applied sciences

MDPI

Article

Using Technology for the Efficient and Precise Assessment of Cognitive Skills in Countries with Limited Standardized Assessment Instruments: A Report on the Case of Saudi Arabia

Mei Tan [1], Nan Li [1], Catalina Mourgues [2], Lesley Hart [1], Abdullah Qataee [3], Mark Grossnickle [4], Chris Errato [4] and Elena Grigorenko [1,5,6,*]

[1] Texas Institute of Measurement, Evaluation and Statistics, University of Houston, 4349 Martin Luther King Boulevard, Houston, TX 77204-60220, USA; mei.tan@times.uh.edu (M.T.); nan.li@times.uh.edu (N.L.); lesley.hart.phd@gmail.com (L.H.)
[2] Department of Psychiatry, School of Medicine, Yale University, 300 George Street, New Haven, CT 06511, USA; catalina.mourgues@yale.edu
[3] National Center for Assessment in Higher Education, King Khaled Road, Riyadh 11537, Saudi Arabia; A.QATAEE@etec.gov.sa
[4] MindTrust, Inc., 470 James Street, #11, New Haven, CT 06513, USA; mgrossnickle@mindtrust.com (M.G.); cerrato@mindtrust.com (C.E.)
[5] Center for Cognitive Sciences, Sirius University, 354340 Sirius, Russia
[6] Child Study Center, School of Medicine, Yale University, 230 S Frontage Road, New Haven, CT 06520, USA
* Correspondence: elena.grigorenko@times.uh.edu

Citation: Tan, M.; Li, N.; Mourgues, C.; Hart, L.; Qataee, A.; Grossnickle, M.; Errato, C.; Grigorenko, E. Using Technology for the Efficient and Precise Assessment of Cognitive Skills in Countries with Limited Standardized Assessment Instruments: A Report on the Case of Saudi Arabia. *Appl. Sci.* **2022**, *12*, 1617. https://doi.org/10.3390/app12031617

Academic Editors: Maria Luisa Lorusso, Francesca Borasio and Sara Mascheretti

Received: 21 December 2021
Accepted: 26 January 2022
Published: 3 February 2022

Abstract: In Saudi Arabia, the country's progress toward appropriate and inclusive education programs for children with disabilities is still evolving. A crucial aspect of this evolution has been the development of a comprehensive assessment battery that covers a broad range of cognitive factors for the diagnosis of neurodevelopment disorders and other types of intellectual atypicalities, including giftedness. The Alif–Ya Assessment Battery consists of 47 subtests based on various theories of intelligence. Alif–Ya capitalizes on advanced technologies to enable its delivery either in person or remotely. Moreover, over half of Alif–Ya's subtests are adaptive; items are selected for the test takers based on their previous responses. In this paper, we provide an overview of the Alif–Ya Assessment Battery, describe how it was designed to make the best use of the latest and best features of technology for the appropriate and accurate assessment of children and adolescents in the Kingdom of Saudi Arabia via remote or in-person administration, and present initial data collected with the battery.

Keywords: Saudi Arabia; cognitive assessment; computerized adaptive testing; neurodevelopmental disorders; diagnosis; treatment; Alif–Ya

1. Introduction

In this paper, we will describe the development of and pilot data collection for a comprehensive assessment battery (Alif–Ya) designed to make the best use of the latest features of technology for the appropriate and accurate assessment of children's neuropsychological functioning in the Kingdom of Saudi Arabia (KSA). Such a tool is needed to better support learning disabilities in schools across the spectrum of intellectual profiles. To contextualize Alif–Ya's conception and design, we briefly outline the current issues of identification and diagnosis in the field of special education that Alif–Ya is intended to address, and the progress of computerized assessment in the field of clinical neuropsychology thus far.

1.1. Methods of Identifying Disability in the KSA: Definitions and Assessments

The Ministry of Education in Saudi Arabia has adopted the American Association on Intellectual and Developmental Disability definition of intellectual disability, which characterizes it as significant limitations in both intellectual functioning and adaptive

behavior. However, it has been reported that 73% of special education programs and institutes rely only on intelligence tests for the diagnosis of intellectual disability [1]. Moreover, the assessments most often used, the Wechsler Intelligence Scales for Children [2] and the Stanford-Binet Intelligence Test [3], were Arabic versions originally adapted for Egyptian and/or Jordanian students, not for Saudi Arabian students [1,4]. Importantly, the versions of these assessments currently available in Saudi Arabia are now outdated, as translations are not routinely updated based on new versions of the original assessments. This constitutes another reason to use assessments originally developed for Saudi Arabia by a Saudi Arabian team.

Saudi Arabian researchers and psychologists have over the years evaluated the neuropsychological functioning in children and adults using a range of assessments, including translated and adapted instruments developed in Western contexts, as well as natively developed intelligence scales, such as a 53-item test developed for the Saudi environment by Al-Teriri [5]. In general, their validity has been found to be adequate. Kearney and colleagues [6] examined the usability of several translated and adapted assessment instruments developed in the West to predict mental retardation in Saudi Arabian children and youth (n = 115, 68 males, M_{age} = 9.9 years, SD = 3.3). Three assessments were examined: the Leiter International Performance Scale [7], a nonverbal test of cognitive ability; the Developmental Test of Visual-Motor Integration [8], a perceptual-motor test; and the Vineland Adaptive Behavior Scales [9]. These assessments were administered to evaluate the cognitive status of children who had been diagnosed by Saudi Arabian physicians as developmentally delayed (mild, moderate, severe or profound). The results indicated that all three instruments significantly differentiated three levels of disability severity (mild vs. moderate vs. profound groups). In a multiple regression analysis, all of the scores (of the Leiter, DVMI, and three subdomain scores of the Vineland) together were associated with level of disability severity, with the Leiter and Vineland Communication scores contributing significantly to the prediction [6], thus illustrating some validity for the use of such assessments for diagnostic purposes in the Kingdom.

In a pilot examination [10] of the equivalence of Arabic and English versions of Golden's Standardized Stroop Task [11], native Arabic bilingual speakers (n = 10, 5 males, ages 16–20) were asked to do both versions sequentially (English first). The researchers concluded that individuals' performance across the Arabic and English versions were virtually the same, as shown in a paired sample *t*-test that revealed no significant differences. Norms for adults (ages 16–65) have now been generated for the Stroop task, along with a number of assessments adapted for the Saudi context [12], namely: the Wisconsin Card Sorting Task [13], the vocabulary and picture completion subtests of the Wechsler Adult Intelligence Scale-Revised [14], and the Test of Nonverbal Intelligence [15]. The nonverbal tests were only modified in that their instructions were translated from English into Arabic. For the verbal tasks, the Stroop and vocabulary test of the WAIS-R, translations of the stimulus materials were carried out, and in the latter, language experts were consulted to make an appropriate selection of vocabulary words. That is, little to no new material was generated for the Saudi Arabian versions of these tests. The Raven's Standard Progressive Matrices (SPM) has also been standardized for use in Saudi Arabia, initially in 1977 [16]. Notably, a second standardization exercise carried out with the SPM with 8–15 year-olds in 2010 in Makka Province reflected an average rise in scores across all age groups of 11.7 IQ points, consistent with the Flynn effect [17].

1.2. The Use of Computerized Assessments for Neuropsychological Testing

The use of computers for the administration, scoring and interpretation of neuropsychological assessments was taken up soon after the introduction of the personal computer in the 1970s [18]. There is now a broad range of such assessment tools available, both adapted from already-existing and well-known tests, e.g., the Wisconsin Card Sorting Test [13] and the Raven's Progressive Matrices [19], and those newly developed, such as Conners Continuous Performance Test II [20], designed to assess complex aspects of atten-

tion, response inhibition, and impulsiveness. There are computerized neuropsychological assessments available for broad evaluative purposes across multiple cognitive domains geared for research, such as the NIH Toolbox for the Assessment of Neurological and Behavioral Function [21]. For clinical purposes, Pearson's Q-Interactive application can be used by clinicians for tablet-based administration of the WISC-V or the Delis-Kaplan Executive Function System [22]. NeuroMarker combines the collection of neurophysiological markers—electroencephalogram (EEG) and event response potential (ERP) measures—with neurocognitive tests for use across the lifespan [23]. Other computerized tools address specific referral questions, such as managing sports-related concussion injuries for use with specific patient populations, such as adults at risk for dementia [18].

Over the years, the strengths and limitations of computerized assessment for clinical use have been well-considered, with the listed benefits generally exceeding the number of concerns [18]. Strengths include accurate and consistent timings in item administration and the collection of response times [24], exportability of data for analysis [25], and the ability to reach remote patient populations through various devices [26]. In addition, computerization has enabled the automatization of adaptive tests [18]. First applied by Alfred Binet and his colleagues in the early 1900s in the administration of the Binet intelligence test [27], test adaptivity is where item sequences are delivered variably based on a child's ongoing performance. Now, item selection algorithms may be devised for computerized adaptive testing (CAT), to deliver test items that match the proficiency of the test-taker [28]. The primary benefits of CAT are efficiency (i.e., requiring the delivery of fewer items), greater measurement precision, and greater test security due to varied item delivery [27,29]. Yet the limitations of computerized assessment have also been noted. Documented concerns have included errors that have occurred in assessment administration due to hardware and software connectivity issues [25,30] and the loss of flexibility in test administration and modes of administrator-respondent interaction, which a clinician may adjust based on the test-taker's motivation and cognitive style [31,32]. In addition, for computerized assessments adapted from paper and pencil forms, experiential and psychometric equivalence between the two have frequently posed doubts about upon validity and raised challenges for their equating [25,33], although some better methods for test equating have evolved [34]. Due to these concerns, the adoption of computerized assessments by clinicians has been gradual; a 2011 survey of 495 practicing neuropsychologists in the United States and Canada revealed that just under half (45.5%) had never used a computerized test battery in their work [18]. Yet, the computerized and online-delivered versions of this battery and others represent an essential advantage in the context of school psychology [35] and telepsychology by increasing the accessibility of psychological assessments in rural, underserved populations [36,37] and when physical distancing guidelines are the norm, as during the COVID-19 pandemic. As the field continues to adopt innovative computer assessment approaches, the potential benefits of technology in neuropsychological assessment should be fully explored, particularly in tests designed without the constraints needed to maintain equivalence with paper and pencil test forms.

Alif–Ya is a set of assessments specifically conceived for computerized delivery and designed for and with researchers, educators, and clinicians in the KSA. The purpose of the battery is to identify cognitive profiles across the spectrum of abilities, to capture atypical functioning that may contribute to diagnosing a range of ability, from intellectual giftedness to a variety of forms of neurodevelopmental disorder. It also fills a gap as a computerized assessment that may be used in the school context.

1.3. The Theoretical Foundations and Content of Alif–Ya

Human intelligence has been defined in several ways. Some of these definitions highlight analytical skills to solve problems; others emphasize social skills and holistic algorithms to perform adaptively in daily life, yet others incorporate executive functions as basic processes of intellectual ability [38]. In the field of education, intelligence assessment has a long tradition in which, from Binet-Simon through Weschler, intelligence quotient

scores (IQ) have been generated as a composite score of several sub-tests sharing similar underlying cognitive processes [39,40] as the psychometric analyses have been shown repeatedly [41–43]. These scores have been used to make important educational decisions, such as school placement or specific curriculum accommodations [44]. However, IQ scores have been criticized as not reflective of a child's actual range of cognitive skills. That is, intelligence tests have not always captured the relevant broad spectrum of skills that characterize future performance accurately [45].

Alif–Ya addresses this issue in part by providing a uniquely broad coverage of cognitive processes, various combinations of which can be assembled to produce overall and modality-specific (e.g., verbal and non-verbal) IQ scores. It is anchored in the theoretical psychometric foundations of the Cattell-Horn-Carroll (CHC) model of intelligence, long recognized as one of the most empirically validated structural models of human cognitive abilities [41–43]. Alif–Ya is also founded upon the clinically derived theories of Alexander Luria [46] encompassed in the Planning, Attention, Simultaneous, Sequential (PASS) theory of intelligence [47]. To encompass the abilities emphasized by both of these theories, Alif–Ya consists of 47 subtests that cover elements of memory (Short-term, Working, Learning, Retrieval), executive control (Attention, Meta-awareness), and reasoning (Fluid, Verbal, Quantitative, Visual, Sociocultural). The content and presentation of each subtest of Alif–Ya was designed to be culturally and linguistically appropriate, reflective of the distinctive features of the Arabic language and its dialects, and of the demographic and geographical variability in the KSA and other Gulf countries. All content was developed collaboratively by the US-based and Saudi Arabian teams of researchers; all subtest visuals were created by artists. Figure 1 maps all of Alif–Ya's assessments and the broad cognitive indices they are designed to evaluate.

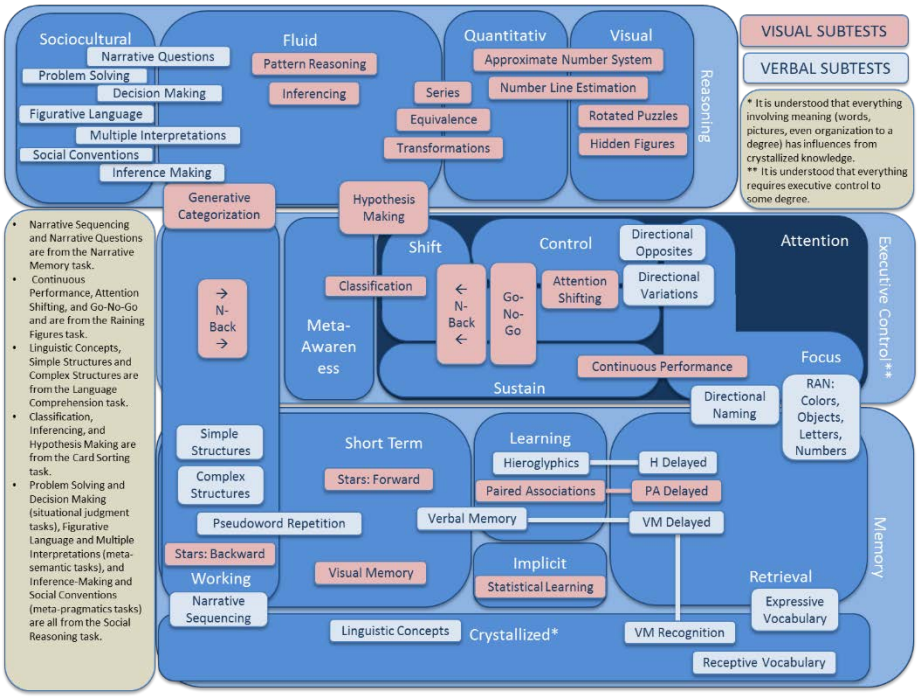

Figure 1. Alif–Ya's subtests and their contributions to the assessment of memory, executive control, and reasoning.

Alif–Ya is designed for the assessment of individuals aged 5–18. Although its primary intended purpose is to generate an overall IQ score, it is multidimensional and provides information about a number of neuropsychological domains that may contribute to the identification of giftedness and intellectual disability, and provides indicators of autism spectrum disorder, attention deficit disorder, language impairment, and impaired brain functioning (e.g., traumatic brain injury, epilepsy). Thus, Alif–Ya can give clinicians a broad

perspective for the interpretation of an individual's abilities by calculating a wide range of various indices of cognitive functioning. The battery includes a number of subtests for constructs that have only relatively recently been recognized as importantly contributing to academic and/or general intellectual functioning. Some examples are: number line estimation, the ability to translate quantities to estimate their relative positions in a linear representation, an indicator of children's representations of numerical magnitude [48]; the approximate number system (ANS), the rapid evaluation of relative quantities of objects, foundational to symbolic learning in mathematics [49]; and statistical or implicit learning, the domain-general ability to detect statistical regularities in information provided by the environment, a potentially key player in language acquisition [50]. In this aspect, Alif–Ya is in line with current innovative neuropsychological assessment development, such assessments for semantic memory [51,52]. In addition, Alif–Ya addresses a component of intelligence that is not part of most traditional assessment batteries yet has been defined as a key skill in non-Western world contexts: social reasoning. Knowing when an action should be taken, its social intent, and recognition of the action's cultural appropriateness are considered skills consistent with high intelligence in many non-Western cultures [53].

1.4. Technological Innovations of Alif–Ya

We briefly describe several of Alif–Ya's technologically innovative features here. A dual device system of delivery (originally 2 iPads) was devised for ease of administration, requiring a relatively limited amount of training for the clinician, and for portability. The system is usable in classrooms, clinical settings, or any other appropriate locations equipped with a wireless connection. The two tablet devices were originally paired locally using a Bluetooth connection. Later versions added remote capabilities to enable the clinician and student to communicate over an internet connection; specifically, using the user datagram protocol (UDP), messages can be sent between geographically distant devices. This allows for connectivity between various tablet, laptop, and desktop devices.

A number of innovations of item delivery have been implemented in Alif–Ya. Computerization ensures the precise and consistent delivery of items in terms of timings, clarity and tone of voice and pronunciation. In addition, a wide range of item types can be delivered, including multiple choice, multiple response, drag-and-drop open-ended (e.g., the child composes a response using shapes), and audio-delivered items that require a voiced response that is manually scored by the clinician.

There are multiple mechanisms for the delivery of clinical information to the clinician during testing. The dual device delivery accommodates the optimal choice of in-person delivery for the highest level of clinical observation during subtest administration. However, the range of device connections available also allows for the use of a third party connection (i.e., Zoom, Teams) so that screen sharing and child observation via a web-capable camera are possible. Computerization also allows the application to deliver real-time information to the clinician about a child's performance at the item level so that the child's performance can be monitored. Directly after an evaluation session, performance information is immediately available to the clinician in the form of scale and standard scores by subtest, and the usual composite (IQ) and index scores (memory, executive control, and reasoning).

Twenty-one of AlifYa's 47 subtests are constructed to deliver CAT. As noted above, CAT is characterized by its efficiency and accuracy in estimating test takers' ability over the traditional linear test. In order to enhance the accuracy of the ability estimate, the error related to the ability estimate needs to be reduced. The error associated with ability estimate is a function of item information introduced in item response theory (IRT), such that the higher information an item provides, the lower error the ability estimate would be. One way to increase the information yielded by a measure without increasing the length of the test is to tailor its delivery of items. Specifically, in a CAT framework, items are selected and administered one by one in a sequential order specific for each test taker. Each selected item represents the most informative item at the current stage of estimating. The item selection depends on previously administered items and the conditional ability

estimate. CAT involves an iterative administration of test items such that tailored items are selected and the ability estimate is continuously updated after each item is administered until reaching certain predefined stopping rules (e.g., reaching certain predefined accuracy level or certain number of items). The tailored testing allows for the use of shorter test length (efficiency) while maintaining precise information of ability estimate (accuracy).

The digital collection of data allows for a wide range of data to be collected with a high level of precision, such as timings of item response (in milliseconds) and response selection types. As all items and their responses may be characterized by a number of parameters, a wide range of process and contrast scores can be generated. These, along with the broad range of constructs that may be evaluated, enable unique skill profiles to be easily generated.

Finally, multiple instances of child performance data collected over time can generate longitudinal datasets. Additionally, ongoing collections create the potential for unlimited improvement through the generation, piloting and adding of new items to increase the precision and accuracy of the parameter estimates.

To illustrate the diagnostic potential of this newly developed battery and its programmed capabilities, we will present data collected with seven subtests as case examples. These subtests were applied to the same broadly sampled group of children (aged 5–18, from all parts of the Kingdom) and feature some of Alif–Ya's unique conceptual and technological innovations.

2. Materials and Methods

2.1. Sample and Materials

In the larger pilot study on Alif–Ya, a sample representative of the Kingdom was recruited in seven regions covering the central, eastern, western, northern and southern areas of the country. Several schools, kindergartens, and universities were selected for participation within both rural and urban areas of each region. Participant ages ranged from 5–18, as per the design of the test battery. Table 1 shows the number of students tested using Alif–Ya in each region.

Table 1. Alif–Ya sample of tested students by region.

Regions	Male	Female	Total
Riyadh	1928	2882	4810
Makkah	1686	1440	3126
Eastern Province	1710	1504	3214
Almadinah	710	593	1303
Aseer	365	266	631
Jazan	285	222	507
Aljawf	117	178	295
Total	6801	7085	13,886

The 47 subtests of Alif–Ya were divided into 9 batteries, each containing 3–8 subtests. Each test battery was designed to take about one hour. Subtests that had been designed to be adaptive and did not need all items to be taken by all participants were subdivided into pilot paths such that each student completed a set of items that had been determined by matching age groups to estimated item difficulty levels. Thus, in this planned missing design, younger students were assigned paths containing, broadly, easy to medium problems; older students were assigned paths containing medium to difficult items. In this report, we will examine the pilot data collected with Battery 1, which contained 7 subtests: approximate number system, equivalence, attention shifting, card-sorting classification, card-sorting hypothesis testing, card-sorting inference making, and narrative memory. Approximate number system, equivalence, card-sorting inference making, and narrative memory were delivered in a planned missing design. We will describe each subtest briefly here (subtest summaries are presented in Table 2).

Table 2. Subtest case examples.

Subtest Name	Broad Construct	Brief Description	Technological Features
Approximate number system	Reasoning	Student views two sets of dots simultaneously and decides which one contains more.	Precise timing of item delivery and measurement of response time
Equivalence	Reasoning	Given a set of shape equivalencies, student equates two sets of shapes by adding the needed shapes to one set.	Drag and drop response; precise measurement of response time
Attention shifting	Executive control	Student touches shapes in a moving stream of raining shapes according to different rules.	Animated delivery; precise measurement of response times
Card sorting-classification	Executive control	Student sorts a series of butterfly and fish cards by physical characteristics (e.g., color, shape and size).	Drag and drop response; precise measurement of response time
Card sorting-hypothesis testing	Executive control/reasoning	Student infers sorting rules for a series of butterfly and fish cards.	Drag and drop response; precise measurement of response time
Card sorting-inference making	Reasoning	Student determines the sorting rules for sets of butterfly and fish cards.	Drag and drop response; precise measurement of response time
Narrative memory	Reasoning/memory	Student listens to brief stories then arranges a set of 2–6 pictures in time order according to the narrated sequence of events. Two comprehension questions are asked about each story.	Digital audio delivery; drag and drop response; precise measurement of response times

2.2. Approximate Number System

In each trial of this task, the student is shown two fields of dots (4–20 each)—one of yellow dots, the other blue—for a brief period of time (600–200 ms). The student must then indicate which field showed more dots by touching the proper target button. The task contains 440 (220 × 2) trials or items. Item difficulty increases along a few parameters: (1) volume of dots; pairs less than 10 dots each are easier than those of more than 10 dots each; (2) the ratio of the dots, with lower ratios representing easier items (e.g., it is easier to distinguish 10 vs. 20 dots, 1:2, than 10 vs. 12 dots, 1:1.2); (3) area: number congruence, when fewer dots take up less area, vs. incongruence, when fewer dots take up more area; and the (4) time dots are shown on screen, which goes from moderately fast to rapid. There are three sizes of dots, close in size but distinguishable, in every trail. There are 440 possible items, delivered in an adaptive fashion.

2.3. Equivalence

The student is given a set of equivalencies (e.g., x = y, x = 2z), expressed not in numerals but in abstract shapes. Based on these given values, the student must deduce an incomplete equivalency (z = ?).

2.4. Attention Shifting

This task presents a variation of a continuous performance task by periodically changing the desired target. Thus, the student must monitor the appearance of a different shape or shapes in accordance with the designated time periods. In this case, shifting periods are designated by day-time and night-time indicators on the screen. During the "day" period, the student must search for a specific target; during the "night" period, the target changes to a completely different shape or set of shapes. The cue for this change is subtle and unannounced.

2.5. Card-Sorting

In these card-sorting tasks, students will be using the same set of cards that vary in two ways on 5 dimensions: shape (butterflies and fish), color (blue and red), size (big and small), stripes (stripes and no stripes), quantity (one and two). The students will use these cards to do three different tasks—classification, hypothesis-testing (based on the Wisconsin Card-Sorting Task, WCST) and inference-making (based on Concept Formation in the Woodcock Johnson). They will start with classification. If they proceed to the end of classification without reaching a ceiling, they will go on to hypothesis-testing. If they proceed to the end of hypothesis-making without reaching a ceiling, they will go on to inference-making. Each of these tasks will generate a sub-score for the student.

2.6. Narrative Memory

In this task, the student listens to a series of progressively longer and more complex readings that are structured as narratives (~20 total; 4–5 per child). The lowest level of difficulty starts with brief sentences, proceeding to 2–3 sentence vignettes, to longer stories (~200 words long). After listening to each story, the student will sequence a set of pictures in time order—that is, the order in which the events occurred in the time period of the story (not the order presented in the text)—to represent what they recall of the story. In addition, two main idea questions will be asked of the child to assess global comprehension of the text (highly related to memory). The difficulty level (complexity) of the text, the number of pictures that need to be sequenced, the type of distractor pictures (always 2), and the level of inference/prediction in the global questions constitute the parameters for the difficulty of the task.

2.7. Procedure

In the large-scale pilot exercise carried out with Alif–Ya, each student was administered one Alif–Ya test battery by a trained clinician. All clinicians had completed their Bachelor's degrees in Psychology. They were recruited from all test regions and were required to take a 4-day training course (6 h per day) to thoroughly understand the purpose and design of Alif–Ya, and to learn how to administer all 47 of Alif–Ya's subtests. Ultimately, 531 clinicians participated in the data collection (288 females). Clinicians then worked in the schools in their home regions. All tests were administered at the school sites. Testing was carried out in person using the dual iPad system. Clinicians selected children of the required ages at random from school lists, then contacted the parents of these children for consent.

2.8. Analytical Plan

Here we report only on the analyses and results for the subtests in Battery 1: approximate number system, equivalence, attention shifting, card-sorting classification, card-sorting hypothesis testing, card-sorting inference making, and narrative memory. We scored these subtests in different ways depending on whether the subtest was adaptive or not. Approximate number system, equivalence, card sorting-inference making, and narrative memory, had all been administered using a planned missing approach, in which different age groups received age-appropriate items according to their estimated ability, with overlap of items between age groups when possible. Items were calibrated by conducing IRT analysis. IRT is built upon the assumption that the probability of a respondent passing an item is a function of the respondent's position on the latent trait continuum and the item properties [54]. In the current study, we employed the 2PL model which includes two item parameters—slope and difficulty. The slope parameter represents the degree to which an item distinguishes individuals of different estimated ability. The difficulty parameter reflects the trait level where the probability of an individual passing that item is 0.50. We computed IRT-based person scores using the expected a posteriori estimator (EAP). The 2PL IRT models and personal score estimates were conducted using the 'mirt' package [55] embedded in the R environment [56].

The attention-shifting test generated three types of measures including number of targets (the shape that should be touched) touched, number of commission errors, and number of omission errors. Errors of commission occur when the test-taker touches a shape that is not the target. Errors of omission occur when the test-taker fails to touch the target. These counts constitute the scores.

Regarding card sorting-classification, when the test-taker moves a card (sorts it) to the correct pile, one point is earned. The sum of the correct responses for the 24 items constitutes the classification score. In the card sorting-hypothesis testing task, two types of scores are reported: the total number of incorrect sorts and the number of perseverative errors. Perseverative errors occur when the test-taker continues to sort according to the same rule (for example, by color) even after a rule change has been indicated.

Descriptive statistics, including the sample mean, standard deviation, minimum, maximum, skewness, and kurtosis, were derived using the base program embedded in R [56]. We also plotted score distribution for each measure in Battery 1. Pearson correlation coefficients were generated to examine the associations between age and the 10 measures.

We use the equivalence subtest to demonstrate how to prepare a subtest for adaptive administration through a simulation study. Simulation research for CAT allows researchers to evaluate the efficiency of the CAT approach as compared to the traditional linear test. The elements of CAT include an item bank with pre-calibrated items, the algorithm for selecting items, the process of estimating ability after each item is administered, the stopping rule, and the final estimation of ability. As described above, the equivalence subtest was calibrated with the 2PL IRT model. We then conducted the simulation using the catR package [57] embedded in the R environment [56]. We first generated the true abilities for 1500 cases according to the population parameters (M = 0, SD = 1). The first item was selected optimally at the ability level of 0 that is driven by Maximum Fisher information (MFI). The algorithm for selecting the next item given the current ability estimate and a set of previously administered items was MFI. We used EAP to estimate the ability after each item was administered. We manipulated the stopping criteria in different ways. In the first set of designs, the CAT process stopped when 10, 15, 20, 25, and 30 items were administered. Additionally, we manipulated the standard errors of measurement (SEM) as the stopping rule such that the CAT process would stop when reaching certain levels of SEM (0.15, 0.22, 0.28, and 0.31). The smaller the SEM, the greater precision the ability estimate would be. SEM can be converted to their corresponding Cronbach's alpha through $\alpha = 1 - \text{SEM}^2$ [58]. Hence, the corresponding Cronbach's alpha for the four SEM as stopping rule was 0.98, 0.95, 0.92, and 0.90. The final ability was estimated with EAP.

To evaluate the performance of different stopping criteria, we examined the overall ability estimation bias, root mean squared error (RMSE), and accuracy. The estimation bias referred to average difference between the CAT estimated and true ability levels. RMSE referred to the square root of the average of squared differences between the CAT estimated and true ability levels. Accuracy referred to the correlation between the CAT estimated and true ability levels. The code and data (both real and simulated) are available in an open-data repository hosted in the KSA.

3. Results

A total of 1931 participants (52.6% girls; M_{age} = 12.07 years, SD = 3.95) completed at least one subtest of Battery 1. Four hundred and thirty-one participants had no missing data on Battery 1; 433 had one missing subtest, 286—two, 233—three, 140—four, 79—five, 85—six, 51—seven, 156—eight, and 47 had nine missing subtests.

Table 3 presents the sample descriptive statistics for each measure in Battery 1. Figure 2 shows the distribution of each subtest score. It is notable that there are multiple extreme values either on the lower tail or the higher tail of the distributions, reflecting a broad range of ability among our sample. For example, one student scored 3.84 standard deviations below the average on the approximate number system subtest, indicating possible deficits

in quantitative reasoning. In contrast, another student scored 2.64 standard deviations above the average on equivalence, indicating potential giftedness.

Table 3. Descriptive statistics by subtest case examples.

Measures	n	M (SD)	Min	Max	Skewness	Kurtosis
1. Approximate number system	1636	0.00 (0.79)	−3.86	1.29	−0.90	0.47
2. Equivalence	1297	0.00 (0.91)	−1.88	2.64	0.10	−0.73
Attention shifting						
3. Target	1432	131.01(20.43)	28.00	154.00	−1.92	3.83
4. Commission	1380	11.63 (15.21)	0.00	141.00	3.34	15.24
5. Omission	1451	15.55 (9.93)	0.00	74.00	1.69	4.27
6. Card sorting-classification	1608	22.57 (1.75)	11.00	24.00	−2.00	5.65
Card sorting-hypothesis testing						
7. Errors	1341	9.22 (4.04)	1.00	23.00	0.43	−0.33
8. Perseverative responses	1341	15.99 (5.70)	5.00	33.00	0.49	−0.52
9. Card sorting-inference making	1573	0.00 (0.88)	−1.56	2.65	0.47	−0.48
10. Narrative memory	1084	0.00 (0.65)	−1.01	2.12	0.83	0.24

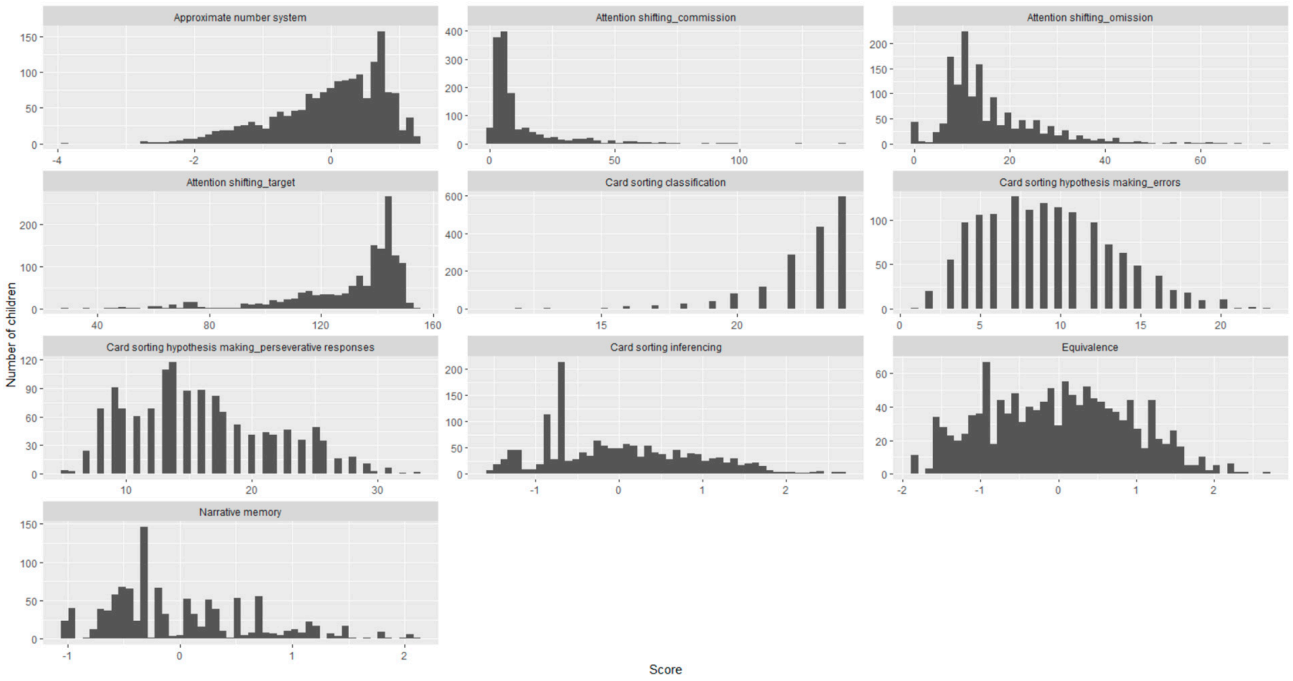

Figure 2. Score distributions of subtest case examples.

Table 4 presents the Pearson correlation coefficients between age and subtest scores. Age was positively correlated with approximate number system, equivalence, attention shifting-target, card sorting-classification, and card sorting-inference making scores. Age was negatively correlated with attention shifting-commission, attention shifting-omission, and card sorting-hypothesis testing scores. Notably, no correlations were found between age and performance on the card-sorting hypothesis-testing scores, nor between age and narrative memory. Regarding the associations between the 10 measures, the small-to-moderate effect sizes of correlations provided evidence of discriminant validity.

Finally, we present the results of the simulation study carried out for one of the planned adaptive tests, equivalence (see Table 5). This illustrates how the data collected in the pilot exercise was used to generate the adaptive algorithm for item selection, to be implemented in the final adaptive version of the test. The results showed that compared to

using test length as the stopping rule, accuracy was higher and the RMSE lower when using precision as the stopping rule. The simulation results favored stopping the CAT process when a certain pre-defined precision level of ability estimate had been reached. Moreover, as compared to the stopping rule of SEM = 0.22 and 0.28, the accuracy was identical, and the RMSE was slightly lower when the stopping rule was set to SEM = 0.31. Importantly, only 14.30 items were needed on average to reach a SEM of 0.31 (Cronbach's alpha = 0.90), suggesting the efficiency and accuracy of developing CAT for the equivalence subtest.

Table 4. Pearson correlation coefficients between age and subtest case examples.

Variable	1	2	3	4	5	6	7	8	9	10
1. Age										
2. Approximate number system	0.24 **									
3. Equivalence	0.41 **	0.24 **								
4. Attention shifting-target	0.38 **	0.20 **	0.27 **							
5. Attention shifting-commission	−0.47 **	−0.24 **	−0.25 **	−0.60 **						
6. Attention shifting-omission	−0.41 **	−0.21 **	−0.27 **	−0.72 **	0.54 **					
7. Card sorting-classification	0.43 **	0.29 **	0.35 **	0.28 **	−0.31 **	−0.27 **				
8. Card sorting-hypothesis testing-errors	−0.08 **	−0.11**	−0.20 **	−0.11 **	0.11 **	0.15 **	−0.19 **			
9. Card sorting-hypothesis testing-perseverative responses	−0.05	−0.08 **	−0.16 **	−0.09 **	0.09 **	0.13 **	−0.15 **	0.78 **		
10. Card sorting-inference making	0.21 **	0.14 **	0.46 **	0.17 **	−0.17 **	−0.15 **	0.20 **	−0.17 **	−0.15 **	
11. Narrative memory	0.04	0.11**	0.28 **	0.08 *	−0.05	−0.03	0.09 **	−0.09 **	−0.13 **	0.28 **

Note. * $p < 0.05$, ** $p < 0.01$.

Table 5. Accuracy and efficiency of CAT for the equivalence subtest.

Design Number	Test Length	SEM	Accuracy	RMSE	Bias	Items Used
The stopping rule was determined by test length.						
1	10		0.86	0.54	−0.02	
2	15		0.85	0.58	−0.04	
3	20		0.85	0.59	−0.03	
4	25		0.85	0.60	−0.04	
5	30		0.85	0.60	−0.05	
The stopping rule was determined by SEM.						
6		0.15	0.91	0.47	−0.07	60.80
7		0.22	0.88	0.53	−0.04	29.86
8		0.28	0.88	0.54	−0.06	19.17
9		0.31	0.88	0.52	−0.05	14.30

Note. SEM = standard errors of measurement, RMSE = root mean squared error.

4. Discussion and Conclusions

To illustrate the diagnostic potential of this newly developed battery and its programmed capabilities, we have presented data collected with seven subtests (generating 10 meaningful scores) as case examples of the utility of some of Alif–Ya's unique conceptual and technological innovations. The analyses of these preliminary data demonstrate some promising results. First, Table 3 and Figure 2 show the descriptive statistics and score frequency distributions for each subtest. These distributions reflect generally expected, reasonable values, indicating appropriate digital delivery of the subtests. We purposely did not remove outliers as we expect future data from more cognitively diverse representative samples to be added to these accumulating data. That is, we assumed that the current range of abilities found in our sample are part of the Saudi Arabian population range, although we did not deliberately include children with known diagnoses. Thus, the presence of

outliers in the card-sorting inference-making and narrative memory subtests also warrant further exploration of children's performance across subtests, when such data are available.

Second, most subtests showed low to moderate correlations with age, in the appropriate direction; significant absolute values ranged from 0.08 to 0.47 ($p < 0.01$), a positive indicator of the digitized subtest performance. Two scores that reflect no correlation with age were perseverative responses in the card-sorting hypothesis-testing subtest and narrative memory (correct sequencing of pictures that illustrate the narrated story). Regarding the former result, a similar lack of correlation was found in the norming exercise of the translated paper version of the WCST in Saudi Arabia. Specifically, the WCST was administered to 198 native Saudi Arabians aged 16–65, and results showed a similar lack of correlation between age and number of correct responses [12]. Thus, the mode of delivery of the subtest may be less a concern than the difficulty of the task itself across ages. Both possibilities will be examined in follow-up studies. Similarly, the narrative memory subtest showed mixed difficulty across ages. A post-hoc correlation study of performance by age groups receiving the exact same items revealed a persistent lack of association between age and performance, indicating a potential mis-estimation of item difficulty on our part, or operational difficulties of the subtest itself. These should be addressed in a further study in which all items are taken by all students, so that difficulty levels can be better estimated; this was not possible in our time-constrained pilot exercise.

Third, the simulation study showed some proof of the concept that data collected with the programmed subtest on equivalence could be used to yield conceptually valid parameters for the accurate and efficient determination of an individual's ability using CAT. Simulations conducted for all of Alif–Ya's 21 potentially adaptive tests have shown largely similar results. More importantly, as more data are collected, item parameters based on cumulative data will become increasingly refined for more accurate performance of the CAT.

In addition, while limited by the lack of diagnosed individuals at both ends of the ability spectrum in our sample and by the implemented missing data approach, the initial results reported here reflect the positive potential of our computerized assessment, Alif–Ya, to assess a broad range of cognitive processes accurately, efficiently, and easily with trained clinicians, in multiple clinical arrangements. Beyond the immediate and clear benefits of computerized administration, data collection and scoring, Alif–Ya presents several notable innovations prescient for the future of assessment. Regarding its ability to support the proper characterization and diagnosis of children, and monitor their development under intervention, Alif–Ya's 47 subtests can be composed and administered in a versatile array of combinations to support many clinical and other applied purposes. These include the characterization of autism spectrum disorder, learning difficulties, and thought disturbances in the domains of language, reading, and mathematics. This is in addition to the standard sets of subtests that contribute to full-scale, verbal and non-verbal IQ, and the various typically derived index scores for reasoning, executive control, and memory and their subcomponents. Alif–Ya additionally includes assessments of six components of social reasoning, and several assessments novel to the computerized environment, including approximate number system, number line estimation, and statistical learning.

In conclusion, Alif–Ya's capability to accommodate multiple assessment configurations—in-person and remote, using a combination of hardwares and operating systems—extends our capability to work with children under new and emerging circumstances. As we have learned through the experiences of the SARS-CoV-2 pandemic, children with special learning and clinical needs should be able to maintain access to their clinical providers for assessment and treatment across distances and unexpected constraints. Clinical assessments originally designed with such versatility in evaluation and administration help us address new conditions with flexibility without compromising quality of care. Such assessments may be particularly valuable in countries where the testing industry is not well-developed and translated assessments might not be culturally suitable. Altogether, Alif–Ya allows the utilization of more fine-grained, accurate information to inform clinicians of children's

ability levels in multiple dimensions and under diverse conditions. In these respects, Alif–Ya may be considered a pioneer in the computerized administration of assessments for the diagnosis and treatment of neurodevelopmental disorders.

Author Contributions: Conceptualization, E.G. and A.Q.; methodology, E.G. and A.Q.; structure of the assessment battery, L.H.; software, C.E. and M.G.; formal analysis, N.L.; writing—original draft preparation, M.T.; writing—review and editing, C.M. and N.L.; figure preparation, L.H. and N.L.; project administration, E.G.; funding acquisition, C.E., E.G. and A.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ministry of Higher Education of the Kingdom of Saudi Arabia, grant number G0500419.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from the parents/caregivers of all subjects involved in the study.

Data Availability Statement: The raw data and R scripts for the descriptive statistics, correlations, and simulation exercises described here can be found at the ETEC website maintained for open data. Please use this link to access this data (https://etec.gov.sa/en/Researchers/Pages/OpenData.aspx).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Alnahdi, G.H. The Extent of Applying the Rules and Guidelines of Assessment and Diagnosis Set Forth in the Regulations of Special Education Institutes and Programs for Students with Intellectual Disability. Master's Thesis, King Saud University, Riyadh, Saudi Arabia, 2007.
2. Wechsler, D.; Kodama, H. *Wechsler Intelligence Scale for Children*; Psychological Corporation: New York, NY, USA, 1949.
3. Terman, L.M.; Merrill, M.A. *Stanford-Binet Intelligence Scale: Manual for the third revision, Form L-M*; Houghton Mifflin: Oxford, UK, 1960; p. xi, 363.
4. Alnahdi, G.H. Special education programs for students with intellectual disability in Saudi Arabia: Issues and recommendations. *J. Int. Assoc. Spec. Educ.* **2014**, *15*, 83–91.
5. Al-Teriri, A. The intelligence scale for children in the Saudi environment. *Psychol. Stud. Ser.* **2004**, *8*, 109–139.
6. Kearney, C.A.; Smith, P.A.; Tillotson, C.A. Assessment and prediction of mental retardation in Saudi children. *J. Dev. Phys. Disabil.* **2002**, *14*, 77–85. [CrossRef]
7. Leiter, R.G. *General instructions for the Leiter International Performance Scale*; Stoelting: Chicago, IL, USA, 1969.
8. Beery, K.E. *Manual for the Developmental Test of Visual-Motor Integration*, 3rd ed.; Modern Curriculum Test: Cleveland, OH, USA, 1989.
9. Sparrow, S.S.; Balla, D.A.; Cicchetti, D.V. *Vineland Adaptive Behavior Scales*; American Guidance Service: Circle Pines, MN, USA, 1984.
10. Al-Ghatani, A.; Obonsawin, M.; Al-Moutaery, K. The Arabic version of the Stroop test and its equivalency to the English version. *Pan Arab J. Neurosurg.* **2010**, *14*, 112–115.
11. Golden, C.J. A group version of the Stroop colour and word test. *J. Person. Assess.* **1975**, *39*, 386–388. [CrossRef] [PubMed]
12. Al-Ghatani, A.; Obonsawin, M.C.; Binshaig, B.A.; Al-Moutaery, K.R. Saudi normative data for the Wisconsin card sorting test, Stroop test, test of non-verbal intelligence-3, picture completion and vocabulary (subtest of the Wechsler adult intelligence scale-revised). *Neurosci. J.* **2011**, *16*, 29–41.
13. Parker, D.M.; Crawford, J.R. Assessment of front lobe dysfunction. In *A Handbook of Neuropsychological Assessment*; Crawford, J.R., Parker, D.M., McKinlat, W.W., Eds.; Lawrence Erlbaum Association: Hillsdale, NJ, USA, 1992; pp. 267–291.
14. Wechsler, D. *WAIS-R Manual*; The Psychological Corporation: New York, NY, USA, 1981.
15. Brown, L.; Sherbenou, R.J.; Johnsen, S.K. *Test of Nonverbal Intelligence*, 3rd ed.; PRO-ED: Austin, TX, USA, 1997.
16. Abu-Hatab, F.; Zahran, H.; Mousa, A.; Khedr, A.; Yousef, M.; Sadek, A. The standardization of the standard progressive matrices in a Saudi sample. In *Studies on the Standardization of Psychological Tests*; Abu-Hatab, F., Ed.; Anglo-Egyptian Library: Cairo, Egypt; Volume 1, 1977; pp. 191–246.
17. Betterjee, A.A.; Khaleefa, O.; Ali, K.; Lynn, R. An increase of intelligence in Saudi Arabia, 1977–2010. *Intelligence* **2013**, *41*, 91–93. [CrossRef]
18. Rabin, L.A.; Spadaccini, A.T.; Brodale, D.L.; Grant, K.S.; Elbulok-Charcape, M.M.; Barr, W.B. Utilization rates of computerized tests and test batteries among clinical neuropsychologists in the United States and Canada. *Prof. Psychol. Res. Pract.* **2014**, *45*, 368–377. [CrossRef]
19. Raven, J.C.; Court, J. *Raven's Progressive Matrices*; Western Psychological Services: Los Angeles, CA, USA, 1938.

20. Conners, C.K.; Staff, M.; Connelly, V.; Campbell, S.; MacLean, M.; Barnes, J. Conners' continuous performance Test II (CPT II v. 5). *Multi-Health Syst Inc.* **2000**, *29*, 175–196.
21. Gerson, R.C.; Cella, D.; Fox, N.A.; Havlik, R.J.; Hendrie, H.C.; Wagster, M.V. Assessment of neurological and behavioural function: The NIH Toolbox. *Lancet Neurol.* **2010**, *9*, 138–139. [CrossRef]
22. Daniel, M. *Equivalence of Q-interactive™-Administered Cognitive Tasks: CVLT-II and Selected D-KEFS Subtests*; Pearson: Bloomington, MN, USA, 2012.
23. Williams, L.; Simms, E.; Clark, C.; Paul, R.; Rowe, D.; Gordon, E. The test-retest reliability of a standardized neurocognitive and neurophysiological test battery:"neuromarker". *Int. J. Neurosci.* **2005**, *115*, 1605–1630. [CrossRef]
24. Bilder, R.M. Neuropsychology 3.0: Evidence-based science and practice. *J. Int. Neuropsychol. Soc.* **2011**, *17*, 7–13. [CrossRef] [PubMed]
25. Bauer, R.M.; Iverson, G.L.; Cernich, A.N.; Binder, L.M.; Ruff, R.M.; Naugle, R.I. Computerized neuropsychological assessment devices: Joint position paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology. *Arch. Clin. Neuropsychol.* **2012**, *27*, 362–373. [CrossRef]
26. Witt, J.-A.; Alpherts, W.; Helmstaedter, C. Computerized neuropsychological testing in epilepsy: Overview of available tools. *Seizure* **2013**, *22*, 416–423. [CrossRef] [PubMed]
27. Weiss, D.J. Adaptive testing by computer. *J. Consult. Clin. Psychol.* **1985**, *53*, 774. [CrossRef] [PubMed]
28. Sereci, S.G. Computerized Adaptive Testing: An Introduction. In *Measuring Up: Assessment Issues for Teachers*; Wall, J.E., Walz, G.R., Eds.; Pro-Ed, Inc.: Austin, TX, USA, 2003; pp. 685–694.
29. Moreno, K.E.; Hetter, R.D. Full Scale CAT-ASVAB Implementation: Challenges and Outcomes. In Proceedings of the 39th Annual Conference of the International MIlitary Testing Association, Sydney, Australia, 14–16 October 1997; p. 312.
30. Creeger, C.P.; Miller, K.F.; Paredes, D.R. Micromanaging time: Measuring and controlling timing errors in computer-controlled experiments. *Behav. Res. Methods Instrum. Comput.* **1990**, *22*, 34–79. [CrossRef]
31. Adams, K.M.; Heaton, R.K. Computerized neuropsychological assessment: Issues and applications. In *Computerized Psychological Assessment: A Practitioner's Guide*; Basic Books: New York, NY, USA, 1987; pp. 355–365.
32. Hoskins, L.L.; Binder, L.M.; Chaytor, N.S.; Williamson, D.J.; Drane, D.L. Comparison of oral and computerized versions of the word memory test. *Arch. Clin. Neuropsychol.* **2010**, *25*, 591–600. [CrossRef]
33. Daniel, M.; Wahlstrom, D. Raw-score equivalence of computer-assisted and paper versions of WISC–V. *Psychol. Serv.* **2019**, *16*, 213. [CrossRef]
34. Gonzalez, J.; Wiberg, M. *Applying Test Equating Methods*; Springer International Publishing: Geneva, Switzerland, 2017.
35. Na, S.D.; Burns, T.G. Wechsler Intelligence Scale for Children—V: Test review. *Appl. Neuropsychol. Child* **2016**, *5*, 156–160. [CrossRef]
36. Barak, A. Psychological applications on the internet: A discipline on the threshold of a new millenium. *Appl. Prev. Psychol.* **1999**, *8*, 231–245. [CrossRef]
37. Luxton, D.D.; Pruitt, L.D.; Osenbach, J.E. Best practices for remote psychological assessment via telehealth technologies. *Prof. Psychol. Res. Pract.* **2014**, *45*, 27–35. [CrossRef]
38. Floyd, R.G.; Farmer, R.L.; Schneider, W.J.; McGrew, K.S. Theories and measurement of intelligence. In *APA Handbook of Intellectual and Developmental Disabilities: Foundations*; Glidden, L.M., Abbeduto, L., McIntryre, L.L., Tasse, M.J., Eds.; American Psychological Association: Washington, DC, USA, 2021; pp. 385–424.
39. Boake, C. From the Binet–Simon to the Wechsler–Bellevue: Tracing the history of intelligence testing. *J. Clin. Exp. Neuropsychol.* **2002**, *24*, 383–405. [CrossRef] [PubMed]
40. Wasserman, J.D. A history of intelligence assessment: The unfinished tapestry. In *Contemporary Intellectual Assessment: Theories, Tests, and Issues*; Flanagan, D.P., McDonough, E.M., Eds.; The Guilford Press: New York, NY, USA, 2018; pp. 3–55.
41. Cattell, R.B. *Intelligence: Its structure, growth, and action*; Elsevier: New York, NY, USA, 1987.
42. Horn, J.L. Measurement of intellectual capabilities: A review of theory. In *Woodcock-Johnson Technical Manual*; McGrew, K., Werder, J.K., Woodcock, R.W., Eds.; Riverside Publishing: Itasca, IL, USA, 1991; pp. 197–232.
43. McGrew, K.S. The Cattell-Horn Theory of Cognitive Abilities: Past, present, and future. In *Contemporary Intellectual Assessment: Theories, Tests, and Issues*; Flanagan, D.P., Harrison, P.L., Eds.; Guilford Press: New York, NY, USA, 2005; pp. 136–181.
44. Nettelbeck, T.; Wilson, C. Intelligence and IQ: What teachers should know. *Educ. Psychol.* **2005**, *25*, 609–630. [CrossRef]
45. Elliott, J.G.; Resing, W. Can intelligence testing inform educational intervention for children with reading disability? *J. Intell.* **2015**, *3*, 137–157. [CrossRef]
46. Luria, A.R. *Human Brain and Psychological Processes*; Harper & Row: New York, NY, USA, 1966.
47. Naglieri, J.A.; Das, J.P.; Goldstein, S. Planning, attention, simultaneous, successive: A cognitive-processing-based theory of intelligence. In *Contemporary Intellectual Assessment: Theories, Tests, and Issues*; Flanagan, D.P., Harrison, P.L., Eds.; The Guilford Press: New York, NY, USA, 2012; pp. 178–196.
48. Campbell, J.I.D. Development of numerical estimation: A review. In *Handbook of Mathematical Cognition*; Campbell, J.I.D. Psychological Press: New York, NY, USA, 2005; pp. 197–212.
49. Dehaene, S. Origins of mathematical intuitions. In *The Year in Cognitive Neuroscience 2009*; New York Academy of Sciences: New York, NY, USA, 2009; pp. 232–259.
50. Saffran, J.R.; Aslin, R.N.; Newport, E.L. Statistical learning by 8-month-old infants. *Science* **1996**, *274*, 1926–1928. [CrossRef]

51. Catricala, E.; Della Rosa, P.A.; Ginex, V.; Mussetti, Z.; Plebani, V.; Cappa, S.F. An Italian battery for the assessment of semantic memory disorders. *Neurol. Sci.* **2013**, *34*, 985–993. [CrossRef]
52. Navarro, M.C.; Marmolejo-Ramos, F.; Vásquez, V.; Carrea, B.; Vélez, J.I.; Mebarak Chams, M. An exploratory study for assessment of multimodal semantic memory in Colombian children. *Int. J. Psychol. Res.* **2020**, *13*, 49–58. [CrossRef]
53. Saklofske, D.H.; Van de Vijver, F.J.; Oakland, T.; Mpofu, E.; Suzuki, L.A. Intelligence and culture: History and assessment. In *Handbook of Intelligence: Evolutionary Theory, Historical Perspective, and Current Concepts*; Goldstein, S., Princiotta, D., Naglieri, J.A., Eds.; Springer: New York, NY, USA, 2015; pp. 341–365.
54. Embretson, S.E.; Reise, S.P. *Item Response Theory*; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2013.
55. Chalmers, R.P. mirt: A multidimensional item response theory package for the R environment. *J. Stat. Softw.* **2012**, *48*, 1–29. [CrossRef]
56. R Core Team. R: A language and environment for statistical computing. In *R: Foundation for Statistical Computing*; R Core Team: Vienna, Austria, 2021.
57. Magis, D.; Barrada, J.R. Computerized adaptive testing with R: Recent updates of the package catR. *J. Stat. Softw.* **2017**, *76*, 1–19. [CrossRef]
58. Moore, T.M.; Calkins, M.E.; Reise, S.P.; Gur, R.C.; Gur, R.E. Development and public release of a computerized adaptive (CAT) version of the Schizotypal Personality Questionnaire. *Psychiatry Res.* **2018**, *263*, 250–256. [CrossRef]