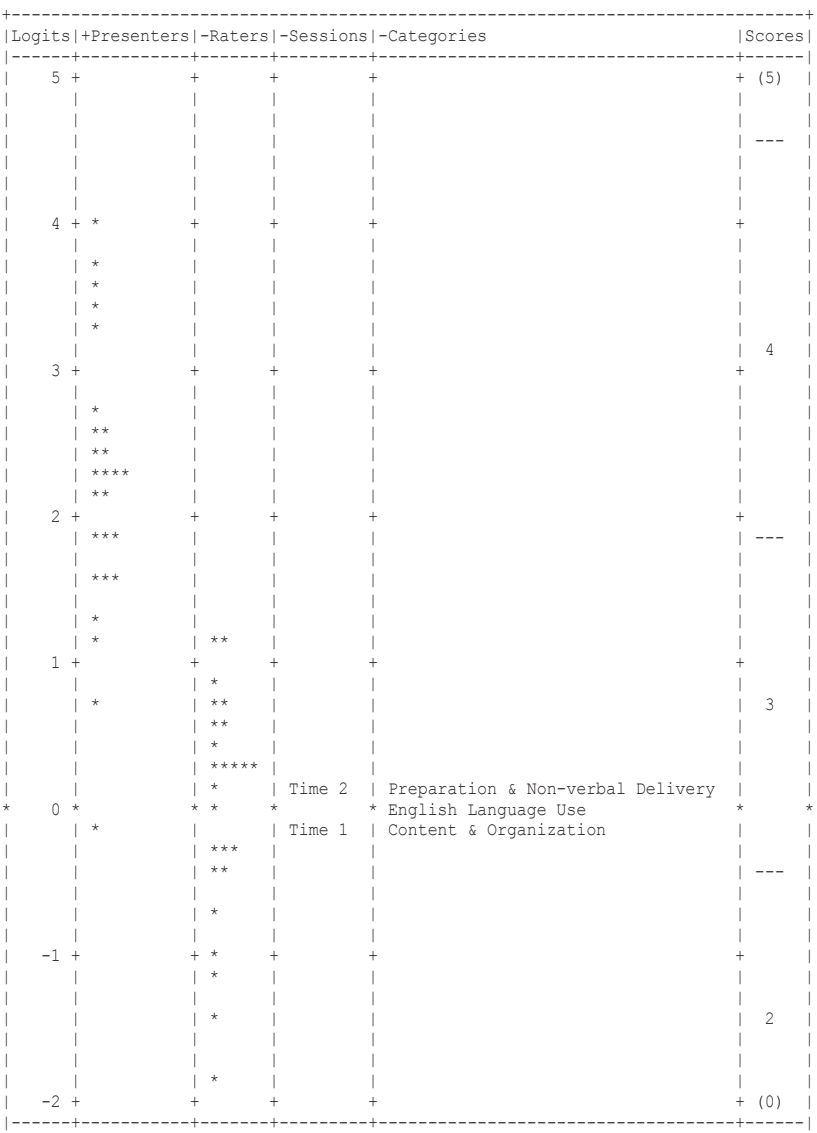

Figure 1. The FACETS Wright map for the presenter ability, rater severity, session difficulty, and category difficulty. Each asterisk (*) indicates one student.

1.00, *SD* = 0.30) and between 0.67 and 1.89 (*M* = 1.01, *SD* = 0.30) for infit and outfit values, respectively. Taken together, most students were consistent in scoring their peer presentations. The mean of the peer assessment also correlated highly with the teacher assessment based on the raw scores at Time 1 ($r = .82$, $p < .001$).

Personality Traits of Student Raters

Table 2 shows the descriptive statistics for the four personality variables in logits. The participants at the group level generally asserted their opinions (relatively high dogmatism; *M* = 0.69) but tended not to stick to their beliefs when people around them had different ideas (low individuality; *M* = -0.94). They had a tendency to value relatedness and harmony with others (high dependency on others; *M* = 1.13), and cared about being evaluated by others (high evaluation apprehension; *M* = 0.90).

Table 2. Descriptive Statistics for Personality Traits (*N* = 21)

			95% CI		
	M	*SE*	LL	UL	*SD*
Dogmatism	0.69	0.30	0.06	1.32	1.39
Individuality	-0.94	.30	-1.57	-0.31	1.38
Dependency on others	1.13	.22	0.68	1.59	1.00
Evaluation apprehension	0.90	.43	-0.01	1.80	1.99

Note. All the estimates are based on Rasch logits. Cl = confidence interval; LL = lower limit, UL = upper limit.

The Effect of Personality Traits on Rater Severity

The research question concerned to what extent personality traits influence rater severity as student raters become familiar with their classmates and with peer assessment. In order to examine the effects when students are less familiar with their classmates and the assessment, a multiple stepwise regression analysis was performed with rater severity at Time 1 as a dependent variable. The results showed that none of the four personality factors significantly predicted rater severity at Time 1 (Table 3). When student raters were relatively new to their classmates and to peer assessment, personalities were not associated with the rater severity of peer assessment.

Table 3. The Regression Analysis of Personalities Predicting Rater Severity at Time 1 (N = 21)

Predictors	F	R^2	B	$SE\ B$	β
Step 1	.75	.16			
Individuality			0.16	0.22	.31
Evaluation apprehension			0.06	0.13	.18
Dogmatism			0.08	0.20	.15
Dependency on others			−0.09	0.25	−.13
Step 2	1.02	.15			
Dogmatism			0.12	0.16	.23
Individuality			0.11	0.17	.21
Evaluation apprehension			0.03	0.09	.08
Step 3	1.54	.15			
Dogmatism			0.13	0.15	.25
Individuality			0.09	0.15	.17
Step 4	2.87	.13			
Dogmatism			0.19	0.11	.36

Note. All variables were nonsignificant. B = unstandardized regression coefficient; β = standardized regression coefficient.

In order to examine the effects when student raters are more familiar with their classmates and peer assessment, another multiple stepwise regression analysis was conducted with rater severity at Time 2 as a dependent variable. The results showed that two of the four predictors (i.e., dependency on others and evaluation apprehension) were significant predictors of rater severity at Time 2 (Table 4). In line with the initial hypotheses, whereas student raters who valued relatedness and harmony with others were more lenient in peer assessment, the personality traits of being independent and assertive did not influence rater severity. Furthermore, students who cared about being evaluated by others were more severe in peer assessment. Taken together, although some personality traits (i.e., dogmatism and individuality) do not have a systematic impact on the rater severity, it appears that certain personality traits (i.e., dependency on others and evaluation apprehension) influenced rater severity when students were more familiar

with their classmates and peer evaluation. However, further research is needed to verify these results, as the confidence intervals of the means of the four independent variables were wide as shown in Table 2.

Table 4. The Regression Analysis of Personalities Predicting Rater Severity at Time 2 ($N = 21$)

Predictors	F	R^2	B	$SE\ B$	β
Step 1	2.57	.39			
Evaluation apprehension			0.25	0.14	.58
Dependency on others			-0.49	0.26	-.56
Dogmatism			0.17	0.20	.28
Individuality			0.05	0.23	.07
Step 2	3.61*	.39			
Evaluation apprehension			0.23	0.10	.53*
Dependency on others			-0.45	0.20	-.52*
Dogmatism			0.20	0.12	.33
Step 3	3.63*	.29			
Dependency on others			-0.52	0.21	-.59*
Evaluation apprehension			0.23	0.10	.53*

Note. B = unstandardized regression coefficient; β = standardized regression coefficient.

*$p < .05$

Conclusion

To the best of my knowledge, this is the first study to investigate the effect of personality traits on rater severity in peer assessment of EFL oral presentations. The present study found that rater personalities tended to cause rater bias in peer assessment under certain circumstances and may jeopardize the precision of peer assessment. However, this study was only a preliminary study conducted with a very small sample size (N = 21). It should be replicated with a larger sample to generalize the findings. As there is a dearth of research investigating rater bias in peer assessment of EFL oral presentations, more research on this issue is also needed.

Acknowledgment

I would like to thank the anonymous reviewers for their detailed and constructive comments and suggestions.

Mitsuko Tanaka is a lecturer at Ritsumeikan University. She holds a PhD in education from Temple University. Her current research interests include individual differences in SLA (e.g., motivation and self-construal) and language assessment.

References

AlFallay, I. (2004). The role of some selected psychological and personality traits of the rater in the accuracy of self- and peer assessment. *System, 32,* 407-425. https://doi.org/10.1016/j.system.2004.04.006

Azarnoosh, M. (2013). Peer assessment in an EFL context: Attitudes and friendship bias. *Language Testing in Asia, 3*(11). https://doi.org/10.1186/2229-0443-3-11

Cheng, W., & Warren, M. (2005). Peer assessment of language proficiency. *Language Testing, 22,* 93-121. https://doi.org/10.1191/0265532205lt298oa

Cross, S. E., Hardin, E. E., & Gercek-Swing, B. (2011). The what, how, why, and where of self-construal. *Personality and Social Psychology Review, 15,* 142-179. Retrieved from http://journals.sagepub.com/

Dörnyei, Z., & Ryan, S. (2015). *The psychology of the language learner revisited*. New York, NY: Routledge.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2,* 197-221. https://doi.org/10.1207/s15434311laq0203_2

Eckes, T. (2015). *Introduction to many-facet Rasch measurement* (2nd ed.). Frankfurt am Main, Germany: Peter Lang.

Linacre, J. M. (2012). *A user's guide to WINSTEPS: Rasch-model computer program.* Beaverton, OR: Winsteps.com.

Linacre, J. M. (2013a). FACETS (Version 3.71.2) [Computer software]. Beaverton, OR: Winsteps.com.

Linacre J. M. (2013b). Winsteps (Version 3.80.0) [Computer software]. Beaverton, OR: Winsteps.com.

Linacre, J. M. (2013c). *A user's guide to FACETS: Rasch-model computer program.* Beaverton, OR: Winsteps.com.

Otoshi, J., & Heffernan, N. (2007). An analysis of peer assessment in EFL college oral presentation classrooms. *The Language Teacher, 31*(11), 3-8. Retrieved from http://jalt-publications.org/tlt/archive

Takata, T. (2000). On the scale for measuring independent and interdependent view of self. *Bulletin of Research Institute of Nara University, 8,* 145-163.

Topping, K. J. (2009). Peer assessment. *Theory Into Practice, 48,* 20-27. https://doi.org/10.1080/00405840802577569

Appendix

English Translation of the Questionnaire Items for Takata's (2000) Self-Construal

To what extent do you agree with each of the following statements?

Factor 1: Dogmatism (DOG)	
DOG1	I always try to have opinions of my own.
DOG2	I always know what I want to do.
DOG3	I always express my opinions clearly.
DOG4	I always speak and act with confidence.
Factor 2: Individuality (IND)	
IND1	The best decisions are the ones I make by myself.
IND2	When I believe in an idea, I do not care what others think of it.
IND3	Even if people around me have different ideas, I stick to my beliefs.
IND4	In general, I make my own decisions.
IND5	Whether something is good or bad depends on how I think about it.
IND6	I do not care when my opinions and behaviors are different from others.

Factor 3: Dependency on Others (DEP)	
DEP1	It is important to maintain harmony with others.
DEP2	It is important for me to be liked by others.
DEP3	How I feel depends on who I am with and what circumstances I am in.
DEP4	I avoid having conflicts with my group's members.
DEP5	When I differ in opinions from others, I often accept their opinions.
DEP6	I sometimes change my attitudes and behaviors depending on who I am with and what circumstances I am in.
Factor 4: Evaluation Apprehension (EVA)	
EVA1	I care about what others think of me.
EVA2	Sometimes I am worried about how things will turn out and have difficulty in getting started.
EVA3	I care about how others evaluate me.
EVA4	When interacting with others, I care about my relationships with them and their social status.

Note. All the questionnaire items are randomly ordered 6-point Likert-scale items. 1 = strongly disagree, 6 = strongly agree.