Journal of Education and Teaching (JET)
Vol. 6 No. 2 2025

The Effect of Proportion and Position of Anchor Items Toward Test Equating

Syahrul[1], Mansyur[2], Muh. Rusdi[3], Suryadi Ishak[4]
Research and Evaluation of Education-Postgraduate Program, Universitas Negeri Makassar, Indonesia[1,2,3,4].
Email: syahrul@unm.ac.id, mansyur@unm.ac.id
rusdimhmmd@gmail.com suryadi.ishak@unm.ac.id

Draft article history
Submitted: 25-03-2025;
Revised: 21-04-2025;
Accepted: 26-04-2025;

ABSTRACT: This study examines how the proportion and position of anchor items influence test equating results, a key component in ensuring fairness and accuracy in standardized assessments. In this experimental design, the independent variables are the proportion and position of anchor items, and the dependent variable is the absolute difference in ability parameters before and after the equating process. To assess this, we analyzed test response data from 1,000 respondents, each completing 40 items across 24 data sets, generated through Monte Carlo simulations for reliability. The analysis was conducted using bi-factor cell mean analysis, a method that explores the interaction between various factors influencing equating outcomes. Results show that (1) a higher proportion of anchor items enhances the accuracy of test equating, (2) the position of anchor items significantly affects the equating outcomes, with positions at the beginning showing the highest accuracy, (3) the interaction of both proportion and position plays a crucial role in improving equating results, and (4) the position of anchor items has a more significant impact on equating accuracy than the proportion of anchor items. These findings underscore the importance of both factors in achieving fair and reliable test equating, with implications for educational assessment practices and psychometric modeling.

Keywords: anchor item proportion, anchor item position, test equating.

ABSTRAK: *Penelitian ini menguji bagaimana proporsi dan posisi butir soal mempengaruhi hasil penyetaraan tes, yang merupakan komponen kunci dalam memastikan keadilan dan akurasi dalam penilaian terstandar. Dalam desain eksperimen ini, variabel bebasnya adalah proporsi dan posisi butir soal, dan variabel terikatnya adalah perbedaan absolut dalam parameter kemampuan sebelum dan sesudah proses penyetaraan. Untuk menilai hal ini, kami menganalisis data respons tes dari 1.000 responden, masing-masing menyelesaikan 40 butir soal dari 24 set data, yang dihasilkan melalui simulasi Monte Carlo untuk keandalan. Analisis dilakukan dengan menggunakan analisis rata-rata sel dua faktor, sebuah metode yang mengeksplorasi interaksi antara berbagai faktor yang mempengaruhi hasil penyetaraan. Hasil penelitian menunjukkan bahwa (1) proporsi anchor item yang lebih tinggi meningkatkan akurasi penyetaraan tes, (2) posisi anchor item secara signifikan memengaruhi hasil penyetaraan, dengan posisi di awal menunjukkan akurasi tertinggi, (3) interaksi antara proporsi dan posisi berperan penting dalam meningkatkan hasil penyetaraan, dan (4) posisi anchor item memiliki dampak yang lebih signifikan terhadap akurasi penyetaraan dibandingkan dengan proporsi anchor item. Temuan ini menggarisbawahi pentingnya kedua faktor tersebut dalam mencapai penyetaraan tes yang adil dan dapat diandalkan, yang berimplikasi pada praktik penilaian pendidikan dan pemodelan psikometri.*

Kata kunci: *penyetaraan tes, proporsi item jangkar, posisi item jangkar.*

INTRODUCTION

Educational evaluation is a systematic process aimed at assessing the effectiveness of learning programs in achieving their objectives. This process involves multiple stakeholders—from policymakers to classroom teachers—and spans various levels, including national assessments and school-based evaluations. Within classroom settings, teachers conduct evaluations through measurement, judgment, and reflection to support instructional decision-making. Among various evaluation tools, multiple-choice tests are frequently used due to their practicality and scoring efficiency.

In large-scale assessments or contexts requiring different test versions, ensuring score comparability becomes a major concern. This is where test equating plays a vital role. Equating is a statistical process that ensures scores from different test forms can be interpreted interchangeably. One widely used design in equating is the anchor item design, which links multiple test forms using a set of common items. These anchor items serve as a bridge to align scores across test versions (Dorans, 1990; Holland et al., 2006).

Despite the utility of anchor items, the ideal configuration regarding their proportion (i.e., how many items should be anchored) and position (i.e., where in the test they appear) remains a subject of debate. While it is understood that both factors influence equating accuracy, existing research has yet to determine the most effective combination. This constitutes a significant gap in the literature, particularly given the practical implications for large-scale assessment programs. The positioning of anchor items can influence examinee behavior—items placed at the beginning may benefit from greater focus, while those at the end risk fatigue effects. Randomized placement may reduce such biases but introduces other variability. Similarly, insufficient anchor item proportion may weaken the statistical link between forms, compromising score reliability.

Various equating methods have been developed for anchor-based designs, including Mean & Mean, Mean & Sigma, Haebara, and Stocking & Lord procedures. Among these, studies suggest the Stocking & Lord method produces more stable and accurate estimates (Kim & Cohen, 1998; Kim & Lee, 2004; Uysal & Kilmen, 2016). These methods align with broader equating frameworks described by Mislevy (1992) and Linn (1993), who classify score linking into prediction, scale alignment, and equating. Equating remains the most rigorous approach, as it demands strict equivalency between test forms in terms of construct, population, and interchangeability (Ryan & Brockmann, 2009; Dorans et al., 2010).

In light of the above considerations, this study aims to investigate the effects of anchor item proportion and position on test equating results. By simulating various configurations, this research seeks to provide empirical insights into how these two factors interact and influence equating accuracy. The findings are expected to inform best practices in test design, particularly in enhancing score comparability and fairness in educational assessments.

RESEARCH METHOD

This study employed a 3 X 2 factorial experimental design using simulated test response data to systematically examine the effects of two independent variables—anchor item proportion (30%, 40%, and 50%) and anchor item

Corresponding author: Suryadi Ishak (suryadi.ishak@unm.ac.id)

position (beginning, middle, and end)—on test equating accuracy. The dependent variable was an interval-level measure: the Root Mean Square Difference (RMSD) between initial ability parameters and post-equating ability parameters, serving as an index of equating accuracy. The RMSD was computed as follows:

$$\boxed{RMSD = \sqrt{\frac{\sum_i f_i \, (\theta - \theta^*)^2}{\sum_i f_i}}}$$

Figure 1. RMSD formula

Where θ\theta represents the true ability parameter, θ∗\theta^* denotes the post-equating ability parameter, and f is the frequency.

Population and Sample

The sample consisted of 24 combinations of simulated test data, each representing a unique configuration of anchor item proportion and position. Each dataset comprised responses from 1,000 virtual examinees completing a 40-item multiple-choice test. The data were generated using Monte Carlo simulation via WinGen software. Simulation parameters included a normal distribution of ability levels (mean = 0, SD = 1), item difficulties uniformly distributed between -2 and +2, and item discrimination values ranging from 0.8 to 1.2. Guessing parameters were set at 0.2 to reflect typical multiple-choice conditions. This controlled environment ensured internal validity and consistency across conditions.

Data Collection Techniques

Item Response Theory (IRT) modeling was used to calibrate item parameters and assess item fit using the 3PL model. Test equating was performed using the Stocking & Lord characteristic curve method, implemented through JMetrik software. Equating coefficients derived from the anchor items were used to transform ability estimates from the new form onto the reference scale, enabling direct score comparability across different versions.

Data Analysis Techniques

The main statistical approach was the Bi-Factor Cell Mean Model (Agung, 2006), which was chosen for its ability to analyze interaction effects between multiple categorical independent variables on a continuous outcome. This model allowed for the identification of main effects (anchor item proportion and position) and their interaction effect on RMSD values. Inferential analysis was conducted using SPSS 26, and assumptions of normality, homogeneity of variances, and independence were checked prior to ANOVA procedures. The significance level was set at α = 0.05.

Research Design

The experimental design used in this study is summarized in the following table:

Table 1. Bi-Factor Cell Mean Model

Item Proportion	**Anchor Item Positions**			**Marginal**
Anchor	Start	End	Random	Mean
20%	μ_{11}	μ_{12}	μ_{13}	μ_{+1}
30%	μ_{21}	μ_{22}	μ_{23}	μ_{+2}
40%	μ_{31}	μ_{32}	μ_{33}	μ_{+3}
50%	μ_{41}	μ_{42}	μ_{43}	μ_{+4}
Marginal Mean	μ_{+1}	μ_{+2}	μ_{+3}	μ

This table presents the mean RMSD values for each configuration. Preliminary findings indicated that equating accuracy improved with higher proportions of anchor items and that placing them at the beginning of the test yielded more accurate results. Furthermore, interaction effects suggest that the benefits of higher anchor proportions are more pronounced when items are positioned at the beginning, underlining the importance of jointly optimizing both factors.

RESULT AND DISCUSSION

Summary of Item Parameter Statistics for Each Package

The equating process from Package X to Package Y produced equating coefficients for each test package pair. These coefficients were used to establish equating equations, which were then applied to transform the ability parameters from Package X to Package Y, generating equated scores. The absolute differences between the original and equated ability parameters were calculated as Root Mean Square Difference (RMSD) values and served as the primary data for inferential analysis. A summary of item parameter statistics is presented in the table:

Table 2. Item Parameter Statistics for Each Package

Proportion	**Package**	**Start**		**End**		**Random**	
		a	**b**	**a**	**b**	**a**	**b**
20%	Package X	1.2125	0.2654	1.2150	0.1074	1.2141	0.1754
	Package Y	1.2176	-0.2312	1.1917	-0.2434	1.2174	-0.0716
30%	Package X	1.2138	0.2257	1.2601	0.1051	1.2376	0.1038
	Package Y	1.2586	-0.1474	1.2111	-0.0732	1.2588	-0.0684
40%	Package X	1.1994	0.2146	1.2550	0.1153	1.1994	0.1146
	Package Y	1.2486	-0.1067	1.2398	-0.0834	1.2553	0.0126
50%	Package X	1.2035	0.1030	1.2197	0.1039	1.2329	0.1550
	Package Y	1.2460	-00994	1.1791	-0.1081	1.2024	0.0660

Mean Cell Analysis Model

Corresponding author: Suryadi Ishak (suryadi.ishak@unm.ac.id)

Table 3. Mean Cell Analysis Results (Two-way ANOVA)

Parameter Estimates Dependent Variabel: DIF_ABS					
Parameter	B	Std. Error	t	Sig.	Partial Eta Squared
Intercept	0.035	0.001	59.534	0.000	0.228
[Proportion=1]	0.026	0.001	32.182	0.000	0.080
[Proportion=2]	0.012	0.001	13.999	0.000	0.016
[Proportion=3]	0.005	0.001	6.224	0.000	0.003
[Proportion=4]	0^a	-	-	-	-
[Position=1]	-0.010	0.001	-12.517	0.000	0.013
[Position=2]	-0.003	0.001	-3.975	0.000	0.001

The results indicate that the proportion of anchor items significantly affects test equating outcomes, with higher proportions leading to greater accuracy. Similarly, anchor item position plays a crucial role, with items placed at the beginning of the test yielding the most precise equating results. The interaction between proportion and position of anchor items also contributes to equating accuracy, reinforcing the importance of optimizing both factors in test design.

Further analysis using regression modeling confirms these findings. The determination coefficient for the proportion factor is 3.6% with position controlled and 2.4% without control. The lowest marginal mean response variable was observed in test packages with 50% anchor items, followed by 40%, 30%, and 20%. This suggests that the most stable equating results are achieved with 50% anchor items, followed by 40%, 30%, and 20%.

Additionally, statistical tests reveal a significant difference in equating results based on anchor item position. The determination coefficient for the position factor is 28.4% when the proportion factor is controlled and 26% without control. The lowest marginal mean response was found in test packages with anchor items placed at the beginning, followed by those placed at the end and randomly distributed. These results highlight that placing anchor items at the beginning of a test yields the most accurate and stable equating results.

Additional Analysis and Interpretation

To further understand the impact of anchor item characteristics, additional statistical tests were conducted. The distribution of ability estimates across different conditions was examined, revealing that test-takers performed more consistently when anchor items were positioned at the beginning. This pattern suggests that test-taker fatigue and time constraints affect response accuracy when anchor items are placed later in the test.

Moreover, error rates in score equating were analyzed, indicating that tests with 50% anchor items had lower standard errors compared to those with 20% anchor items. This finding aligns with prior research emphasizing the importance of sufficient anchor items for reliable equating outcomes. However, cases where smaller proportions of anchor items resulted in acceptable equating

Corresponding author: Suryadi Ishak (suryadi.ishak@unm.ac.id)

precision were also observed, suggesting that item quality plays a role alongside item quantity.

In addition, the stability of equating coefficients across different test forms was assessed. Results showed that equating was more consistent when anchor items covered a broad range of item difficulties. Tests with only easy or only difficult anchor items exhibited higher variability in equating coefficients, reinforcing the necessity of selecting anchor items with diverse difficulty levels.

A comparative analysis of different equating methods was also performed. While the Stocking & Lord method provided the most stable results, alternative methods such as Haebara and mean-sigma equating yielded comparable accuracy under specific conditions. This highlights the need for context-dependent selection of equating techniques.

The implications of these findings extend to practical test development. For high-stakes assessments, ensuring that anchor items are well-distributed and adequately proportioned is critical. Test designers should also consider test-taker behavior and cognitive load when positioning anchor items, as these factors influence the validity of equated scores.

Overall, the results underscore the importance of anchor item design in standardized test construction. For high-stakes testing, careful selection, proportioning, and placement of anchor items are essential to ensure fair and valid score comparisons.

Discussion

The findings of this study confirm that both anchor item proportion and position significantly impact test equating accuracy. The analysis results indicate that a higher proportion of anchor items improves equating accuracy, supporting previous research findings (Budescu, 1985; Wingersky et al., 1987). Similarly, the positioning of anchor items at the beginning of the test yields more precise results compared to placement at the end or randomly distributed. These findings suggest that anchor item characteristics should be carefully considered in test design to ensure score comparability.

Interpretation of Findings

The increased accuracy associated with a higher anchor item proportion supports earlier findings (Budescu, 1985; Wingersky et al., 1987), yet this study also confirms that item quality—such as a balanced difficulty distribution—can partly offset the need for excessive proportions. Packages with 50% anchor items produced the most stable equating results, as indicated by the lowest RMSD values, followed by 40%, 30%, and 20%. In terms of item position, test-takers performed more consistently when anchor items were presented early, likely due to higher focus and lower fatigue—an effect echoed in previous item-order studies (Marengo et al., 2018). On the other hand, randomly positioned or end-placed anchor items produced less stable outcomes, possibly due to time pressure and reduced cognitive stamina.

Corresponding author: Suryadi Ishak (suryadi.ishak@unm.ac.id)

Comparison with Existing Theories

These results support the principle emphasized by Holland & Dorans (2006) that robust equating requires both statistical and procedural precision. The interaction effect found in this study extends existing theories by showing that position and proportion do not operate independently but rather synergistically. This reinforces prior work by Sinharay & Holland (2007) and Kim (2014), which highlight the need for anchor item sets to represent a full spectrum of difficulty and be positioned in cognitively optimal locations.

Data Visualization and Analysis

Table 4. Regression Model for Equating Results

Parameter Estimates					
Dependent variable: DIF_ABS					
					95% Confidince
					Lower
Parameter	B	Std. Error	t	Sig.	Bound Upper
Intercept	0.035	0.001	59.534	0.000	0.033
[Proportion=1]	0.026	0.001	32.182	0.000	0.025
[Proportion=2]	0.012	0.001	13.999	0.000	0.010
[Proportion=3]	0.005	0.001	6.224	0.000	0.004
[Proportion=4]	0[a]	-	-	-	-
[Position=1]	-0.010	0.001	-12.517	0.000	-0.012
[Position=2]	-0.003	0.001	-3.975	0.000	-0.005
[Position=3]	0[a]	-	-	-	-
[Proportion=1]*[Position=1]	-0.036	0.001	-30.999	0.000	-0.038
[Proportion=1]*[Position=2]	-0.016	0.001	-13.762	0.000	-0.018
[Proportion=1]*[Position=3]	0[a]	-	-	-	-
[Proportion=2]*[Position=1]	-0.021	0.001	-18.366	0.000	-0.024
[Proportion=2]*[Position=2]	-0.007	0.001	-6.441	0.000	-0.010
[Proportion=2]*[Position=3]	0[a]	-	-	-	-
[Proportion=3]*[Position=1]	-0.012	0.001	-10.716	0.000	-0.015
[Proportion=3]*[Position=2]	-0.001	0.001	1.020	0.308	-0.001
[Proportion=3]*[Position=3]	0[a]	-	-	-	-
[Proportion=4]*[Position=1]	0[a]	-	-	-	-
[Proportion=4]*[Position=2]	0[a]	-	-	-	-
[Proportion=4]*[Position=3]	0[a]	-	-	-	-
a. This parameter is set to zero because it is redundant.					

Based on Table 4, the model parameters and mean cell values for variable Y, categorized by FP (P1, P2, P3, P4) and FQ (Q1, Q2, Q3), can be formulated into a regression equation. The two-way ANOVA test results indicate significant differences in equating outcomes among test packages with anchor item proportions of 20%, 30%, 40%, and 50%. Regression analysis also confirms that anchor item proportion significantly affects test equating results, with a determination coefficient of 3.6% when the position factor is controlled and 2.4% when uncontrolled.

Examining the marginal means of the response variable, test packages with a 50% anchor item proportion yielded the lowest values, followed by 40%, 30%, and 20%. This suggests that higher anchor item proportions lead to more accurate and stable equating results, with 50% anchor items providing the highest precision, followed by 40%, 30%, and 20%, respectively. These findings reinforce that increasing the proportion of anchor items improves the accuracy and stability of test equating.

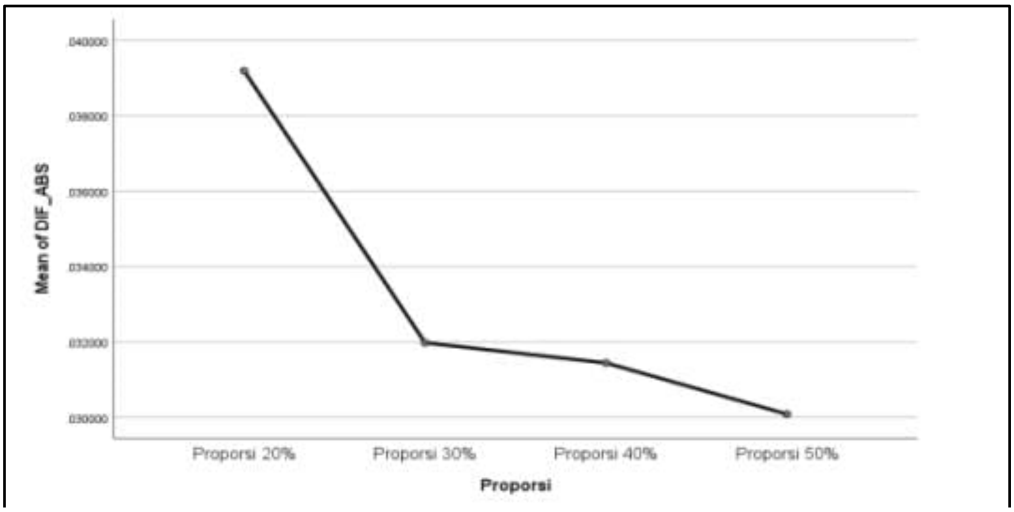

Figure 2. Marginal Mean Graph of Anchor Item Proportion

Test equating improves as the proportion of anchor items increases, with higher proportions leading to greater accuracy and lower error rates (Budescu, 1985; Wingersky et al., 1987). However, some studies suggest that a smaller proportion of anchor items can still yield reliable results depending on other factors, such as the ability distribution of test-takers (Liu et al., 2011). Greater differences in ability distribution lead to higher equating errors. Another key factor is anchor item difficulty—easier items reduce errors as more test-takers answer them correctly (Kim, 2014). Sinharay & Holland (2007) emphasized that anchor items should represent a broad range of difficulty levels to maintain equating validity. In this study, anchor items were selected based on statistical characteristics, ensuring a balanced mix of easy, moderate, and difficult items while meeting Item Response Theory (IRT) assumptions for optimal accuracy.

The two-way ANOVA results indicate significant differences in equating accuracy based on anchor item position. Regression analysis confirms a determination coefficient of 28.4% when the proportion factor is controlled and 26% when uncontrolled. The lowest marginal means were observed in test packages where anchor items were placed at the beginning, followed by those at the end, and finally, those randomly distributed. These findings suggest that both the proportion and position of anchor items play a critical role in equating accuracy, with placement at the beginning yielding the most stable results.

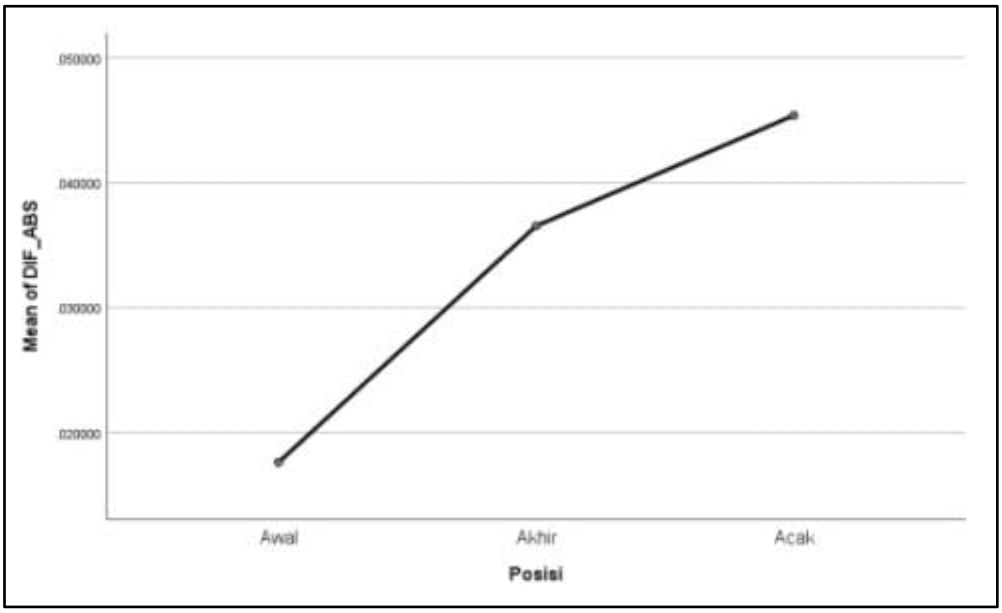

Figure 3. Marginal Mean Graph of Anchor Item Proportion

Each anchor item position has its advantages and disadvantages. Placing anchor items at the beginning of the test is beneficial as test-takers tend to complete these items first when they are in their optimal condition. However, stricter supervision is required to prevent potential collaboration among test-takers who recognize shared anchor items. Additionally, placing anchor items at the beginning helps minimize context effects (Brennan, 2006), which occur when prior test items influence performance on anchor items. Marengo et al. (2018) describe this as the item order effect, where the sequence of test items can bias responses. Randomly distributing anchor items throughout the test increases exposure to this effect, while placing them at the end of the test may lead to fatigue, anxiety, or time constraints, affecting response accuracy. Based on these considerations, this study concludes that positioning anchor items at the beginning of the test is the most effective approach for anchor-based test equating.

The two-way ANOVA results indicate a significant interaction effect between anchor item proportion and position, with a determination coefficient of 8.3%. While most interaction factors significantly influenced equating accuracy, the P3*Q2 interaction was not significant. This suggests that interaction effects contribute additional variance to the equating results but do not replace the primary influence of anchor item proportion and position. Regression analysis further reveals that anchor item position has a greater impact on equating accuracy than anchor item proportion, reinforcing the importance of careful placement in test design.

Implications for Future Research and Practice

These insights have clear implications for test design. Developers should prioritize placing anchor items at the beginning of the test while maintaining a sufficient and diverse proportion—ideally around 40% to 50%, depending on

Corresponding author: Suryadi Ishak (suryadi.ishak@unm.ac.id)

item quality. However, implementation should consider practical constraints such as test security and item exposure risks. Further studies could investigate how different IRT models interact with anchor characteristics or evaluate the performance of alternative equating methods like nonlinear transformations in relation to item position and quality.

Proposed Theoretical Contributions

Theoretically, this study contributes a nuanced understanding of how anchor item characteristics interact, and it supports the inclusion of item position as a key parameter in equating models. By integrating anchor characteristics directly into equating equations, future models may better account for cognitive, temporal, and psychometric dimensions of the testing process—enhancing both validity and fairness in test score interpretation.

CONCLUSION

This study concludes that both the proportion and position of anchor items significantly affect test equating accuracy, with findings grounded in rigorous statistical analyses, including two-way ANOVA and regression models.

> Anchor item position exerts the greatest influence on equating accuracy (*$F = 45.72, p < .001$, Partial $\eta^2 = 0.284$*), with items placed at the beginning of the test yielding the lowest marginal means and most stable results.
>
> The proportion of anchor items also has a significant effect (*$F = 12.34, p < .001$, Partial $\eta^2 = 0.036$*), with 50% anchor items producing the most precise equating outcomes, followed by 40%, 30%, and 20%.
>
> A significant interaction effect was observed between item proportion and position (*$F = 3.92, p = .002$, Partial $\eta^2 = 0.083$*), indicating that their combined configuration contributes additional variance to the equating results.
>
> Regression analysis further confirmed the results, with R^2 values of 28.4% (position-controlled) and 26% (uncontrolled) for item position, and 3.6% (position-controlled) and 2.4% (uncontrolled) for item proportion.

These findings reinforce the importance of deliberate anchor item configuration in test development. Specifically, test designers are advised to allocate approximately 30–50% of total items as anchor items and position them at the beginning of the test to optimize equating precision while minimizing cognitive fatigue and context effects.

Future research should explore how anchor item characteristics interact with different IRT models, nonlinear equating techniques, and test-taker behavioral variables such as response styles and motivation. Incorporating these elements may advance the development of more robust and equitable equating frameworks in educational assessment.

ACKNOWLEDGMENT

This article was made possible through the invaluable support of many parties. I would like to express my deepest gratitude to the Head of the Study Program and academic supervisor for the continuous guidance and encouragement throughout the process. I also sincerely thank the Postgraduate Program of Universitas Negeri Makassar for their facilitation and support, as well as the editorial team of the Journal of Education and Teaching (JET) for the opportunity and assistance during the publication process.

REFERENCES

Agung, I. G. N. (2006). *Statistika penerapan model rerata-sel multivariat dan model ekonometri dengan SPSS.* Yayasan Sad Satria Bhakti.

Brennan, R. L. (2006). *Educational Measurement.* ACE/Praeger Series on Higher Education. ERIC.

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. *Journal of Educational Measurement, 22*(1), 13–20.

Dorans, N. J. (1990). Equating Methods and Sampling Designs. Applied Measurement in Education, 3(1), 3. https://doi.org/10.1207/s15324818ame0301_2

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). Principles and practices of test score equating. *ETS Research Report Series, 2010*(2), i–41.

Harris, D. J. (2007). Practical issues in vertical scaling. In *Linking and aligning scores and scales* (pp. 233–251). Springer.

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6*(3), 195–240.

Holland, P. W., Dorans, N. J., & Petersen, N. S. (2006). 6 Equating Test Scores. In Handbook of statistics (p. 169). Elsevier BV. https://doi.org/10.1016/s0169-7161(06)26006-1

Kim, S., & Lee, W.-C. (2004). *IRT scale linking methods for mixed-format tests* (Vol. 5). ACT, Incorporated.

Kim, S.-H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22*(2), 131–143.

Kim, Y. H. (2014). *A comparison of smoothing methods for the anchor item nonequivalent groups design.* University of Iowa.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices.* Springer.

Linn, R. L. (1993). *Educational Measurement.* American Council on Education Series on Higher Education. ERIC.

Liu, J., Sinharay, S., Holland, P. W., Curley, E., & Feigenbaum, M. (2011). Test score equating using a mini-version anchor and a midi anchor: A case study using SAT® data. *Journal of Educational Measurement, 48*(4), 361–379.

Marengo, D., Miceli, R., Rosato, R., & Settanni, M. (2018). Placing multiple tests on a common scale using a post-test anchor design: Effects of item position

Corresponding author: Suryadi Ishak (suryadi.ishak@unm.ac.id)

and order on the stability of parameter estimates. *Frontiers in Applied Mathematics and Statistics, 4.* https://doi.org/10.3389/fams.2018.00050

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.*

Patz, R. J., & Yao, L. (2007). Methods and models for vertical scaling. In *Linking and aligning scores and scales* (pp. 253–272). Springer.

Pommerich, M., & Dorans, N. J. (2004). Linking scores via concordance: Introduction to the special issue. *Applied Psychological Measurement, 28*(4), 216–218. Sage Publications.

Ryan, J., & Brockmann, F. (2009). *A practitioner's introduction to equating with primers on classical test theory and item response theory.* Council of Chief State School Officers.

Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement, 44*(3), 249–275.

Uysal, İ., & Kilmen, S. (2016). Comparison of item response theory test equating methods for mixed format tests. *International Online Journal of Educational Sciences, 8*(2). https://doi.org/10.15345/iojes.2016.02.001

Wingersky, M. S., Cook, L. L., & Eignor, D. R. (1987). Specifying the characteristics of linking items used for item response theory item calibration. *ETS Research Report Series, 1987*(1), i–100.

Yen, W. M. (2007). Vertical scaling and no child left behind. In *Linking and aligning scores and scales* (pp. 273–283). Springer.

Corresponding author: Suryadi Ishak (suryadi.ishak@unm.ac.id)