PEL Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana
2017, 54(2), 1-16

Pontificia Universidad Católica de Chile
Facultad de Educación

Using a Multifocal Lens Model and Rasch Measurement Theory to Evaluate Rating Quality in Writing Assessments

Evaluación de calidad de calificaciones en exámenes escritos a través del modelo multifocal de lente y la teoría de medición de Rasch

Jue Wang & George Engelhard, Jr.
The University of Georgia, USA

Abstract

Lens models have been used extensively for the examination of human judgments. Applications of lens model have occurred in a variety of contexts including judgments of reading achievement, clinical diagnosis, and personality characteristics. Rater-mediated writing assessments involve human judgments toward student essays. The judgmental process determines the quality of ratings, and this directly affects validity and fairness of ratings. Lens models provide a theoretical framework to evaluate rater judgment in rater-mediated assessments. Rasch measurement theory offers an alternative methodological approach for studying human judgments, and we propose combining Rasch measurement theory with a multifocal lens model. In order to illustrate our approach, the ratings of three expert raters and 20 operational raters from a state writing assessment program in the United States are examined in this study. There are only 35% of the students with comparable writing proficiency measures based on a comparison of ratings from operational raters and expert raters. The combination of a multifocal lens model with Rasch measurement theory offers a new pathway for understanding the quality of ratings for rater-mediated assessments.

Keywords: Multifocal lens model; Rasch measurement theory; rater-mediated assessments; writing assessments

Post to:
Jue Wang
126C Aderhold Hall,
110 Carlton St., Athens, GA 30602
Email: cherish@uga.edu
Tel: 1.706.255.5379

ISSN:0719-0409 DDI:203.262, Santiago, Chile
doi: 10.7764/PEL.54.2.2017.3

Resumen

Los modelos de lente han sido usados extensivamente para examinar los juicios humanos. Las aplicaciones de estos modelos han ocurrido en una variedad de contextos que incluyen juicios de logros de lectura, diagnósticos clínicos y características personales. Los exámenes escritos sujetos a calificaciones mediadas involucran juicios humanos hacia los ensayos de los estudiantes. Los procesos de juicio determinan la calidad de las calificaciones, y esto afecta directamente la imparcialidad y validez de las calificaciones. El modelo de lente provee un marco teórico con el que se puede evaluar los juicios en exámenes sujetos a calificaciones mediadas. La teoría de medición de Rasch ofrece un acercamiento metodológico alternativo con el modelo multifocal de lente. Para poder ilustrar nuestro acercamiento a esta teoría, fueron examinadas las calificaciones de tres expertos calificadores y 20 calificadores operacionales de un programa estatal de apoyo a la escritura en los Estados Unidos. Hay solo un 35% de estudiantes con competencias comparables en escritura, medidas en base a la comparación de calificaciones de calificadores operacionales y calificadores expertos. La combinación de un modelo de lente multifocal con la teoría de medición de Rasch ofrece nuevas maneras de entendimiento sobre la calidad de calificaciones mediadas por evaluadores.

Palabras clave: Modelo multifocal de lente; Teoría de medición de Rasch; evaluaciones escritas; evaluaciones mediadas por evaluadores

Human judgments play at key role in rater-mediated writing assessment, and the quality of the judgments determines the reliability, validity and fairness of the ratings. Lens models can provide a theoretical framework to evaluate rater judgment in rater-mediated assessments. The original lens model was introduced by Brunswik (1952, 1955a, 1955b, 1956). In his early research, Brunswik (1955a) focused on the analysis of perceptual constancy. The lens model was adapted by Hammond (1955) to evaluate clinical judgment, and this line of research established the potential for using the lens model to investigate human judgments including social judgment theory (Doherty, 1996). A meta-analysis by Kaufmann, Reips, and Wittmann (2013) documented the extensive use of lens models for the examination of human judgment and decision making in a variety of contexts.

A defining feature of lens models is that they include a comparison between two systems: an *ecological system* and a *judgmental system* (Cooksey, 1996; Hammond, Hursch, & Todd, 1964; Hursch, Hammond, & Hursch, 1964; Tucker, 1964). For example, Cooksey (1986) used a lens model approach to compare teacher judgments of student reading achievement to the scores of student reading achievement on a standardized test via a set of cues. Cooksey (1986) defined the ecological system in terms of the relationship between standardized test scores of reading test and three cues (*i.e.*, social economic status, reading ability, and oral language ability). In a similar fashion, he defined the judgmental system to establish the relationship between teacher judgments and the same set of cues. Regression methodology is typically applied in lens-model studies, and regression-based indices are used to compare ecological and judgmental systems. Cooksey, Freebody, and Wyatt-Smith (2007) also applied a lens model to study teacher's judgment on writing achievement.

In this study, we are comparing the judgments of expert and operational raters using a multifocal lens model. In this case, we have two judgmental systems that are being examined. Specifically, two lens models are presented with each lens model including the judgmental system of

each rater group. These two lens models are shown in Figure 1. Model 1 represents the lens model for expert raters. These expert raters define the criterion system based on a set of cues (domains and rating scale). The experts focus on the cues to make their judgments (expert ratings) showing how student writing proficiency should be defined. Model 2 shown in the right panel is a mirror image of Model 1. Model 2 reflects the judgments of the operational raters in the writing assessment system indicating how student writing proficiency are actually assessed. These models are called lens models because they resemble the way that light passes through a lens with the cues serving to focus on the judgments about writing proficiency.

There is a set of student essays in the center of Figure 1 that form the focal points for the expert and operational raters to estimate student writing proficiency. We are calling this a multifocal lens model because the focus is on student essays with two judgment systems (expert and operational raters) that need to be brought into congruence based on the writing proficiency of these students which are reflected in their essays. Engelhard (2013) proposed a lens model for examining rating quality. In his earlier work, observed ratings were used to examine facets of rater judgments based on Rasch measurement theory (Engelhard, 1992, 1994). Our proposed lens model in this study can be viewed as an extension of his work in that we conceptualize two separate judgment systems reflected in Wright Maps that are separately estimated for expert and operational raters.

Previous lens model studies used correlation and multiple regression analyses to examine the lens models. Hammond (1996) suggested that current research studies using lens model approach overemphasized the role of multiple regression techniques, and that the "lens model is indifferent-a priori-to which organizing principle is employed in which task under which circumstances; it considers that to be an empirical matter" (p. 245). In this study, we encourage the use of Rasch measurement theory as an *organizing principle*. There are several advantages of using Rasch measurement theory.

First, observed ratings on writing assessments are ordinal. It is more appropriate to use categorical data analysis techniques, such as Rasch models (Andrich, 1988; Linacre, 1989; Rasch, 1980; Wright & Masters, 1982). Second, Rasch measurement models provide analyses at the item, rater and person levels that offer detailed information for each domain and category usage in the lens model as well as linear estimates of writing proficiency. This reflects a transition from classical to modern measurement theory that can utilize the new rules of measurement (Embretson, 1996). Third, the invariant properties of Rasch measurement models provide the opportunity to examine the comparability across two Wright Maps of the judgments of expert and operational raters. If an appropriate level of model-data fit is obtained, then the invariant properties of Rasch measurement theory, such as cue-invariant measurement and rater-invariant cue calibration, can be obtained.

Purpose

The purpose of this study is to describe a multifocal lens model that can be used for evaluating rating quality in the rater-mediated writing assessments. We investigate the correspondence between experts and operational raters in conceptualizing writing proficiency based on Wright Maps. Wright Maps are visual displays for representing the lens models used by raters in

two different groups (expert and operational raters) to evaluate student proficiency in writing. In this study, Rasch measurement models are used to calibrate the cues (*e.g.*, domains and rating scales) for the multifocal lens model, and to compare the writing proficiency estimates obtained from expert and operational raters' ratings.

Methodology

Participants

Data were collected from a statewide writing assessment program for Grade 7 students in the southeastern United States. A random sample of 100 essays was used in this study. Operational ratings were obtained with 20 well-trained raters on this sample of 100 student essays. A panel of three expert raters assigned the criterion ratings for these 100 essays. Experts usually are content specialists who have many years of experiences and a deep understanding of the content area of writing. They design and provide the training and practice tests for the raters before going into the operational scoring stage. Operational raters are those who receive training from expert raters and perform operational scoring.

Procedures

Two separate many-facet Rasch (MFR) models are used to examine expert ratings and operational rater ratings separately in the Facets computer program (Linacre, 2015). These two MFR models represent a multifocal lens model with lens models estimated separately for expert and operational raters. The invariance of judged difficulties of the domains and rating-scale structures were compared across the two models. The first model analyzed the ratings of a panel of three experts, and the second model analyzed the operational ratings. In this study, each MFR model included three facets: Writing proficiency, raters, and domains. The equation for Models 1 and 2 can be expressed as follows:

$$\ln\left[\frac{P_{injk}^{(g)}}{P_{inj(k-1)}^{(g)}}\right] = \lambda_i^{(g)} - \theta_n^{(g)} - \delta_j^{(g)} - \tau_{jk}^{(g)} \qquad (1)$$

where:

$P_{injk}^{(g)}$ = probability of Student *n* receiving a rating *k* in Domain *j* by Rater *i*.

$P_{inj(k-1)}^{(g)}$ = probability of Student *n* receiving a rating *k-1* in Domain *j* by Rater *i*.

$\lambda_i^{(g)}$ = logit-scale location (i.e., severity) of Rater *i,*

$\theta_n^{(g)}$= logit-scale location (i.e., judged writing proficiency) of Student *n,*

$\delta_j^{(g)}$ = logit-scale location (i.e., judged difficulty) of Domain *j,* and

$\tau_{jk}^{(g)}$ = threshold parameter indicating the difficulty of Category *k* relative to Category *k*-1 for each domain;

(g) = group indicator that has two values 1 and 2. Value of 1 refers to **Model 1** with expert raters, and a value of 2 represents **Model 2** with operational raters.

The log of the odds that a student receives a rating in Category k rather than in Category $k-1$ given their location on the writing proficiency, the severity of the rater, and the difficulty of the domain, are calculated. The threshold parameter reflects the structure of the rating scale, and it is not considered as a facet in the model.

Instrument

The instrument used in this study is a writing assessment for Grade 7 students. Students were asked to write an essay based on the prompt. An analytic rating scale was used for each domain: (a) idea development, organization, and coherence (IDOC Domain), and (b) language usage and conventions (LUC Domain). The rating scale for IDOC Domain has four categories, and the LUC Domain has three categories.

Results

The two MFR models are displayed using Wright Maps to reflect two separate lens models in Figure 2. The Wright Map for Model 1 with expert ratings is shown in the left panel, and the Wright Map for Model 2 with ratings from operational raters is shown in the right panel. Each Wright Map displays the distribution of writing proficiency and domain locations, as well as the usage of rating categories for each domain. In the writing proficiency (WP) column, the frequencies of examinees at each location are displayed, and examinees are ordered from less proficient (higher logit measure) to more proficient (lower logit measure). In the domain column, the locations of the IDOC Domain and LUC Domain are shown, and the locations are ordered from more difficult (higher logit measure) to less difficult (lower logit measure) for examinees to get a higher score. The structure of category usage is provided in two columns -- RS.1 for IDOC Domain and RS.2 for LUC Domain on the Wright Maps. By comparing the two Wright Maps and the MFR model measures, we can explore the correspondence between expert and rater judgments. This forms the basis for the comparison between the two judgmental systems.

The summary statistics for each facet are displayed in Table 1. The rater facet and domain facet are centered at 0. The estimated writing proficiency measures obtained from experts have a mean of .91 logits with a standard deviation 3.06. The mean of the writing proficiency measures for operational raters is .55 logits with a standard deviation of 2.95 logits. The weighted (Infit) and unweighted (Outfit) mean square errors (MSE) are used to evaluate model fit (Linacre, 2015). For both Infit and Outfit MSE statistics, the closer the statistics are to the expected value of 1.00, the better model fits. The mean values of Infit and Outfit measures for all facets are either .98 or .97 as shown in Table 1, indicating good model data fit. A significant chi-squared statistic for essay separation in Model 1, $\chi^2(99)=797.2$ with $p<.05$, indicates that student writing proficiency has statistically significant variability (Linacre, 2015). Similarly, the operational raters have a significant chi-squared statistic, $\chi^2(99)=5483.2$ with $p<.05$, and this also indicates significant variability in writing proficiency among students.

Based on the multifocal lens model, we define accuracy as the correlation of the estimated writing proficiencies obtained from the two lens models. These raters exhibit a high level of accuracy

with a correlation coefficient of .94. Panel A in Figure 3 shows an approximate linear relationship between the estimated writing proficiencies from the two models. Even though the correlation is high, Panel B shows important differences between the estimated writing proficiencies from expert and operational raters. Using a difference band of ±.50 logits to determine the substantive significance of the differences (Tennant & Pallant, 2007), there are non-negligible discrepancies between the two models for some of the estimated writing proficiency measures for students. There are only 35% of the students who have relatively small difference in writing proficiency measures between -.50 and .50 logits. The vast majority of students have fairly large differences: 44% of students received lower writing proficiency measures from operational ratings compared to measures based on expert ratings, while 21% of students received higher writing proficiency measures obtained from operational ratings compared to measures from expert ratings. Ideally, the goal of training is for operational raters to behave like expert raters with negligible differences in estimated writing proficiency. The deviations of student writing proficiency measures based on operational ratings from those obtained with expert raters indicate that operational ratings may lead to biased measures of student writing proficiency even with well-trained professional raters. This may be due to the difference between judgmental systems and lens models utilized by operational and expert raters which is not fully resolved during training.

The analyses also show differential functioning between domains and category usages information between the two models. From the Wright Maps (Figure 2), we can see that the operational raters weigh the difficulty of two domains differently from the experts. Table 2 shows the location measures of IDOC Domain (.92 logits) and LUC Domain (-.92 logits) in Model 1. A higher measure on a domain facet indicates that it is more difficult for students to achieve high scores; therefore, the IDOC Domain is harder than LUC Domain. The location measures in Model 2 are .50 and -.50 logits for IDOC Domain and LUC Domain respectively. Even though the IDOC Domain also appears as more difficult than LUC Domain in Model 2, the locations of two domains are closer so that operational raters weigh the difficulty of two domains differently from the experts. The Infit and Outfit MSE are both close to 1.00, and this indicates good model fit.

The category information provided in the two models also vary. Table 3 compares the category statistics and structure in two models for IDOC Domain. Experts used a rating of 2 more frequently than operational raters (49% versus 39%). Operational raters gave a rating of 4 more than experts (10% versus 4%). The distance between adjacent threshold estimates should be within 1.40 and 5.00 logits in order to view them as distinctive (Linacre, 1998; Engelhard, 2013). In our analyses, the distances of thresholds for IDOC Domain in two models are all within this range; therefore, we can say that thresholds are distinctive, and that each category carries unique information regarding writing proficiency. As suggested by Engelhard and Wind (2013), we plotted the structure of the rating scale for IDOC Domain using the estimated categories coefficients (Table 3, Panel B). This shows different usages of rating categories between experts and operational raters. For our analyses, the estimated thresholds in Model 1 are more spread out than those in Model 2. The probability category curves (Figure 4) and category information curves (Figure 5) of IDOC Domain between two models confirm that the category usage by experts is more spread out than it is by operational raters.

Table 4 compares the category information for LUC Domain in two models. The proportions of usage for each category are comparable between expert and operational raters. The estimated thresholds for the same category in both models are also very close. The distances between adjacent threshold estimates for LUC Domain in two models are between 1.40 and 5.00 logits so that all three categories are important for the entire scale. The structures of the rating scale of LUC Domain (Table 4, Panel B) shows very similar usages of rating categories between expert and operational raters. The probability category curves (Figure 4) and category information curves (Figure 5) of LUC Domain between two models are also comparable, indicating similar usages of the scale for LUC Domain between expert raters and operational raters. Overall, the category usage of operational raters is relatively accurate as compared to expert raters in LUC Domain; however, more discrepancies between experts and raters appeared in IDOC Domain.

Discussion

The Multifocal Lens Model provides a promising perspective and approach for evaluating rating quality within the context of rater-mediated assessments. Rasch measurement theory brings a new methodology for examining the multifocal lens model and comparing the judgmental systems of expert and operational raters. As Karelaia and Hogarth (2008) said, "the simple beauty of Brunswik's lens model lies in recognizing that the person's judgment and the criterion being predicted can be thought of as two separate functions of cues available in the environment of the decision" (p. 404). In this study, we compared writing proficiency measures evaluated by operational and expert raters via the same set of cues (*i.e.*, domain measures and category usages) in two lens models. The analyses of the cues provided information for exploring rating quality and rater judgements in greater details.

The results of this study indicate that operational raters provided some different ratings from experts' criterion ratings, and that these ratings can be viewed as inaccurate ratings which did result in different estimates of writing proficiency for some students. Only 35% of the writing proficiency measures were comparable under the two judgmental systems used by expert and operational raters. Therefore, inaccurate ratings may bias the estimates of student writing proficiency for the rest of students. The analyses of the cues revealed different judgmental weights for the difficulty of domains and different category usages between operational and expert raters. First, the operational raters viewed the two domains as being closer in terms of difficulty as compared with the expert raters. Second, operational raters had different category usages of the rating scale for IDOC Domain from the expert raters. The structure of the rating scale for IDOC Domain had wider categories in Model 1 (Expert raters) than in Model 2 (Operational raters). For LUC Domain, operational raters were more consistent in using the rating scale as expert raters.

Earlier research, such as Sulsky and Balzer (1988), suggested several ways to obtain "true scores" in rater-mediated assessments, including using the ratings provided by a panel of experts and the average ratings of all the operational raters. In our analyses, we have the estimated writing proficiency measures in Model 1 based on expert ratings, and the estimated writing proficiency measures in Model 2 based on all operational ratings. In other words, we compared these two ways of creating "true scores" for writing proficiency, and concluded that they did not provide comparable results. The judgmental systems used by the expert and operational raters are not the same even after thorough training.

A major advantage of Rasch measurement theory, compared to multiple regression techniques, is to obtain estimated writing proficiency measures on an interval scale. The local independence property of Rasch measurement theory ensures the comparisons to be valid and reasonable based on the invariant writing proficiency measures. Of course, this local independence property is contingent on obtaining good model-data fit which was warranted in our study. In addition, Rasch model, as a technique to analyze categorical data, are more appropriate in dealing with responses with an ordinal measurement level. More importantly, the rating data in human judgment studies are categorical in structure.

Hammond (1996) stressed that various methodologies can be used with the lens model framework. We believe that Rasch measurement theory should join the family of current methods used with lens models. In this study, Rasch measurement theory was used for rating data with different categories for each domain. A family of Rasch measurement models (Andrich, 1988; Linacre, 2015; Rasch 1980; Wright & Masters, 1982) can be used to perform the analyses using the proposed framework for various research purposes. Future research should explore other methodological approaches based on Rasch measurement theory. Nestler and Back (2015) proposed a cross-classified structural equation model to evaluate a lens model for personality judgment. Structural equation modeling (SEM) can incorporate some measurement models based on item response theory (IRT; Baker & Kim, 2004; Muthén & Asparouhov, 2013). Future research can focus on estimating the multifocal lens model all at once based on IRT within SEM framework.

In this study, we did not focus on individual raters. Cooksey (1986) in an earlier study evaluated the overall teachers' prediction on children's reading achievement, and then examined each individual teacher. Future research could extend our framework to identify inaccurate raters by comparing each individual rater's ratings with the expert ratings. This approach can provide detailed information for each rater including rating accuracy, judged difficulty of domains, and structure of category usage.

In summary, this study described the use of Rasch measurement theory to evaluate a rater judgment model framed by using a lens model approach. The multifocal lens model as a theoretical framework for rater judgment can be used to compare rating quality and judgments between expert and operational raters. The discrepancies between two lens models can be identified by Rasch models as an empirical methodology. The combination of Rasch measurement theory and lens model approach can help us understand rater judgments and provide guidance for improving rater training; therefore, improve rating quality as well as the reliability, validity and fairness of ratings within the context of rater-mediated writing assessments.

The original article was received on October 3rd, 2016
The revised article was received on March 16th, 2017
The article was accepted on October 18th, 2017

References

Andrich, D. (1988). *Rasch Models for Measurement.* Newbury Park, CA: Sage.

Baker, F. B., & Kim, S. H. (Eds.). (2004). *Item response theory: Parameter estimation techniques*. CRC Press.

Brunswick, E. (1952). *The conceptual framework of psychology.* Chicago: University of Chicago Press.

Brunswik, E. (1955a). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62*(3), 193-217.

Brunswik, E. (1955b). In defense of probabilistic functionalism: A reply. *Psychological Review, 62*(3), 236-242.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley, CA: University of California Press.

Cooksey, R. W., Freebody, P., & Davidson, G. R. (1986). Teachers' predictions of children's early reading achievement: An application of social judgment theory. *American Educational Research Journal*, *23*(1), 41-64.

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications*. Academic Press.

Cooksey, R.W., Freebody, P., & Wyatt-Smith, C. (2007). Assessment as judgment-in-context: Analyzing how teachers evaluate students' writing. *Educational Research and Evaluation, 13*(5), 401-434.

Doherty, M. E., & Kurz, E. M. (1996). Social judgement theory. *Thinking & Reasoning*, *2*(2-3), 109-140.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341–349.

Engelhard, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education,* 5(3), 171-191.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93-112.

Engelhard, G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement*, *33*(1), 56-70.

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge Academic.

Engelhard, G., & Wind, S.A. (2013). *Rating Quality Studies using Rasch Measurement Theory.* College Board Research Report 2013-3.

Hammond, K.R. (1955). Probabilistic functioning and the clinical method. *Psychological Review*, 62, 255–262.

Hammond, K. R., Hursch, C. J., & Todd, F. J. (1964). Analyzing the components of clinical inference. *Psychological review*, *71*(6), 438.

Hammond, K. R. (1996). Upon reflection. *Thinking & Reasoning*, *2*(2-3), 239-248.

Hursch, C. J., Hammond, K. R., & Hursch, J. L. (1964). Some methodological considerations in multiple-cue probability studies. *Psychological review*, *71*(1), 42.

Karelaia, N., & Hogarth, R. (2008). Determinants of linear judgment: A meta-analysis of lens studies. *Psychological Bulletin, 134*(3), 404-426.

Kaufmann, E., Reips, U. D., & Wittmann, W. W. (2013). A critical meta-analysis of lens model studies in human judgment and decision-making. *PloS one*, *8*(12), e83528.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA press.

Linacre, J. M. (1998). Investigating rating scale category utility. *Journal of outcome measurement*, *3*(2), 103-122.

Linacre, J. M. (2015) *Facets computer program for many-facet Rasch measurement, version 3.71.4.* Beaverton, Oregon: Winsteps.com.

Muthén, B., & Asparouhov, T. (2013). Item response modeling in Mplus: a multi-dimensional, multi-level, and multi-time point example. *Handbook of Item Response Theory: Models, Statistical Tools, and Applications*.

Nestler, S., & Back, M. D. (2015). Using cross-classified structural equation models to examine the accuracy of personality judgments. *Psychometrika*, 1-23.

Rasch (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research. (Expanded edition, Chicago: University of Chicago Press, 1980).

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology*, *73*(3), 497.

Tennant, A., & Pallant, J. F. (2007). DIF matters: A practical approach to test if differential item functioning makes a difference. *Rasch measurement transactions, 20*(4), 1082-1084.

Tucker, L. R. (1964). A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch, and by Hammond, Hursch, and Todd. *Psychological Review, 71*(6), 528-530.

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis. Rasch Measurement*. Chicago: MESA Press.

Appendix

Table 1.

Summary statistics for Facets models

	Model 1 (Experts; n=3)			Model 2 (Raters; n=20)		
	Essays	Experts	Domains	Essays	Rater	Domains
Measure						
Mean	.91	.00	.00	.55	.00	.00
SD	3.06	.86	1.31	2.95	.51	.70
N	100	3	2	100	20	2
Infit MSE						
Mean	.98	.98	.98	.98	.98	.97
SD	.64	.11	.04	.26	.11	.04
Outfit MSE						
Mean	.97	.97	.97	.97	.97	.97
SD	.69	.18	.07	.28	.12	.02
Separation statistics						
Reliability of separation	.89	.96	.99	.98	.91	>.99
Chi square (χ^2)	797.2*	56.8*	98.1*	5483.2*	217.3*	213.0*
df	99	2	1	99	19	1

Note: *p<.05; MSE represents the mean square error.

Table 2.

Domain statistics for Facets models

	Model 1 (Experts; n=3)			Model 2 (Raters; n=20)		
Domains	Measure (SE)	Infit MSE	Outfit MSE	Measure (SE)	Infit MSE	Outfit MSE
Idea Development, Organization, and Cohesion (IDOC)	.92 (.13)	1.01	1.02	.50 (.04)	1.00	.99
Language Usage and Convention (LUC)	-.92 (.14)	.95	.93	-.50 (.05)	.95	.96

Note: SE represents standard errors; MSE represents the mean square error.

Table 3.

Category statistics for Idea Development, Organization, and Cohesion Domain

Panel A. Category statistics

	Model 1 (Experts; n=3)			Model 2 (Raters; n=20)		
Rating category	Proportion of usage	Rasch-Andrich Threshold measure	Distance	Proportion of usage	Rasch-Andrich Threshold measure	Distance
1	21%			23%		
2	49%	-4.27		39%	-3.47	
3	27%	.17	4.44	28%	.25	3.72
4	4%	4.10	3.97	10%	3.22	2.97

Panel B. Estimated rating category structure

Ratings	1 (Lowest)	2	3	4 (Highest)
Estimated thresholds with expert rater ratings		τ_1 (-4.27 logits)	τ_2 (.17 logits)	τ_3 (4.10 logits)
Estimated thresholds with operational rater ratings		τ_1 (-3.47 logits)	τ_2 (.25 logits)	τ_3 (3.22 logits)

Note: Distance refers to the difference between the threshold measures for two adjacent categories.

Table 4.

Category statistics for Language Usage and Convention Domain

Panel A. The category statistics

	Model 1 (Experts; n=3)			Model 2 (Raters; n=20)		
Rating category	Proportion of usage	Rasch-Andrich Threshold measure	Distance	Proportion of usage	Rasch-Andrich Threshold measure	Distance
1	22%			25%		
2	50%	-2.26		48%	-2.29	
3	28%	2.26	4.52	27%	2.29	4.58

Panel B. Estimated rating category structure

Ratings	1 (Lowest)	2	3 (Highest)
Estimated thresholds with expert rater ratings		τ_1 (-2.26 logits)	τ_2 (2.26 logits)
Estimated thresholds with operational rater ratings		τ_1 (-2.29 logits)	τ_2 (2.29 logits)

Note: Distance refers to the difference between the threshold measures for two adjacent categories.

Model 1 Model 2

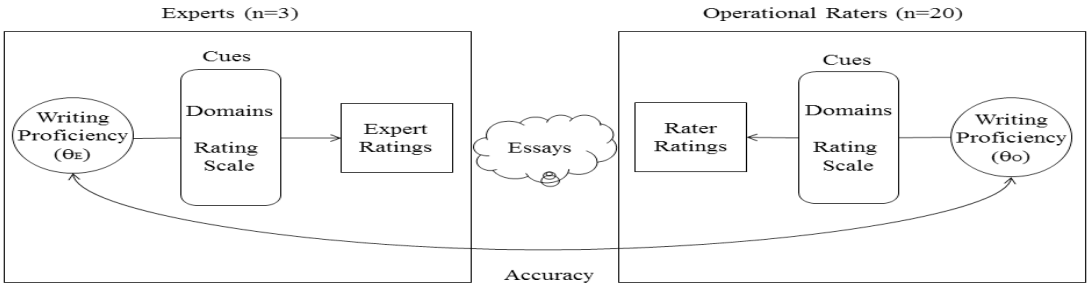


Figure 1. Multifocal Lens Model for Writing Assessment

Model 1
Experts (N=3)

Model 2
Operational Raters (n=20)

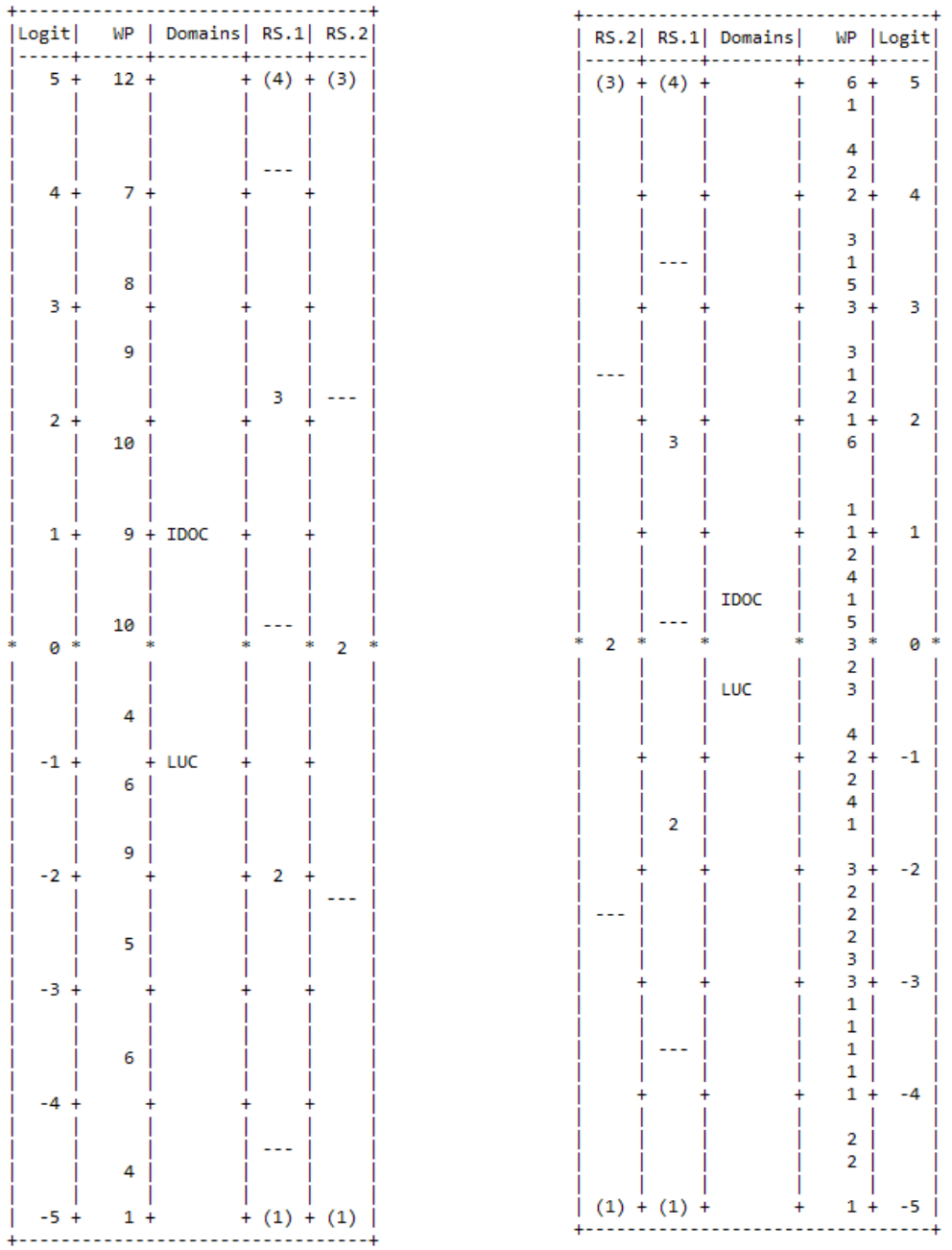
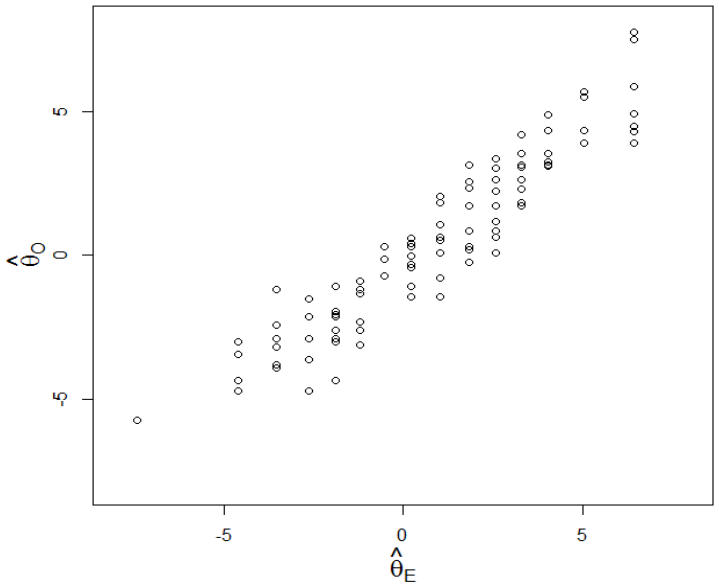


Figure 2. Wright Maps for Experts and Operational Raters

Note: The numbers displayed in WP column represents counts of essays at each location. WP refers to the estimated writing proficiency measures. RS.1 indicates the category usage by raters of the IDOC Domain, and RS.2 is for the LUC Domain.

Panel A. Plot of estimated writing proficiency for Models 1 and 2



Panel B. Plot of differences between estimated writing proficiency for Models 1 and 2

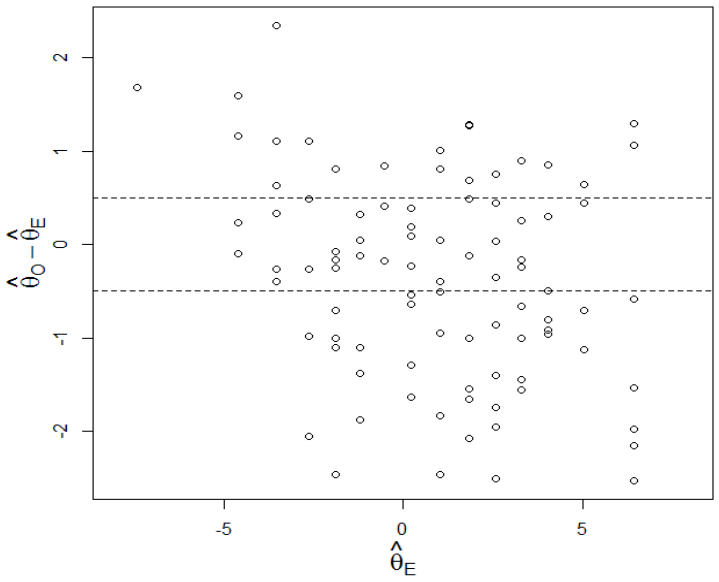


Figure 3. Comparison of Writing Proficiency based on Models 1 and 2

Note: the dotted lines in Panel B indicate a band of ±.5 logits.

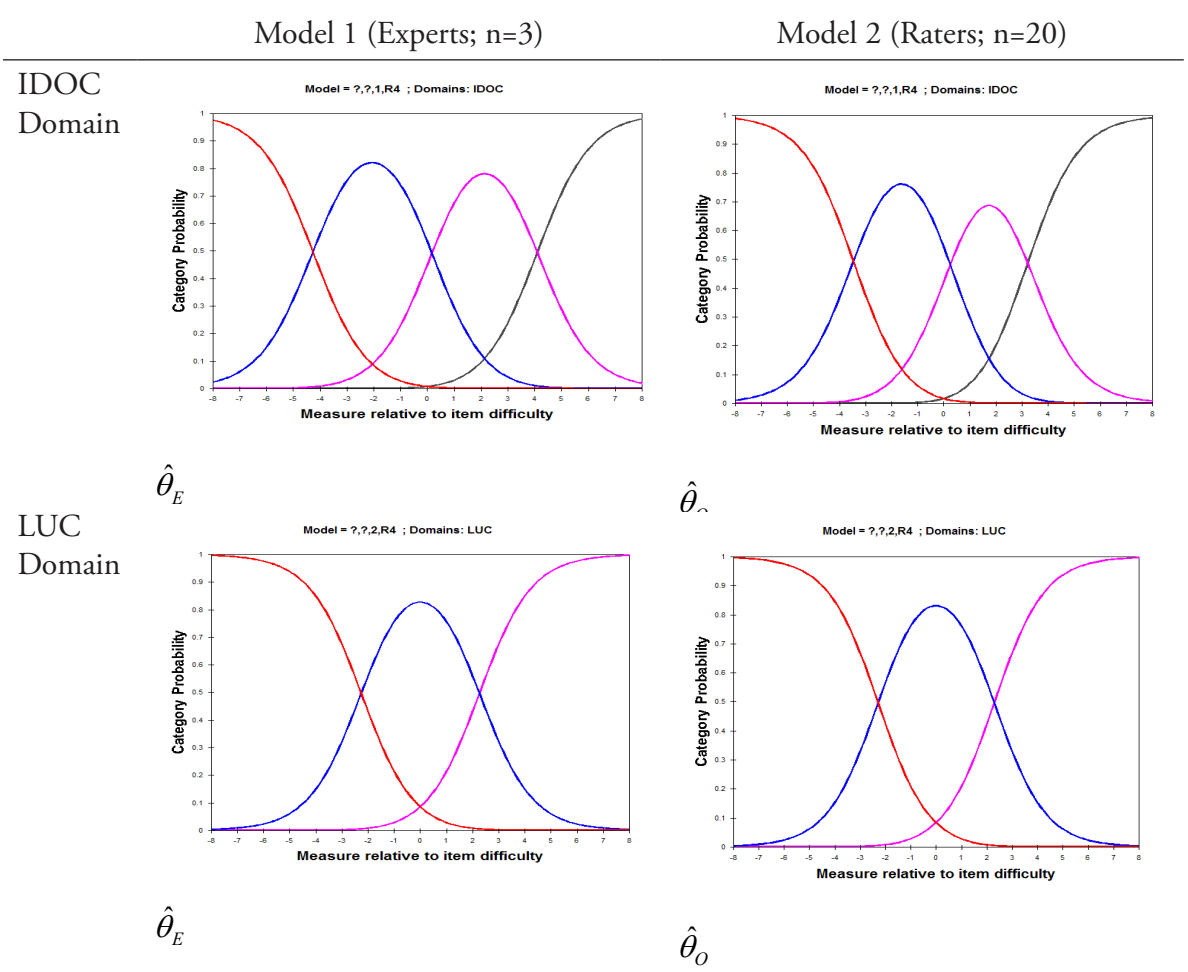


Figure 4. Category characteristic functions

Note: each curve represents a category. Red indicates a category rating of 1, blue is the category rating of 2, purple refers to the category rating of 3, and black line for IDOC Domain is for category rating of 4.

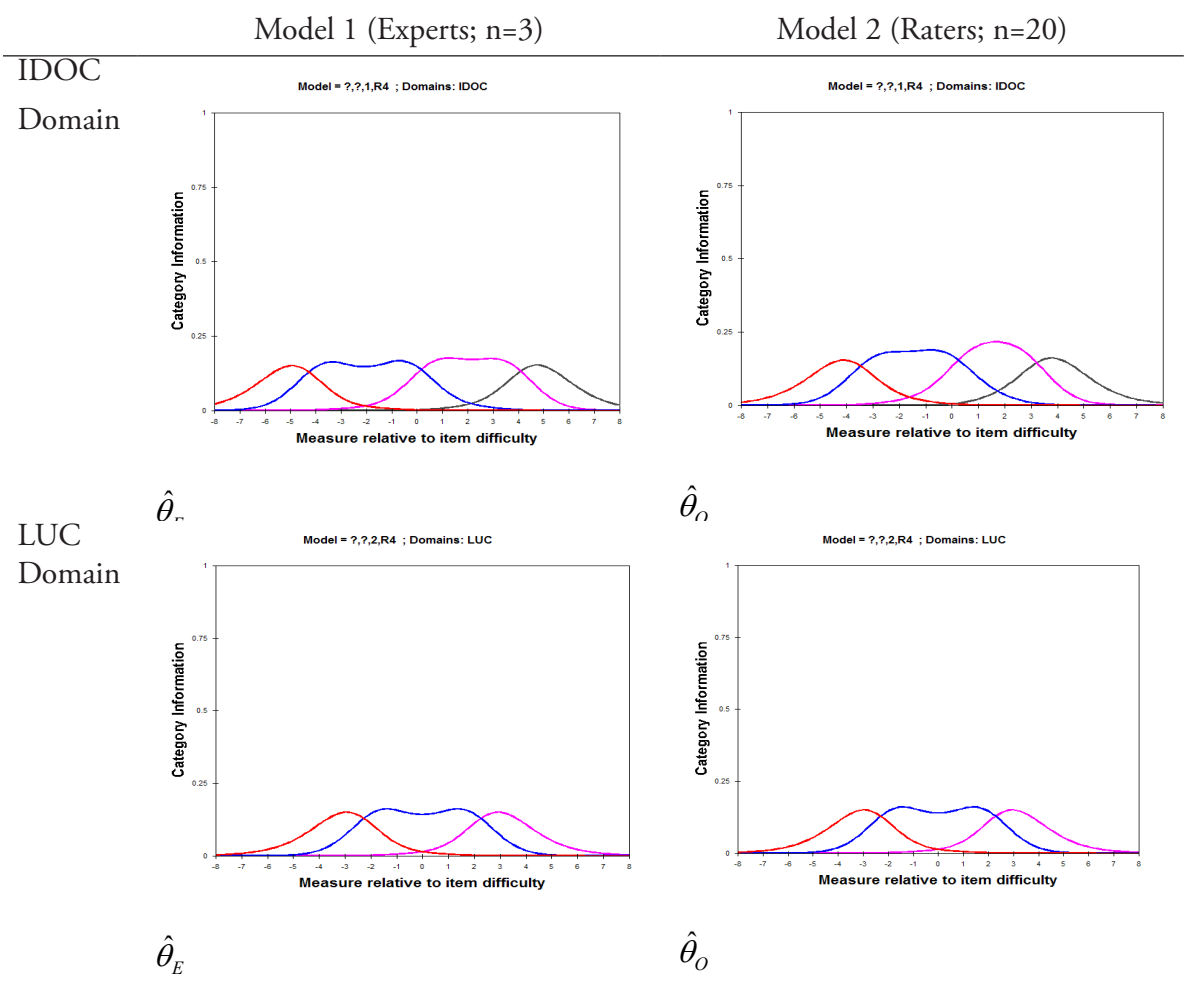


Figure 5. Category information functions

Note: Each curve represents a category. Red indicates a category rating of 1, blue is the category rating of 2, purple refers to the category rating of 3, and black line for IDOC Domain is for category rating of 4.