Available online at ijci.wcci-international.org

International Journal of Curriculum and Instruction 13(3) (2021) 3344–3358

IJCI
International Journal of Curriculum and Instruction

Applying generalizability theory in language testing: Comparing nested and crossed scoring designs in the assessment of speaking skills

Murat Polat [a] *, Nihan Sölpük Turhan [b]

[a] *Anadolu University, Tepebaşı, Eskişehir, 26400, Turkey*
[b] *Fatih Sultan Mehmet Vakıf University, Üsküdar, İstanbul, 34200, Turkey*

Abstract

Scoring language learners' speaking skills is open to a number of measurement errors since raters' personal judgements could involve in the process. Different grading designs in which raters score a student's whole speaking skills or a specific dimension of the speaking performance could be settled to control and minimize the amount of the error in grading foreign language speaking skills. Therefore, the present study aimed to compare G and Phi coefficients gained from the scores of a full factorial (fully crossed) model versus a nested model where rubric components were nested in graders. Four experienced raters and 116 intermediate level language learners studying at a Turkish state university's language school voluntarily participated in this exploratory study. Findings revealed that the G and Phi values obtained with a full factorial grading model were higher. In addition, checking the variance components according to the source of variation, the variance associated with the student's main effect of the full factorial pattern was higher, while the variance value of the residual effect was lower. These findings revealed that full factorial designs could generate more reliable results in speaking exams, thus, it is recommended for language schools to implement the full factorial design in speaking exams when practical conditions such as enough time or sufficient number of graders are available.

Keywords: Generalizability theory; full factorial grading model; nested grading model; intra- rater reliability

1. Introduction

Most language education programs require a combination of interviews and speaking skills assessment, because there is no classical test (including either multiple-choice or open-ended questions) that can provide valid and reliable data to measure the speaking ability of language learners. In this sense, administering speaking exams is useful not only to test basic qualities regarding how language learners use the target language but also to provide evidence that communicative language teaching, which has been trendy for a long

* Corresponding author Murat Polat. Phone.: +90-222-3350580-6120
E-mail address: mpolat@anadolu.edu.tr

time, really works (Bachman and Palmer, 1990). Weir (1990) proposed the use of oral interviews in testing language proficiency by which a substantially high degree of content and construct validity in communicative language assessment could be achieved. Therefore, in foreign language assessment and evaluation practices, it will be appropriate to use open-ended questions in performance tasks, the answers of which are structured by the student itself for measuring high-level cognitive skills of language learners, in addition to objectively scored multiple choice tests. However, the deficiency of using the open-ended questions structured by the student in oral or written language tests when compared to objective tests is the reliability concern in scoring (Connor-Linton, 1995). Salaberry (2000) underlined this important aspect in performance assessment since subjective opinions of raters may come into play in grading process and may result in a certain amount of scoring error, so inter-rater reliability is of greater importance in such measurement and evaluation activities.

Using pre-determined scoring rubrics is one of the commonly used tools to minimize the involvement of raters' subjective opinions in a scoring process and to make more objective scoring in assessment and evaluation. Bacha (2001) defined those rubrics as assessment tools used in scoring oral or written tests which are structured by students and their answers are graded in terms of their semantic and syntactic qualities within specific component bands, and finally the total score is the sum of these component scores. Thus, Popham (1997) underlined the fact that rater expertise and agreement (if more than one rater is involved) should be considered crucial in this type of scoring. However, a substantial amount of studies related to rater agreement in language testing reveal the results of unreliability and rater disagreement in scores (Brown, 1995; Hamp-Lyons, 1995; Heaton, 1988; Kondo-Brown, 2002; Linacre et.al., 1990; Sweedler-Brown, 1993; Turner-Upshur, 2002) Most of these studies report how much unreliable raters are – not only in their own inconsistency in grading students' performances but also in their failure to score consistently with the other graders.

This rater disagreement and amount of scoring error in grading learners' performances are important issues worth investigating since those raters who could be the sources of probable grading error are mostly experienced, trained and guided by scoring rubrics. These rubrics are one of the tools that are commonly used to prevent subjective opinions of raters from entering scoring process in assessment and evaluation activities by which those raters are guided to make more objective scorings. Besides, Goodrich (1996) stated that scales are scoring tools which could assist students according to which criteria their work is going to be evaluated and which score their performance will correspond to. Thus, these tools not only guide students about the way their performances are scored but also contribute to graders' carrying out the scoring task more objectively. The literature presents two main types of scoring scales in terms of the theory they are based on: analytic and holistic rubrics (Nunn, 2000). In holistic scales, a unique grade is assigned to the overall performance of the student and there are definitions in the scale which determine

the quality of student performance at each level. This type of scale is mostly used when some minor errors in a student's speech are disregarded and his/her overall performance is basically focused on (Arter & McTighe, 2001). On the other hand, more commonly used analytical rubrics are scoring tools that provide information about achievement levels in various dimensions of student performance. Thus, it can be possible to present an outline of a learner's mistakes or language errors in particular components of the scale prepared to test the target skills (Marcoulides & Simkin 1992). Sasaki and Hirose (1999) presented four noteworthy reasons that testing via analytic method is more advantageous than holistic style since it is highly comparable, proven to be more reliable, reinforces the raters put emphasis on the task rather than the ideas and finally provides useful diagnostic information which is rare in other testing methods. Taking those advantages of analytic rubrics into consideration, the analytical rubric type was preferred in this study to evaluate students' English-speaking skills.

Nested versus crossed design in speaking tests

Using valid and reliable scoring rubrics in performance assessment may increase rater agreement and could enable the raters to generate scores that could be diagnosed. Although it is expected that the use of these rubrics will increase compatibility between raters and contribute to objective scoring, the reliability of the tests whose answers are structured by the student should be tested by a number of methods (Güler, 2009). In this process, various determinations can be made via a number of methods based on GT (generalizability theory), CT (classical test theory) and IRT (item response theory) (Aytuğ & Toraman, 2020). Moreover, while testing the agreement between raters in classical test theory, the in-class correlation coefficient, Kappa coefficients of Cohen and Fleiss, Kendall's W coefficient and the level of agreement between raters can be used (Verbeke & Molenberghs, 2001). Based on item response theory, Multi-Facet Rasch Model can be used. Like other item response theory models, implicit test fit is used as the probability of a test response in a Multi-facet Rasch model (Macmillan, 2000). This is basically an expansion of the classic rasch model, which adds the parameters of rater rigidity, test adversity, and any other reasons of random error that affect test performance to the model, and defines these sources of variability in the measurement (Congdon, 2007; Iramaneerat et al.,2008).

In most performance tests, while raters' judgements are an important source of error in scoring students' language skills, it is also known that these judgements are personal and could be affected by many variables which can interfere with actual measurement outcomes. Nevertheless, in models based on classical test theory, most of the times only a single source of error is considered and studied in graders' final scores' reliability analyses (Gelman, 2005). Thus, this method will not let simultaneous estimation of reliability grounded on diverse reasons of variance. In these cases, generalizability theory is used for the control of a single reliability quantity through simultaneously evaluating the errors that may stem from various sources of variability including rater, duration, measurement

form, tasks or other independent variables (Güler & Gelbal, 2010). Therefore, generalizability theory was preferred for the analysis of the relevant data in this study.

Depending on the number of variability sources in generalizability theory, studies can be carried out on single-facet or multi-facet universes (Engelhard & Myford, 2003). In this case, one of the sources of variability, most often the source of student variability, is considered as the measurement object. To illustrate, a measurement case in which the student's performance, rubric and rater variable sources are included is called a two-faceted universe, since two surfaces remain (rubric, rater) when the student's performance is taken as the measurement object.

There are different patterns that can be created in two-facet universes. Hoyt (2000) defines the pattern in which all students' answers are graded according to specific rubric components and all raters score all students' performances as full factorial design (sxcxr). However, because of practicality concerns, it is not possible to use full factorial patterns under all testing circumstances. In such cases, nested patterns are generally used. Although there are different nested patterns, one of the most frequently used ones due to practical conditions is the sx(c:r) pattern (Kolko, 1993). In this design, while each rater scores different rubric components in the test, all students answer all components and all raters score all student performances. In other words, while the rubric components are nested in the raters, student performances are crossed with rubric components and raters. Although there are different nested patterns in generalizability theory that can be applied, this study is limited to the two separate patterns described above.

In foreign language assessment, the use of full factorial designs in scoring open-ended questions is sometimes not possible for practical reasons (Lakes & Hoyt, 2008). However, the determination and comparison of the reliability coefficients obtained in the case where the aforementioned nested design and/or full factorial design is used will provide important contributions to the testing and evaluation applications.

Finally, when the studies in the literature are examined, it is possible to see studies in which methods grounded on generalizability theory, classical test theory and multi-faceted rasch model are used while determining reliability of classical or performance-based language tests (Akın & Baştürk, 2012; Borkenau et al., 2001; 2003; Çetin et al., 2016; Iramaneerart et al., 2009; Parlak & Doğan 2014; Reynolds et al., 2009; Ruetten, 1994; Sudweeks et al., 2004). However, when the prominent studies in the literature are examined, there are very few studies comparing the cases in which completely crossed and nested designs are applied for two-facet universes in generalizability theory. Consequently, results of this study will be noteworthy since the aim of the research is to use the Generalizability Theory in cases where the patterns are completely crossed and the scoring

rubric components are nested in the raters in the scoring process of testing English speaking skills. Considering this objective, following questions will be answered:

1. What will be the G and Phi values gathered from the full factorial pattern (sxcxr) in which all sources of variability are crossed?

2. What will be the variance components gathered from the full factorial pattern (sxcxr) where all sources of variability are crossed?

3. What will be the G and Phi values driven from the nested pattern (sx(c:r)) in which the rubric components are nested in the raters?

4. What will be the variance components driven from the nested pattern (sx(c:r)) in which the rubric components are nested in the raters?

5. Are there significant differences between G and Phi coefficients and variance components when different scoring designs are employed?

2. Method

The present study is an exploratory study and it aims to reveal if there is a difference between the G and Phi values gained when the full factorial model (sxcxr) and the model in which the items are nested in the rater (sx(c:r)) are used in the scoring of the English speaking skills. This research model best fits to the aim of this study since Karasar (2005) defines exploratory studies as basic research techniques, which are used to gain deeper insight, develop a theory or test existing theories.

2.1. Participants

The study group of this research consists of 116 intermediate level language learners studying at a Turkish state university's language school in 2018-2019 academic year and 4 experienced graders working in the same institution who had at least 10 years of experience in making oral interviews. All the participants participated in the study voluntarily. The ages of the students ranged from 18 to 20 and majority of them were female students (61%).

2.2. Instruments

A speaking interview question set and an analytic rubric were the data collection instruments of this study. The speaking interview question set included four questions for each student at intermediate language level. The questions which mainly covered general issues about the students' education lives, their general preferences or their plans about the future were prepared by the testing team of the language preparatory

school. Next, the analytic rubric which was developed by the language school's testing team has four separate components. These components are content, fluency, grammatical competence and lexical competence and the score weights of these components are equal (namely 25 points for each that forms 100 points in total).

2.3. Procedure

To be able to compare nested and crossed scoring designs, two different grading settings were planned for the study. First, all the interviews were video-recorded and grading sessions were held by raters' watching and scoring those records. Next, a full factorial design was implemented where graders scored all student performances by scoring all the 4 components of the rubric. In this context, 4 raters scored 4 components and a total of 116 students' speaking performances (sxcxr). Subsequently, a nested design was created in which the rubric components were nested in the raters but the students were crossed with the components and raters (sx(c:r)). At his phase, the same raters scored the same speaking performances; yet, rater 1 only scored the first component of the rubric (content), rater 2 only scored the second component of the rubric (fluency), rater 3 only scored the third component of the rubric (grammatical competence) and rater 4 only scored the fourth component of the rubric (lexical competence). In both designs, the same rater group scored the speaking performances.

2.4. Data collection

In the initial phase of the data collection procedure, necessary official permissions were taken from the language schools' administration. Next, the intermediate level students of the language school were invited to contribute to the study and 116 out of 684 students accepted to participate voluntarily. The students took the speaking tests in groups of four and each student answered 4 questions about general speaking topics. The rater group carried out all the interviews and a total of 29 interview sets were all video-recorded. Each record lasted about 20-25 minutes. In the final step, the rater group watched and scored all the students' speaking performances individually according to two different designs and submitted their scores to the researchers. The data collection process took three weeks in total.

2.5. Data analysis

During the analyses of the research data, means of squares, components of variance and percentages were calculated for the main and common sources of variability in both scoring designs which were carried out by the same graders but in different grading methods. Furthermore, the relative / absolute error variances and the G and Phi values

were analyzed separately including the scores gathered from both research designs. Edu G 6.0 package program was used in the process of making the necessary calculations.

3. Results

This study aimed to compare the G and Phi values gained from two different speaking exam settings where the scoring designs were completely crossed and the scoring rubric components were nested in the graders in scoring foreign language speaking skills of a group of language learners. In this section, findings for both scoring designs were presented, respectively.

3.1. Findings of the full factorial pattern (sxcxr)

The variance components obtained through the G study for the fully factorial design created for 116 students, 4 components and 4 graders are revealed in Table 1. As for the legend of Table 1, the symbol "s" represents the student, the symbol "c" the rubric component and the symbol r represents the rater variability sources.

Table 1. Variance components obtained for the full factorial (sxcxr) pattern

Source of Variance	Total Square	df	Mean Square	Variance	%
s	350.7074	115	3.0496	0.1672	34.7
c	158.9788	3	52.9929	0.0868	20.2
r	6.9845	3	2.3282	0.0028	0.3
sxc	119.5729	345	0.3466	0.0167	4.1
sxr	67.7957	345	0.1965	0.0112	3.2
cxr	0.8962	9	0.0996	-0.0039	0.0
sxcxre	173.6816	1035	0.1678	0.1678	37.5
Total	879.6171	1855			

When the results of the analysis presented in Table 1 were examined, it was seen that the variance component (0.167) estimated for the student (s) main effect variance had the second highest share (34.7%) in the total variance when the full factorial design is used in testing speaking skills. This finding can be interpreted as the possible performance differences between students' English-speaking skills in the actual dimension obtained by the measurement design.

When the main effect of the rubric component was examined, its estimated variance (0.086) had the third highest share in the total variance and explains 20.2% of the whole scoring variance. According to this finding, it could be understood that raters' grading

tendencies and the difficulty levels of each component in the analytic rubric used for speaking assessment differ a lot.

When the main effect of the rater (r) is examined, it was seen that a very small variance component (0.0028) is estimated. The main rater effect explained only 0.3% of the total variance. In the light of this finding, it could be stated that the stringency and leniency levels of the raters' scores to students' performances did not differ significantly. Next, the joint effects were analyzed and the student-component joint effect (sxc) revealed that its variance component (0.0167) explains 4.1% of the total variance. This finding could indicate that the difficulty levels of the rubric components slightly differ from student to student. In addition, the results revealed that the variance component (0.011) calculated for the student-rater joint effect (sxr) explains 3.2% of the whole score variance. Thus, it could be concluded that the scores given by the raters differ slightly for each student's performance.

It was determined that the variance component (-0.0039) calculated for the component-rater joint effect had a negative value and had the smallest share in the total variance. Considering this finding, Cronbach et al. (1972) suggested that negative variance values could be considered as critical values. Brennan (1983), on the other hand, emphasized that negative variance values should be used in the calculation of variance components, but zero should be assigned instead of negative values after the computation process is completed, thus; the method suggested by Brennan was considered in the present study. Although Shavelson and Webb (1991) stated that Brenan's method may prevent biased estimation of variance components, they also emphasized that both methods have certain limitations. In this regard, it can be concluded that the variance component obtained for the component rater joint effect does not make any significant influence to the total variance.

This result indicated that the scores given by the raters do not differ according to the rubric components. Finally, when the residual effect source of variance (sxcxr,e) was examined, it was determined that it had the highest variance component (.0167). The remaining impact component explains 37.5% of the total variance and it can be summarized that the test technique used at this phase of the study contained different variance sources including the students, components, raters and random errors that were not measured in this study. The reliability levels calculated for the full factorial design mentioned above were presented in Table 2.

Table 2. The G and Phi coefficients for sxcxr pattern

$N_{student}$	=116
$N_{component}$	= 4
N_{rater}	= 4

G coefficient	0.85
Phi coefficient	0.78

When the G and Phi values obtained from the (sxcxr) pattern were examined, it was determined that the G coefficient calculated based on the relative error variance was 0.85 and the Phi coefficient calculated on the absolute error variance was 0.78. Unlike the relative error variance, when calculating the absolute error variance, all components (main and common effects) were considered, so the Phi coefficient was always lower than the G coefficient. Since the reliability coefficients vary between 0 and 1 and coefficients of 0.70 and above are acceptable (Crocker & Algina, 2008), it could be stated that both reliability coefficients obtained from sxcxr pattern were within acceptable limits.

3.2. Findings of the (sx(c:r)) pattern (rubric components nested in raters)

The variance components obtained from the nested design where the rubric components were assigned to specific raters but the students were crossed with the components and raters (sx(c:r)) are presented in Table 3.

Table 3. Variance components obtained for the full factorial sx(c:r) pattern

Source of Variance	Total Square	df	Mean Square	Variance	%
s	108.9212	115	0.9471	0.1322	31.0
r	7.8265	3	2.6089	-0.0312	0.0
c:r	40.6235	1	40.6235	0.1135	26.7
sxr	48.3958	345	0.1402	0.0101	3.4
sxc:re	61.4215	115	0.5341	0.1678	38.9
Total	267.1885	579			

(s stands for students, c for rubric components and r for raters)

When the findings in Table 3 were examined, the variance component related to the student variability source had the second highest variance value with a value of 0.13. The "student" variance component explained 31% of the total score variance. Accordingly, differences among students could be determined in the dimension obtained by the measurement process. However, it was detected that the variance component value of students was higher in the full factorial design (presented in Table 1) and explained 34,7% of the total variance. Thus, in this context, it could be stated that the full factorial design

revealed the difference between student performances better than the design in which the rubric components were nested in the raters.

Moreover, it could also be stated that the variance component calculated for the main rater effect (-0.031) had a negative value and this result should be interpreted as the variance component did not make any contribution to the total variance. In addition, there was almost no variability between the raters' scores in the (sx (c: r)) pattern. The variance value for the component rater common interaction (.113) explained 26.7% of the total score variance. This value was the third largest variance value and showed that there were differences between the scores obtained from each rubric component scored by raters.

The variance component obtained for the student rater joint effect was 0.01 and explained 3.4% of the total score variance. It could be interpreted as the difficulty levels of the rubric components differed slightly for each student in the (sx (c: r)) pattern. Finally, when the residual effect source of variability sx(c:r)was examined, it was determined that it had the highest variance component (.0167). The remaining impact component explained 38.9 of the total score variance. Thus, these findings revealed that the interaction among students, rubric components, raters and random errors that were not measured in the study were all included in this scoring process. Moreover, when compared to the full factorial model, it was possible to state that the residual effect variance in (sx (c: r)) pattern in which the rubric components were nested in the raters was higher, and therefore, random errors were more involved in this measurement process. The reliability coefficients calculated for the nested design in which the above-mentioned rubric components nested in the raters were presented in Table 4.

Table 4. The G and Phi coefficients for sx(c:r) pattern

$N_{student}$	=116	
$N_{component}$	= 4	
N_{rater}	= 4	
G coefficient		0.79
Phi coefficient		0.72

When the reliability coefficients obtained for the nested design, in which the rubric components were nested in the raters and the students were crossed with the rubric components and the raters (sx(c:r)), were examined, it was found that the G coefficient was 0.80 and the Phi coefficient was 0.71. Considering that the reliability coefficients vary between 0 and 1 and that coefficients of .70 and above were acceptable, it could be summoned that both reliability coefficients in the (sx(c: r)) pattern were within acceptable limits. However, it must also be specified that both G and Phi values obtained from the full

factorial design were higher than those in the design in which the rubric components were nested in raters.

4. Discussion & Conclusion

Within the scope of this exploratory research, 116 intermediate level language learners' speaking performances in a Turkish state university's language school were tested by four experienced raters via an analytic rubric made up of 4 distinctive components including content, fluency, grammatical competence and lexical competence. Finally, results of the total scores' variance components, reliability values gained from the full factorial scoring pattern and the nested pattern where the rubric components were nested in the graders were all compared and analyzed.

As a result of the findings, first, it was determined that "student" main effect variance component was higher in the full factorial scoring pattern (0.167 for the sxcxr design and 0.132 for sx(c:r) design). In other words, it is possible to infer that the actual speaking skill differences between language students were more successfully revealed in full factorial scoring models. The student variability source stemmed from different scores obtained from the responses of each student whose language speaking skill is the targeted measurement object, and it is desirable that those scores have higher variance values (Güler et al., 2012).

Next, looking at the source of residual effect variability, it was seen that it had the highest variance value in both designs. Similar findings were also obtained in studies conducted on a similar subject (Lynch & McNamara, 1998, Yılmaz & Başusta, 2015). However, the variance component obtained from the pattern in which the rubric components were nested in the raters and the ratio it explained in the total variance were higher than the values obtained from the full factorial scoring pattern. In this case, it was possible to predict that more random errors were involved in the measurement process while using the model in which the components were nested in the raters compared to the full factorial scoring model.

Finally, when the reliability levels calculated in both designs were compared, it was found that the G and Phi values (G: 0.85, Phi: 0.78) determined for the full factorial scoring design were higher than the values of the design in which the rubric components were nested in the raters (G: 0.79, Phi: 0.72). This finding led us to the conclusion that more reliable scores could be obtained when a full factorial scoring design is organized while scoring learners' English-speaking skills with an analytical rubric when compared to the design in which the rubric components are nested in the raters.

It should also be noted that reliability of the performance scores obtained from the types of measurement tools, in which a student constructs the whole answer independently

and where different correct answers can be found, is more problematic to calculate than objective tests whose answers are predetermined. Many test-relevant or irrelevant sources of error may be involved in the grading process such as the halo effect or rater bias. Thus, the Generalizability Theory, which allows the inclusion of different error sources in the process at the same time, might provide a significant advantage in such cases.

However, the use of nested patterns is mandatory for some institutions for practical reasons. For instance, in some testing settings, different questions are prepared by different members of the juries and those who prepare separate questions are invited to score their own questions at the end of the testing process and the pieces of scores each grader assign are eventually brought together to form the final grade. The findings obtained in this study are significant since they reveal that more reliable measurements can be made when the full factorial design is used when compared to the scoring model where the rubric components are nested in the speaking graders. In this context, it is recommended to language schools use a full factorial scoring design in order to gain more valid, reliable and sound test scores in in-class assessment settings and evaluation practices.

On the other hand, some other researchers (Lakes & Hoyt, 2008, Yılmaz & Gelbal, 2011) compared the full factorial design (sxcxr) to the design (sx(c:r)) in which the students were nested in the raters and stated that the G and Phi values were estimated higher in the design where the students were nested in the raters when compared to the full factorial design. This result contradicted the present research's findings. At this point, it can be stated that nested designs are more advantageous than full factorial designs in terms of time, labor and economy in speaking tests performed in large groups. The alternative scoring of rubric components or student performances may contribute to less fatigue of raters and a healthy evaluation. When a full factorial scoring design, where raters score all students and rubric components, is used in testing settings including large groups, it may be possible for raters to make erroneous assessments due to fatigue or carelessness. However, using the full factorial design in in-class testing applications that are not performed on large groups will contribute to obtaining more reliable results.

To conclude, the problem of scoring open-ended questions either in written or spoken form, a common problem of language testing which has increased especially in large-scale exams, is on the agenda nowadays. In such cases, it is not feasible to use the full factorial scoring design where the number of participants is too high. However, it is recommended to researchers who will study similar topics to conduct studies including different nested designs (in which rubric components, interview questions and raters are nested in students, etc.) which might provide a higher level of reliability in speaking exams.

References

Akın, Ö., & Baştürk, D. (2012). Keman Eğitiminde Temel Becerilerin Rasch Ölçme Modeli ile Değerlendirilmesi. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 31*, 175-187. Retrieved from: https://dergipark.org.tr/tr/pub/pauefd/issue/11112/132860

Arter, J. A., & Mctighe, J. (2001). *Scoring Rubrics in The Classroom: Using Performance Criteria for Assessing and Improving Student Performance*, Thousand Oaks, CA: Corvin Press.

Aytuğ, A.K.M., & Toraman, Ç. (2020). Development and application of the Commitment to Profession of Medicine Scale using classical test theory and item response theory. *Croatian Medical Journal. 61(5)*,391-400. DOI: 10.3325/cmj.2020.61.391

Bacha, N. (2001). Writing evaluation: what can analytic versus holistic essay scoring tell us? *Science Direct, System 29,* 371-383. Retrieved from: https://doi.org/10.1016/S0346-251X(01)00025-2

Bachman, L. F., & Palmer, A. S. (1990). The construct of the FSI Oral Interview. *Language Learning, 31(1)*, 67-86.

Borkenau, P., Riemann, R., Angleitner, A., & Spinath, F. (2001). Genetic and environmental influences on observed personality: Evidence from the German Observational Study of Adult Twins. *Journal of Personality and Social Psychology. 80*, 655–668.

Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City, IA: ACT, Inc.

Brown, A. (1995). The effect of rater variables in the development of an occupation specific language performance test. *Language Testing, 12*, 1-15.

Cetin, B., Guler, N., & Sarica, R. (2016). Using generalizability theory to examine different concept map scoring methods. *Eurasian Journal of Educational Research, 66,* 211-228 http://dx.doi.org/10.14689/ejer.2016.66.12

Congdon, P. (2007). *Bayesian Statistical Modelling*, 2nd Ed. Wiley, Chichester.

Connor-Linton, J. (1995). Looking behind the curtain: What do L2 composition ratings really mean? *TESOL Quarterly 29(4),* 762-765.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory.* U.S.A: New York: Chapman & Hall/CRC.

Cronbach, L.J., Glaser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles.* New York: Wiley.

Engelhard, G., & Myford, C. M. (2003). Monitoring faculty consultant performance in the advanced placement English Literature and composition program with a many-faceted Rasch model. *ETS Research Report Series (1),* 1-60.

Gelman, A. & Hill, J. (2007). *Data analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press, Cambridge, UK.

Goodrich, H. (1996). Students Self-Assessment: At the intersection of metacognition and authentic assessment. Doctoral dissertation. Cambridge, MA: Harvard University.

Güler, N. (2009). Generalizability Theory and Comparison of the Results of G and D Studies Computed by SPSS and Genova Packet Programs. *Education and Science. 34*, 154.

Güler, N., & Gelbal, S. (2010). Studying reliability of open-ended mathematics items according to the classical test theory and generalizability. *Educational Sciences: Theory & Practice, 10 (2),* 989–1019. Retrieved from: https://hdl.handle.net/20.500.12619/44393

Güler, N., Uyanık, G. K., & Teker, G. T. (2012). *Genellenebilirlik Kuramı*. Pegem Akademi: Ankara, Türkiye.

Hamp-Lyons, L. (1990). Rating non-native writing: the trouble with holistic scoring. *TESOL Quarterly, 29 (4),* 753-758.

Heaton, J.B. (1988). *Writing English Language Tests*. Longman Handbooks for Language Teachers. Longman Group UK Limited.

Hoyt, W.T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods. 5,* 64–86.

Iramaneerat, C., Yudkowsky, R., Myford, C. M., & Downing, S. M. (2008). Quality control of an OSCE using generalizability theory and many-faceted Rasch measurement. Advances in Health *Sciences Education, 13(4),* 479–493.

Iramaneerat, C., Myford, C. M., Yudkowsky, R., & Lowenstein, T. (2009). Evaluating the effectiveness of rating instruments for a communication skills assessment of medical residents. *Advances in Health Sciences Education, 14 (4),* 575–594.

Karasar, N. (2005). *Bilimsel Araştırma Yöntemi.* Ankara, Nobel Yayın.

Kolko, D.J. (1993). Further evaluation of child behavior ratings: Consistency across settings, time, and sources. *Journal of Emotional and Behavioral Disorders. 1,* 251–259.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias measuring Japanese second language writing performance. *Language Testing, 19 (1),* 3-31.

Lakes, K.D., & Hoyt, W.T. (2008). What sources contribute to variance in observer ratings? Using generalizability theory to assess construct validity of psychological measures. *Infant and Child Development. 17,* 269–284.

Linacre, J. M., Wright, B. D., & Lunz, M. E. (1990). *A facets model for judgmental scoring.* MESA Memo, 61. Chicago, IL: MESA.

Lynch, B. K., & McNamara, T. F., (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*, 158-80.

Macmillan, P. D. (2000). Classical, generalizability, and multifaceted Rasch detection of inter-rater variability in large, sparse data sets. *The Journal of experimental education, 68 (2),* 167-190.

Marcoulides, G. & Simkin, M. G. (1992). Evaluating Student Papers: The Case for Peer Review. *Journal of Education for Business, 67(2),* 80-83.

Nunn, R. (2000). Designing rating scales for small-group interaction. *ELT Journal, Volume 54/2*, 111-132.

Parlak, B., & Doğan, N. (2014). Comparison of answer key and scoring rubric for the evaluation of student performances. *Hacettepe University Journal of Education, 29 (2),* 189–197.

Popham, J. W. (1997). What's Wrong and What's Right with Rubric. *Educational Leadership, 55 (2),* 72-75.

Reynolds, C. R., Livingston, R. L., & Wilson, V. L. (2009). *Measurement and Assessment in Education.* Upper Saddle River, NJ: Pearson/Merrill Publishers.

Ruetten, M.K. (1994). Evaluating ESL Students' Performance on Proficiency Exams. *Journal of Second Language Writing, 3 (2),* 85-96.

Salaberry, R. (2000). Revising the revised format of the ACTFL Oral Proficiency Interview. *Language Testing, 17 (3),* 289-310.

Sasaki, M., & Hirose, K. (1999). Development of an analytic rating scale for Japanese L1 writing. *Language Testing, 16 (4),* 457-478.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A Primer*. Newbury Park CA: Sage.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9 (3),* 239–261.

Sweedler-Brown, C.O. (1993). ESL Essay Evaluation: The influence of sentence-level and Rhetorical Features. *Journal of Second Language Writing, 2 (1),* 3-17.

Turner, C.E., & Upshur, J.A. (2002). Rating scales derived from student samples: Effects of the scale Marker and the Student Sample on Scale Content and Student Scores. *TESOL Quarterly, Vol. 36 (1*), 27-38.

Verbeke, G., & Molenberghs, G. (2001). *Linear Mixed Models for Longitudinal Data*. Springer, New York.

Weir, C. J. (1990). *Communicative Language Testing*. New York: Prentice Hall.

Yılmaz, N.F., & Başusta, N. B. (2015). Genellenebilirlik Kuramıyla Dikiş Atma ve Alma Becerileri İstasyonu Güvenirliğinin Değerlendirilmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi. 6 (1)*. DOI: 10.21031/epod.49284

Yılmaz, N.F., & Gelbal, S. (2011). İletişim becerileri istasyonu örneğinde genellenebilirlik kuramıyla farklı desenlerin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi. 41,* 509-518.
