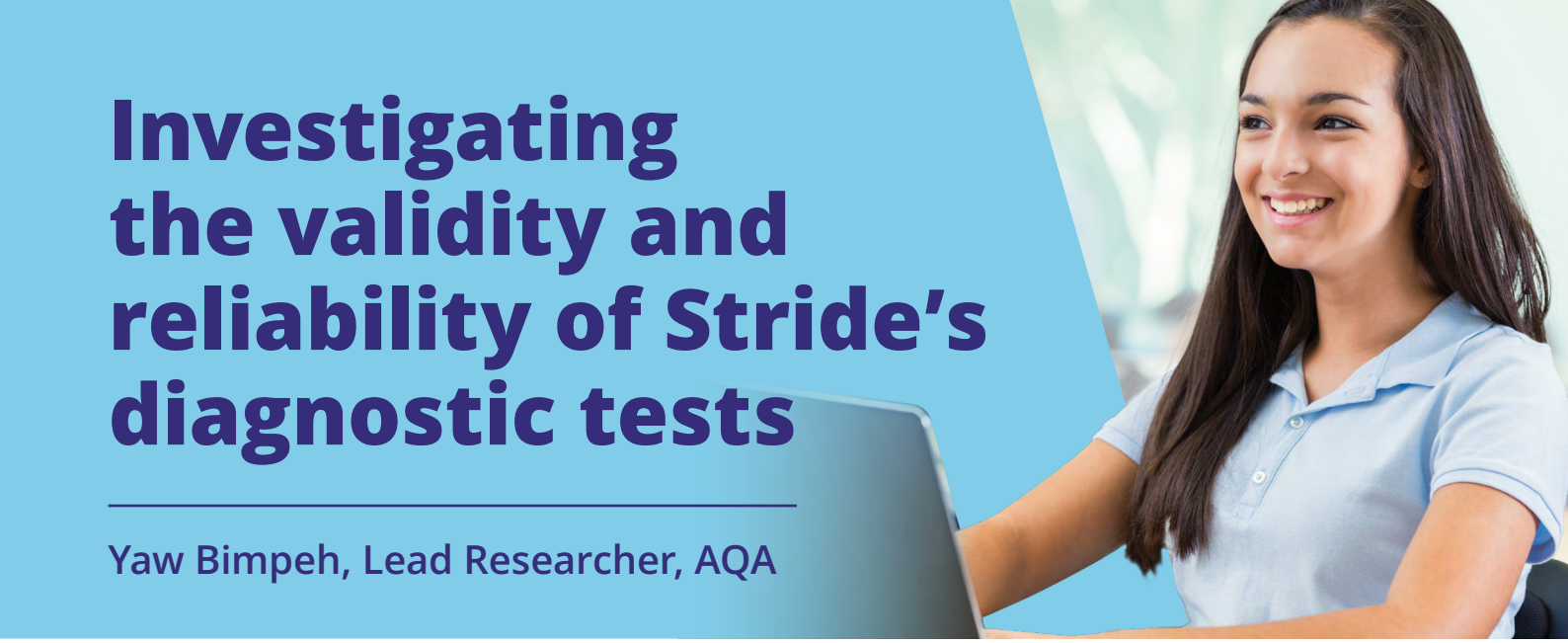
Investigating the validity and reliability of Stride's diagnostic tests



Yaw Bimpeh, Lead Researcher, AQA

Executive summary

Stride is AQA's new adaptive assessment, providing diagnostic tests to help students prepare for GCSE Maths.

This study investigated the validity and reliability of the diagnostic tests by trialling Stride multiple times, administering user-experience surveys and interviewing students from a range of schools. The research provided valuable insights into the reliability of the information about students' knowledge and skills, as well as the accuracy of inferences drawn from the diagnostic tests.

The findings indicate that the tests accurately gauge students' proficiency across a wide spectrum of abilities and effectively distinguish between individuals with varying levels of skill. Additionally, the difficulty level of the tests appears to align appropriately with the abilities of the students, ensuring accessibility for all individuals within the tested group.

Moreover, there is ample evidence supporting the validity of the diagnostic tests in relation to test content, internal structure and student response processes. The reliability of the tests is rated as excellent or good, indicating the suitability of the tests for diagnostic or formative purposes.

The results also indicate that items in the tests have varying levels of power to distinguish between individuals with high and low abilities, ie certain items are more effective at distinguishing than others.

When asked about their overall test-taking experience, most respondents (78%) rated it positively. About 73% of respondents said they agreed with Stride's identification of their strong points or competencies. According to 75% of respondents, tests like these would be beneficial for their learning.

This research underscores the significance of Stride's ability to deliver personalised and adaptive feedback to students, including explanations, hints, and resources tailored to individual learning needs. The testing process is helpful for both students and teachers and is effective at tracking students' progress over time.

This study accompanied qualitative research to gather student experiences of trialling the tests.[1]

[1] *Investigating students' experiences of piloting Stride*

Questions matter

Introduction

The focus of this work was to evaluate Stride's adaptive diagnostic tests, which employ learning objectives, cognitive mapping and a range of factors – such as competency, metacognition and self-awareness – to precisely assess each learner's areas of weakness and strength. The Stride platform provides personalised and adaptive feedback to students, including explanations, hints and remediation resources tailored to individual learning needs. It also offers teachers real-time analytics and reports on student performance to quickly identify gaps in a student's understanding of fundamental maths concepts. This can help teachers to tailor their instruction and interventions.

A key objective of Stride is to help students to reach their full potential in GCSE Maths, potentially improving results and saving teachers valuable time. Each of the diagnostic tests is underpinned by one of five key concepts: numbers, algebra, proportions, graphs, and shapes.

Research was carried out to evaluate how well these key concepts align with maths teachers' pedagogical practice and students' learning and understanding. A qualitative strand of research gathered and analysed data on user feedback relating to the platform and its use in the classroom. This report describes how empirical evidence was used to assess the validity and effectiveness of the diagnostic tests in identifying a student's areas of strength and weakness. The study investigated the system's ability to:

- differentiate between students of differing ability levels
- provide consistent and dependable information about students' knowledge and skills
- demonstrate concurrent validity with mathematics competency tests
- support accurate inferences based on the diagnostic formative tests.

Many research studies have highlighted the value of adaptive learning. For over 30 years, computer scientists and cognitive researchers have worked on adaptive learning systems to replicate human tutoring interactions. As highlighted by Calhoun Williams (2019), adaptive learning is a key technology that can support education by delivering content, posing questions, assigning tasks, providing hints and encouraging attitude adjustments. The core design of adaptive learning systems involves a 'closed loop' mechanism that collects learner data to assess progress, suggest learning activities and offer personalised feedback.

These adaptive learning systems utilise various algorithms, such as item response theory and machine learning, to personalise the learning experience. Research suggests that these systems can enhance student learning, with many studies showing positive outcomes. For example, a systematic review by Xie et al. (2019) found that 86% of studies on the impact of adaptive learning on learning outcomes reported positive results. Sahoo et al. (2023) found that a continuous periodic formative assessment model had a valid educational impact. A complex adaptive framework can be employed to address the multifaceted challenges and enhance the sustainability of continuous learning.

Illustration of test materials and concept map

A learning-pathway structure underpins the adaptivity of the Stride tests. The concept map in Figure 1 illustrates the key concept 'Numbers' and its related learning objectives, which a student must have mastered by the end of Key Stage 3 or beginning of Key Stage 4 (when students are 14–15 years old). Each node in Figure 1 corresponds to a specific learning objective and an arrow from a node *A* to a node *B* indicates that understanding *A* is a prerequisite for understanding *B*. This relational structure between learning objectives underpins the adaptivity of the test, enabling it to route learners through different paths to mastery. More precisely, each node comprises a set of items designed to assess a particular aspect of understanding and knowledge related to that learning objective.

The most fundamental ideas, which are a prerequisite for understanding, are shown in green on the map. These green nodes reflect what all students should understand by the time they begin GCSE Maths. The red nodes are the key ideas that students really need to master to make satisfactory progress through the course. The blue nodes represent more advanced ideas that will be developed at GCSE.

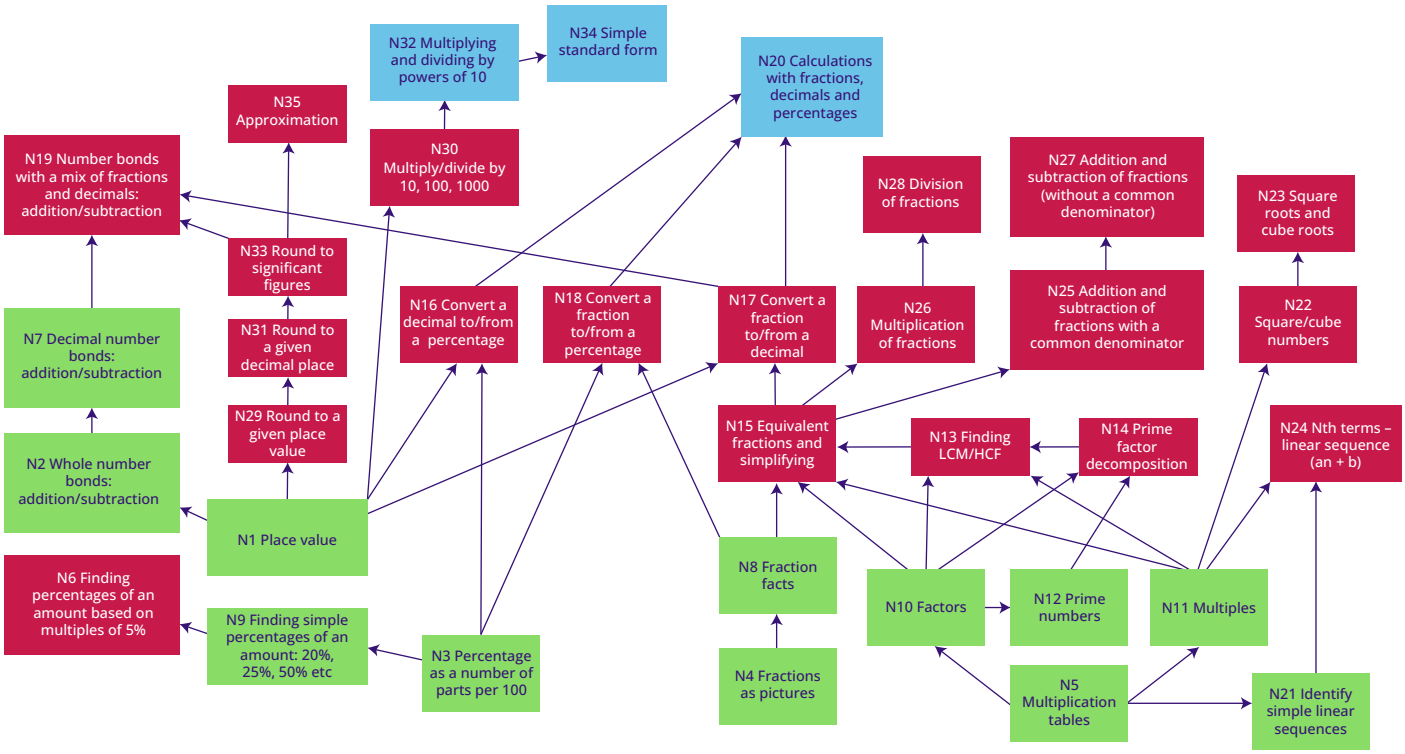


Figure 1: Concept map for Numbers

The different colours represent the three assessment objectives (green = prerequisite for understanding; red = conceptual knowledge; blue = applications of basic concept).

Method

Stride was trialled with a sample of 15 secondary schools in England. Following the initial recruitment of 1,500 students aged 14 to 15 years, two groups (or cohorts) were formed: Cohort 1 had access to Stride's remediation features – tools and functionalities designed to address students' learning gaps or misunderstandings in real-time – while Cohort 2 had access to the assessments only, not the remediation features. After the test, participants completed a survey about their experience. A mixed-methods research approach was used. This combined quantitative and qualitative analyses of test performance, user experience, engagement and satisfaction with the adaptive learning platform, with thematic analysis of semi-structured interviews.

Additionally, an iterative evaluation method enabled refinement of the Stride platform, based on feedback and performance data. The item response data from diagnostic tests, platform interactions and feedback from teachers and students helped to assess the platform's appropriateness and usefulness. Subsequent trials were conducted with sample sizes ranging from 500 to 1,000 students per key concept to validate platform improvements.

Data analysis

A structural equation modelling framework (Fornell & Larcker, 1981) was used to analyse the reliability of measures of students' proficiency with respect to the learning objectives and to explore construct validity, ie that the tests measure what they are purported to measure. This involved first relating each assessment objective to test items intended to measure it, taking into account measurement error, and then examining the covariance between measures of the assessment objectives.

This method enabled appraisal of the validity of assessment objectives, the reliability of measurement (assessed by comparing the degree of variance explained by each measure with the amount attributable to measurement error) and the relationships between various nodes (sub-concepts). Rasch-based analysis was also used for the six aspects of Messick's validity for the adaptive test in the framework of evaluation used by Wolfe and Smith (2007) and Beglar (2010).

Results

Discrimination and reliability of test items

The findings in this study are based on the most recent trials of the Stride platform.

The results suggest that the test items across all five key concepts typically show a strong positive relationship with the total score and effectively distinguish between varying levels of performance. On average, the discrimination of the items (their effectiveness in distinguishing between students with different overall performance levels) ranges from 22% to 33% (Figure 2). Items that did not provide substantial insights were either adjusted or removed.

There was some variation in item difficulty across the key concepts. The average item scores varied between 55% and 70% depending on the specific key concept under assessment. Numbers and graphs were areas where individuals tended to score more highly. These results indicate that the tests are appropriately challenging for most students.

The reliability of the tests, which indicates the consistency of its results across different applications, ranges from 74% to 90% (Figure 2). This seems sufficiently high for a diagnostics and formative assessment, meaning there would be no need to make the test longer or extend the test time, which would increase the burden on examinees. Each test can contain a different number of items, ranging between 34 and 45 on average per student.

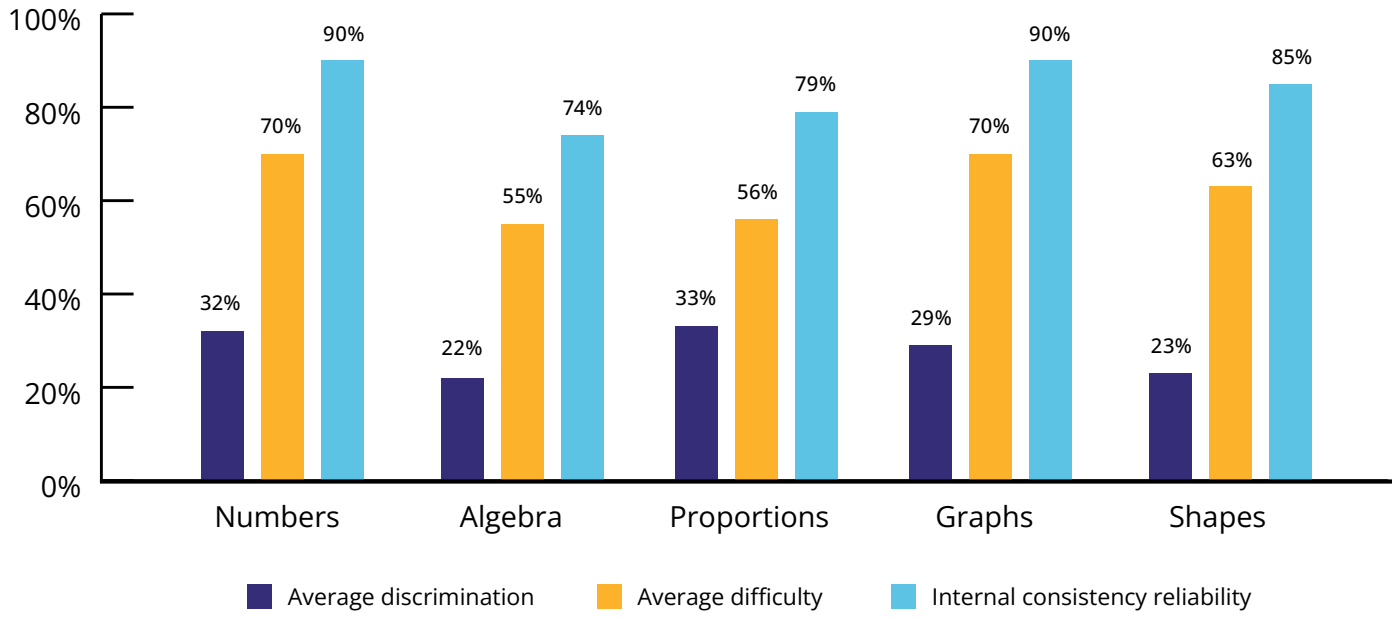


Figure 2: Summary of test performance indicators

Discrimination:
'good' = above 30%;
'fair' = between 10% and 30%;
'poor' = below 10%.

Difficulty:
'hard' = below 20%;
'good' = between 20% and 50%;
'best' = between 50% and 80%;
'very easy' = above 80%.

Reliability:
'poor' = below 40%;
'fair' = between 40% and 60%;
'good' = between 60% and 75%;
'excellent' = above 75%.

Figure 3 provides the person reliability statistics, which evaluate how effectively a set of items can differentiate between individuals based on their abilities or traits. The person reliability values measure the reliability of the test in ranking individuals based on their abilities or performance. They show the consistency of a test in estimating students' proficiencies with respect to one of the five key concepts. This indicates the reliability of the test items and the precision of the proficiency measurement, offering a more nuanced understanding of students' proficiency level within the tested domain.

As depicted in Figure 3, the reliability of predicted proficiency levels based on individuals' responses to test items was 86% on average; that is, above the cut-off guideline of 80% (Linacre, 2007), which indicates good person reliability. These results provide evidence to support construct validity and reasonable confidence of replicability of the person and item ordering across similar samples.

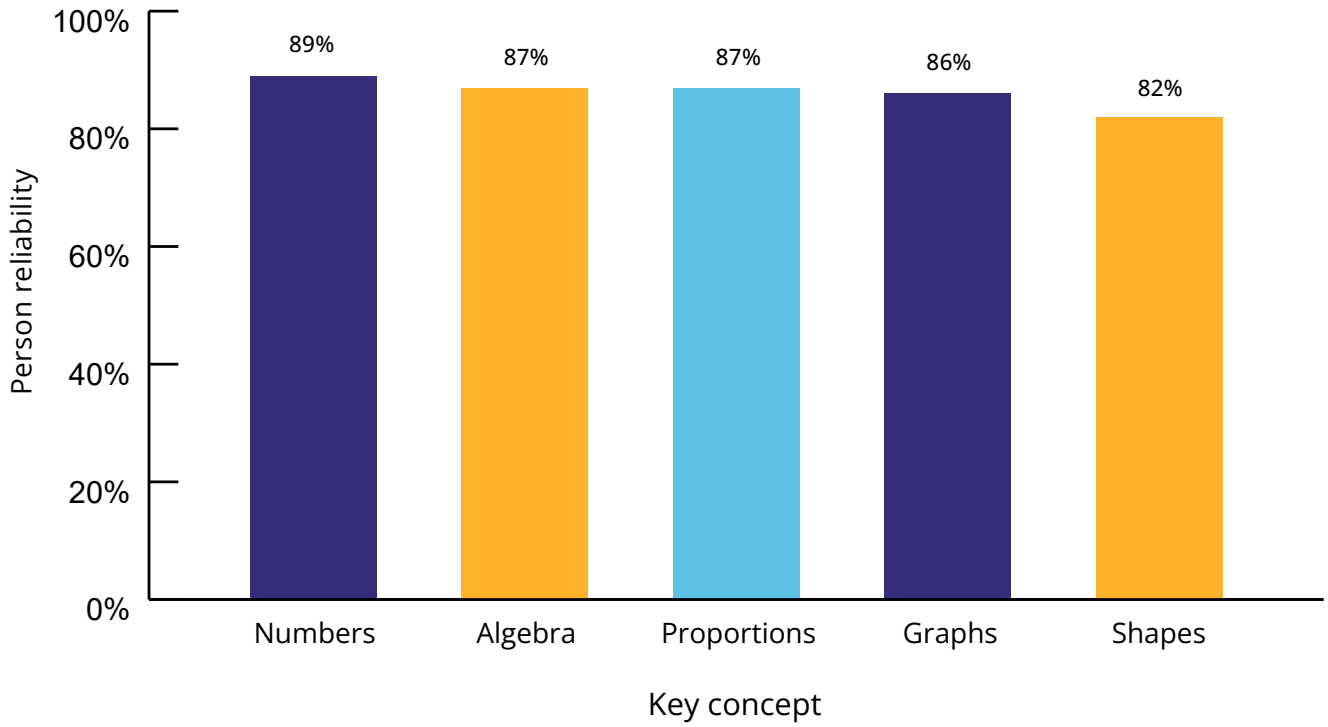


Figure 3: Person reliability for each key concept

Content aspect of construct validity

Wright maps, or item-person maps, were produced for each of the diagnostic tests.[2] For example, Figure 4 depicts the item-person map for the Numbers test. The item-person maps show the distribution of test takers' ability estimates alongside estimates of the item difficulties.

Guided by the assessment framework established by Wolfe and Smith (2007) for, what Messick (1989) calls, the 'content aspect of construct validity', the following were analysed: gaps and redundancy in the person-item map along the vertical line, the mismatch between item and person means, and infit and outfit statistics (measures of how well individual items measure the target construct and whether test takers' performances align with model expectations; Bond & Fox, 2015).

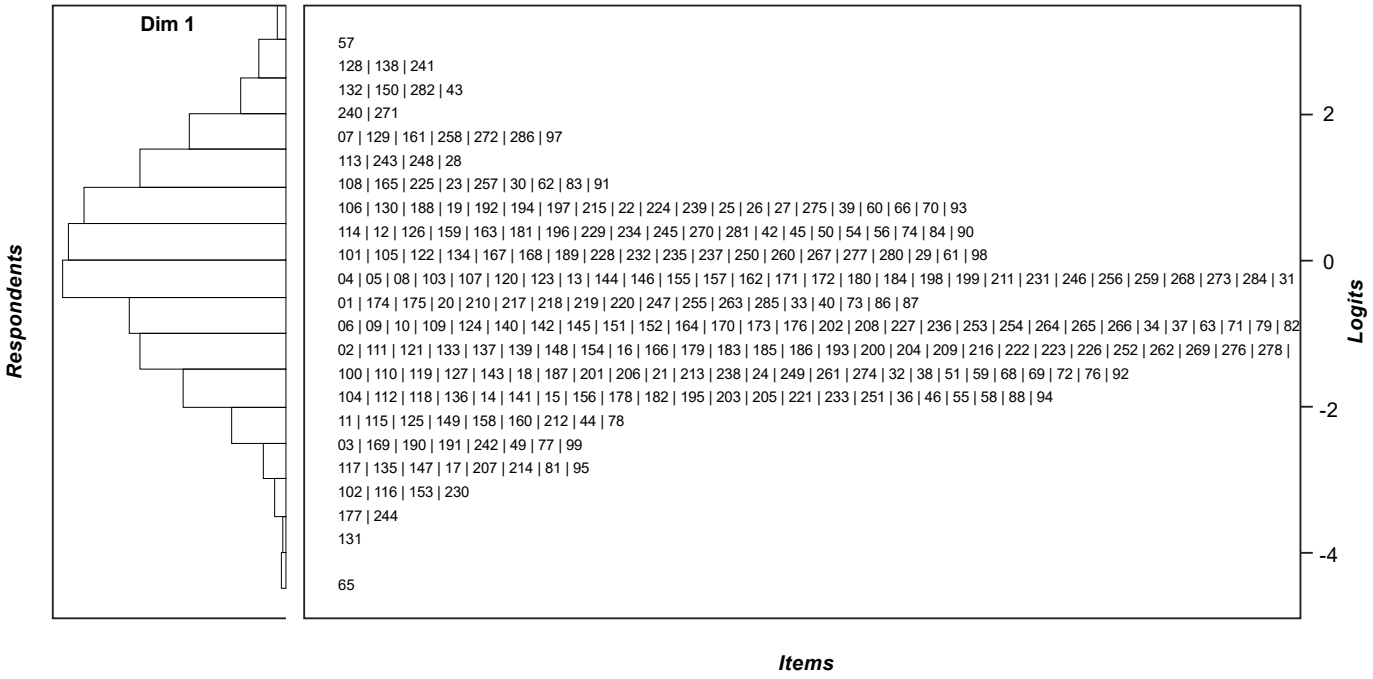


Figure 4: Item-person map for Numbers

Rasch analysis was used to evaluate the diagnostic tests as a measurement tool. Rasch-based methods are particularly suitable for assessing pedagogical tests. These analyses produce measures of proficiency and item difficulty that are independent of the specific items on the test and the sample of test takers.

The item-person maps present a concise overview of item statistics for the diagnostic tests. Along the vertical axis of the figure, items are positioned based on their level of difficulty, with easier items placed at the bottom and more challenging ones at the top. This mapping indicates a satisfactory level of representativeness in the spread of item difficulties. It aids in examining some of the aspects of test validity summarised by Messick (1989), including the so-called content, structural and substantive validity of the test. It facilitates the evaluation of how well a student's proficiency aligns with the difficulty of the learning objectives.

[2] A Wright map presents both item difficulties and person abilities arranged along the same logit scale. This enables visualisation in terms of the targeting of the test to the sample, as well as the targeting of individual items to persons.

For a well-targeted test, the average ability levels of the test takers should align with the average difficulty levels of the test items, typically around 0 on the so-called logit scale. This scale indicates that an item with a difficulty of 0 logits is expected to be answered correctly about 50% of the time by students with average proficiency. As shown in Figure 4, in the case of Numbers, the average abilities of the individuals are positioned around 0, which aligns with mean item difficulty measures.

The Rasch model's item fit statistics for the five key concepts were close to the ideal value of 1.00, with mean item infit values ranging from 0.97 to 1.00 and outfit values from 1.00 to 1.07. Since these values do not significantly deviate from the Rasch model's expected value of 1.00, the measurement of student proficiency is considered to have minimal random noise (Linacre, 2018). The item infit statistics show how much items contribute to measuring the underlying construct (Bond & Fox, 2015). These values help in evaluating and ensuring the accuracy and reliability of the measurement model in assessing student performance.

As depicted in Figure 4, there is no significant gap between items or learning objectives for the Numbers test, suggesting that the test's difficulty level matches well with the students' abilities, making it accessible to all individuals within the tested group. Comparable trends were observed for the other key concepts assessed (algebra, proportions, graphs, and shapes) in terms of the degree of alignment between a student's proficiency and the difficulty of the items assessing the learning objectives.

The results of the Rasch analysis contribute evidence to support the construct validity of the tests.

Assessing convergent validity of constructs

The study also assessed convergent validity to confirm that the items together measure the same latent constructs through construct reliability (CR) and average variance extracted (AVE) values. AVE measures the average proportion of variance in the items that is explained by the learning objectives. All five key concepts showed satisfactory construct reliability (above 70%), although the average variances were slightly below 0.50. Convergent validity was confirmed using the Fornell and Larcker (1981) criterion, which requires CR to be over 70% and AVE to be 0.50 or higher.

Various statistics were utilised to assess how well the empirical data aligned with the theoretical model. The confirmatory factor analysis demonstrated a good fit for the five key concepts. All goodness-of-fit indicators, such as GFI, CFI, and NFI,[3] met the threshold of 0.90, which is considered essential for model fit. Hu and Bentler (1999) suggested that model fit assessments should be based on the combined evaluation of multiple fit indices. The goodness-of-fit metrics for the structural equation modelling showed satisfactory results, supporting the validity of the associations among the concept maps. These statistics help provide an understanding of how closely the observed response patterns of the test takers correspond to the model's predictions.

The evidence strongly suggests that the diagnostic tests consistently yield insights into students' competencies and knowledge. The findings show that both empirical evidence and theoretical justifications support the accuracy of conclusions drawn from test scores, indicating the validity of the underlying construct.

[3] GFI (goodness of fit index) assesses the goodness of the approximation of the model rather than its correctness; NFI (normed fit index) compares the chi square of the empirical model with the chi square statistics of the null model; and CFI (comparative fit index) takes the same approach as the NFI but is adjusted by the degrees of freedom.

Conclusions and discussion

The evaluation of Stride's diagnostic tests revealed that they accurately gauge students' proficiency and effectively distinguish between individuals with varying levels of skill. Moreover, the tests' difficulty levels appear to align appropriately with the abilities of the students in the samples selected, ensuring accessibility for virtually all individuals within the tested group.

There is strong evidence supporting the validity of the diagnostic tests in terms of test content, internal structure and student response processes.The reliability of the tests is evaluated to be excellent or good, making them appropriate for diagnostic or formative use.

The findings provide strong support for the hypothesised relationship between the learning objectives. Hence, the learning-pathway structures are confirmed. As a result, the maths content related to numbers, algebra, proportions, graphs, and shapes depends on learners having the required prior knowledge and understanding, along with the ability to apply key concepts. The results show that all three assessment objectives met the construct reliability minimum threshold of 70%.

The results provide sound evidence that the diagnostic tests give consistent information about students' knowledge and ability. The testing process is helpful for both students and teachers, and it is effective at tracking students' progress over time.

When asked about their overall test-taking experience, most respondents (78%) rated it positively. About 73% of respondents said they agreed with Stride's identification of their strong points or competencies. According to 75% of respondents, tests like these would be beneficial for their learning.

References

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27*(1), 101–118. https://doi.org/10.1177/0265532209340194

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

Calhoun Williams, K. (2019). *Prepare for AI's new adaptive learning impacts on K-12 education*. Gartner. https://www.gartner.com/en/documents/3947160

Draba, R. E. (1977). *The identification and interpretation of item bias* (Research Memorandum No. 25). The University of Chicago.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39–50. https://doi.org/10.2307/3151312

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Sahoo, S., Tirpude, A. P., Tripathy, P. R., Gaikwad, M. R., & Giri, S. (2023). The impact of periodic formative assessments on learning through the lens of the complex adaptive system and social sustainability principles. *Cureus, 15*(6), e41072. https://doi.org/10.7759/cureus.41072

Wolfe, E. W., & Smith, E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II–Validation activities. *Journal of Applied Measurement, 8*(2), 204–234.

Xie, H., Chu, H.-C., Hwang, G.-J., & Wang, C.-C. (2019). Trends and development in technology-enhanced adaptive/personalized learning: A systematic review of journal publications from 2007 to 2017. *Computers & Education, 140*, Article 103599. https://doi.org/10.1016/j.compedu.2019.103599

Yan, H. (2020). *Using learning analytics and adaptive formative assessment to support at-risk students in self-paced online learning*. 2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT). https://doi.org/10.1109/ICALT49669.2020.00125

@AQA_Research

aqa.org.uk/about-us/our-research

research@aqa.org.uk

G02587_01a

AQA Questions matter