Kasetsart Journal of Social Sciences 41 (2020) 592–597

Kasetsart Journal of Social Sciences

journal homepage: http://kjss.kasetsart.org

Examining and controlling rater severity and leniency effects on alignment evaluation between science items and science learning indicators using many-facets Rasch modeling

Budsayarat Janprasert[a], Nuttaporn Lawthong[a,*], Sungworn Ngudgratoke[b]

[a] *Department of Educational Research and Psychology, Faculty of Education, Chulalongkorn University, Pathumwan, Bangkok 10330 Thailand*
[b] *School of Education, Sukhothai Thammathirat Open University, Nonthaburi 11120, Thailand*

Article Info

Article history:
Received 12 December 2018
Revised 2 July 2019
Accepted 3 July 2019
Available online 1 December 2020

Keywords:
alignment,
many-facets Rasch modeling,
rater severity and leniency effects

Abstract

This research aimed to (1) examine rater severity and leniency effects on alignment evaluation between the science items and the science learning indicators in junior secondary education, and (2) investigate the alignment evaluation between the science items and the learning indicators when rater severity and leniency effects were controlled. Research subjects were (1) 1,089 in-class science items collected from junior secondary schools under the Office of the Basic Education Commission in Bangkok, and (2) 20 expert panelists participating on alignment evaluation process. Research instrument was a rating scale questionnaire. The inter-rater reliability examined by intra-class correlation was .94 (95% confident interval [CI]: .91–.97). Data analysis was performed using the application of Many-facet Rasch Modeling (MFRM) in FACETS software, version 3.80.3 (Linacre, 2018) and paired-samples t-test. The findings revealed that (1) there were rater severity and leniency effects on alignment evaluation between science items and learning indicators; and (2) when the rater effects were controlled by the MFRM, there was statistically significant difference at the .01 level between non-controlled and controlled mean alignment scores, and alignment evaluation scores of some items were shifted when the effects were controlled.

Introduction

The alignment between learning standards and assessments is the key to ensure precision and reliability of learning expectations in standard-based education (La Marca, 2001; Rothman, 2003). Thus, educational assessment should be aligned with learning standard to maintain the quality of schools and students. Accordingly, studying alignment in educational context is crucial for educational development at any level (from classroom level to state or national level). Alignment evaluation can be helpful to maintain the quality and effectiveness of teaching, measurement and evaluation complying with the learning standards and the learning indicators.

The alignment between test items and learning indicators is typically conducted by expert panelists. Expert panelists carefully evaluate each test item to ensure that they align with learning standards and indicators (Porter & Smithson, 2001, 2002; Rothman, Slattery, Vranek, & Resnick, 2002; Webb, 1997, 1999, 2007). The number of expert panelists on alignment evaluation process has been widely reported such as greater than 2 panelists (Porter & Smithson, 2002), 5–7 panelists (Webb, 2007), and 8 panelists (Mallinson, Stelmack, & Velozo, 2004).

* Corresponding author.
E-mail address: bjanprasert@gmail.com (N. Lawthong).

https://doi.org/10.34044/j.kjss.2020.41.3.22
2542-3151/© 2020 Kasetsart University.

An expert panelist is the key person in evaluating the alignment between test items and learning indicators. Therefore, eliminating or reducing errors based on expert panelists evaluation is crucial. Expert panelists with severe severity and leniency evaluation are likely to provide inaccurate alignment evaluation results (Anderson, Irvin, Alonzo, & Tindal, 2015; Wolfe, 2004). False negative items are wasteful item consumption and false positive items affect overall validity. Rater severity and leniency effects on alignment evaluation process should be controlled. Many-facet Rasch Modeling (MFRM) is an appropriate approach that can explain the rater severity and leniency effects and control them effectively (Engelhard, 1994; Myford & Wolfe, 2000).

This research focuses on the application of MFRM to examine and control rater severity and leniency effects on alignment evaluation process between science items and science learning indicators. The reliability of alignment evaluation results is crucial to maintain quality of teaching, measurement and evaluation. As a result, teachers and higher-level executives can classify test items accurately according to the learning standards and indicators. The collection of test items can be further developed to database or item storage. The information from alignment evaluation can be used in future development to elevate school-made test items to be aligned with learning standards and indicators according to the Basic Core Curriculum.

Literature Review

Alignment

Recently, there have been 3 conventional methods used to study the alignment between test items and learning indicators; (1) Webb's methodology (Webb, 1997, 1999, 2007), (2) survey of enacted curriculum (SEC) or so-called Porter's methodology (Porter & Smithson, 2001, 2002), and (3) methodology of Achieve Inc. (Rothman et al., 2002). Expert panelist is mandatory for all 3 methods. Accordingly, expert panelist should be a master in related fields such as academic matter, technical matter, curriculum, attribute of examinee, and learning standard and indicator (La Marca, Redfield, Winter, Bailey & Hansche, 2000). It can be concluded that alignment evaluation is solely based on expert panelist, so subjectivity of rater's judgment should be reduced (Song & Wolfe, 2015). There are many conventional approaches to reduce errors from subjectivity of rater's judgment, for example, to conduct rigorous selection process to procure expert panelist, to hold informative meeting and intensive training for expert panelist before alignment evaluation process, and to oversee alignment evaluation process precisely. Other than conventional approaches, statistical approach is an interesting approach to control rater effect and reduce error from subjectivity of rater's judgment (Wolfe, 2004). In this research, MFRM is applied for controlling rater severity and leniency effects on alignment evaluation process between test items and learning indicators.

Severity and Leniency Effects

Severity and leniency effects can be defined as "rater's likelihood to proceed with overly positive or overly negative evaluation" (Barrett, 2005; Engelhard, 1994; Myford & Wolfe, 2003, 2004). Rater with severity effect is likely to underestimate ratee while rater with leniency effect is likely to overestimate ratee. Both rater with severity effect and rater with leniency effect are problematic for any kind of evaluation. As a result, severity and leniency effects of rater should be controlled. Regarding raw scores and rater logit scores in Rasch model, rater with lower rater logit score indicates severity of rater while rater with higher rater logit score indicates leniency of rater.

Many-facets Rasch Modeling

Many-facets Rasch modeling (MFRM) is the statistical model intended to explain and control rater severity and leniency effect (Engelhard, 1994; Myford & Wolfe, 2000). Linacre (1994) further developed MFRM to be able to employ it with order category test item or partial credit test item. The amended version of MFRM not only focuses on 2 components or facets but any component can be placed in MFRM logistic function such as test item difficulty, severity effect, and rating scale structure. MFRM can identify rater's likelihood of severity and leniency so MFRM is one of the approaches to analyze effect of various sources of error variation to validity of assessment. The researcher employed MFRM to analyze data and finalize main pattern effect of rater, ratee, trait, item and so forth. Moreover, MFRM can analyze each component or facet separately (Myford & Wolfe, 2003) and MFRM is able to assess individual–level effects for each facet such as rater, and item.

Based on MFRM, all components are analyzed simultaneously but they are independent from each other. All components are calibrated on the same rater scale. As a result, rater's severity, ratee's performance, and trait difficulty can be measured on the same scale. Application of MFRM on alignment evaluation process between test items and learning indicators can be helpful to examine alignment among raters and control rater's severity and leniency. With alignment among rater and controlling rater severity and leniency effects, evaluation is more reliable as well as error from subjectivity of rater's judgment being reduced. Based on MFRM, item parameter is evaluated instead of person parameter and rater parameter is evaluated instead of item parameter. Further information can be found in Data Analysis section.

Learning Standards and Indicators

Learning standards are descriptions of what or how well a student can master a certain knowledge and ability. Learning standards are typically divided into two categories: (1) content standards and (2) performance standards (Hambleton, 2001). Content standards refer to what the students are expected to know or be able to do. Performance standards describe how well the students are expected to know or be able to do according to the content standards. The educational process aims to ensure that the students can reach the standards and the process of assessment aims to measure the standards related to the curriculum.

Learning indicators specify what students should know and be able to do as well as their characteristics for each level so learning indicators reflect the learning standards. This then means learning indicators can be utilized for prescribing contents, determining learning units and organizing teaching-learning activities. Learning indicators serve as essential criteria for evaluation in order to verify the student's quality.

Methodology

Samples

1. One thousand and eighty-nine (1,089) classroom test items were collected from departments of science at junior high schools under the Office of the Basic Education Commission: OBEC in Bangkok metropolitan area. All classroom test items were teacher-made items in academic year of 2016 and aligned with 40 learning indicators of the National Institute of Educational Testing Service: NIETS. The 40 learning indicators were applied to Ordinary National Educational Test (ONET) for ninth grade students as of academic year of 2016. Samples of classroom test items were collected by multi-stage sampling. Initially, 4 junior high schools in Bangkok metropolitan area were chosen by simple random sampling. Then, 48 sets of classroom test items were collected. Finally, only items which aligned with 40 learning indicators of NIETS qualified so 1,089 classroom test items were acquired.

2. Twenty expert panelists were chosen by purposive sampling. The qualifications for expert panelist are listed as following; (1) master in standards and indicators of department of science in junior high school, (2) possess bachelor's degree of education (science) and master's degree of educational measurement and assessment, and/or (3) possess more than 3 years of experience in department of science in any junior high school.

Fifteen (75 %) of the expert panelists were females and the other five (25 %) were males. Fourteen (70 %) of them were teachers in secondary schools while four (20%) were lecturers in universities and the other two (10%) were academics. Most expert panelists possessed more than 6 years experience. Nine (45%) of them possessed 6–8 years of experience and eight (40%) of them possessed more than 8 years of experience and the remaining three (15%) possessed 3–5 years of experience.

Research Instrument

Research instrument was rating scale questionnaire. Alignment rating was divided into 5 levels from 0–4 (0 = item was totally unaligned with certain learning indicator, 1 = item was partially unaligned with certain learning indicator, 2 = panelist was unsure that item was aligned or unaligned with certain learning indicator, 3 = item was partially aligned with certain learning indicator, 4 = item was totally aligned with certain learning indicator) Expert panelists needed to attend the informative meeting held by researcher as stated in "Collection of Data" section before beginning evaluation process. Expert panelists were required to evaluate the alignment between classroom test items and science learning indicators. Classroom test items with 3.00 or greater alignment average score should be aligned with science learning indicator.

Collection of Data

1. A meeting was held for 20 expert panelists to give instruction and conduct a workshop. Then, expert panelists attended consensus building training, and finally, practiced alignment evaluation. Expert panelists were able to discuss freely in their group for more accurate evaluation. The meeting finished within the same day.

2. After the meeting was held, expert panelists did their own alignment evaluation independently based on instruction and assigned items. There were 1,089 items, which were too many for expert panelists, so the researcher divided them into 25 subsets by simple sampling. Each subset comprised of 41–46 items. Each expert panelist was assigned to evaluate 7 subsets (287–312 items). Expert panelists had a period of 1 month to complete the task. The researcher provided them with a copy of a set of assigned items so they were able to complete this task online as well. In the meantime, the researcher did gentle follow-up from time to time.

Data Analysis

1. FACETS Version 3.80.3 (Linacre, 2018) was employed to analyze rater severity and leniency effects based on MFRM. Generally, MFRM consists of 4 parameters as following; (1) person, (2) item, (3) item threshold, and (4) rater. However, in this research, there were only 3 parameters for MFRM, which were (1) item, (2) rater, and (3) rater threshold. Thus, item parameter replaced person parameter for evaluation, and rater parameter replaced item parameter [row of data matrix (subscript *n*) = item parameter, and column of data matrix (subscript I) = rater parameter]. The application was based on Andrich's rating scale model (Andrich, 1978) under 2 conditions; (1) all raters were able to evaluate alignment between item and learning indicator in the same manner, and (2) probability of rater changing alignment level for each item was equal, for example, changing from 3 (partially aligned) to 4 (totally aligned). Raters were required to evaluate alignment which is latent trait of item based on severity and leniency of each rater. Logistic function was 2-facet model. Dependent variable was rater logit and independent variable were components or facets (items, and raters) (Equation 1).

$$\ln (P_{\text{nik}}/P_{\text{nik-1}}) = B_n - D_i - F_k \quad (1)$$

Where P_{nik} is the probability that item n is rated into category k on by rater j, P_{nik-1} is the probability that item n is rated into category $k-1$ on by rater j, B_n is the latent alignment of item n, D_i is the rater-specific severity, and F_k is a category-specific step parameter. F_k is not considered as parameter of model.

2. To identify rater severity and leniency of raters, initially, Infit MNSQ and Outfit MNSQ were employed to determine alignment between model and observed value. The acceptable value was 0.50 to 1.50 (Myford & Wolfe, 2003). Then, rater logit was checked to identify rater's severity and leniency. Thus, the more positive rater logit, the more rater severity while the more negative rater logit, the more rater leniency (Barrett, 2005; Myford & Wolfe, 2004).

3. Comparison of alignment evaluation between learning indicator and item: before severity and leniency effect were controlled and after severity and leniency effect were controlled.

3.1 A comparison was made of observed average (Obs. Avge) and fair-mean average (Fair-M Avge) by paired-samples t-test.

3.2 The difference between Obs. Avge. and Fair-M Avge. was observed and recorded for notable items based on the following incident (1) initially aligned (Obs. Avge. ≥ 3) to be unaligned (Fair-M Avge < 3), and (2) initially unaligned (Obs. Avge < 3) to be aligned (Fair-M Avge ≥ 3).

Results

Examination of Rater Severity and Leniency Effects

1. Fixed χ^2 test was conducted via FACETS Version 3.80.3 (Linacre, 2018) to determine the difference among facets (Engelhard, 1994). This research assumed that all rater severity was aligned after adjustment and fixed χ^2 was at statistical significance level of .01 (χ^2 = 975.70, d.f. =19, p = .00). In conclusion, there were 2 or more raters with different level of severity at statistical significance (Myford & Wolfe, 2003).

2. Raters were likely to exhibit severity rather than leniency since most of rater logit were positive as seen in Table 1. Rater 4 was the most severity (rater logit = 1.83) and rater 20, rater 19, and rater 16 respectively (rater logit = 1.03, 0.93, and 0.92). Rater 13 was the most leniency (rater logit = -3.24) and rater 12, rater 15, and rater 14 respectively (rater logit = -1.83, -1.47, and -1.15).

In Figure 1 (Column 3), variable map revealed that rater 4 was the most severity with highest positive rater logit, thus rater 4 was at the top of column. On the other hand, rater 13 was at the bottom of column since rater 13 possessed highest negative rater logit.

3. Internal consistency of raters (intra-rater reliability) was determined by Infit MNSQ and Outfit MNSQ. Observed value of 17 out of 20 raters (85%) fitted the model expectation well (Infit MNSQ and Outfit MNSQ ranged from 0.70 to 1.31). Rater 13 possessed Outfit MNSQ below 0.50 which was overfit. Rater 12 and 15 possessed MNSQ greater than 1.50 which was misfit.

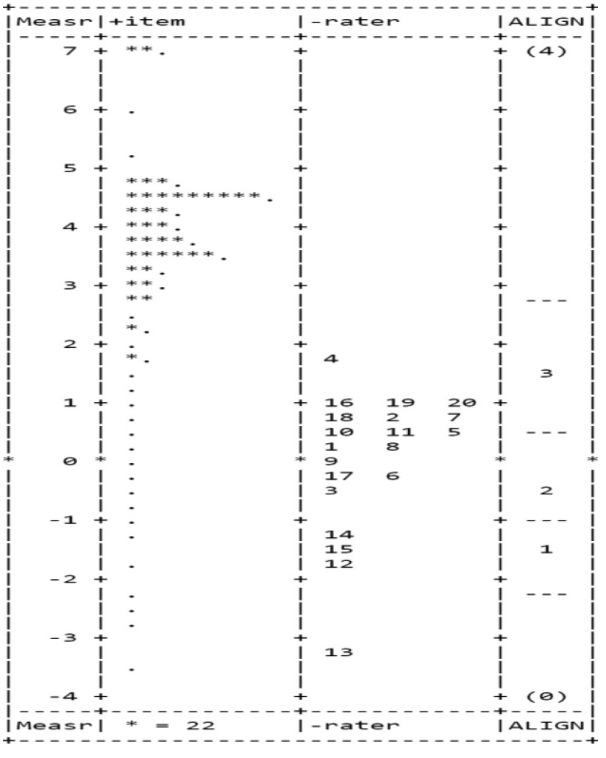

Figure 1 Variables map of MFRM using FACETS2. Examination of the alignment between test items and learning indicators with controlled rater severity and leniency effects

Examination of Alignment between Test Items and Indicators with Controlled Severity and Leniency Effects.

1. Item facet analysis was employed to determine Obs. Avge. which is raw mean without rater severity and leniency effect control and Fair-M Avge. which is adjusted mean with rater severity and leniency control. Item with Fair-M Avge. of 3.00 or greater (Fair-M Avge ≥ 3) was aligned with learning indicator. Accordingly, 1,013 items or 93.02 percent were aligned but the other 76 items or 6.98 percent were unaligned (Fair-M Avge < 3). Analysis of 14 items can be found in Table 2.

2. To determine difference of Obs. Avge. and Fair-M Avge., paired-samples t-test was employed. As a result, Obs. Avge. and Fair-M Avge. were different at statistical significance level of .01 (t = 17.04, p = .00). According to Cohen (1988), effect size was found 0.52 which was moderate.

3. When rater severity and leniency effects were controlled, 11 initially "aligned" items (Fair-M Avge ≥ 3) were subsequently "unaligned" (Fair-M Avge < 3) (see item 14 in Table 2). In contrast, other 10 initially "unaligned" items were subsequently "aligned" (see item 8 in Table 2). In conclusion, there were 21 items that were shifted as a result of controlling rater severity and leniency effects.

4. Alignment between observed value and test items was conducted by Infit MNSQ and Outfit MNSQ as seen in 4.1 and 4.2

4.1 Eight hundred and eighty-three (883) out of 1,089 items (81.80%) possessed infit MNSQ and outfit MNSQ value range from 0.50–1.49, which fitted the model expectation.

4.2 One thousand and thirteen (1,013) items possessed greater than Fair-M Avge of 3.00 but 21 items (2.07%) possessed infit MNSQ and Outfit MNSQ beyond 1.50 which were misfit (see item 3, 7, and 12).

Conclusion and Discussion

The finding revealed that rater severity and leniency effects truly presented on alignment evaluation process. This may reflect that personal experience of raters influenced their tendency toward severity or leniency (Bowsiripon, 2000; Saenplue & Naiyapatana, 2013). Raters were likely to exhibit severity rather than leniency as well. Anderson et al. (2015), also reported that raters were likely to exhibit severity. This study also found that raters were likely to exhibit severity rather than leniency. According to rater experience, raters with more experience were likely to exhibit more severity than raters with less experience. Alignment evaluation required high analytical skill and familiarity with learning standards and indicators. Raters with less experience may not be able to perform well in some complicated items or indicators so they were likely to exhibit leniency. MFRM is appropriated approach to control rater severity and leniency effect on alignment evaluation process. With rater effect under control, educational evaluation should be more reliable and equitable (Turner, 2003). According to Infit MNSQ and Outfit MNSQ, three raters (15%) were unfit with model. Consequently, instruction for raters should be more concise and informative and the workshop session should be more intensive. If the problem persists, problematic rater should be replaced.

Table 1 Analysis of rater facet using MFRM

Rater	Obs. Avge	Fair-M Avge	Measure	*SE*	Infit	Outfit
1	3.63	3.62	0.24	0.12	0.93	0.86
2	3.43	3.46	0.80	0.11	1.17	1.25
3	3.70	3.78	-0.47	0.14	1.20	0.79
4	3.07	3.09	1.83	0.09	0.71	0.73
5	3.53	3.58	0.40	0.11	1.10	1.09
6	3.71	3.74	-0.26	0.13	1.10	1.12
7	3.54	3.50	0.66	0.11	0.99	1.02
8	3.64	3.63	0.22	0.12	0.77	0.70
9	3.67	3.71	-0.09	0.13	0.98	0.90
10	3.54	3.55	0.49	0.11	0.99	0.96
11	3.53	3.55	0.51	0.11	0.99	0.98
12	3.90	3.94	-1.83	0.22	1.79	0.51
13	3.93	3.98	-3.24	0.27	0.98	0.18
14	3.75	3.88	-1.15	0.15	1.01	0.76
15	3.85	3.91	-1.47	0.18	1.69	0.55
16	3.40	3.42	0.92	0.10	1.27	1.12
17	3.68	3.74	-0.23	0.13	1.28	0.83
18	3.46	3.49	0.71	0.11	1.31	1.30
19	3.39	3.41	0.93	0.10	0.98	0.92
20	3.25	3.38	1.03	0.10	1.09	1.08

Note: Obs. Avge: Raw mean before rater severity and leniency effect controlled
Fair-M Avge: Adjusted mean after rater severity and leniency effect controlled
Measure: Rater logit

Table 2 Item facet analysis by MFRM

Item	Grade	Obs. Avge	Obs. Median	Fair-M Avge	Measure	*SE*	Infit	Outfit
1	7	3.85	4.00	3.89	4.79	0.64	0.82	0.66
2	8	2.30	2.50	2.41	0.23	0.28	0.61	0.63
3	8	3.60	4.00	3.74	3.75	0.90	1.67	1.82
4	9	3.60	4.00	3.74	3.75	0.90	0.64	0.65
5	7	3.40	3.00	3.56	3.04	0.79	0.36	0.39
6	8	3.00	3.00	3.12	1.77	0.70	0.62	0.63
7	8	3.60	4.00	3.68	3.52	0.86	2.01	1.74
8	8	2.80	3.00	3.07	1.64	0.65	1.12	1.32
9	9	3.60	4.00	3.76	3.85	0.88	0.67	0.73
10	7	3.00	3.00	2.74	0.86	0.70	0.74	1.16
11	8	3.40	4.00	3.22	2.03	0.84	1.17	0.90
12	9	3.40	3.00	3.49	2.83	0.80	1.90	1.65
13	9	1.40	4.00	1.60	-0.88	0.53	0.59	0.60
14	7	1.00	2.00	0.44	-2.42	0.65	0.95	0.92

Note: Obs. Avge: Raw mean before rater severity and leniency effect controlled
Fair-M Avge: Adjusted mean after rater severity and leniency effect controlled
Measure: Rater logit

The alignment of 21 items (1.90%) were shifted when rater severity and leniency effects were controlled. MFRM is the key for that circumstance since MFRM is statistical approach intended to reduce rater severity and leniency effects. As a result, educational assessment, especially scoring-based system, is even more reliable. It can be concluded that raw mean should be adjusted before evaluation to reduce rater severity and leniency effects.

Most items (93.02%) were aligned with learning standards and indicators. According to the Bureau of Educational Testing, Office of the Basic Education Commission (2016), 98.54 percent of items for ninth grade students were aligned with science learning standards and indicators. Typically, lecturers need to design their own teaching plan, so they should acknowledge and do extensive research for NIETS learning standards and indicators. These learning standards and indicators were applied to ONET as well as participating in ONET being mandatory for students. The government also provides excellent support for lecturers to encourage them to make properly progress.

Regarding 76 items unaligned with learning indicators, most were the items used in “force and motion” strand in relation to learning indicators that required students to know and be able to explain acceleration and effects of resultant force acting to objects. The content in this strand is complicated. Teachers need some time to understand the content and the

learning indicators. As a result, teacher-made test items may not align with learning indicators. Furthermore, 21 test items were misfit. All of them possessed more than 1.50 Infit MNSQ and Outfit MNSQ because their observed values were too different from expected value (Myford & Wolfe, 2003). This may reflect disagreement among raters. The following solution should work; (1) to re-evaluate with newly replaced raters, and (2) to reject that test item.

It can be concluded that severity and leniency particularly influence alignment evaluation process. In addition to alignment evaluation, severity and leniency can be problematic in any kind of evaluation, especially score-based evaluation that is typically used in educational context. Consequently, rater's subjectivity should be controlled. MFRM is statistical approach that can oversee such an issue effectively. MFRM can extensively enhance quality of component and facet in assessment as well as being cost-effective issue of educational measurement.

Recommendation

1. The test items aligned with learning standards and indicators should be reserved to further measurement and assessment. Eventually, collection of test items may be developed into test item database.

2. As this research is retrospective study, further study should consider application of MFRM with prospective study. Prospective study should be conducted before lecturer has designed their test items, so any information gained from prospective study should be advantageous for validity of instrument.

3. MFRM should be applied in another dimension other than content match presented in this study, for example, cognitive complexity dimension (depth match), which was found in Webb's alignment methodology or Porter's alignment methodology.

4. Future research should focus on studying other rater effects on alignment evaluation process such as central tendency effect, restriction of range effect, randomness effect, and differential rater functioning over time.

Conflict of Interest

There is no conflict of interest.

Acknowledgments

The research was funded by the 90th anniversary of Chulalongkorn University Fund (Ratchadaphiseksomphot Endowment Fund).

References

Anderson, D., Irvin, S., Alonzo, J., & Tindal, G. A. (2015). Gauging item alignment through online systems while controlling for rater effects. *Educational Measurement: Issues and Practice, 34*(1), 22–33.

Andrich, D. A. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573.

Barrett, S. (2005). Raters and examination. In S. Alagumalai, D. D. Curtis, & N. Hungi (Eds.), *Applied Rasch measurement: A book of exemplars* (pp. 159–178). Dordrecht, The Netherlands: Springer.

Bowsiripon, U. (2000). *A Comparison of generalizability coefficients of mathematics tests with different scoring methods, numbers of raters and experience of raters* (Unpublished master's dissertation). Srinakharinwirot University, Bangkok, Thailand. [in Thai]

Bureau of Educational Testing, Office of the Basic Education Commission. (2016). *Monitoring and examining the quality of test items used in classroom measurement and evaluation of school under office of the basic education commission* (Research report). Bangkok, Thailand: The Agricultural Cooperative Federation of Thailand. [in Thai]

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

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

Hambleton, R. K. (2001). Setting performance standards on educational assessment and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89–116). Mahwah, NJ: Lawrence Erlbaum.

La Marca, P. M. (2001). Alignment of standards and assessments as an accountability criterion. *Practical Assessment, Research & Evaluation, 7*(21). Retrieved from https://pareonline.net/getvn.asp?v=7&n=21.

La Marca, P. M., Redfield, D., Winter, P. C., Bailey, A., & Hansche, D. (2000). *State standards and state assessment systems: A guide to alignment*. Washington, DC: Council of Chief State School Officers.

Linacre, J. M. (1994). *Many-facet Rasch measurement.* Chicago, IL: MESA Press.

Linacre, J. M. (2018). *FACETS* (Version 3.80.3). Chicago, IL: MESA Press.

Mallinson, T., Stelmack, J., & Velozo, C. (2004). A comparison of the separation ratio and coefficient alpha in the creation of minimum item sets. *Medical Care, 42*, 117–124.

Myford, C. M., & Wolfe, E. W. (2000). *Monitoring sources of variability within the test of spoken English assessment system* (TOEFL Research Report No. 65). Princeton, NJ: Educational Testing Service.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386–422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189–277.

Porter, A. C. & Smithson, J. L. (2001). *Defining, developing, and using curriculum indicators* (CPRE Research report series: RR-048). Philadelphia, PA: Consortium for Policy Research in Education (CPRE): University of Pennsylvania, Graduate School of Education.

Porter, A. C., & Smithson, J. L. (2002, April). *Alignment of assessments, standards and instruction using curriculum indicator data*. Paper presented at the Annual Meeting of American Educational Research Association, New Orleans, LA.

Rothman, R. (2003, March). *Imperfect matches: The alignment of standards and tests.* Commissioned paper prepared for the National Research Council's Committee on Test Design for K–12 Science Achievement, Washington, DC.

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing* (CSE Technical Report No. CSE-TR-566). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Saenplue**,** B, & Naiyapatana, O. (2013). Study of agreement of rater's severity/leniency and differential rater functioning over time on essay composition by the elementary schooling grade 3. *Silpakorn Educational Research Journal, 5*(2), 335–347. [in Thai]

Song, T. & Wolfe, E. W. (2015). *Distinguishing Several Rater Effects with the Rasch Model*. Chicago, IL: National Council of Measurement in Education Annual Meeting.

Turner, J. (2003). *Examining an art portfolio assessment using a many-facet Rasch measurement model.* (Unpublished doctoral dissertation), Chestnut Hill, MA: Boston College.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Research Monograph No. 6). Washington, DC: Council of Chief State School Officers.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states* (Research Monograph No. 18). Washington, DC: Council of Chief State School Officers.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education, 20*, 7–25.

Wolfe, W. E. (2004). Identifying rater effects using latent trait models. *Psychology Science, 46*(1), 35–51.