The Rater Performance Categorization System (RPCS) for Intensive English Programs

OPEN ACCESS

Manuscript ID:
EDU-2021-09033986

Volume: 9

Issue: 3

Month: June

Year: 2021

P-ISSN: 2320-2653

E-ISSN: 2582-1334

Received: 06.04.2021

Accepted: 15.05.2021

Published: 01.06.2021

Citation:
Şahin, Alper. "The Rater Performance Categorization System (RPCS) for Intensive English Programs." *Shanlax International Journal of Education*, vol. 9, no. 3, 2021, pp. 225-241.

DOI:
https://doi.org/10.34293/education.v9i3.3986

Alper Şahin
Atılım University, Ankara, Turkey
https://orcid.org/0000-0001-7750-4408

Abstract
There are several student performances assessed in Intensive English Programs (IEPs) worldwide in each academic year. These student performances are mostly graded by human raters with a certain degree of error. However, the accuracy of these performance assessments is of utmost importance because they feed data into some high stakes decisions about the students and such performance assessments constitute a large number of students' scores. Therefore, the accuracy of these performance assessments should be given priority by the IEPs. However, the current rater performance monitoring systems which can help the administrators of IEPs to monitor rater performance in performance assessment are away from practicality because they require the use of complex mathematical models and specialized software. A practical and easy to maintain rater performance categorization system is proposed in this paper and it was accompanied by a sample study. Its benefits to the administrators of IEPs and their raters are also discussed besides its practical considerations.

Keywords: Rater performance categorization system, Performance assessment, Assessing writing, Assessing speaking, Language testing

In English language teaching, constructed-response items are commonly used and these constructed response items are mostly graded by human raters. Similarly, the students' performance in two productive skills: speaking and writing, is widely assessed. These student performances are also graded by human raters and this, as expected, constitutes a degree of error because error-free grading is not possible; however, most of these assessments are done by different raters for different students and the scores devised by those raters are used to make some high-stakes decisions. Therefore, the high grading quality of these assessments should be maintained. To improve the rating quality in performance assessment, raters' rating quality should first be identified and compared. Rater bias may always be there as the raters vary in their background information (McNamara, 1996). Such a problem can only be revealed after the performance evaluation of the raters, with the help of which the rating quality can be monitored. Therefore, it should not be wrong to state that the raters who rate in a standardized manner bring the rating quality to their program.

Performance Assessment in Intensive English Programs (IEPs)

Intensive English Programs (IEPs) are 6 to 24-month pre-faculty programs which aim to improve students' competence and performance in the English language before they start their actual programs at the university where all departmental courses are held in the English language. Such IEPs are also called English Preparatory Schools.

Due to the intensity of these programs, students are subjected to writing and speaking exams more frequently than an ordinary person within short periods because such programs generally assess student performance at least 4-8 times a month. This may mean that instructors rate student performance (writing and speaking) at an IEP as frequently as at least once or twice a week and a rater rates student performance around 48-60 times during an academic year. This number of highly frequent subjective performance assessment are prune to some rater bias and consistency problems if not handled properly due to the excessive number of raters rating student performance. Because speaking and writing assessments are prune to multiple sources causing errors (Sebok & Syer, 2015). As this problematic issue was combined with the anxiety and the feeling of being trapped during the covid-19 pandemic decreasing the motivation and probably the accuracy of the raters was one of the key concerns of the EIP administrators. Therefore, the accuracy in performance assessment and rating quality should be the priority in these programs much more than ever. Because these scores assigned to student performance are used as a part of the high-stakes decisions (pass-fail).

Rating Quality Indicators

When a rater assigns a score to student performance, this score also includes the rater effects. Rater effects constitute the errors in rater judgments (Myford & Wolfe, 2009). Leniency / Severity, Centrality / Extremity, and Accuracy / Inaccuracy are the three common types of rater effects (Wolfe & McVay, 2012). According to Wolfe & McVay (2012), leniency refers to a rater's overrating the student performance systematically. On the contrary, severity refers to a rater's underrating the student performance systematically over time, that is, being stringent while scoring (Wolfe & McVay, 2012). Centrality and extremity are also two rater effects with opposite meanings. At the same time, the former refers to a rater's continuously using central score categories and refraining from using the extreme score categories in the rating scale, the latter refers to just the opposite (Wolfe & McVay, 2012). It is a rater's using only the extreme score categories (lowest or the highest) on the rating scale. Finally, inaccuracy refers to a rater's deviating from the true score randomly and accuracy refers to a rater's having high correlations with true scores (Wang, et al., 2017). As can be seen, rater effects can manifest themselves in a variety of ways. Thus, to identify these various rater effects on student performance scores and monitor rater performance over time, some rater performance monitoring systems and models were developed.

The Current Rater Performance Monitoring Systems and Models

There are a couple of rater monitoring systems and models that have been developed and proposed in varying complexity levels by different scholars in the field. A summary of the most well-known systems and models that are currently available will be briefly presented here.

Myford & Wolfe (2009) proposed a framework for monitoring rater performance over some time. They studied some statistical indices to monitor the change in the rater performance over time. For this purpose, they selected 51,233 student essays from Advanced Placement English Literature and Composition (AP ELC) examination and asked 101 raters to rate these papers. They created a panel of scoring leaders to grade the 28 validity essays through consensus rating. As a result of their study, they found out that the rater performance drift over time existed and they could impact the rating quality in nontrivial ways. Moreover, they also found out that the raters exhibited changes in their rating accuracy and their scale category use over time.

In their study, Cao, et al., (2010) offered a ranking order of the raters' performance through the use of the Bayesian Approach and considering the data as ordinal data. In their model, they were able to identify the raters' bias, measurement error, and discrimination power as in the three-parameter model of the Item Response Theory (IRT). They used the data of 10 raters for 39 proposals and they could identify the raters' measurement error, discrimination ability,and bias successfully. They also presented the correlation of each rater's scores with other scores. They concluded that small bias and measurement error accompanied by large discrimination yields a qualified rater.

DeCarlo, et al., (2011) proposed a hierarchical rater model (HRM) based on the signal detection rater model for rater performance monitoring. They claimed that the true level of an examinee's performance was not observed. Rather, it is a latent category. Moreover, a score assigned to student performance by a rater does not directly indicate the examinee's ability. In addition to this, they thought that the rating task did not have one level. They thought a rater's putting student performance into performance categories constitutes just the first layer of their HRM. The second layer constitutes the ordinal indicators of the examinee's ability determined by an IRT model. Only after determining these two layers, the examinees' true performance (θ) could be reached. Therefore, they stated that examinee data in constructed response items should be handled accordingly. They used data of 2,350 examinees answering two essay items. They made each essay graded by 2 raters. 34 raters graded the first item and 20 raters graded the second item among the 54 raters available. 13 raters scored both items. With the help of the model they used, they could identify the lenient and severe raters and the raters following centrality by not using extreme values in the rating scale. However, they reported that the system they developed tagged some raters as lenient, although they were not when all response categories were not used, or they were used differently in the rating scale. As a result, they could separate the item characteristics from the rater characteristics with the help of the model they employed which is a hierarchical model based on signal detection theory.

Wang, et al., (2017) studied the efficiency of two essay selection methods for an adaptive rater monitoring system they developed. They claimed that the validity essays used as a base to monitor examinee performance might not be giving valid information regarding the raters. For example, a lenient rater consistently assigning high scores to examinees may be categorized as an accurate rater when a student performance with the already high score assigned was used as validity paper. To get over this problem, they proposed adapting the validity papers to the rater effect and detecting rater severity/leniency and centrality/extremity based on these adaptive essays selected according to the rater's rating behavior. For this purpose, they used conventions of computerized adaptive testing and proposed two validity essay selection methods: the single fisher information method and the D-optimal method. To test the accuracy of these essay selection methods, they used two simulation studies, one with data generated by the simulation software from a normal distribution N(0,4) with a rater sample size of 1,000 and essay bank size of 600 and one with real data of 400 essays graded by 131 professional raters. Their results suggested that both essay selection methods worked around similar accuracy and rater parameters could be yielded with fewer essays when adaptive rater monitoring methods were used.

Shin, et al., (2019) targeted the difficulty and cost of finding appropriate validity essays/papers in appropriate numbers for standardization training and finding experts to grade them for standardization and rater monitoring purposes in their study. Therefore, they proposed a rater monitoring system that would use automated scoring engines to grade the validity papers leaving consensus scoring by the expert raters obsolete. They used 131 human raters to rate 189 essays by middle school students. Each of these essays was assigned a score by an expert panel of human raters. The essays were also scored by an automated scoring engine. They compared the feasibility of using scores assigned by an automated scoring engine and an expert panel of human raters when monitoring rater performance. They reached 100% identical decisions on leniency / severity, 66.4% on accuracy / inaccuracy, and 93.1% on centrality / extremity of the raters. They concluded that Automated scoring engines could only be used for leniency/severity decisions. They also suggested using larger data in similar studies as they thought this outcome could be due to the small sample data they utilized.

Wang, et al., (2020) examined the raters' performance in the Canadian English Language Benchmark Assessment for Nurses (CELBAN) exam speaking component in terms of raters consistency and severity, use of rating scales, and rating bias using Many-facets Rasch Measurement (Linacre & Wright, 1989). They used 115 raters and 2,698 examinations in four parallel forms. They included five facets to their study; examinee measurement report (Facet 1), rater measurement report (Facet 2), test site (Facet 3), test version (Facet 4), and criterion measurement report (facet 5). They concluded that the vocabulary is the easiest, grammar is the most difficult criterion for the examinees to get a high score from. With the help of the system they developed, they were able to report the rater bias on a logit scale individually. They could reach detailed information about the rater performance under each category of the rubric with the expected score, observed scores, standard error, t and p values.

As can be seen there are many ways of identifying the rater effects on student performance assessment. Each program, whether it is an IEP or not should pay attention to the rater training and should set up a system to monitor their raters' performance.

Problem

It is not a common practice to evaluate and monitor the teacher's or rater's performance in IEPs (Huang, et al., 2018). It is thought by the author of this paper that there are some reasons behind it. To elaborate, as mentioned earlier most popular rater monitoring systems currently available are based on Many Facets Rasch Measurement (MFRM; Wang et al., 2020; Myford & Wolfe, 2009; Wigglesworth, 1993; Davis, 2016), Bayesian approach (Cao, et al., 2010), Rasch Partial Credit model (Wang, et al., 2017), Hierarchical rater model (DeCarlo, et al., 2011), and automated scoring engines (Shin, et al., 2019). These monitoring systems provide the administrators with highly detailed and robust systems which can be attained by using complex mathematical models, and methods like the Bayesian method (Cao, et al., 2010), Maximum likelihood estimation (Shin, et al., 2019), log-ratio test (Wang, et al., 2017), time facet model (Myford & Wolfe, 2009), Signal detection rater model (DeCarlo, et al., 2011), generalized partial credit model (DeCarlo, et al., 2011), mixed-effects ordinal probit model (Shin, et al., 2019) and some specialized software like Facets (Linacre, 2014), Winsteps (Linacre, 2018), Stata (Stata Corp., 2013), latent gold (Vermunt & Magidson, 2005), glam (Rabe-Hesketh, et al., 2004), and WinBUGS (Lunn, et al., 2009). The use of these mathematical models and methods which can only be implemented by using some specialized software mentioned above causes extra cost, energy, time, training, and expertise to organizations. More importantly, managers and instructors of the programs like IEP are social sciences (mostly English Language Teaching) graduates. It is not expected for them to be good at math to comprehend these models and methods. Therefore, IEP managers and raters would just be confused about mathematical models and methods. However, they are the ones who needed such rater performance monitoring systems much more than most other programs due to the frequency of the performance assessment they do each academic year. Therefore, a rater categorization system that won't need such complex mathematical calculations, models or specialized software; that could be easily understood and interpreted by IEP raters and administrators and that would fit the nature of the work done in IEPs was needed.

Significance of the Study

The problem stated above leaves the IEP raters and administrators with no information regarding the rater's performance. If IEP raters knew how well they did on a performance assessment task, they would try to develop their performance in that task, and it would provide the raters with a goal in their professional development. In addition to this, if the IEP administrators knew how well their raters performed, they would take individual or to-the-point precautions like doing some training targeting individuals or a specific group of raters. If such a practical and easy to maintain performance categorization system existed, it would also decrease the face-to-face training needs and expenses of the large programs like IEPs dramatically. They would choose to call only the raters below a certain performance level to their onsite or offsite rater training activities. Therefore, the Rater Performance Categorization System (RPCS) was developed to close this gap in the literature and provide the IEP raters and administrators with a practical and easy to establish system to monitor rater performance. A research study on the RPCS was developed to use RPCS in a real-life setting and to answer the following questions:

- What are the benefits of the RPCS to the administrators in real-life settings?
- What are the benefits of the RPCS to the raters in real-life settings?
- Are there some considerations of the use of the RPCS in real-life settings?

Method

In this part of the paper, the participants of the study, the RPCS, data collection procedure and data analysis methods will be presented.

Participants of the Study

The participants of the study consisted of 87 raters(among a total of 101 raters available) who responded to the questionnaire regarding the RPCS and 86 raters who participated in all stages of the rating tasks at EIP of an English medium university. The information obtained from the background information survey responded by the 87 instructors can be found below in Table 1.

As shown in Table 1, most of the participants (82.7%) had over 5 years of English language teaching experience. Moreover, the participants had mostly bachelor's degrees in English Language Teaching (41.4%) and English / American Literature (39.1%). Apart from this, 16.1% of the participants had B.A. degrees in linguistics and 3.4% had B.A. degrees in Translation and interpretation. Last but not least, 41.4% of the participants were working for the institution for around 3 to 5 years, 14.9% were working for 0 to 2 years, 19.5% were working for the institution for 6 to 10 years and 24.1% were working for the institution for 11+ years. This shows that most of the participants (85.1%) worked for the institution for more than 3 years.

Table 1: Descriptive Background Information about the Participants

Year of Experience	f	%	BA Degree	f	%	Institutional Experience	f	%
0-2 years	5	5.7	English Language Teaching	36	41.4	0-2 years	13	14.9
3-5 years	10	11.5	English / American Literature	34	39.1	3-5 years	36	41.4
6-10 years	22	25.3	Translation and interpretation	3	3.4	6-10 years	17	19.5
11-15 years	25	28.7	Linguistics	14	16.1	11+ years	21	24.1
16+ years	25	28.7						
Total	87	100.0	Total	87	100.0	Total	87	100.0

The Rater Performance Categorization System (RPCS)

The RPCS is the umbrella term that covers the whole system which involves some sub-scores and categories with the help of which raters' performance scores are calculated and their corresponding categories are determined.

There are two components of the RPCS. They are the Rater Performance Score (RPS) and Rater Performance Category (RPC). RPS is simply the absolute distance of an individual raters' score from the Estimated True Score (TE) of the performance task calculated as shown in equation 1:

$$\mathrm{T_E} = \frac{(\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{t_j}) + \mathrm{t_k}}{2} \qquad (1)$$

where t_j is the score assigned to the performance by the rater j, and t_k is the score assigned to the performance by the testing office members (considered as senior raters and these scores are determined by their consensus as a group) and N is the total number of raters.

As can be seen, there is an assumption in estimating the true score taking the two common assumptions currently available into consideration: 1) majority opinion or general will (Cao, et al., 2010) 2) Expert opinion (Shin, et al., 2019). In the present study, the estimated true score has been reached by averaging these two scores. In IEPs expert raters are not always the expert raters in industrial standards and taking their scores as true scores might be problematic. More importantly, as the number of the raters rating a student's performance increases, the accuracy of the rating increases (Mariano, 2002). Therefore,taking an average of the general will and expert scores was thought to increase the validity of the estimated true scores used to monitor rater performance.

After calculating the TE, the RPS is calculated for each performance task (paragraph, essay or speaking) for rater j as in equation 2:

$$\mathrm{RPS}_{j(a)} = \mathrm{T}_{\mathrm{E}} - \mathrm{t}_{j} \qquad (2)$$

where t_j denotes the score assigned to the performance task "a" by the rater j.

After the RPS is calculated for each rater for each performance task, the RPC can be determined. The RPC consists of summative performance categories where the rater performance is labeled as A+, A, B, C, D, E based on RPS intervals are listed in Table 2.

Table 2: RPC and RPS Correspondence in the RPCS

RPC	RPS Interval	Example (20 pts as maximum score)	Example (100 pts as maximum score)
A+	+/-0.00 – +/- 4.99%	+/-0.00 – +/-0.99 RPS	+/-0.00 – +/-4.99 RPS
A	+/-5.00% – +/-7.49%	+/-1.00 – +/-1.49 RPS	+/-5.00 – +/-7.49 RPS
B	+/-7.50% – +/-9.99%	+/-1.50 – +/-1.99 RPS	+/-7.50 – +/-9.99 RPS
C	+/-10.00% – +/-12.49%	+/-2.00 – +/-2.49 RPS	+/-10.00 – +/-12.49 RPS
D	+/-12.50% – +/-14.99%	+/-2.50 – +/-2.99 RPS	+/-12.50 – +/-14.99 RPS
E	+/-15.00% and above	+/-3.00 and above RPS	+/-15.00 and above RPS

As shown in Table 2, raters are put into the RPCS according to the RPS intervals provided in the form of percentages calculated using the maximum score. For example, if the maximum score of a performance task is 20, RPS Interval for Category "A" raters is between +/-1.00 to +/-1.49 RPS, that is the RPS between +/-5% and +/-7.49% of the maximum score. Therefore, a rater with RPS of +1.4 receives "A" as RPC for that performance task.

While determining the RPC intervals, the double blind-review policies of some reputable universities and testing organizations were reviewed to find out the commonly accepted deviance between two scores assigned by two raters at these institutions. As a result of this review, it was found that acceptable differences between two scores were found to change between 5% to 10%. For example, while the University of Edinburgh (n.d.) allows for 5% difference, the University of Southampton (n.d.) allows for 6% difference,and the Cambridge assessment (Rodeiro, 2007) allows for 10% difference between the two blind ratings by two raters. Taking these figures as a base, the gold standard of 5% which was clearly accepted by all these institutions, was taken as the lowest acceptable distance from the TE and the raters within this distance were categorized as "A+" category raters. The highest acceptable level, that is 10%, was taken as the "B" level rater category and a mid-point between these two figures, that is 7.5%, was determined as "A" level RPC. Then, the following RPCS were determined in 2.5% intervals as in the first three categories up to 15% distance and above from TE.

An average RPS and a corresponding Average RPC based on this average RPS can be determined. However, to calculate average RPS, it is necessary that the absolute values of the individual RPS be used before they are averaged as in equation 3:

$$RPS_{j\,(average\,)} = \frac{(\sum_{\mathrm{i=1}}^{\mathrm{N}} \left|RPS_{j\,(a)}\right|)}{N} \qquad (3)$$

where $\mathrm{RPS}_{\mathrm{j(a)}}$ denotes the RPS for the rater j in performance task a and N is the total number of the RPS that will be averaged.

The absolute RPS values should be used while calculating $\mathrm{RPS}_{\mathrm{average}}$ and determining corresponding $\mathrm{RPC}_{\mathrm{average}}$ because the aim is to calculate the average distance from the TE. If the absolute values of RPS are not used, some RPS values may neutralize each other. To elaborate, let's suppose that there are three paragraphs rated, and the individual RPS for each of them are 1.4, -1.4 and 0.7 respectively. While calculating $\mathrm{RPS}_{\mathrm{average}}$, if absolute scores were to be taken, $\mathrm{RPS}_{\mathrm{average}}$ of 1.16 and corresponding $\mathrm{RPC}_{\mathrm{average}}$ of "A" would be obtained. However, if absolute RPS values were not used, $\mathrm{RPS}_{\mathrm{average}}$ of 0.23 and corresponding $\mathrm{RPC}_{\mathrm{average}}$ of "A+" would be reached and it would be totally misleading.

Similarly, RPS$_{average}$ scores can also be averaged and RPS$_{overall}$ for rater j can be obtained using the equation 4 :

$$RPS_{j\,(overall\;)} = \frac{(\sum_{i=1}^{N} RPS_{j\,(average\;\;)})}{N} \qquad (4)$$

where N is the total number of the RPS$_{average}$ for rater j. It should be noted that RPC$_{overall}$ can also be determined after RPS$_{overall}$ is calculated for each rater.

Data Collection Procedure & Analysis

To implement the RPCS in areal-life setting and to identify what benefits and considerations it provides the administrators and raters with, the data collection procedure and analysis of this study was planned to have three stages details of which are presented below.

Stage 1: First, nine sample student performances in three performance task types, that is 3 paragraphs, 3 essays, 3 speaking performance have been selected from a previously administered midterm exam. Paragraphs and essays were anonymized before being shared with the raters and the sample student speaking performances were shared as voice recordings. While selecting the sample student performance for each task type, the samples were chosen in three performance levels (1 low, 1 medium, 1 high) based on the scores assigned to them by the raters after the midterm exam they were taken from. Then, 101 raters were asked to grade these sample student performances without knowing their rater performance will be assessed using the rating scales that have been used for around two years in the department each of which had 20 points as the maximum point.

Stage 2: RPS and RPC for each performance task (e.g. Paragraph 1, Paragraph 2, Paragraph 3) were calculated using MS Excel 365 (Microsoft Corporation, 2011). In addition, RPS$_{average}$ was also calculated for each type of performance task (paragraph, essay, and speaking) and their corresponding RPC$_{average}$ was also determined. Moreover, RPS$_{overall}$ and RPC$_{overall}$ were also calculated and determined and shared with the raters participated the study in stage 1. The announcement of all these RPSs and RPCs was made with a detailed explanation of what those numbers/categories mean, and a presentation was made to the raters on how the RPSs were calculated and RPCs were determined. Following the presentation, a standardization training of up to two hours was conducted for each performance task type (paragraph, essay and speaking) using sample student performances graded. The RPSs and RPCs calculated at this stage were shared with the participants as "Pre-training RPSs / RPCs"

Stage 3: Another set of nine student performance task types (3 paragraphs, 3 essays, 3 speaking performance) from the same midterm exam with similar performance levels (1 low, 1 medium, 1 high) were selected, and raters were asked to grade those sample student performances as well. At this stage, the individual RPSs for each performance (e.g. essay 1, essay 2, essay 3), RPS average for each performance task (Paragraph, Essay, and speaking), were calculated and their the corresponding RPCs and RPC$_{average}$ were determined. Moreover, RPS$_{overall}$ and RPC$_{overall}$ were also calculated and determined. Then, they were shared with the participants as "Post-Training RPSs / RPCs". It was found out at this stage that a total of 86 out of 101 available raters participated in both pre and after training scoring activities.

Moreover, based on the RPC$_{average}$ for different types of tasks, the percentage of raters with better and worse RPC$_{average}$ before and after the trainings were identified, and the average change in their RPS$_{average}$ was calculated. For this purpose, Root mean squared difference (RMSD) for pre and post-training RPS$_{average}$ values for each task type was also calculated based on equation 5:

$$\mathrm{RMSD_a} = \sqrt{\frac{\sum_{i=1}^{N}(t_j - T_E)^2}{N}} \qquad (5)$$

where a denotes the task type, t$_j$ denotes the score assigned to the student performance by the rater j and T$_E$ is the Estimated True Score as mentioned earlier. Last but not the least, as a follow-up, participants were asked to respond to a 12-item survey about the RPCS. Their responses were subjected to a frequency analysis.

Results & Findings

Results of the study will be presented under this title in three parts: Pre-Post training RPCS, The change in RPS and RPC after the announcement of RPCS and the training, and the RPCS feedback questionnaire.

Pre and Post-Training RPCs

To see how the RPCS works and what conclusions can be drawn from it, the number of raters in each RPC$_{average}$ for different tasks were counted, a table showing pre and post-training RPC$_{average}$ of the raters and their percentages out of 86 raters was prepared and presented in Table 3.

As can be seen in Table 3, pre-training RPC$_{average}$ indicated that for paragraph tasks, the highest percentage of the raters were placed in the "A+" category (24.4%). For the essay task, the highest percentage of the raters were placed in "A+" and "A" categories with 27.9% of the raters in each. Moreover, for the speaking task, similarly, the RPC with the highest number of raters was the "A+" (27.9%) as well. However, the pre-training RPC$_{overall}$ indicates that the "A" category is the RPC with the most raters placed with 40.7%. It is important to note that 31.5% of the raters were placed into the "C", "D" and "E" categories combined according to the pre-training RPC$_{average}$ for paragraph task. This figure decreases to 21% in essay task. However, it increases to 36% for speaking task and 29.2% of the raters were placed in "C", "D" and "E" combined according to the pre-training RPC$_{overall}$. It is also interesting to note that 18.6% of the raters were placed in the "E" category for the speaking task.

Table 3: Pre and Post Training RPC$_{average}$ and RPC$_{overall}$

	Task	A+	%*	A	%	B	%	C	%	D	%	E	%	Total
Pre-Training	Paragraph	21	24.4	17	19.8	21	24.4	9	10.5	12	14.0	6	7.0	86
	Essay	24	27.9	24	27.9	20	23.3	10	11.6	4	4.7	4	4.7	
	Speaking	24	27.9	17	19.8	14	16.3	8	9.3	7	8.1	16	18.6	
	Overall	9	10.5	35	40.7	17	19.8	12	14.0	9	10.5	4	4.7	
Post-Training	Paragraph	37	43.0	11	12.8	13	15.1	7	8.1	9	10.5	9	10.5	86
	Essay	16	18.6	27	31.4	23	26.7	13	15.1	4	4.7	3	3.5	
	Speaking	31	36.0	22	25.6	9	10.5	11	12.8	10	11.6	3	3.5	
	Overall	20	23.3	24	27.9	21	24.4	13	15.1	7	8.1	1	1.2	
Difference	Paragraph	16	18.6	-6	-7.0	-8	-9.3	-2	-2.3	-3	-3.5	3	3.5	86
	Essay	-8	-9.3	3	3.5	3	3.5	3	3.5	0	0.0	-1	-1.2	
	Speaking	7	8.1	5	5.8	-5	-5.8	3	3.5	3	3.5	-13	-15.1	
	Overall	11	12.8	-11	-12.8	4	4.7	1	1.2	-2	-2.3	-3	-3.5	

*Percentages may not total to 100 as only one decimal is used due to space constraints in the table

The post-training RPC$_{average}$ listed in Table 3 indicated that a large percentage of the raters were placed in the "A+" category (43%) for the paragraph task. Similarly, the largest percentage of the raters (36.0%) were placed in the "A+" category for the speaking task. The largest percentage (31.4%) of raters were placed into the "A" category for essay task. Similarly,the "A" category has the largest percentage (27.9%) of the raters for the RPC$_{overall}$. The percentages of the raters in "C", "D" and "E" categories combined for paragraph, essay, speaking tasks and RPC$_{overall}$ are 29.1%, 23.3%, 27.9%, and 24.4% respectively.

When the difference in the RPC$_{average}$ for the paragraph task is considered, a shift towards the higher categories can be observed. It seems like raters changed their categories towards a higher category after the training making the highest change in the percentage of raters in the "A+" category with 18.6% increase. A similar pattern is observed for the speaking task in which it can be said that the raters in the "E" category decreased 15.1% and these raters moved to the higher categories. It can also be seen from the difference part of Table 3 that there was also a shift towards the medium categories ("A", "B" and "C") from the lower categories ("D", and "E") for essay task. However, the shift towards a higher category trend seems to be broken for the essay task after the training as 9.3% of the rater lost their "A+" level after the training. It seems like the RPC change for the post-training essay task took place towards the medium categories from the boths ends of the RPCS contrary to the obvious shift from lower to higher categories in other tasks. Last but not least, although it seems like 11 raters moved from the "A" category to the "A+" category based on their $\text{RPC}_{\text{overall}}$, it should be considered that it may be just a coincidence as many raters changed their categories and raters can change their categories more than one category above if they correct their RPSS dramatically. It can be still said that there is also a shift towards the higher categories observed when the change in the number of raters in $\text{RPC}_{\text{overall}}$ is analysed.

The Change in $\text{RPS}_{\text{average}}$ and $\text{RPC}_{\text{average}}$ after the Announcement of RPCS

To get more detailed insights into the magnitude of the change between the raters' performance after the RPCS was announced and the training was given, some more analyses were done. The first analysis done for this purpose was the Average $\text{RPS}_{\text{average}}$ for all raters $\text{RPC}_{\text{average}}$ combined and corresponding $\text{RPC}_{\text{average}}$ for different types of tasks used in the study. This information was expected to give a clear picture of in what task types the training worked for or against the rating quality.

Table 4: $\text{RPS}_{\text{average}}$ and $\text{RPC}_{\text{average}}$ for Different Types of Tasks

	Paragraph		Essay		Speaking		Overall	
	$\text{RPS}_{\text{average}}$	$\text{RPC}_{\text{average}}$	$\text{RPS}_{\text{average}}$	$\text{RPC}_{\text{average}}$	$\text{RPS}_{\text{average}}$	$\text{RPC}_{\text{average}}$	$\text{RPS}_{\text{Overall}}$	$\text{RPC}_{\text{Overall}}$
Pre-Training	1.74	B	1.47	A	1.97	B	1.73	B
Post-Training	1.57	B	1.60	B	1.40	A	1.52	B

As can be seen in Table 4, there is an improvement in $\text{RPS}_{\text{average}}$ for paragraph, speaking, and $\text{RPS}_{\text{overall}}$ because RPS decreased from 1.74 to 1.57 for paragraph, from 1.97 to 1.40 for speaking, and 1.73 to 1.52 for overall RPS. However, the change was towards the negative direction (from 1.47 to 1.60) for Essay RPS after the training. The magnitude of the change can be understood better when it is considered that these figures are averages of the $\text{RPS}_{\text{average}}$ from 86 raters.

In order to get more detailed information about the change in $\text{RPS}_{\text{average}}$ and $\text{RPC}_{\text{average}}$ after the training, the percentage of the raters moved into a better or worse $\text{RPS}_{\text{average}}$ after the training, and the average change in their $\text{RPS}_{\text{average}}$ was calculated. The findings of this analysis can be found in table 5.

Table 5: Direction and the Magnitude of the Change in $\text{RPS}_{\text{average}}$ and $\text{RPC}_{\text{average}}$

	Paragraph		Essay		Speaking		Overall	
	$\text{RPS}_{\text{average}}$	$\text{RPC}_{\text{average}}$	$\text{RPS}_{\text{average}}$	$\text{RPC}_{\text{average}}$	$\text{RPS}_{\text{average}}$	$\text{RPC}_{\text{average}}$	$\text{RPS}_{\text{Overall}}$	$\text{RPC}_{\text{Overall}}$
Better	60%	-0.90	45%	-0.71	66%	-1.39	65%	-0.35
Worse	40%	0.96	55%	0.83	34%	1.03	35%	0.37

As can be seen from Table 5, most of the raters improved their $\text{RPC}_{\text{average}}$ for paragraph tasks (60%), Speaking tasks (66%), and Overall (65%). However, in the essay task as previously mentioned, only 45% of the raters improved their $\text{RPC}_{\text{average}}$ and 55% of them decreased it.

It can also be seen from the data in Table 5 that the average $\text{RPS}_{\text{average}}$ difference was -0.90 for the ones who improved their RPC for the paragraph task. This amount corresponds to around 5% of the total score as the total score for the ratings was 20 points. Moreover, it can also be seen in Table 5 that the highest positive change (-1.39) occurred in the average $\text{RPS}_{\text{average}}$ of all the raters who improved their RPC in speaking tasks (66%) combined. An interesting point that can be drawn out of Table 5 is that the difference values are mostly (Paragraph, Essay, Overall) highly close to each other regardless of the direction of the change.

In order to get more detailed insights about how the change took place, RMSD for each performance task was also calculated. This time, in order to distinguish between different levels of performance in each task was taken into account and RMSD was calculated based on the different performance levels in each task type separately. The findings in terms of pre and post-training RMSDs are shown in Table 6.

Table 6: Change in RMSD Values for Each Task Type After the Training

	Paragraph			Essay			Speaking		
	Low	Medium	High	Low	Medium	High	Low	Medium	High
RMSD-Pre	2.05	2.39	2.25	1.57	1.89	2.08	2.67	2.28	2.87
RMSD-Post	1.62	2.34	2.11	1.98	1.69	2.36	1.40	2.09	2.24
Difference	-0.43	-0.05	-0.14	0.41	-0.20	0.28	-1.27	-0.19	-0.63

As can be seen in Table 6, such an analysis provides a deeper insight into the problematic performance level that the raters have difficulty in rating. As the table suggests, the highest change happened in the low-level paragraph graded by the raters for the paragraph task. It is important to note that RMSD for the mid-level performance for paragraph task is still high and training didn't help it that much. More training on the medium-level paragraph can be organized based on this finding.

When the essay part of Table 6 is analyzed, the reason why essay RPCs were decreased can be seen easily. However, there is an interesting finding emerges here. It can be seen in Table 6 that the training positively affected the rating in medium level essays. However,it also indicates that the actual problematic performance levels for essay tasks were low- and high-level essays. This may be taken as an indicator of the confusion among the raters in these two levels of essays and need for more training of the raters for these levels.

It was identified previously that the speaking task was the task which raters most benefited from the training and the RPCS. The RMSD change observed in Table 6 also reflects this improvement in scoring with -1.27, -0.19 and -0.63 change in the RMSD of low, medium and high level speaking performance respectively.

The RPCS Feedback Questionnaire

As mentioned earlier, a 12-item questionnaire was shared with the participants and was responded to by 87 raters. The questions asked and the mean scores obtained in this questionnaire can be found in Table 7.

Table 7: 12-Item RPCS Feedback Questionnaire Questions and their Means

#	Question	Mean
Q1	I think the rater performance categorization system should be used in institutions where multiple raters are used to gradea large number of student performance (writing and/or speaking)	3.39
Q2	Seeing my rater performance category is helpful because I know how I am doing among other raters	3.71
Q3	Seeing my rater performance category affects my rater performance in a positive way	3.45
Q4	I will do my best to improve or maintain my rater performance category	4.08
Q5	The Rater Performance Categorization System has made me follow the standardization trainings more carefully	3.47
Q6	I think the rater performance categorization system is fair	3.33
Q7	I think my performance as a rater is reflected correctly by my rater performance category	3.34
Q8	The Rater Performance Categorization System has made me more aware of what I am doing while marking	3.44
Q9	The Rater Performance Categorization System has made me grade student performance more carefully	3.22

Q10	I am anxious about my rater performance category	2.01
Q11	I think the rater performance category system should be connected to the annual performance review of the raters	2.29
Q12	I think the rater performance categorization system will be helpful to us to grade the student performance better as a group	3.45

As can be seen in Table 7. The items endorsed most by the raters are Q2 and Q4. The least endorsed items are Q10 and Q11. The frequency table for the items can be found in Table 8.

Table 8: Frequency and Percentages* Table for 12-Item Feedback Questionnaire

Q1		Q2		Q3		Q4		Q5		Q6	
f	%	f	%	f	%	f	%	f	%	f	%
8	9,2	6	6,9	9	10,3	5	5.7	9	10.3	13	14,9
8	9,2	10	11,5	11	12,6	3	3.4	11	12.6	7	8,0
32	36,8	15	17,2	23	26,4	16	18.4	19	21.8	26	29,9
20	23,0	28	32,2	20	23,0	19	21.8	26	29.9	20	23,0
19	21,8	28	32,2	24	27,6	44	50.6	22	25.3	21	24,1
87	100	87	100	87	100	87	100	87	100	87	100

Q7		Q8		Q9		Q10		Q11		Q12	
f	%	f	%	f	%	f	%	f	%	f	%
12	13.8	14	16.1	19	21.8	45	51.7	35	40.2	10	11.5
6	6.9	7	8.0	7	8.0	17	19.5	12	13.8	8	9.2
25	28.7	17	19.5	19	21.8	11	12.6	26	29.9	25	28.7
28	32.2	25	28.7	20	23.0	7	8.0	8	9.2	21	24.1
16	18.4	24	27.6	22	25.3	7	8.0	6	6.9	23	26.4
87	100	87	100	87	100	87	100	87	100	87	100

*Percentages may not total to 100 as only one decimal is indicated due to space constraints in the table

The raters rated their agreement to the statements in the questionnaire with a 5-point likert scale ((1) I strongly disagree, (2) I disagree, (3) Not sure, (4) I agree, (5) I strongly disagree). As can be seen in Table 8, 44.8 (4 and 5 combined) of the raters agree and strongly agree with the statement that RPCS should be used by organizations where multiple raters are used to grade a large number of student performance (Q1). Only 18.4% (1 and 2 combined) were against this statement, and 36.8% were not sure about it.

64.4% of the raters agreed that seeing their RPC was helpful because it showed how they were doing among other raters (Q2). 17.2% of the raters were not sure about it and only 18.4% of the raters were against this statement. Responses to this item can be taken as a clear endorsement of the RPCS by the raters because this is considered as one of the most important functions of the RPCS. Similarly, 50.6% of the raters endorsed the statement that seeing their RPC affects their performance in a positive way (Q3). 26.4% were unsure and 22.9% were against the idea. The responses to this item can also be interpreted as the endorsement of another main benefit of RPCS to the raters.

72.4% of the raters agreed with the statement that they would do their best to improve or maintain their RPC (Q4). 18.4% of the raters were not sure about it. Only 9.1% disagreed with this statement. This may be taken as an indicator that the RPCS was a source of motivation to rate better for the raters.

55.2% of the raters agreed that the RPCS had made them follow the standardization training sessions more carefully (Q5). 21.8% of the raters were unsure about it and 22.9% were against the statement. This may indicate that RPCS is also helpful to motivate the raters to follow the standardization training sessions more carefully.

47.1% of the raters agreed with the statement that the RPCS is fair (Q6). 29.9% were unsure about it. 22.9% were against the statement. The high number of raters being unsure about the statement may be since they did not fully comprehend how their scores RPSS and RPCS were calculated although it was explained before the training with a separate presentation. Still, the figures indicate that a large percentage (47.1%) of the raters think that RPCS is fair.

50.6% of the raters endorsed the statement that their performance was reflected correctly by their RPCS (Q7). 28.7% were unsure about it and 20.7% of the raters didn't think that their performance was reflected correctly by their $RPC_{average}$. It is interesting to note that 20.7% of the raters were against the statement. This may be due to several reasons one of which may be that they could not fully reflect their performance because of a source of distraction like noise while they were grading their validity tasks.

56.3% of the raters agreed with the statement that the RPCS had made them more aware of what they were doing (Q8). 19.5% of the raters were unsure about it and 24.1% were against the statement. It is important to note that another benefit of the RPCS was endorsed by the raters.

48.3% of the raters endorsed the statement that the RPCS made them grade student performance more carefully (Q9). 21.8% were not sure about it and 29.8 % of the raters were against the statement. This may be an indicator of around 30% of the raters did not take the RPCS carefully as it was a new system, or their disagreement may be because they already take rating business seriously enough before the introduction of the RPCS. However, it is nice to see that the RPCS functioned as a source of motivation for around 50% of the raters.

16% of the raters agreed with the statement that they were anxious about their RPC (Q10). 12.6% were unsure about it. 71.2% of the raters disagreed with the statement. It is nice to see RPCS didn't trigger any kind of anxiety among most of the raters.

Only 16.1% of the raters endorsed the use of RPCS in terms of the annual performance evaluation of the raters (Q11). 29.9% were unsure about it and It was interesting to find out that 54% of the raters were against the idea to connect the RPCS with the annual performance evaluation of the raters. This may be interpreted as although the raters were not anxious of having their performance monitored by the RPCS, they were against the use of RPCS for administrative purposes and decisions.

50.5% of the raters agreed with the statement that RPCS would be helpful to the department to grade the student performance as a group (Q12). 28.7% were unsure and 20.7 % of the raters were against the statement. However, it can be said that most of the raters think that RPCS is helpful to increase the rating quality as a whole.

Discussion

In this part of the paper, the findings regarding the research questions will be discussed. First, the benefits of using the RPCS in real life settings to administrators and raters will be discussed. Then, some considerations regarding the use of RPCS in real-life settings will be discussed further and some suggestions will be made.

RQ1. What are the Benefits of the RPCS to the Administrators in Real-Life Settings?

The findings suggest that the RPCS can provide the administrators with deep insights regarding the rater performance provided by no other tool that existed to IEP before the RPCS was introduced because the practical calculations that could be done by common office software without requiring specialized software and relatively easier complexity of the formula used to rate the rater performance puts the RPCS a step ahead of the current rater monitoring systems in the eyes of the IEP instructors and administrators.

As noted in the results and findings section, the change in the RPSs and RPCs for paragraph, essay, and speaking tasks, as well as the overall RPSs and RPCs, could be identified before and after the training thanks to the RPCS. If the RPCS was not introduced, it would not be possible to measure the change in the rater performance in this detail. Therefore, it can be said that the RPCS helps the administrators to benchmark the change in each rater's performance or a group of raters' performance or even the change in all raters' performance for individual tasks or multiple tasks before and after trainings or over time.

Another benefit of the RPCS is that it helps the administrators to pinpoint organizational or individual training needs more successfully. As findings suggest, the RPCS can give detailed information about the performance tasks that were graded successfully and unsuccessfully by all raters as a whole or a group of raters or even a rater individually. This enables the administrators to envision and plan the training needs as a group or individually within the organization. With the help of this information, some training sessions specifically matching the needs of individual raters, a group of raters, or all rater as a whole. This would save time, energy, and some budget to the organizations.

Last but not the least, the RPCS provides the administrators with objective proof of performance with which some administrative decisions can be made during the annual review of the raters. Although not desired by the raters (based on their responses to Q11) , sometimes, administrative decisions can be taken if the training cannot help the raters improve their performance. RPCs and RPSs can provide the administrators with solid proof of performance in this regard upon which some administrative decisions can be made. However, it should be noted that using RPCS in this way should be kept to some extreme situations as much as possible.

RQ2. What are the Benefits of the RPCS to the Raters in Real-Life Settings?

As drawn out of the responses of the raters to the statements in the 12-item questionnaire, RPCS may:

- help the raters see how they are doing among the other raters.
- positively affect the raters as it provides the raters with a fair performance indicator.
- motivate the raters to maintain or get better RPSs and RPCs.
- make the raters more aware of what they are doing.
- motivate the raters to rate more carefully.

All these benefits listed above suggest that the RPCS provides the raters with some solid performance indicators and help them identify their weaknesses in grading student performance. This naturally provides the raters with a source of motivation to improve themselves. Therefore, the RPCS directly or indirectly may foster the professional development of the raters and has the potential to increase the rating quality within the organization in the long run.

RQ3. Are there Some Considerations of the use of RPCS in Real-Life Settings?

There have some considerations emerged regarding the use of RPCS for rater monitoring purposes.

Rater Effects

The RPCS helps the administrators in terms of identifying only the lenient or severe raters through the individual RPSs and RPCs that are calculated for each validity tasks. However, it doesn't give direct numerical information about the centrality/extremity or accuracy/inaccuracy rater effects. However, accuracy/inaccuracy of raters' scores can be obtained by correlating them with the TE. In order to identify centrality/extremity, "conditional formatting" menu of MS Excel 365 (Microsoft Corporation, 2011) can be used and raters' continuously using central values or extreme values of the scale can be colored differently and such raters can be identified visually by the IEP administrators.

Validity Tasks

As RPCS does not reflect the rater effects like centrality/extremity directly, it is critical to use validity papers or tasks that reflect a varying performance levels (low, medium, high) as was the case in this sample study. This would help the RPCS to balance the $\text{RPS}_{\text{average}}$ and $\text{RPC}_{\text{average}}$ scores and categories because if validity papers with only one performance level (For ex.: high level), is chosen, lenient raters consistently overrating student performance may be superfluously and wrongly categorized in higher RPCs. Therefore, level variability in validity tasks should be maintained.

Suggested use of RPCs

It is suggested that the raters within A+ category can be considered as "Proficient Raters". Raters in A category can be considered as "Skilled Raters". Raters in "B" category can be considered as "Emerging Raters", Raters in category "C" can be considered as

"Developing Raters" and raters in categories "D" and "E" can be considered as "Struggling Raters". Which RPC categories to use as acceptable level of accuracy can be left to the sole decision of the organizations using the RPCS. However, supporting rewarding and praising raters in "A+" and "A" categories, putting these categories in front of "B" and "C" category raters as a goal may be a good idea. It should also be noted that raters in "D" and "E" categories may require individual support and it is suggested that they should be provided with individual support if resources of the organization permits.

Using the RPCS for Administrative Decisions

The findings of this research study indicated that basing decisions on a single RPC for one performance task may not be practical and may be misleading at times. Therefore, $RPC_{average}$ is suggested to be determined as a summative performance indicator (to both the raters and the administrators) based on the $RPS_{average}$ calculated after multiple performance tasks are graded. For example, if there are three paragraphs graded, RPS and RPC for each of them can be calculated and the $RPC_{average}$ should be determined based on $RPS_{average}$ calculated as well. This was how the RPCS was used in this introductory study of the RPCS. An example showing RPS and RPC of each performance task, and $RPS_{average}$ and $RPC_{average}$ of five raters rated paragraph 1 (P1), paragraph 2 (P2), and paragraph 3 (P3) can be found in Table 9.

Table 9: An Example of Average RPC Determined based on Average RPS

	Paragraph 1		Paragraph 2		Paragraph 3		Average	
	RPS 1	RPC 1	RPS 1	RPC 1	RPS 3	RPC 3	$RPS_{average}$	$RPC_{average}$
Rater 1	-0.58	A+	-0.80	A+	-0.30	A+	0.56	A+
Rater 2	-0.83	A+	0.45	A+	-1.55	B	0.94	A+
Rater 3	-0.58	A+	-2.30	C	-2.80	D	1.89	B
Rater 4	-0.08	A+	0.20	A+	0.20	A+	0.16	A+
Rater 5	0.42	A+	-3.80	E	0.20	A+	1.47	A

As can be seen in Table 9, basing decisions on one RPS or RPC can be misleading. $RPS_{average}$ and $RPC_{average}$ can be more informative regarding the rater performance as each individual student performance has its own challenges and/or some raters may be just unlucky at the time of the rating. Therefore, increasing the number of performance tasks rated and calculating $RPS_{average}$ and $RPC_{average}$ based on multiple performance task scores can be more robust against such considerations. However, it is not the case in the current example in Table 9. These figures reflect the real scores of five raters in the current study and as can be seen, some rater's RPS and RPC change dramatically between P1, P2, and P3 (see raters 2, 3 and 5), although some raters perform stably well (See raters 1 and 4). This may not be because raters 2 and 3 were incompetent or unlucky, but it may be because P1, P2, and P3 were paragraph samples with three different performance levels (low, medium, and high). For example, RPC for P1 may indicate these raters' performance in scoring a student paragraph with a low performance level. Therefore, these RPC for P1 may only be interpreted as the competence of these raters in scoring student paragraphs with a low performance level. Therefore, difference between the RPC of the raters in this example may be indicating the raters' competence in rating paragraphs in different performance levels.

As can be seen in the example above, RPCS may be used to envision rater performance as much detail as rater performance on tasks with different student performance levels if used appropriately; however, it also shows that the reliance on the RPCS scores to take some administrative decisions should be done conservatively. To elaborate, according to the $RPC_{average}$ that came out as a result of the RPCS, the reason for the rater's distance from the TE should be determined before a final decision is made about the rater. For example, those who score leniently may be overrating because of being too pro-student or because they don't take their job seriously. It may be necessary to distinguish these two groups of raters. Moreover, those who have overrated may be rating different levels of student performance (valid for instructors teaching English Language Learners) with short intervals in between like rating student papers in B1 level shortly after rating papers in A2 level. In this case, they may be overestimating the performance of students with higher-level solely due to this reason. At the same time, there may be parts in the rating scale that are misinterpreted by a group of raters, and errors in rubrics may also cause negative or positive bias against the raters in this group.

Similarly, it may be just a problematic task driving raters to overrate or underrate. For these reasons, briefly, before RPCS is used for administrative decisions, it should be examined closely why the raters in the lower categories are in those lower categories. Last but not least, it should also be noted that supporting raters through training based on their $\text{RPC}_{\text{average}}$ is an integral part of RPCS because it is known that training increases the inter-rater reliability (Davis 2016). Therefore, first, raters in lower RPCS should be supported with training and final decisions should be made in case of no improvement afterward. It should also be kept in mind that studies indicated that raters' severity levels can be changed after a while (Hoskens & Wilson, 2001; Myford & Wolfe, 2009). Similarly, it can be assumed that leniency (Wolfe et al., 2007) or other rater effects can be changed over time because each item has their unique challenges and rater trainings and the monitoring systems like RPCS can help the raters develop their rating skills. Therefore, raters should be monitored for a certain period of time before a final administrative decision can be made about them.

More importantly, it should be kept in mind that RPCS was created to motivate and support raters in the first place, not to punish them. This was why it was called the Rater Performance "Categorization" System not "Evaluation" system. In connection with this, while the raters in the lower RPC are supported with pieces of training, it is highly recommended to praise or reward the raters in the upper categories as it would be a highly valuable contribution to the RPCS' efficiency in the organization.

Conclusion

Covid-19 pandemic has caused problems in all human beings' life. However, students continue to take tests results of which are used for some high-stakes decisions. This makes what raters grading student performance do much more critical than ever because rater performance is prone to drastic changes in human psychology. Moreover, determining the quality of the rater performance is vital especially at institutions where performance results are used for high stakes decisions because such ratings are done subjectively (Koizumi, et al., 2017). This can only be realized by setting up a rater performance monitoring system. However, the rater performance monitoring systems that are currently available require the use of advanced mathematical models and specialized software for calculations. They were too confusing for the raters and administrators of IEP administrators and raters. A rater performance monitoring system which doesn't require the use of complex mathematical models, easy to calculate and be understood by IEP administrators and raters was necessary. In addition, a rater performance categorization system to motivate the raters by benchmarking their performance, providing them with detailed feedback, and giving them the path to developing themselves professionally during this pandemic period was also needed. In order to cater for these needs, the RPCS was developed.

A research study was designed as a sample study of how the RPCS could be used in real-life settings. For this purpose, 101 raters were given a set of nine sample paragraphs, essays, and speaking tasks (3 samples with three performance levels -low, mid, high- from each) from a previous midterm exam and the raters were asked to grade these tasks without knowing their performance would be categorized. When the participant raters completed their tasks, their RPSs and RPCs were determined and shared with them. They were taken into a standardization training session. Then, they were given another set of nine sample student performances. Out of 101 raters, 86 raters participated in all stages of the study. Then their RPSs, RPCs, and average RPSs and average RPCs were determined and shared with them. Their feedback was taken through a 12-item questionnaire.

It was identified in this sample study to use the RPCS in a real-life setting that the RPCS provides administrators with an easy to calculate, practical and fair performance indicators about the raters if used appropriately. Similarly, the RPCS provides the raters with detailed feedback about their strengths and weaknesses in different task types, shows them how they are doing among other raters, motivates them to follow training sessions more carefully, makes them more aware of what they are doing, and allows them to develop their rating skills.

As can be seen, the RPCS not only benefits the administrators in terms of benchmarking their raters' performance but also motivates the raters to develop themselves professionally. Therefore, the RPCS is suggested to be used in large organizations or departments where a large number of raters rate student performance many times throughout the year taking the suggestions listed under discussion part of this article into consideration. The findings suggested that the RPCS had the potential to contribute to such organizations/departments to elevate the rating quality as a whole.

References

Cao, Jing, et al. "A Bayesian Approach to Ranking and Rater Evaluation: An Approach to Grant Reviews." *Journal of Educational and Behavioral Statistics*, vol. 35, no. 2, 2010, pp. 194-214.

Davis, Larry. "The Influence of Training and Experience on Rater Performance in Scoring Spoken Language." *Language Testing*, vol. 33, no. 1, 2016, pp. 117-135.

DeCarlo, Lawrence T., et al. "A Hierarchical Rater Model for Constructed Responses, with a Signal Detection Rater Model." *Journal of Educational Measurement*, vol. 48, no. 3, 2011, pp. 333-356.

Double-Blind Marking and Moderation Policy. University of Southampton.

Hoskens, Machteld, and Mark Wilson. "Real-time Feedback on Rater Drift in Constructed-Response Items: An Example from the Golden State Examination." *Journal of Educational Measurement*, vol. 38, 2001, pp. 121-145.

https://www.ed.ac.uk/institute-academic-development/learning-teaching/staff/assessment/moderation-guidance

Huang, Lan-fen, et al. "Evaluating CEFR Rater Performance through the Analysis of Spoken Learner Corpora." *Language Testing in Asia*, vol. 8, 2018.

Koizumi, Rie, et al. "A Multifaceted Rasch Analysis of Rater Reliability of the Speaking Section of the GTEC CBT." *ARELE: Annual Review of English Language Education in Japan*, vol. 28, 2017, pp. 241-256.

Linacre, J.M., and B.D. Wright. "The "Length" of a Logit." *Rasch Measurement Transactions*, 1989, pp. 54-55.

Linacre, John M. *A User's Guide to Winsteps® Ministep Rasch-Model Computer Programs.*

Linacre, John M. *Many-Facet Rasch Measurement* MESA Press, 1992.

Lunn, David, et al. "The BUGS Project: Evolution, Critique, and Future Directions." *Statistics in Medicine*, vol. 28, 2009, pp. 3049-3067.

Mariano, Louis T. *Information Accumulation, Model Selection and Rater Behavior in Constructed Response Student Assessments*. Carnegie Mellon University, 2002.

McNamara, Tim. "Applied Linguistics and Measurement: A Dialogue." *Language Testing*, vol. 28, no. 4, 2011, pp. 435-440.

"Microsoft Excel 365." https://www.microsoft.com/en-in/microsoft-365/excel

Myford, Carol M., and Edward W. Wolfe. "Monitoring Rater Performance Over Time: A Framework for Detecting Differential Accuracy and Differential Scale Category Use." *Journal of Educational Measurement*, vol. 46, no. 4, 2009, pp. 371-389.

Rabe-Hesketh, Sophia, et al. *GLLAMM Manual*. University of California, 2004.

Rodeiro, Carmen L. "Agreement between Outcomes from Different Double-Marking Models." *Research Matters*, vol. 4, 2007, pp. 28-34.

Sebok, Stefanie S., and Mark D. Syer. "Seeing Things Differently or Seeing Different Things? Exploring Raters' Associations of Noncognitive Attributes." *Academic Medicine*, vol. 90, no. 11, 2015, pp. 50-55.

Shin, Hyo Jeong, et al. "Human Rater Monitoring with Automated Scoring Engines." *Psychological Test and Assessment Modeling*, vol. 61, 2019, pp. 127-148.

Stata Statistical Software: Release 13. StataCorp, 2013.

Vermunt, Jeroen K., and Jay Magidson. *Technical Guide for Latent Gold 4.0: Basic and Advanced*. Statistical Innovations, 2005.

Wang, Chun, et al. "Essay Selection Methods for Adaptive Rater Monitoring." *Applied Psychological Measurement*, vol. 41, no. 1, 2017, pp. 60-79.

Wang, Peiyu, et al. "Examining Rater Performance on the CELBAN Speaking: A Many-Facets Rasch Measurement Analysis." *Canadian Journal of Applied Linguistics*, vol. 23, no. 2, 2020, pp. 73-95.

Wigglesworth, Gillian. "Exploring Bias Analysis as a Tool for Improving Rater Consistency in Assessing Oral Interaction." *Language Testing*, vol. 10, no. 3, 1993, pp. 305-319.

Wolfe, Edward W., and Aaron McVay. "Applications of Latent Trait Models to Identifying Substantively Interesting Raters." *Educational Measurement: Issues and Practice*, vol. 31, no. 3, 2012, pp. 31-37.

Wolfe, Edward W., et al. *Monitoring Reader Performance and DRIFT in the AP® English Literature and Composition Examination Using Benchmark Essays*. College Board. 2007.

Author Details

Alper Şahin, *Atılım University, Ankara, Turkey,* ***Email ID****: alpersahin2@yahoo.com.*