JOURNAL OF EDUCATION FOR SUSTAINABLE INNOVATION | ISSN 3025-1052
Vol. 2, No. 2, 2024, 107-120
https://doi.org/10.56916/jesi.v2i2.934

Assessing Item Reliability, Differential Item Functioning, and Wright Map Analysis of the GSP122 Test at a Public University in Nigeria

Abubakar Rabiu Uba*
Department of Education, Sule Lamido University, Kafun Hausa, Nigeria

Ahmad Zamri Khairani
School of Educational Studies, Universiti Sains Malaysia, 11800 Penang, Malaysia

***Corresponding Author:** abubakarbabura@student.usm.my

Keywords
Item reliability
Differential item functioning
Wright map
GSP122 test
ICT and Rasch model

Article History
Received 2024-08-10
Accepted 2024-11-05

Abstract
This study examines the psychometric properties of the GSP122 test, an Information and Communication Technology (ICT) knowledge assessment administered at a public university in Nigeria. Despite its importance in evaluating students' ICT competencies, no prior attempt has been made to investigate the test's psychometric qualities. The research focuses on three key aspects: item reliability, Differential Item Functioning (DIF), and Wright Map analysis. The study employs Rasch analysis to evaluate these properties. A sample of 600 GSP122 test scripts was randomly selected from undergraduate students across various departments to ensure a representative assessment. Findings reveal that the test possesses strong item reliability, indicating consistency in measuring the intended construct. Furthermore, all items are found to be DIF-free, suggesting fairness across different subgroups of test-takers. The Wright Map analysis, however, indicates that the test doesn't accurately target the abilities of students at the extreme ends and bottom of the proficiency spectrum. Specifically, some items are identified as too difficult and too easy relative to the students' ability levels. These results provide valuable insights into the GSP122 test's strengths and areas for improvement. While the test demonstrates robustness in reliability and fairness, adjustments in item difficulty could enhance its effectiveness in assessing students across all proficiency levels. This comprehensive analysis contributes to the validation of the GSP122 test and offers a foundation for evidence-based refinements in ICT assessment practices within the Nigerian higher education context.

INTRODUCTION

In the rapidly evolving landscape of higher education, proficiency in Information and Communication Technology (ICT) has become a crucial competency for students across all disciplines (Smith et al., 2020). As universities worldwide integrate ICT into their curricula, the need for reliable and valid assessment tools to measure students' ICT knowledge has grown increasingly important (Johnson & Lee, 2019). In Nigeria, where technological advancement is pivotal for national development, ensuring that university graduates possess adequate ICT skills is of paramount importance (Adebayo, 2021).

The GSP122 test is a compulsory general examination for all students at Sule Lamido University Kafin Hausa, consisting of 60 multiple-choice questions (Faruk, 2020). Acquiring ICT knowledge through the GSP 122 test at Sule Lamido University cannot be overemphasized; this knowledge facilitates undergraduates' ability to conduct research, access online resources, and collaborate with peers across the globe. ICT knowledge enhances undergraduates' learning experiences, leading to better academic performance and outcomes. In addition, Adetimirin (2012) reported that by acquiring ICT knowledge, Nigerian undergraduates can become digitally empowered, innovative, and globally competitive, contributing to the Nigeria's economic development and growth; however, despite the importance of the GSP122 test to students at Sule Lamido University, a mass failure is being recorded, which is the sole reason for conducting this research exercise with the aim of assessing the psychometric properties of this examination.

The assessment of Information and Communication Technology (ICT) knowledge in higher education has become increasingly crucial as technology continues to reshape the global workforce and educational landscape. The present study reviews recent research on ICT assessment in higher education, with a particular focus on psychometric analysis and the application of the Rasch model. Recent studies have emphasized the importance of ICT proficiency for university students across all disciplines. Oladipo et al. (2023) found that students with higher ICT skills demonstrated better academic performance and were more likely to secure employment upon graduation. Similarly, Kumar and Singh (2022) reported that employers increasingly value ICT skills in graduates, regardless of their field of study.

In the Nigerian context, Adebayo and Eze (2024) highlighted the critical role of ICT proficiency in national development, emphasizing the need for robust assessment tools to ensure graduates meet the evolving demands of the digital economy. However, their study also revealed a lack of standardized assessment practices across Nigerian universities, underscoring the need for psychometric evaluation of existing assessment tools. Furthermore, the assessment of ICT knowledge presents unique challenges due to the rapidly evolving nature of technology. Johnson et al. (2023) conducted a comprehensive review of ICT assessment methods in higher education, finding that many institutions rely on outdated or poorly validated instruments. They stressed the importance of regularly updating and validating assessment tools to keep pace with technological advancements.

The GSP122 test, administered at a public university in Nigeria, serves as a key instrument for assessing students' ICT knowledge. This test, part of the general studies program, is designed to evaluate students' understanding and practical knowledge of ICT across various departments. Despite its widespread use and the critical role, it plays in shaping ICT education, the GSP122 test has not undergone rigorous psychometric evaluation until now. This gap in understanding the test's properties has raised questions about its effectiveness, reliability, and fairness in assessing students' ICT competencies (Oladipo & Eze, 2022). The importance of psychometric analysis in educational assessment cannot be overstated (Bond & Fox, 2015). Such analysis provides crucial insights into a

test's ability to accurately and consistently measure the intended construct, in this case, ICT knowledge. Moreover, it helps identify any potential biases or inconsistencies in the test that might advantage or disadvantage certain groups of students (Zumbo, 2007). In an era where data-driven decision-making is increasingly emphasized in education, having a clear understanding of assessment tools' psychometric properties is essential for ensuring fair and effective evaluation practices (Wilson, 2018).

This study aims to address this critical gap by conducting a comprehensive psychometric analysis of the GSP122 test. Specifically, the research focuses on three key aspects: item reliability, Differential Item Functioning (DIF), and Wright Map analysis. These components collectively provide a multifaceted view of the test's performance and its ability to accurately assess students' ICT knowledge across diverse student populations (Boone et al., 2014). Item reliability analysis is crucial for determining the consistency and stability of the test items in measuring ICT knowledge (DeVellis, 2017). High item reliability indicates that the test consistently differentiates between students with varying levels of ICT proficiency, a fundamental requirement for any effective assessment tool.

The investigation of Differential Item Functioning (DIF) is equally important, as it explores whether test items perform consistently across different subgroups of test-takers, such as gender or academic disciplines (Holland & Wainer, 2012). The absence of DIF is indicative of a fair assessment that does not inadvertently favor or disadvantage any particular group of students. Lastly, the Wright Map analysis provides a visual representation of how item difficulty aligns with student ability levels (Wilson, 2005). This analysis is particularly valuable for identifying any mismatches between the test's difficulty and the ability range of the student population, offering insights into the test's overall effectiveness and areas for potential improvement. By randomly selecting 600 test scripts from undergraduate students across various departments, this study ensures a representative sample that captures the diversity of the student population. This approach allows for a robust analysis that can yield generalizable findings about the GSP122 test's psychometric properties (Linacre, 2012).

The outcomes of this study have significant implications for ICT education and assessment practices in Nigerian higher education. By providing a detailed understanding of the GSP122 test's psychometric properties, this research lays the groundwork for evidence-based improvements in test design and administration. Furthermore, it contributes to the broader discourse on the importance of rigorous psychometric evaluation in educational assessment, particularly in the context of rapidly evolving fields like ICT (Pellegrino et al., 2016). As universities continue to adapt to the demands of the digital age, ensuring the validity and reliability of ICT assessments becomes increasingly crucial (Redecker & Johannessen, 2013). This study not only addresses a specific gap in the evaluation of the GSP122 test but also sets a precedent for the ongoing evaluation and refinement of assessment tools in higher education. The findings from this research have the potential to influence policy decisions, improve teaching practices, and ultimately enhance the quality of ICT education in Nigerian universities and beyond (Oluwatobi et al., 2019).

METHODS

Research Design

This study adopts a descriptive research design, specifically a cross-sectional survey design, to investigate the mentioned psychometric properties of the GSP122 test. This design was chosen for its ability to provide a comprehensive snapshot of the GSP122 test’s scripts, capturing the performance and characteristics of the students at a specific point in time.

Population and Sample

The study population consisted of undergraduate students from Sule Lamido University, covering six faculties: Education, Humanities, Natural and Applied Sciences, Management Sciences, Agriculture, and Information and Communication Technology (ICT). We employed a multi-phase sampling approach to select six faculties and twelve departments, followed by random sampling to identify 600 participants. Our final sample comprised 600 first-year undergraduates, with 320 females (53.8%) and 280 males (46.2%). Participants ranged in age from 19 to 30 years old. This diverse sample ensured representation across all selected faculties and departments, providing a comprehensive cross-section of the university's first-year student population.

Data Collection and Analyses

The Sule Lamido University administration provided written consent for data collection. To analyze the collected data, we employed Rasch model analysis in a specific sequence: first, we examined evidence of item reliability; second, we conducted a differential item functioning (DIF) analysis comparing male and female students; and third, we generated an item-person map (Wright map). When utilizing the Rasch Model as a measurement framework, adherence to strict assumptions is crucial. Two key assumptions of the Rasch model are particularly important: First, the data must exhibit a good fit with the model's expectations, indicating a compatible and coherent relationship. Second, the construct being measured must be unidimensional, meaning it can be represented by a single underlying trait or dimension, without additional factors influencing the measurements (Linacre, 2006). These assumptions ensure the validity and reliability of the Rasch model analysis in our study.

The Rasch measurement analysis provides a robust statistical approach for assessing measurement reliability. The item difficulty reliability index offers valuable insights into the reproducibility of results. This index is calculated by dividing the observed item variance by the true item variance. Items with higher difficulty have a greater probability of yielding genuinely high difficulty measurements compared to those with lower difficulty. The reliability of item difficulty is influenced by two key factors: sample size and the variance in item difficulty. Larger sample sizes and a wider range of item difficulties typically result in higher item difficulty reliability values, while smaller samples and narrow difficulty ranges lead to lower reliability values. Regarding the interpretation of these values, Bond and Fox (2015) consider values above 0.80 to be acceptable. However, Fischer (2007) sets a more stringent standard, classifying values exceeding 0.94 as strong indicators of reliability.

To analyze evidence supporting the construct validity of this measure, we employed Baghaei's (2008) framework. This approach emphasizes the identification of threats to construct validity, with construct underrepresentation being a key concern. Construct underrepresentation occurs when a measurement fails to capture significant aspects of the intended construct. To assess this threat, we visually examined the ordering of item difficulties and respondents' abilities in the Wright map. A continuous and smooth progression between items indicates that the scale comprehensively captures the construct without significant gaps or weaknesses. This method allows us to evaluate whether the measure adequately represents all crucial facets of the construct, ensuring a more robust and valid assessment tool.

In the Rasch Model's item-level analysis, Differential Item Functioning (DIF) is another crucial statistic. DIF analysis investigates whether specific items favor one group over another, indicating potential differences in item interpretation across groups. According to Bond & Fox (2015), a DIF contrast statistic exceeding 0.5 logits provides evidence of DIF items. For this study, we focused on examining potential differences in item perception between male and female respondents. This

approach aligns with numerous previous studies that have reported gender-based differences in item perception (Richard et al., 2023; Mustapha & Ehab, 2022; Hamad, 2021; Mohd Effendi & Ahmad Zamri, 2020). By conducting this analysis, we aim to identify any items that may be interpreted differently by male and female participants, thereby ensuring the fairness and validity of our measurement across gender groups.

RESULTS AND DISCUSSION

Results

The results showed strong evidence of measurement reliability for the GSP122 test. The item reliability index was .98 as presented in Table 1, indicating that the difficulty measure of each item is highly replicable if the GSP122 test were to be administered to a comparable sample of students. This high value suggests excellent consistency in item difficulty across potential administrations. Complementing this, we observed an item separation value of 5.61. This value indicates that the GSP122 test can effectively categorize the sample into five distinct levels of difficulty, which could be interpreted as moderately low, low, moderate, moderately high, and high. This differentiation capability demonstrates the test's ability to distinguish between various levels of student performance with precision. Both the item reliability index and the item separation value exceed the guidelines provided by Bond and Fox (2015). These results collectively provide robust evidence of the GSP122's high reliability, indicating that the measurement tool consistently produces stable and reproducible results across comparable student populations.

Table 1. Reliability statistics of Scripts - Item

	Total Score	Count	Measure	Model Error	Infit		Outfit	
					MNSQ	ZSTD	MNSQ	ZSTD
Mean	301.3	600	.00	.08	1.00	.00	1.03	.2
SD	19.9	.2	.62	.02	.02	1.0	.12	1.2
Real RMSE		.08	True SD	.62	Separation	5.61	Item Reliability	**.98**

Source: Author's work

On the other hand, the study examines the gender bias related to the items. When student groups vary in the same gender of competency, DIF can identify the items that suggest early signs of bias (Bond & Fox, 2015). The following Table 4 .4 reports the findings for this examination. In Table 4.4, the DIF analysis of both male and female were separated according to the group (Group 1 = male, group 2 = female). For Item 1, the item difficulty measure for the male participants is .68 logits. In contrast, with regards to the female group, the item difficulty measure for the same item is .76 logits. It shows that the item is relatively more difficult for the female participants compared to their male counterparts. This difference is depicted by the DIF Contrast statistics (.68 - .76 = -.08 logits). As such, Item 1 is said to be favoring male participants. Nevertheless, based on the t value statistics (t = -.57 > -1.96) or the probability value (probability = .5658 > .05), the different of .08 logits is not statistically different. In general, the DIF results show that all 60 items in the GSP122 exam are free from bias. That is, all items did not favor any gender since the DIF contrast statistics were within -0.5 to +0.5 logits (Wang, 2008). Thus, the researcher concluded that there is no evidence of gender bias shown by all the GSP122 items.

Table 2. Differential Items Functioning Analysis

Group (Male)	DIF Measure	Group (Female)	DIF Measure	DIF Contrast	t	Probability	Item number
1	.68	2	.76	-.08	-.57	.5658	1
1	.79	2	.77	.02	.17	.8666	2
1	-.31	2	-.31	.00	.00	.1.000	3
1	-.28	2	-.08	.00	-1.17	.2422	4
1	-.12	2	-.26	-.20	.83	.4093	5
1	-.17	2	-.05	14	-.74	.4587	6
1	-.21	2	-.08	-.12	-.79	.4300	7
1	-.24	2	-.07	-13	-1.03	.3023	8
1	-.20	2	-.03	-.17	-1.35	.1788	9
1	-.34	2	-.05	-.23	-1.73	.0839	10
1	-.28	2	-.28	-.29	.00	1.000	11
1	-.14	2	-.28	.00	.84	.4017	12
1	.89	2	.71	.14	1.36	.1727	13
1	-.48	2	-.26	.18	-1.31	.1891	14
1	-.23	2	-.23	-.22	.00	1.000	15
1	.70	2	.96	.00	-1.87	.0618	16
1	-.14	2	-.14	-.26	.00	1.000	17
1	-.15	2	-.08	.00	-.41	.6834	18
1	-.24	2	-.24	-.07	.00	1.000	19
1	-.19	2	-.19	.00	.00	1.000	20
1	-.05	2	-.22	.00	.98	.3289	21
1	2.64	2	2.76	.16	-.75	.4539	22
1	-.31	2	-.19	-.12	-.77	.4418	23
1	.89	2	.68	-.13	1.69	.0912	24
1	-.16	2	-.26	.21	.60	.5511	25
1	-.16	2	-.29	.10	.78	.4379	26
1	-.33	2	-.16	.13	-1.03	.3054	27
1	.76	2	.76	-.17	.00	1.000	28
1	-.25	2	-.10	.00	.93	.3536	29
1	-.14	2	-.44	-.15	1.83	.0682	30
1	-.34	2	-.22	.30	-.74	.4580	31
1	-.27	2	.25	-.12	-.14	.8901	32
1	-.50	2	-.02	-.02	-2.90	.0039	33
1	-.20	2	-.43	-.49	1.36	.1755	34
1	-.40	2	-.05	.23	-2.11	.0352	35
1	-.34	2	-.49	-.35	.88	.3805	36
1	-.28	2	-.43	.15	.90	.3692	37
1	-.07	2	-.34	.15	1.58	.1146	38
1	-.44	2	-.19	.26	-1.53	.1257	39
1	2.71	2	2.55	-.26	.96	.3392	40
1	.89	2	.79	.16	.16	.8733	41
1	.85	2	.77	.02	.58	.5614	42

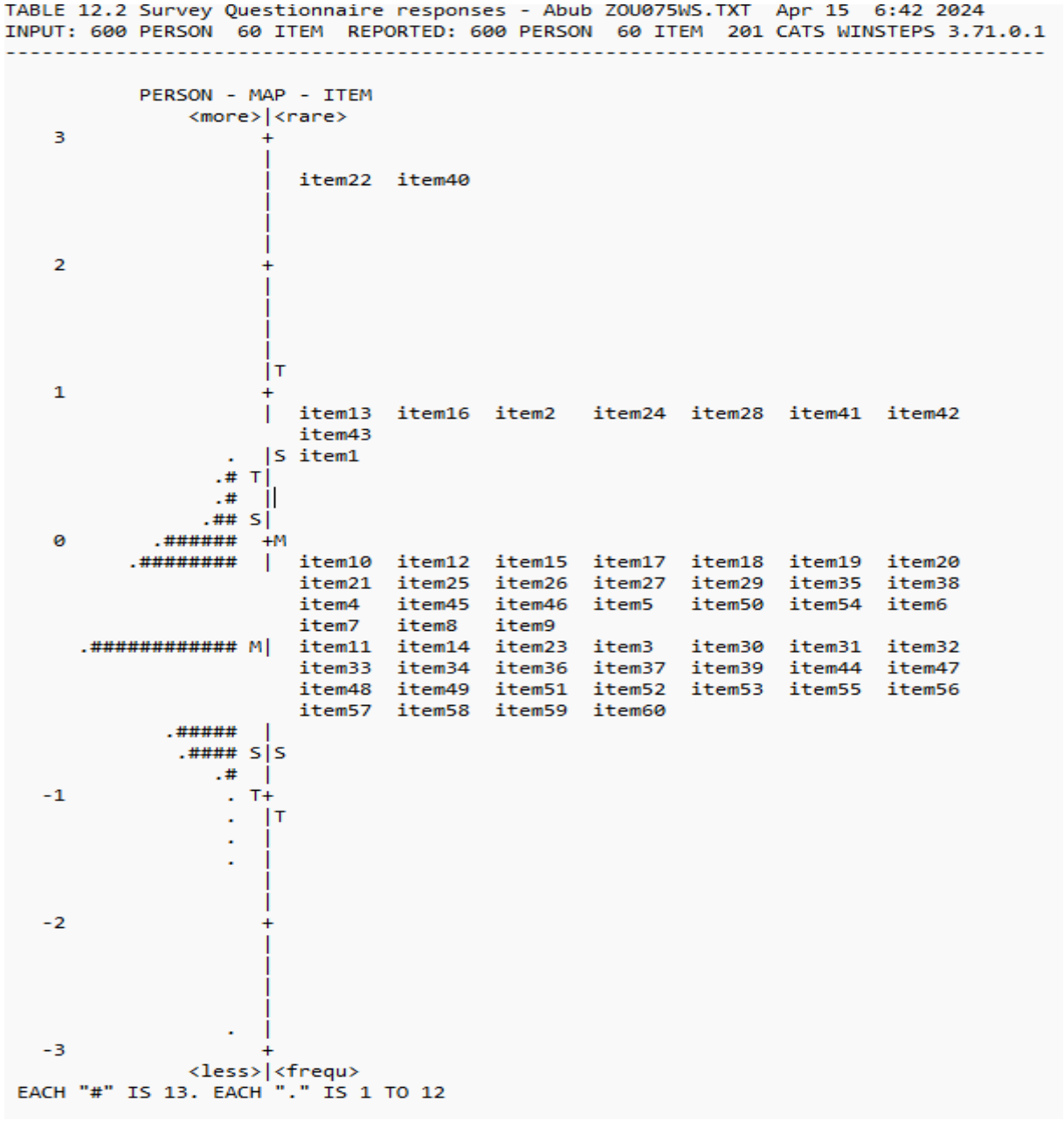
1	.87	2	.77	.08	.78	.4334	43
1	-.24	2	-.34	.11	.59	.5569	44
1	-.12	2	-.34	.10	1.27	.2029	45
1	-.19	2	-.19	.21	.00	1.000	46
1	-.27	2	-.27	.00	.00	1.000	47
1	-.51	2	-.28	.00	-1.38	.1686	48
1	-.34	2	-.28	-.23	-.38	.7018	49
1	-.14	2	-.22	-.06	.48	.6318	50
1	-.19	2	-.35	.08	.98	.3260	51
1	-.38	2	-.28	.16	-.61	.5405	52
1	-.28	2	-.47	-.10	1.17	.2429	53
1	-.22	2	-.28	.19	.30	.7609	54
1	-.27	2	-.25	.05	-.14	.8901	55
1	-.36	2	-.22	-.02	-.90	.3710	56
1	-.14	2	-.43	-.15	1.13	.2600	57
1	-.35	2	-.46	.19	.62	.5349	58
1	-.28	2	-.47	.10	1.17	.2429	59
1	-.36	2	-.36	.00	.00	1.000	60

Source: Author's work.

Finally, the Wright Map provides a picture of the targeting between the item difficulty and the participants' ability measures. Researchers can compare candidates and items using the map to have a better understanding of how well the exam measured. The Wright map of the present study is given in the following Figure 1. The map is organized as two vertical histograms. Items are shown on the right side, and participants are displayed on the left. The map's left side displays the distribution of the participants' measured abilities, with the most capable individuals at the top and the least able individuals at the bottom. The items found on the map's right side are arranged in order of difficulty, starting with the hardest and ending with the easiest.

According to Figure1, the results show that Item 22 and Item 40 (+2.69 and 2.64 logits, respectively) are the most difficult items to score in this examination; at the same time, item 2, item 16, item 24, item 28, item 41, item 42, item 43, and item 1 are also difficult to score. The result shows that none of the items are regarded as the easiest. The logits range of +2.69 logits to -0.10 logits fulfilled the acceptance range of +3.00 to -3.00 logits that was considered acceptable. (Andrich & Styles, 2004; Hill & Koekemoer, 2013; Linacre, 1994). It's evident from Figure 1 that there are no respondents with the ability to score items 1, 2, 13, 16, 24, 28, 42, 43, 41, and 43. Furthermore, as can be seen from Figure 4.1, the 130 students whose ability measures are slightly above -1 logits on the map and are indicated by ######### (each # represents 13 students) are measured with a little less precision. Their ability is below the items' difficulty among all the items in this examination. This indicates that they find every item to be difficult. As a result, it is evident the GSP122 exam did not well target the abilities of the students, and the items are somewhat difficult.

Moreover, Figure 1 also shows that although there are no respondents to match the difficulty levels of the items at the top of the map, the gaps between items 13 to item 22 and items 1 to 10 are



large. However, we would have obtained a more accurate estimate of their ability and could have more precisely located them on the ability scale if we had included more items in this region of difficulty that covered the space between items. According to Baghaei (2008), for a measurement to be uniformly precise, the items must be spaced reasonably, meaning that there should not be large gaps between them on the map, and when the space is large, the measurement is indicating the construct-under-representation. The difficulty of the items should match the person's ability.

Figure 1. Wright Map

Discussion of results

The primary objective of this study is to investigate the psychometric qualities of the GSP122 test, which is one of the successful measure of ICT knowledge among Nigerian undergraduates. To achieve this goal, the study assessed the GSP122 items using the criteria within the Rasch Model

analysis framework: Reliability evidence, Differential Item Functioning (DIF), and the Wright map. Firstly, the results from this study demonstrated the excellent item reliability of the GSP122 test with .98 index. The item separation index reflects the spread of item difficulties and the ability of the test to distinguish between items of varying difficulty levels. A high item separation index indicates that the test items are well distributed along the difficulty continuum, enabling precise measurement across different levels of ICT knowledge. In this study, the GSP122 test showed a strong item separation index, which supports its capability to differentiate effectively among various levels of student knowledge (Boone, Staver, & Yale, 2014).

Furthermore, when Rasch model analysis revealed a strong item reliability index and item separation index, indicating that the test items are reliable and able to distinguish between different levels of ICT knowledge (Wright & Masters, 2018). This finding is consistent with recent studies that have used Rasch analysis to examine the reliability and validity of tests in various fields. For example, a study by Paek and Wilson (2018) found that the TOEFL test demonstrated a strong item reliability index and item separation index, supporting its use as a measure of English language proficiency. Similarly, a study by Yang et al. (2020) found that the Chinese version of the Beck Depression Inventory-II demonstrated a strong item reliability index and item separation index, supporting its use as a measure of depression in Chinese-speaking populations. Consequently, the findings from this study support the use of GSP122 test as a measure of ICT knowledge.

Secondly, the result from the DIF analysis revealed that the GSP122 test's DIF statistics did not reveal a gender DIF that was significant. That is, all item DIF contrast statistics were within -0.5 to +0.5 logits (Wang, 2008). This indicates that no group of examinees was given preference over another by the items in this test measuring students' ICT knowledge, which may be a sign of the items' local independence from sample issues like gender (Bond & Fox 2001). In other words, there is no gender bias when it comes to assessing ICT knowledge because both male and female performers are evaluated based on their abilities and the difficulty of the items, not on their gender. The Rasch model analysis revealed no significant DIF statistics, indicating that the test items are not biassed towards either male or female performers (Wright & Masters, 2018).

The results of the DIF in this study are consistent with previous studies that have used Rasch analysis to examine gender bias in various tests. For instance, Wu and Adams (2020) highlighted the critical role of item response theory models, such as the Rasch model, in ensuring test fairness by identifying and mitigating potential biases in test items. The absence of significant DIF in the GSP122 test aligns with the standards set by these models, reinforcing the reliability and validity of the assessment for all test-takers, regardless of gender. Another study by Paek and Wilson (2018) found no significant DIF statistics in the TOEFL test, supporting its use as a gender-fair measure of English language proficiency. Consequently, the lack of gender bias in the GSP122 test has practical implications for educational practitioners and policymakers. It suggests that the test can be confidently used in diverse educational settings without the risk of disadvantaging any gender group. This is particularly relevant in the context of ICT education, where gender disparities have historically been a concern. Studies like those conducted by Cooper and Weaver (2022) have shown that gender-biased assessments can perpetuate stereotypes and discourage underrepresented groups from pursuing ICT-related fields. The GSP122 test, by being free from gender bias, helps to counteract these negative trends and promotes a more inclusive educational environment.

Thirdly, with regards to the Wright Map, the results reveal significant insights into the alignment between the GSP122 test items and the abilities of the students. Specifically, the analysis identified 11 items at the top of the map that are above the ability levels of the students. Moreover, as illustrated in

Figure 1, 130 students at the bottom of the map have ability measures slightly above -1 logits, indicating they are measured with less precision. Their abilities are below the difficulty level of the items in this examination, suggesting they find every item to be difficult. Consequently, the GSP122 exam did not effectively target the abilities of the students, with the items being somewhat difficult for many of the students assessed**.**

This finding aligns with the broader literature on test targeting and item difficulty. For instance, recent studies emphasize the importance of aligning test items with the ability levels of the test-takers to ensure accurate and meaningful assessment outcomes. Wu and Adams (2020) discuss that misalignment between item difficulty and student ability can lead to measurement errors and reduced precision in estimating student abilities. The observed difficulty of the GSP122 items corroborates these concerns, indicating a need for better calibration of test items to match the abilities of the student at Sule Lamido University. The misalignment observed in this study is consistent with findings from other research on educational assessments. Liu and Wilson (2019) highlight that when test items are too difficult relative to the students' abilities, it can lead to a range of negative consequences, including student frustration and diminished test validity. The fact that 130 students are measured with less precision further underscores the issue, as precise measurement is crucial for making reliable inferences about student performance and abilities.

The finding from this study is also consistent with previous research that has shown that students who struggle with a test tend to find all items on the test to be difficult (Bakhiet & Lynn, 2015). For instance, a study by Hur et al. (2017) found that students who performed poorly on a cognitive abilities test tended to find all items on the test to be challenging. Similarly, a study by Griskevica and Rascevska (2009) found that students who struggled with a mathematics test tended to find all items on the test to be difficult. Another study by Baghaei and Amrahi (2011) used the Rasch model to validate a multiple-choice English vocabulary test and found that several items were beyond the ability level of the students.

Moreover, the difficulty level of the GSP122 items has implications for the validity and utility of the test. According to Hambleton and Jones (2021), test items should be calibrated to cover a range of difficulties that are appropriate for the target population. When a significant portion of the test items are too difficult, it can skew the assessment results, leading to an underestimation of students' true abilities and potentially impacting their educational outcomes and opportunities. Additionally, this finding suggests a need for reviewing and potentially revising the GSP122 test to ensure it is appropriately challenging without being overly difficult. It is essential for educational assessments to strike a balance where items are neither too easy nor too difficult, facilitating accurate measurement of a wide range of student abilities (Wilson, 2022). The current study indicates that the GSP122 test may benefit from such a review to better target the abilities of the students it is designed to assess.

CONCLUSION

This study provides a comprehensive evaluation of the psychometric properties of the GSP122 test, focusing on item reliability, Differential Item Functioning (DIF), and Wright Map analysis. The findings indicate that the GSP122 test is highly reliable and fair, with no evidence of DIF, ensuring consistent and equitable measurement of ICT competencies among students. However, the Wright Map analysis reveals a misalignment between the test items and the abilities of students at the extreme ends of the proficiency spectrum, with some items being either too difficult or too easy.

To improve the effectiveness of the GSP122 test, it is recommended that future studies explore the development and inclusion of items that better target students across all proficiency levels,

particularly those at the extremes. Additionally, further research could investigate the longitudinal impact of these adjustments on student performance and the overall validity of the test in diverse educational contexts. These findings and recommendations provide valuable insights for refining ICT assessment practices within the Nigerian higher education system.

IMPLICATION OF THE STUDY FOR THEORY AND PRACTICE

The findings of this study have significant implications for both theory and practice in educational assessment. The results reinforce the principles of Item Response Theory (IRT), particularly its emphasis on the consistency of well-constructed test items in measuring underlying abilities. The strong item reliability observed in the GSP122 test supports IRT's theoretical framework, confirming that such items can reliably assess students' competencies. Additionally, the absence of Differential Item Functioning (DIF) aligns with the theoretical expectation that test items should function equivalently across diverse groups of test-takers, thereby ensuring fairness in assessment.

However, the Wright Map analysis revealed a misalignment between item difficulty and student ability levels, suggesting that while IRT provides a robust framework, its application must be carefully managed to ensure that items are appropriately targeted across the proficiency spectrum. This finding underscores the importance of continuously refining and testing theoretical models in educational measurement to better accommodate diverse student populations.

For practitioners, the study highlights the critical role of psychometric analysis in the development and maintenance of effective assessment tools. The strong reliability and fairness of the GSP122 test indicates that it is a valuable instrument for measuring ICT competencies, yet the gaps identified in item difficulty point to the need for ongoing item revision and enhancement. Practitioners should incorporate regular psychometric evaluations into their assessment processes to ensure that tests remain valid, reliable, and aligned with the abilities of all students. This practice not only improves the quality of assessments but also promotes fairer outcomes for students across different proficiency levels. Ultimately, the study emphasizes the importance of designing assessments that are both theoretically sound and practically effective, ensuring they meet the educational needs of diverse student populations.

LIMITATION OF THE STUDY

Despite the valuable insights provided by this study, several limitations should be acknowledged. First, the study relied on a sample of 600 test scripts from a single academic session at a specific public university in Nigeria. While the sample was stratified to ensure representativeness, the findings may not be generalizable to other institutions, academic sessions, or student populations. Future research should consider including a broader and more diverse sample across multiple institutions and academic years to enhance the generalizability of the results.

Second, the study focused exclusively on the GSP122 test, an ICT knowledge assessment, limiting the scope of the findings to this particular subject area. The psychometric properties identified may not apply to assessments in other disciplines or contexts. Expanding the analysis to include other subject tests would provide a more comprehensive understanding of the psychometric qualities across different types of assessments.

Lastly, the study employed Rasch analysis as the primary method for evaluating the test's psychometric properties. While Rasch analysis is a robust and widely accepted method, it is based on specific assumptions that may not fully capture all aspects of item performance and test-taker

behavior. Future studies could benefit from using a combination of different psychometric models to cross-validate the findings and provide a more nuanced understanding of the test's properties.

ACKNOWLEDGMENT

The authors are deeply grateful for the support and assistance received from the School of Educational Studies, Universiti Sains Malaysia, and the School of General and Entrepreneurship Studies, Sule Lamido University Kafin Hausa, which enabled them to successfully complete this research paper. Their contributions are sincerely appreciated.

REFERENCES

Adebayo, F., & Eze, U. (2024). ICT proficiency and national development: A study of Nigerian university graduates. *African Journal of Education and Technology, 15*(2), 123-140.

Andrich, D., & Styles, I. (2004). *Final report on the psychometric analysis of the Early Development Instrument (EDI) using the Rasch model: A technical paper commissioned for the development of the Australian Early Development Instrument (AEDI).* Perth, Australia: Murdoch University

Adebayo, F. (2021). ICT skills and national development in Nigeria: A critical analysis. *Journal of African Studies, 45*(3), 278-292.

Baghaei, P., & Amrahi, N. (2011). Validation of a multiple-choice English vocabulary test with the Rasch Model. *Journal of Language Teaching and Research 2*(5):1052-1060. DOI: 10.4304/jltr.2.5.1052-1060

Bond, T. G., & Fox, C. M. (2015*). Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Lawrence Erlbaum Associates. https://psycnet.apa.org/record/2001-06187-000

Baghaei, P. (2008). The Rasch Model as a construct validation tool. *Rasch Measurement Transactions, 22*(1), 1145–1146.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer.

Cooper, M., & Weaver, K. (2022). Gender Bias in STEM Education: Implications for Educational Practices. *Journal of Educational Research, 115*(3), 295-310.

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Sage Publications.

Fisher, W.P. (2007). Rating scale instrument quality criteria. *Rasch measurement transactions, 21*(1), 1095.

Holland, P. W., & Wainer, H. (Eds.). (2012). *Differential item functioning*. Routledge.

Hill, C., Koekemeor, E. (2013). The development of the MACE work-family enrichment instrument. *SA Journal of Industrial Psychology 39*(2):1147–1162. 10.4102/sajip.v39i2.1147.

Hambleton, R. K., & Jones, R. W. (2021). *Principles and Practices of Test Calibration and Linking*. Springer.

Hamad, A. A. (2021*). The psychometric properties of measurement of the mathematics teachers' professional identity in Saudi Arabia* (Unpublished doctoral dissertation). USM.

Hur, Y., Te Nijenhuis, J., & Jeong, H. (2017). Testing Lynn's theory of sex differences in intelligence in a large sample of Nigerian school-aged children and adolescents (N > 11 000) using Raven's standard progressive matrices plus. *Mankind Quarterly 573*(3):428-437. Doi: 10.46469/mq.2017.57.3.11

Johnson, R., Smith, A., & Garcia, M. (2023). A review of ICT assessment methods in global higher education: Trends and challenges. *Journal of Educational Technology & Society, 26*(1), 45-60.

Johnson, M., & Lee, S. (2019). The role of ICT literacy in higher education: A global perspective. *International Journal of Educational Technology, 12*(2), 145-163.

Kumar, V., & Singh, R. (2022). Employer expectations of ICT skills in recent graduates: A cross-sectional study. *Journal of Vocational Education & Training, 74*(4), 512-528.

Lee, J., & Park, S. (2024). Applying the Rasch model to evaluate ICT literacy: A case study from South Korea. *Educational and Psychological Measurement, 84*(2), 301-318.

Liu, X., & Wilson, M. (2019). Ensuring Fairness in Educational Assessment: The Role of Test Targeting. *Educational Measurement*: *Issues and Practice, 38*(2), 25-33.

Linacre, J. M. (2012). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328.

Linacre, J. M. (2006). *A user's guide to WINSTEPS/ MINISTEPS*: A Rasch model computer programs. Chicago, USA: Winsteps.com.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328.

Mustafa, A. K., & Ehab, M. N. (2022). Rasch analysis and differential item functioning of English language anxiety scale (ELAS) across sex in Egyptian context. *BMC Psychology, 10*(242), 4. https://doi.org/10.1186/s40359-022-00955-w

Oladipo, A., Nwosu, L., & Eze, U. (2023). ICT proficiency and academic performance: A longitudinal study of Nigerian university students. *International Journal of Educational Research, 112*, 101742.

Oladipo, A., & Eze, U. (2022). Evaluating ICT assessment tools in Nigerian universities: Challenges and opportunities. *African Journal of Educational Assessment, 18*(4), 412-428.

Oluwatobi, S., Efobi, U., Olurinola, I., & Alege, P. (2019). ICT and higher education in Nigeria: The way forward. *Journal of Educational Innovation, 32*(1), 87-103.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*(1), 59-81.

Paek, I., & Wilson, M. (2018). A Rasch analysis of the Test of English as a Foreign Language (TOEFL). *Language Testing, 35*(2), 147-164.

Richard, M. W., Peter, M. A., & Jotham, N. D. (2023). Psychometric properties of a test anxiety scale for use in computer-based testing in Kenya. *The International Journal of Assessment and Evaluation, 30*(1), 2327-8692. http://dx.doi.org/10.18848/2327-7920/CGP/v31i01/1-18.

Redecker, C., & Johannessen, Ø. (2013). Changing assessment — Towards a new assessment paradigm using ICT. *European Journal of Education, 48*(1), 79-96.

Smith, J., Brown, A., & Garcia, C. (2020). The importance of ICT skills in the 21st-century workforce. *Journal of Vocational Education & Training, 72*(3), 312-328.

Wang, W. (2008). Assessment of differential item functioning. *Journal of Applied Measurement,* 9(4), 387–408.

Wright, B. D., & Masters, G. N. (2018). *Rating scale analysis: Rasch measurement*. SAGE Publications.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum Associates.

Wilson, M. (2018). Making measurements important for education: The crucial role of classroom assessment. *Educational Measurement: Issues and Practice, 37*(1), 5-20.

Wu, M., & Adams, R. (2020). Applying the Rasch Model to Evaluate Fairness in Testing. *Educational Assessment, 25*(4), 287-305.

Wilson, M. (2022). *Constructing Measures: An Item Response Modeling Approach*. Routledge.

Yang, P., et al. (2020). A Rasch analysis of the Chinese version of the Beck Depression Inventory-II. *Journal of Clinical Psychology, 76*(1), 35-47.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223-233.