KARYA ILHAM PUBLISHING
CONNECTING MINDS, BRIDGING IDEAS, INSPIRING READERS

Journal of Advanced Research in Social and Behavioural Sciences

Journal homepage:
https://karyailham.com.my/index.php/jarsbs/index
ISSN: 2462-1951

Validity and Reliability of Artificial Intelligence (AI) Instrument in Instructional Leadership Practice Based on Rasch Model Approach

Jamelaa Bibi Abdullah[1,*], Normarina Abd Rahman[2]

[1] Institut Aminuddin Baki, Genting Highlands Campus, Malaysia

ARTICLE INFO

Article history:
Received 23 October 2025
Received in revised form 17 December 2025
Accepted 20 December 2025
Available online 26 December 2025

Keywords:
Artificial Intelligence (AI); instructional leadership; Rasch Model; construct validity; reliability; item polarity; item fit

ABSTRACT

This pilot study was conducted to validate and examine the reliability of an instrument designed to measure Artificial Intelligence (AI) in Instructional Leadership Practice. The instrument consists of 59 items distributed among 45 principals, headmasters, and senior assistant curriculum teachers under the Ministry of Education Malaysia. The instrument was developed to measure four constructs: (i) the three pillars of AI, (ii) defining the school mission with AI, (iii) managing AI-assisted instructional programs, and (iv) creating a positive AI-assisted climate. The Rasch Model approach was used to examine the validity and reliability of the instrument in this pilot study. The Rasch approach was chosen because it allows for measuring both item and respondent reliability more rigorously than Cronbach's Alpha. It also enables item removal based on item polarity, item fit, and standardized residual correlations. The final analysis revealed that ten items did not meet the criteria and were removed. The final instrument contained 49 items that met validity and reliability standards, making it suitable for measuring the four constructs of AI in Instructional Leadership Practice. As this was a pilot study, future large-scale implementation can be conducted to further measure these constructs among principals, headmasters, and senior assistant teachers.

1. Introduction

Recent developments in Artificial Intelligence (AI) have generated significant global discourse on AI-integrated education, particularly following the launch of ChatGPT, a universal AI model capable of engaging in human-like dialogue to solve complex problems [1,2]. Similarly, the COVID-19 pandemic accelerated the rapid adoption of digital technologies such as virtual classrooms and social media platforms [3], redefining the nature of blended learning and teaching [4].

Principals with low motivation often fail to integrate information technology into teaching and learning [5]. This indicates that principals' motivation needs to be strengthened through digital transformation guided by the International Center for Leadership in Education (ICLE) model [6]

[*] *Corresponding author.*
E-mail address: jamelaa@iab.moe.gov.my

https://doi.org/10.37934/arsbs.41.1.6169

emphasizing three of its seven pillars: student engagement and learning, learning spaces and environment, and professional learning and growth.

Instructional leadership focuses on enhancing teaching quality and learning outcomes by managing the school’s key inputs — teachers and students. Leadership directly influences student academic achievement, which is inherently tied to innovation in teaching. The integration of AI and instructional leadership, therefore, forms the central theme of this study. Given the limited research in Malaysia on the intersection between AI and instructional leadership, this study aims to fill this gap.

To ensure the questionnaire instrument developed is valid and reliable, a pilot study was conducted, followed by an analysis using the Rasch Measurement Model. The Rasch Model enables deeper examination of each item’s functionality beyond Cronbach’s Alpha, identifying the strengths and weaknesses of every item in the instrument.

1.2 Data Analysis Based on the Rasch Measurement Model

Several diagnostic analyses are typically used in the Rasch measurement model to evaluate the validity and reliability of an instrument. These include:

(i) Testing the reliability and separation indices of items and respondents.
(ii) Detecting item polarity in measuring constructs.
(iii) Examining the item fit of each instrument item.
(iv) Determining dependent items through standardized residual correlations.
(v) Identifying item difficulty and respondent ability levels.
(vi) Detecting Differential Item Functioning (DIF) across items.
(vii) Examining the functionality of the rating scale categories.
(viii) Assessing the unidimensionality of the construct.

Beyond verifying construct validity, the Rasch approach also allows researchers to analyze relationships between variables, levels of measurement, and correlation. However, in this paper, the Rasch Model was specifically applied to validate and ensure the reliability of the newly developed AI Instructional Leadership Instrument.

Conventionally, validity and reliability are assessed through Cronbach’s Alpha. However, this study applied Rasch analysis for deeper insight through four diagnostics; (i) item–respondent reliability and separation indices, (ii) item polarity (PTMEA CORR), (iii) item fit statistics, and (iv) standardized residual correlations. These four diagnostics collectively determine whether an instrument meets the psychometric standards necessary for model building.

1.3 Purpose of the Study

The objective of this pilot study was to test the reliability of the developed instrument and detect weaknesses. The Rasch analysis examined item functionality through item–respondent reliability, item polarity, item fit, and standardized residual correlations.

2. Methodology

2.1 Quantitative Approach

This pilot study employed a quantitative approach using a survey instrument distributed to respondents. The sample consisted of 45 principals, headmasters, and senior assistant curriculum teachers leading schools under the Ministry of Education Malaysia. According to [7] an ideal pilot study sample ranges between 25 and 100 respondents, while Johanson and Brooks (2010) recommend a minimum of 30 respondents for preliminary scale development.

Data were analyzed using *Winsteps Version 3.72.3* under the Rasch Measurement Model framework. The AI in Instructional Leadership Instrument consisted of 59 items across four main constructs:

1. The Three Pillars of AI
2. Defining the School Mission with AI
3. Managing AI-Assisted Instructional Programs
4. Creating a Positive AI-Assisted Climate

3. Results

The Rasch analysis examined item functionality through four aspects (i) item–respondent reliability and separation, (ii) item polarity (PTMEA CORR), (iii) item fit, and (iv) standardized residual correlations.

3.1 Item Reliability and Separation

Based on the Rasch measurement model approach, the acceptable value of Cronbach's Alpha (α) is between 0.71 – 0.99 where it is at the best level (71% - 99%) as described in Table 1 [8].

Table 1
Interpretation of the Alpha-Cronbach Score [8]

Score Alpha-Cronbach	Reliability
0.9 – 1.0	Excellent and effective with a high level of consistency
0.7 – 0.8	Good and acceptable
0.6 – 0.7	Acceptable
<0.6	Items need to be repaired
<0.5	Items need to be dropped

To determine the reliability of items in an instrument, statistical analysis with the Rasch measurement model approach is used with reference to the reliability values as well as the isolation of items. The results of the analysis of the pilot study found that the reliability value obtained based on the Alpha Cronbach value (α) was 1.00 as shown in Table 2. It clearly shows that this value means that the instrument used is in excellent condition and effective with a high level of consistency thus can be used in real research.

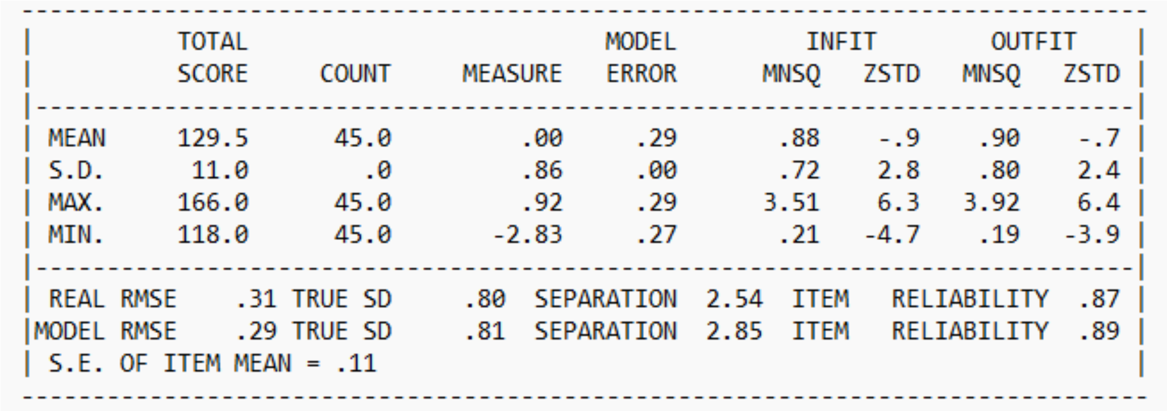
Table 2
Reliability Value (Alpha Cronbach) for Pilot Studies

person RAW SCORE-TO-MEASURE CORRELATION = .97
CRONBACH ALPHA (KR-20) person RAW SCORE RELIABILITY = 1.00

The analysis of the instrument was also conducted comprehensively by examining the reliability and separation indices of both items and respondents. Table 3 shows the reliability and item isolation values where the item reliability value is 0.87, while the item isolation value is 2.54. Based on the item's reliability value, a value of 0.87 indicates the item is in good and acceptable condition [8].

While the item separation value is 2.54 if rounded is equal to 3.0 and that explains the separation value that is more than 2.0 is a good value. According to [9] a good index segregation value is more than a 2.0 value.

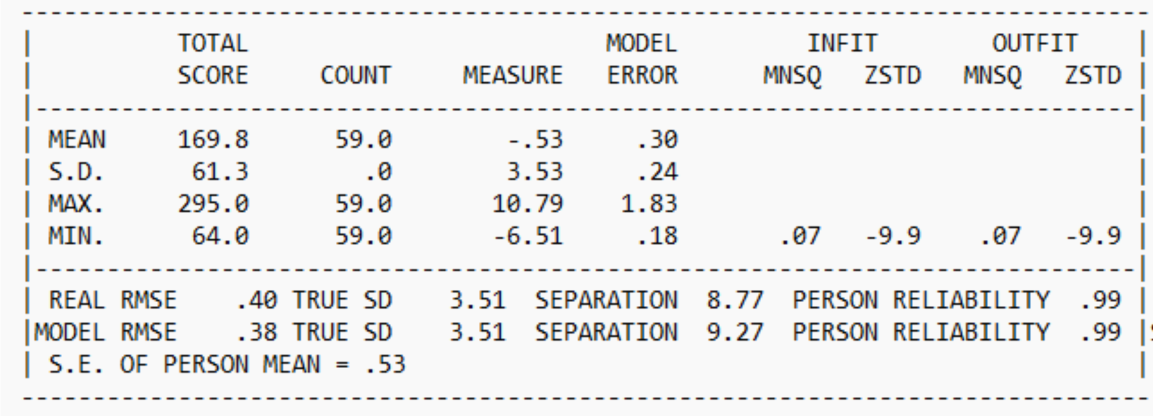
Table 3
Reliability and item isolation values of the entire instrument construct: A pilot study



```
---------------------------------------------------------------------------------
|          TOTAL                         MODEL        INFIT        OUTFIT    |
|          SCORE      COUNT     MEASURE  ERROR     MNSQ   ZSTD  MNSQ   ZSTD |
|-------------------------------------------------------------------------------|
| MEAN     129.5       45.0         .00    .29       .88    -.9    .90    -.7 |
| S.D.      11.0         .0         .86    .00       .72    2.8    .80    2.4 |
| MAX.     166.0       45.0         .92    .29      3.51    6.3   3.92    6.4 |
| MIN.     118.0       45.0       -2.83    .27       .21   -4.7    .19   -3.9 |
|-------------------------------------------------------------------------------|
| REAL RMSE    .31 TRUE SD     .80  SEPARATION  2.54  ITEM   RELIABILITY  .87 |
|MODEL RMSE    .29 TRUE SD     .81  SEPARATION  2.85  ITEM   RELIABILITY  .89 |
| S.E. OF ITEM MEAN = .11                                                     |
---------------------------------------------------------------------------------
```

Meanwhile, based on Table 4, the reliability value of the respondents is 0.99 and the isolation value of the respondents is 8.77. This shows that the reliability value of the respondents is very high and very good. This is because [8] explain that reliability values above 0.8 are good and strongly accepted. Meanwhile, the segregation value of the respondents showed a good segregation value on the level of difficulty of the item in accordance with [9] who explained that an segregation value of more than 2.0 is a good value.

Table 4
Reliability and respondent isolation values for the entire instrument construct: A pilot study



```
---------------------------------------------------------------------------------
|          TOTAL                         MODEL        INFIT        OUTFIT    |
|          SCORE      COUNT     MEASURE  ERROR     MNSQ   ZSTD  MNSQ   ZSTD |
|-------------------------------------------------------------------------------|
| MEAN     169.8       59.0        -.53    .30                                |
| S.D.      61.3         .0        3.53    .24                                |
| MAX.     295.0       59.0       10.79   1.83                                |
| MIN.      64.0       59.0       -6.51    .18       .07   -9.9    .07   -9.9 |
|-------------------------------------------------------------------------------|
| REAL RMSE    .40 TRUE SD    3.51  SEPARATION  8.77  PERSON RELIABILITY  .99 |
|MODEL RMSE    .38 TRUE SD    3.51  SEPARATION  9.27  PERSON RELIABILITY  .99 |
| S.E. OF PERSON MEAN = .53                                                   |
---------------------------------------------------------------------------------
```

3.2 Item Polarity (PTMEA CORR)

The examination of the Point Measure Correlation (PTMEA CORR) values was conducted to detect item polarity and to assess the extent to which each construct aligns with its intended measurement objective. A positive (+) PTMEA CORR value indicates that the item successfully measures the intended construct [8]. Conversely, a negative (–) value suggests that the item does not measure the intended construct. Such items should be revised or removed, as they may not align with the question focus or may be difficult for respondents to interpret.

Based on Table 5, no items were required to be removed. However, the results show that all items demonstrated positive movement in the same direction as their respective constructs, indicating that they effectively measure and are consistent with the intended construct. A higher PTMEA CORR value further signifies that the item has strong discriminative power in differentiating among respondents' abilities.

Table 5
Point measure correlation value

Entry Number	*Point Measure Corr.*	*Item*	*Entry Number*	*Point Measure Corr.*	*Item*	*Entry Number*	*Point Measure Corr.*	*Item*
7	.46	pgb7	51	.90	pgb_galperprog2	47	.92	pgb_insentif3
6	.53	pgb6	58	.90	pgb_insetpembel4	23	.92	pgb_selia4
2	.55	pgb2	12	.90	pgb_mmt3	34	.92	pgb_pantau5
8	.56	pgb8	33	.90	pgb_pantau4	55	.92	pgb_insetpembel1
1	.60	pgb1	21	.90	pgb_selia2	39	.93	pgb_lindungi5
5	.60	pgb5	32	.91	pgb_pantau3	45	.93	pgb_insentif1
3	.60	pgb3	49	.91	pgb_insentif5	44	.93	pgb_nampak5
4	.61	pgb4	46	.91	pgb_insentif2	37	.93	pgb_lindungi3
9	.62	pgb9	59	.91	pgb_insetpembel5	28	.93	pgb_selaras4
10	.83	pgb_mmt1	56	.91	pbg_insetpembel2	38	.93	pgb_lindung4
14	.84	pgb_mmt5	53	.91	pgb_galperprog4	31	.93	pbg_pantau2
11	.86	pgb_mmt2	54	.91	pgb_galperprog5	17	.94	pgb_mrng3
35	.86	pgb_lindungi1	57	.91	pgb_insetpembel3	24	.94	pgb_selia5
36	.87	pgb_lindungi2	26	.91	pgb_selaras2	29	.94	pgb_selaras5
15	.88	pgb_mrgn1	19	.91	pgb_mrng5	18	.94	pgb_mrgn4
40	.88	pgb_nampak1	30	.92	pgb_pantau1	22	.94	pgb_selia3
13	.88	pgb_mmt4	20	.92	pgb_selia1	50	.94	pgb_galperprog1
16	.89	pgb_mrng2	41	.92	pgb_nampak2	25	.95	pgb_selaras1
48	.89	pgb_insentif4	43	.92	pgb_nampak4	27	.95	pgb_selaras3
52	.89	pgb_galperprog3	42	.92	pgb_nampak3			

3.3 Fit of Construct Measuring Items

The suitability (fit) of the items in measuring a construct can be evaluated through the *infit* and *outfit* Mean Square (MNSQ) values. According to [8] acceptable MNSQ infit and outfit values should range between **0.6 and 1.4**, ensuring that the developed items are appropriate for measuring the intended constructs.

However, greater attention should first be given to the *outfit* index compared to the *infit* index when determining whether an item corresponds accurately to the latent variable being measured. An MNSQ value **greater than 1.4 logits** indicate a *misfitting item*, suggesting that the item may confuse respondents or measure unintended aspects of the construct. Conversely, an MNSQ value **less than 0.6 logits** implies that the item is *too predictable* or *too easy* for respondents [9].

In addition, the standardized z-score fit statistics (*ZSTD infit* and *ZSTD outfit*) should also fall within the range of **–2 to +2** [8]. Nevertheless, when MNSQ infit and outfit values are within the acceptable range, the ZSTD indices may be disregarded [9].

If these conditions are not met, the item should be considered for refinement or removal to ensure that only items with acceptable fit values remain. Table 6 below presents the *misfit order*, displaying the ten items with the highest MNSQ values and the thirty-one items with the lowest MNSQ values obtained from the item-level statistical analysis.

Table 6
Item fit based on MNSQ value

Entry	*Infit*		*Outfit*		*Ptmea*	*Items*
number	*MNSQ*	*ZSTD*	*MNSQ*	*ZSTD*	*Corr*	
7	2.40	5.2	3.92	6.4	A .46	pgb7
4	3.51	6.3	3.28	5.3	B .61	pgb4
6	2.98	6.1	2.82	5.0	C .53	pgb6
3	2.39	5.1	2.70	4.8	D .60	pgb3
2	2.24	4.7	2.39	3.7	E .55	pgb2
5	2.39	5.1	2.25	3.8	F .60	pgb5
1	2.12	4.4	2.25	3.5	G .60	pgb1
8	1.92	3.4	1.96	2.2	H .56	pgb8
9	1.79	3.0	1.76	1.9	I .62	pgb9
10	1.07	.4	1.49	1.7	J .83	pgb_mmt1

Based on Table 6, ten items were found to fall outside the acceptable range and therefore required refinement or removal. Items with *outfit MNSQ* values exceeding 1.40 included: **pgb7 (3.92)**, **pgb4 (3.28)**, **pgb6 (2.82)**, **pgb3 (2.70)**, **pgb2 (2.39)**, **pgb5 (2.25)**, **pgb1 (2.25)**, **pgb8 (1.96)**, **pgb9 (1.76)**, and **pgb_mmt1 (1.49)**.

Meanwhile, thirty-one items that exhibited *MNSQ* values below 0.6 were refined based on the researchers' judgment and expert consultation to ensure that the items remained aligned with the construct and suitable for measurement.

3.4 Standardized Residual Correlations Measurement

The measurement of standardized residual correlations is used to detect **local dependence**, which refers to whether one item is dependent on another. Local dependence occurs when two items share a high positive correlation, indicating redundancy in measurement. According to [9], if the correlation value between two items exceeds **0.7**, it suggests that the items are interdependent and not measuring distinct aspects of the construct. Therefore, [9] recommends retaining only one item from each pair to ensure unidimensionality and instrument quality.

He further emphasizes that to develop a high-quality instrument, one of the items in each interdependent pair should be removed. The selection of items to retain should be based on the **MNSQ values**, where values closest to **1.00** indicate better fit and measurement precision [9].

Based on Table 7, ten pairs of items exhibited high correlation values, indicating local dependence. Specifically, there was a correlation value of **0.92** between items *pgb_insentif4* and *pgb_insentif5*; **0.90** between *pgb_insetpembel2* and *pgb_insetpembel3*; **0.88** between *pgb_insetpembel1* and *pgb_insetpembel2*, and also between *pgb_mmt4* and *pgb_mmt5*, and between *pgb1* and *pgb2*; **0.87** between *pgb_selaras4* and *pgb_selaras5*; **0.86** between *pgb_insetpembel4* and *pgb_insetpembel5* and between *pgb_nampak2* and *pgb_nampak3*; **0.85** between *pgb_lindungi4* and *pgb_lindungi5*; and **0.84** between *pgb_nampak1* and *pgb_nampak2*.

These high correlation values suggest that the paired items share the same measurement intention or overlap in content, potentially measuring the same latent dimension. Therefore, careful consideration must be given to such items, and one item from each correlated pair should be removed to maintain instrument unidimensionality and avoid redundancy.

Referring to the MNSQ values for the involved items, the items that needed to be removed were: *pgb_insentif5*, *pgb_insetpembel3*, *pgb_insetpembel1*, *pgb_mmt4*, *pgb1*, *pgb2*, *pgb_selaras5*, *pgb_insetpembel5*, *pgb_nampak3*, *pgb_lindungi5*, and *pgb_nampak2*. The selection of items to be eliminated was also aligned with the items identified in the previous analysis as having negative *PTMEA CORR* values.

Meanwhile, items with *MNSQ* values closest to **1.00** were retained, as they demonstrated the best fit to the construct being measured. These included: *pgb_insentif4*, *pgb_insetpembel2*, *pgb_mmt5*, *pgb_insetpembel4*, *pgb_nampak2*, *pgb_lindungi4*, and *pgb_nampak1*.

Table 7
Largest standardized residual correlation on items

Correlation	***Entry Number***	***MNSQ Outfit***	***Result***	***Entry Number***	***MNSQ outfit***	***Result***
.92	pgb_insentif4	.81	Remain	pgb_insentif5	.70	Eliminate
.90	pgb_insetpembel2	.59	Remain	pgb_insetpembel3	.47	Eliminate
.88	pgb_insetpembel1	.46	Eliminate	pgb_insetpembel2	.59	Remain
.88	pgb_mmt4	.77	Eliminate	pgb_mmt5	1.07	Remain
.88	pgb1	2.25	Eliminate	pgb_2	2.39	Eliminate
.87	pgb_selaras4	.38	Remain	pgb_selaras5	.33	Eliminate
.86	pgb_insetpembel4	.66	Remain	pgb_insetpembel5	.34	Eliminate
.86	pgb_nampak2	.92	Remain	pgb_nampak3	.73	Eliminate
.85	pgb_lindungi4	.50	Remain	pgb_lindungi5	.44	Eliminate
.84	pgb_nampak1	1.33	Remain	pgb_nampak2	1.16	Eliminate

4. Discussion

Following data analysis, each item was reviewed in accordance with the established statistical indices and criteria required to achieve the validity and reliability standards of the instrument, as defined by the Rasch Measurement Model. The process of item removal and refinement was conducted through consultation with field experts and by considering their professional evaluations and judgments.

Based on the findings from the pilot study, ten (10) items did not meet the analytical requirements and were therefore removed from the instrument. Meanwhile, thirty-one (31) items were refined and retained based on their relevance to the study’s objectives and contextual importance. The overall summary of retained and removed items is presented in **Table 8** below.

Table 8
Summary of retained and removed items

No.	Construct	Items Retained	Number Retained	Items Removed	Number Removed
1.	B. The three pillars of AI	pgb3, pgb4, pgb5, pgb6, pgb7, pgb8, pgb9	7	Pgb1, Pgb2	2
2.	C. Defining the school's mission with AI	pgb_mmt1, pgb_mmt2, pgb_mmt3, pgb_mmt4, pgb_mrng1, pgb_mrng2, Pgb_mrng3, pgb_mrng4, pgb_mrng5	9	pgb_mmt5	1

3.	D. Administer AI-assisted instructional programs	pgb_selia1, pgb_selia2, pgb_selia3, pgb_selia4, pgb_selia5, pgb_selaras1, pgb_selaras2, pgb_selaras3, pgb_pantau1, pgb_pantau2, pgb_pantau3, pgb_pantau4, pgb_pantau5,	13	pgb_selaras4 pgb_selaras5	2
4.	E. Creating a positive AI-assisted climate	pgb_lindungi1, pgb_lindungi2, pgb_lindungi3, pgb_lindungi4, pgb_nampak1, pgb_nampak2, pgb_nampak4, pgb_nampak5, pgb_insentif1, pgb_insentif2, pgb_insentif3, pgb_insentif4, pgb_galperprog1, pgb_galperprog2, pgb_galperprog3, pgb_galperprog4, pgb_galperprog5, pgb_insetpembel2, pgb_insetpembel3, pgb_insetpembel4	20	pgb_lindungi5, pgb_nampak3, pgb_insentif5, pgb_insetpembel1, pgb_insetpembel5	5
		SUM	**49**		**10**

The final instrument, therefore, consists of **49 validated items** distributed across four main constructs, after the elimination of **10 misfitting items**. This refinement process ensured that only items with strong construct alignment, acceptable fit indices, and positive item polarity were retained. The revised instrument demonstrates high internal consistency and construct validity, confirming its suitability for measuring the application of Artificial Intelligence (AI) within instructional leadership practices.

These findings are consistent with recommendations by [8] and [9], who assert that instrument refinement through item deletion is necessary to enhance construct clarity and measurement accuracy. The retained items effectively represent the intended domains, reinforcing the instrument's reliability and validity for future large-scale applications in educational leadership research.

5. Conclusion

Based on the findings of this pilot study, it can be concluded that **validity and reliability** are essential aspects that must be emphasized when developing new research instruments. The results of the Rasch analysis revealed that ten (10) items were removed due to concerns regarding their validity and reliability.

The overall analysis indicates that the refined instrument demonstrates strong psychometric properties and can be confidently used as a reliable tool for measuring Artificial Intelligence (AI) integration within instructional leadership practices.

The implications of this analysis are significant, as the validated instrument provides researchers and practitioners with a robust framework for assessing leadership competencies that integrate AI elements. This, in turn, contributes to enhancing instructional quality and supporting academic improvement among students.

The integration of Artificial Intelligence (AI) into instructional leadership represents a key dimension of innovation in educational management. Therefore, further research should be conducted using a larger sample to confirm the instrument's generalizability and to explore its application in promoting digital transformation and effective leadership practices in schools.

Acknowledgement

This research was no funded by any grant.

References

[1] Chen, Lijia, Pingping Chen, and Zhijian Lin. "Artificial intelligence in education: A review." *IEEE access* 8 (2020): 75264-75278. https://doi.org/10.1109/ACCESS.2020.2988510

[2] Fullan, Michael, Cecilia Azorín, Alma Harris, and Michelle Jones. "Artificial intelligence and school leadership: challenges, opportunities and implications." *School Leadership & Management* 44, no. 4 (2024): 339-346.

[3] A.Harris & M.Jones. (2020). COVID 19 – school leadership in disruptive times. School Leadership & Management . https://doi.org/10.1080/13632434.2020.1811479

[4] García-Peñalvo, Francisco José. "The perception of Artificial Intelligence in educational contexts after the launch of ChatGPT: Disruption or Panic?." (2023). https://doi.org/10.14201/eks.31279

[5] Wahyu dan Yudi. (2024). Menyepadukan kepimpinan pengajaran dan alat pengajaran AI dalam kelas; Analisis untuk Dimensi Etika. *Prosiding IPersidangan Kebangsaan mengenai Agama, Sains dan Pendidikan* (2024) 3, 277-283

[6] Sheninger, E. (2019). *Digital Leadership; Changing Paradigms For Changing Times*. California. Corwin-SAGE Company. https://doi.org/10.21125/inted.2019.2528

[7] Cooper, D. and Schindler, P. (2011) Business Research Methods. 11th Edition, McGraw Hill, Boston.

[8] Bond, Trevor G, & Fox, Christine M. (2007). Applying the Rasch Model: Fundamental Measurement in the Human Sciences

[9] Linacre, J. M. . (2019). A user's guide to WINDTEPS Rasch-model computer programs. Chicago, Illinois: MESA Press.