IACAT 2022

8th Conference of the International Association for Computerized Adaptive Testing

Frankfurt am Main, Germany | September 20-22, 2022 | Hosted by Goethe University Frankfurt

Program and Conference Information

https://www.iacat2022.com

@IACAT_org

@adaptivetesting

International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Contents

Welcome Message from the
IACAT President Alina A. von Davier

Dear friends and colleagues,

On behalf of the International Association for Computerized Adaptive Testing, I am delighted to welcome you to the 8th IACAT conference, taking place at the Goethe University, Frankfurt am Main, Germany.

It is wonderful to welcome you all in person after our rhythm of biennial conferences was interrupted! Now we are back in force with an interesting program and a diverse group of attendees!

Over the years, computerized adaptive testing has gained more and more attention from assessment, organizational psychology, and more recently, the learning and EdTech communities. With the advances in technology and the increase in the test takers' expectations for an innovative, personalized, and delightful learning and testing experience, the adaptivity of the tests prove to be the cornerstone of this century's education.

The IACAT conferences have become the premier event where you meet people interested in adaptive learning and testing, learn about advances in theory, research, and applications, and a place where professionals, scholars, and students from all parts of the world share knowledge and experiences. What makes the conference especially powerful is the range of topics and activities—including the pre-conference workshops, keynotes, and all the presentations contributing to IACAT's mission to advance the science of adaptive learning and testing by incorporating innovations, such as computational models for test development and psychometrics. Please explore our rich program for this year's event!

In 2014, IACAT established a prestigious award to further encourage young scholars in sharing their research. We are impressed by the interest of early career researchers and students and the quality of their submissions. Moreover, IACAT's official peer-reviewed electronic journal, the Journal of Computerized Adaptive Testing (JCAT) also provides an excellent avenue to publish research.

Finally, I would like to sincerely thank our host, Professor Andreas Frey (Chair), the local organizing committee, the Board members, and our sponsors for their individual and collective contributions without which this conference would not have been possible.

With gratitude,
Alina A. von Davier
President

Welcome Message from the Conference Host Andreas Frey

Dear Colleagues and Conference Delegates,

I am delighted that we can host the IACAT conference in 2022 as an in-person conference. It is an honor to welcome the world's leading CAT researchers to Goethe University Frankfurt. Goethe University Frankfurt is one of the largest universities in Germany and has a strong international research record. Together with the Leibniz Institute for Research and Information in Education (DIPF), which is also located on Campus Westend, Frankfurt is one of the most significant locations in the German-speaking countries when it comes to educational measurement, psychological measurement, and psychometrics.

It is really excellent to see how many people from a large range of countries are conducting research on CAT at a high level and are willing to share their findings at the conference. It is not at all an exaggeration to state that the 2022 program is the best and most comprehensive account of CAT research you can currently find worldwide. Connected to this, we are pleased to be able to support 10 PhD students from around the globe with stipends so that they can attend the conference. This support of young scholars serves as a basis to further foster CAT research and development in the coming years.

In addition to numerous contributions based on the traditional understanding of CAT as an IRT-based method to measure individual differences, the program also includes several presentations that bridge the gap to artificial intelligence on the one hand and to adaptive systems that combine learning and testing on the other. In my estimation, the future development of CAT as a research and development area will largely depend on the extent to which we succeed in combining the traditional psychometric core of CAT with these current trends. This year's program indicates that the CAT community is moving in this forward-looking direction.

The complete organizing team is thrilled to have you here. We wish you a conference with interesting academic exchanges, stimulating social interactions and nice impressions of Frankfurt.

Andreas Frey
(Conference Host)

Organizing Committees

Local Organizing Committee

Andreas Frey (Chair), Goethe University Frankfurt, Germany

Christoph König, Goethe University Frankfurt, Germany

Anette Stache, Goethe University Frankfurt, Germany

Aron Fink, Goethe University Frankfurt, Germany

Svenja Hammerstein, Goethe University Frankfurt, Germany

Organizing Committee / IACAT Board of Directors

David J. Weiss, University of Minnesota (President Emeritus)

Alina von Davier, Duolingo (President)

Tony Zara, Pearson VUE (Vice President)

John Barnard, EPEC Pty Ltd (Past President)

Duanli Yan, ETS (Secretary)

Kathleen Gialluca, Pearson VUE (Treasurer)

Alper Şahin, Atilim University (Director of Education)

Nathan Thompson, ASC (Membership Director)

Cliff Donath, Philips Healthcare (Executive Director)

Theo Eggen, University of Twente (Board Member)

For your Orientation

Free Public Transportation

You will receive a ticket for free public transportation in Frankfurt including Frankfurt International Airport. This ticket is already covered by your registration fee. With the ticket, you can use all means of transport such as the suburban railway (S-Bahn), the subway (U-Bahn), the bus, and the tram free of charge. Thus, you only have to pay for the public transport journeys you make before you pick up your conference material.

Public Transportation to the Conference Venue

From Frankfurt Airport (FRA)

Take the suburban railway (S-Bahn) lines S8/S9 departing from the station "Frankfurt (Main) Flughafen Regionalbahnhof" (located in Terminal 1, Level 0) to the city center. Ticket vending machines are located on the platform. They accept cash (Euro), credit cards, and Maestro/VPay cards (for people with a European account). Change to the subway (U-Bahn) lines U1, U2, U3, or U8 at the station "Hauptwache" and travel three stops in the northbound direction to "Holzhausenstraße". Then walk approximately 600 meters to the conference venue. The complete journey takes about 30-35 minutes.

From Frankfurt Central Station

Take the suburban railway (S-Bahn) lines S1, S2, S3, S4, S5, S6, S7, S8, or S9 to the stop "Hauptwache" and change there to the subway lines U1, U2, U3, or U8 to the stop "Holzhausenstraße". Then walk approximately 600 meters to the conference venue. The complete journey takes about 20 minutes.

Another option is to take bus line 64 from Frankfurt Central Station in the direction of "Frankfurt (Main) Ginnheim" and disembark at "Bremer Strasse". This bus stop is about 600 meters from the conference venue.

From other places in and around Frankfurt

All public transportation in the Rhein-Main metropolitan area which includes Frankfurt is organized by the Rhein-Main-Verkehrsverbund (RMV). The RMV's trip planning service (including a link to download their app) is available here.

Note that **Google Maps is not very useful for searching public transport connections in Frankfurt** because most of the lines are not listed there.

COVID-19 Regulations

Due to the COVID-19-pandemic, you have to use medical masks to cover your mouth and nose while using public transport (busses and trains). These include surgical masks or virus-filtering masks complying with the FFP2, KN95 and N95 standards.

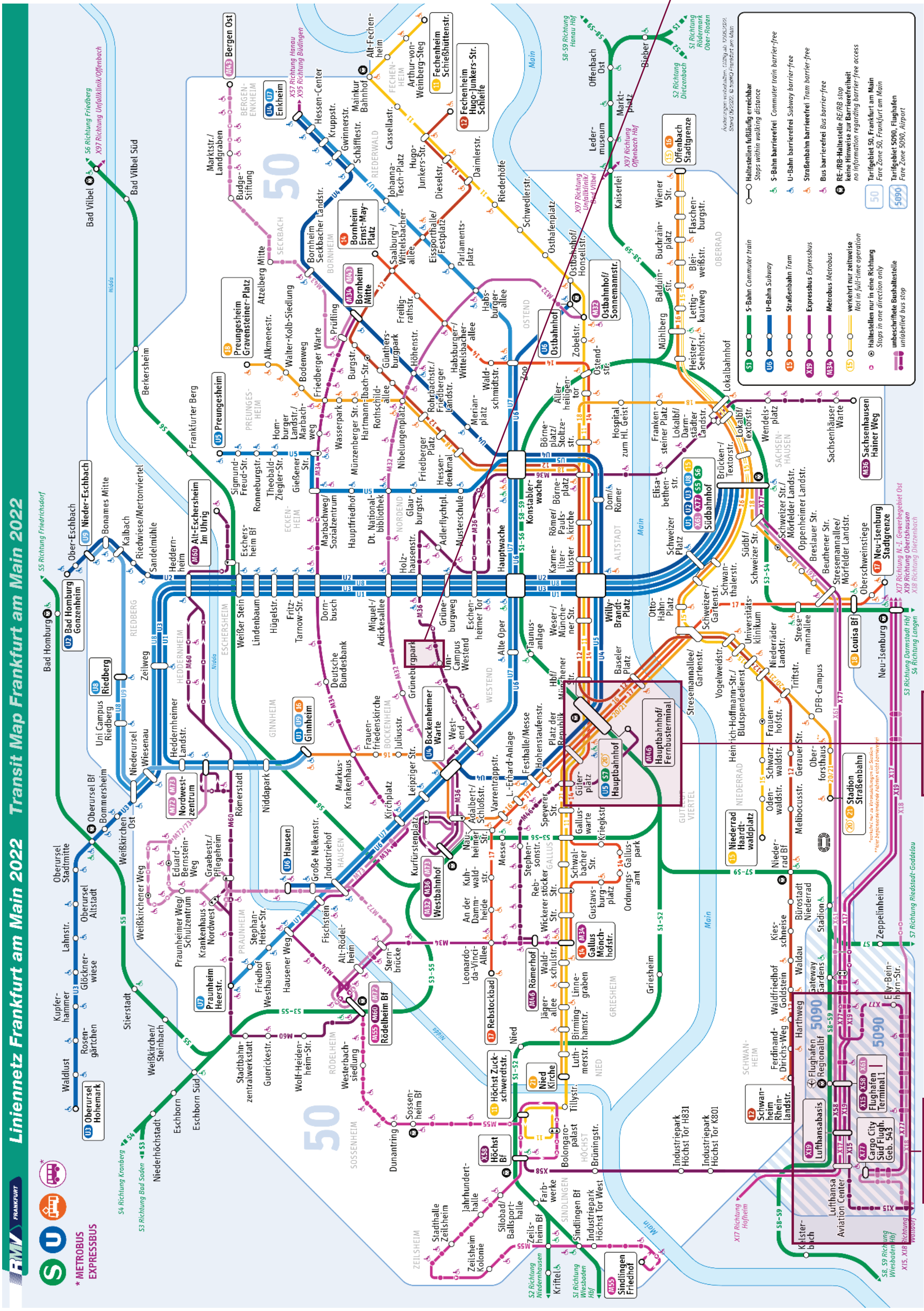
Taxi

TAXI FRANKFURT:	Dial +49.69.23 00 01
TAXI 33 Echo-Funk:	Dial +49.69.23 00 33
App:	Taxi Deutschland, available for iOS and Android

Public Transportation Information

Transit map Frankfurt

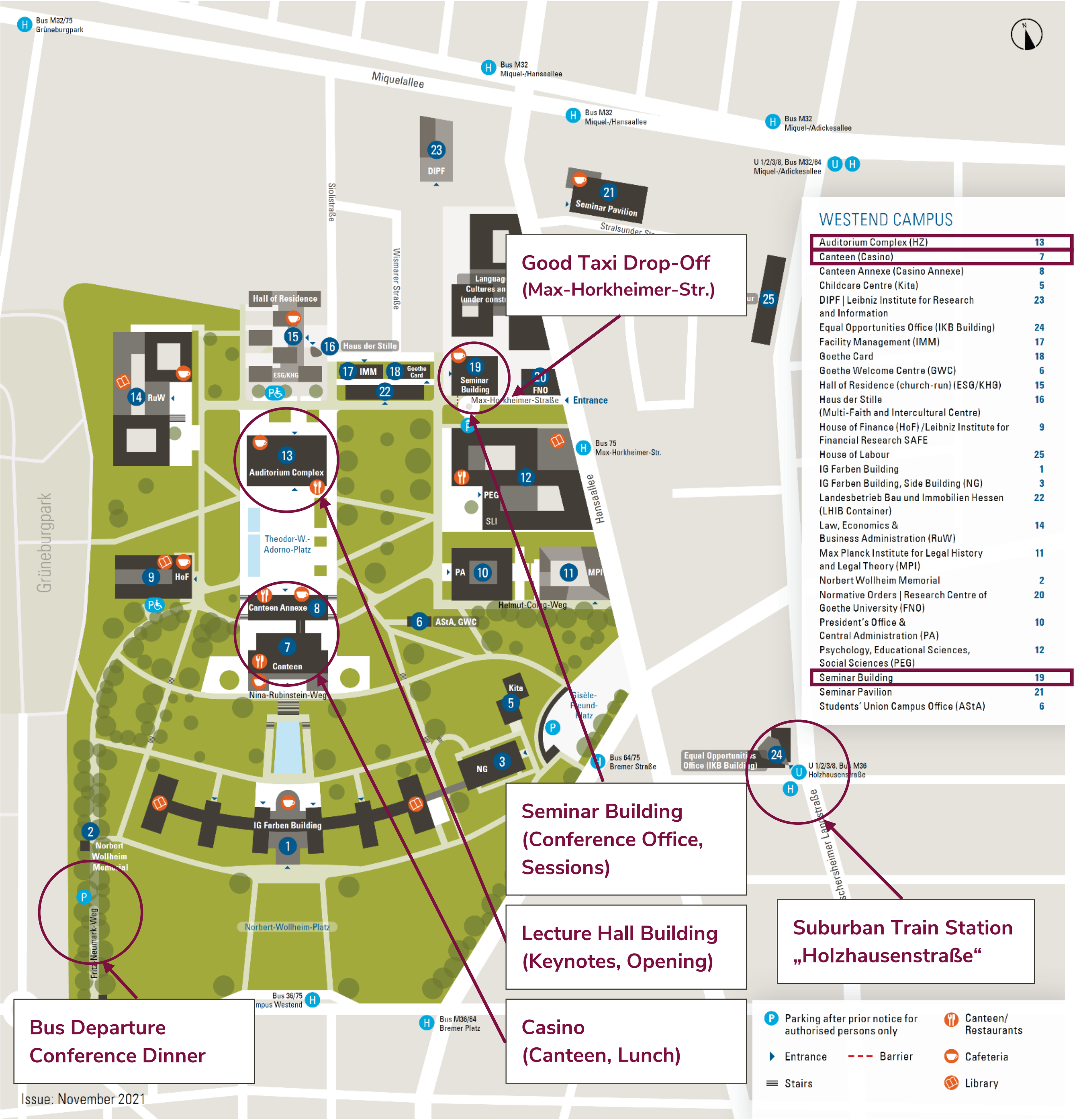
Conference Venue
Campus Westend


Liniennetz Frankfurt am Main 2022
Transit Map Frankfurt am Main 2022
S1 S-Bahn Commuter train
U6 U-Bahn Subway
15 Straßenbahn Tram
X19 Expressbus Expressbus
M34 Metrobus Metrobus
verkehrt nur zeitweise
Not in full-time operation
Haltestellen nur in eine Richtung
Stops in one direction only
unbeschilderte Bushaltestelle
unlabelled bus stop
Haltestellen fußläufig erreichbar
Stops within walking distance
S-Bahn barrierefrei Commuter train barrier-free
U-Bahn barrierefrei Subway barrier-free
Straßenbahn barrierefrei Tram barrier-free
Bus barrierefrei Bus barrier-free
RE-/RB-Haltestelle RE/RB stop
keine Information zur Barrierefreiheit
no information regarding barrier-free access
Tarifgebiet 50, Frankfurt am Main
Fare Zone 50, Frankfurt am Main
Tarifgebiet 5090, Flughafen
Fare Zone 5090, Airport

Frankfurt
Central
Station

Frankfurt
Internat.
Airport

Conference Venue – Campus Westend



Good Taxi Drop-Off
(Max-Horkheimer-Str.)

Seminar Building
(Conference Office,
Sessions)

Lecture Hall Building
(Keynotes, Opening)

Suburban Train Station
„Holzhausenstraße“

Bus Departure
Conference Dinner

Casino
(Canteen, Lunch)

WESTEND CAMPUS

Auditorium Complex (HZ)	13
Canteen (Casino)	7
Canteen Annexe (Casino Annexe)	8
Childcare Centre (Kita)	5
DIPF | Leibniz Institute for Research and Information	23
Equal Opportunities Office (IKB Building)	24
Facility Management (IMM)	17
Goethe Card	18
Goethe Welcome Centre (GWC)	6
Hall of Residence (church-run) (ESG/KHG)	15
Haus der Stille (Multi-Faith and Intercultural Centre)	16
House of Finance (HoF) /Leibniz Institute for Financial Research SAFE	9
House of Labour	25
IG Farben Building	1
IG Farben Building, Side Building (NG)	3
Landesbetrieb Bau und Immobilien Hessen (LHIB Container)	22
Law, Economics & Business Administration (RuW)	14
Max Planck Institute for Legal History and Legal Theory (MPI)	11
Norbert Wollheim Memorial	2
Normative Orders | Research Centre of Goethe University (FNO)	20
President's Office & Central Administration (PA)	10
Psychology, Educational Sciences, Social Sciences (PEG)	12
Seminar Building	19
Seminar Pavilion	21
Students' Union Campus Office (AStA)	6

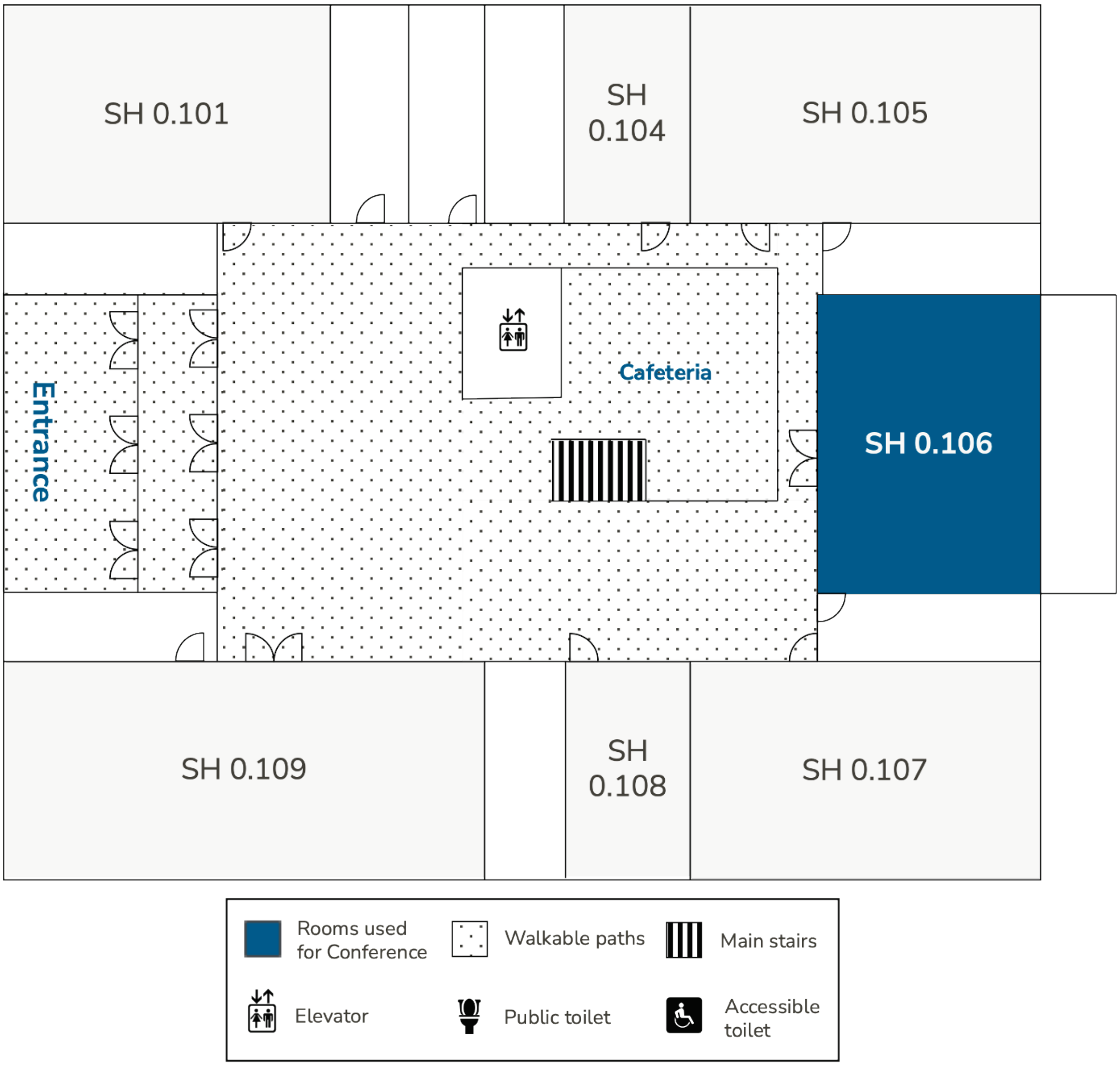
P Parking after prior notice for authorised persons only	Canteen/ Restaurants
▶ Entrance	--- Barrier
Cafeteria	
≡ Stairs	Library

Room Map

Building	Rooms used for Conference
Lecture Hall Building	Lecture Hall HZ 3
Seminar Building	Seminar Room SH 0.106 Seminar Room SH 1.105 Seminar Room SH 1.106 Seminar Room SH 1.107

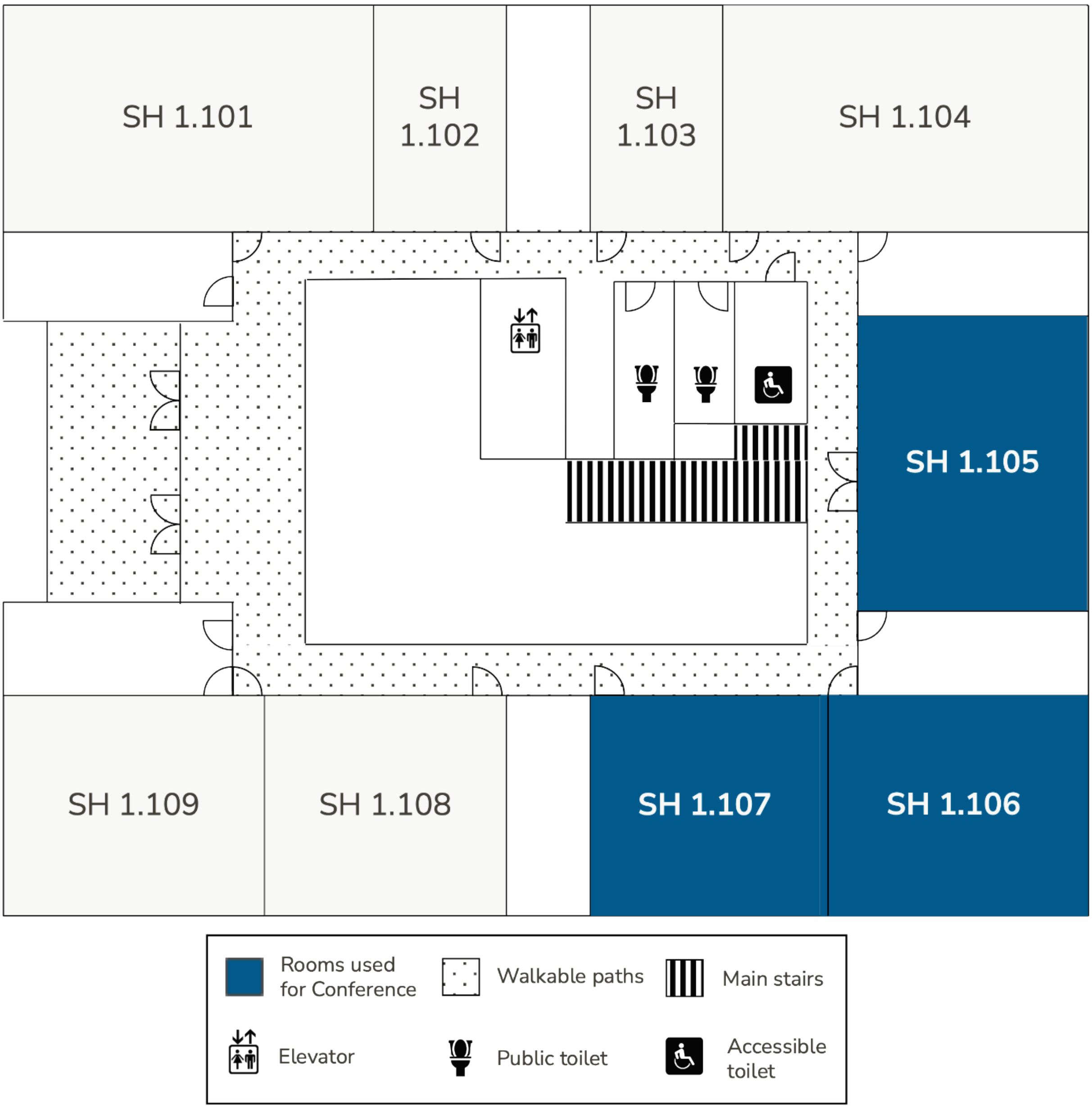
Seminar Building

Ground floor



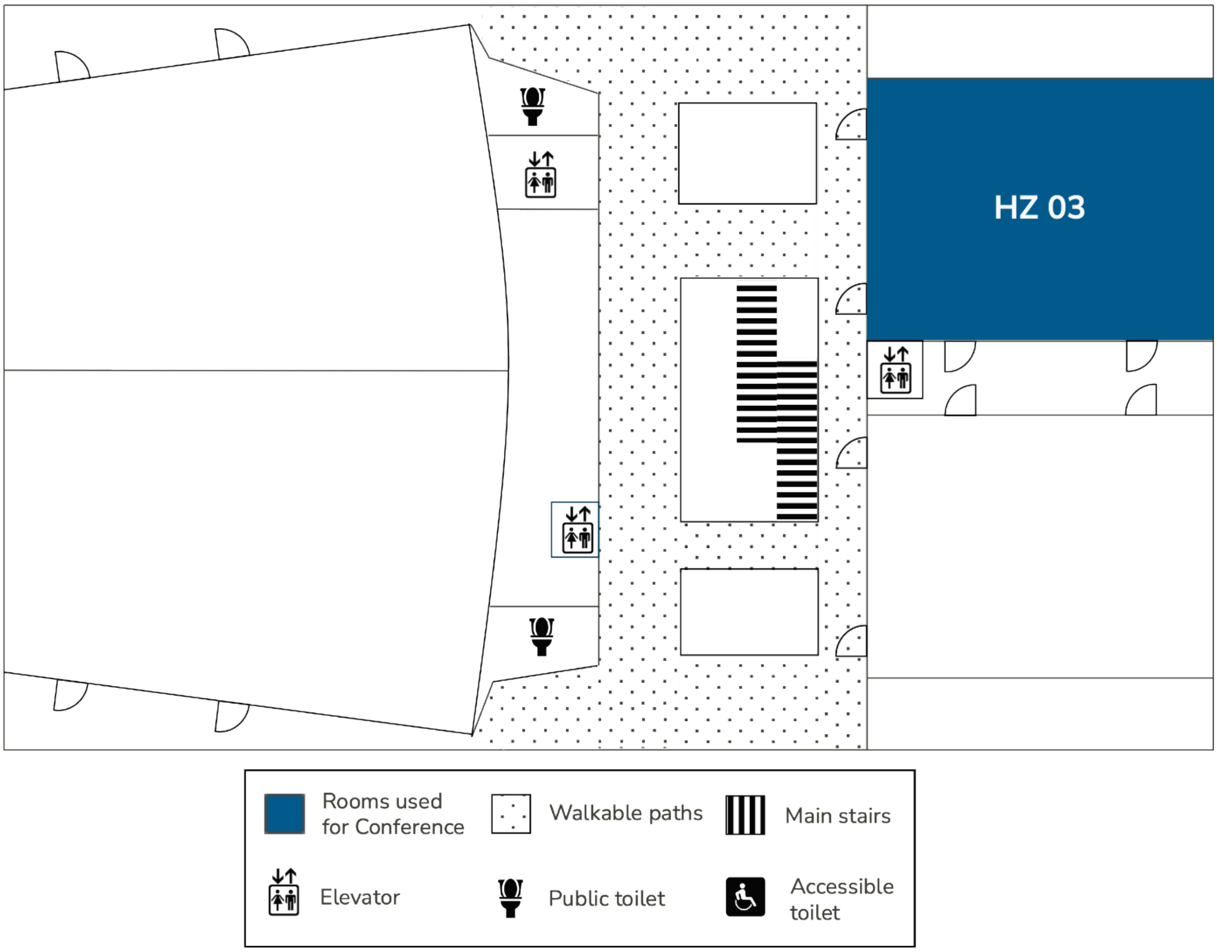
Seminar Building

First floor



Lecture Hall

First floor



General Conference Information

Conference Registration and Information

Ground floor of the Seminar Building, Campus Westend, Goethe University Frankfurt, Max-Horkheimer-Straße.

Desk hours:

Tuesday, September 20	.... 07:00 AM to 07:00 PM
Wednesday, September 21	.... 07:30 AM to 06:00 PM
Thursday, September 22	.... 08:00 AM to 07:00 PM

Wi-Fi

As a participant of the IACAT 2022 conference, you can access the internet wirelessly at the conference venue. The easiest way to connect is by using an existing eduroam account. You can login with the credentials of your eduroam account to all "eduroam"access points. You can find these access points virtually everywhere on Campus Westend of Goethe University Frankfurt. You can check whether your institution is part of the eduroam network here. We recommend that you check your eduroam login information before you leave for Frankfurt.

If you do not have an eduroam account and would like to use Wi-Fi on site, please contact the conference information desk, which is located on the ground floor of the seminar building. The staff will provide you with a conference account.

Name Badges and Tickets

Please collect your name badges together with your conference material at the conference information desk when you arrive at the conference venue. Name badges are required for admission to the conference events and should be worn at all times during the conference. If you registered for social events, please make sure to bring the separate tickets you received during the registration process to these events.

Cell Phone Policy

As a courtesy to all speakers and fellow attendees, please turn off cell phones or set them to silent mode during all sessions. Please do not answer calls in the seminar rooms when a session is in progress.

Catering

Catering is included in the conference fee.

Hot and cold drinks will be served in the foyer (ground floor) of the seminar building on all three conference days from 08:30 AM to 04:30 PM. Additionally, snacks will be available during the breaks, as indicated in the program.

Lunch will be served in the "Casino" building (the canteen). The casino can be reached from the Lecture Hall Building in less than 5 minutes (see Campus Map). You can choose anything you want from the casino's menu. There is always a variety of main and side dishes; desserts and soups can be added as desired. Use a small bowl for each side dish. You will receive coupons together with your conference material at the registration. You can pay with these coupons at the cash desk of the casino.

On Tuesday, September 20, from 5:30 PM to 07:00 PM we look forward to welcoming you to a reception in the foyer and, weather permitting, on the rooftop terrace (3rd floor) of the Lecture Hall Building.

COVID-19 Regulations

Traveling to and from Germany

The latest information about entry restrictions, quarantine regulations, and the COVID-19 testing regime can be found on the homepage of the German Federal Foreign Office. Please follow the link here.

COVID-19 regulations in Germany

The latest information about COVID-regulations in Germany (mask requirements, social distancing etc.) can be found here.

Public transport

Due to the COVID-19-pandemic, you have to use medical masks to cover your mouth and nose while using public transport (busses and trains). These include surgical masks or virus-filtering masks complying with the FFP2, KN95 and N95 standards.

Regulations on conference venue

Except for the recommendation to wear masks within the buildings, there are currently no COVID-19 related restrictions in place at the Goethe University Frankfurt. We will inform you at the conference should this change.

Social events

Welcome Reception

On Tuesday, September 20, 05:30 PM - 07:00 PM we look forward to welcoming you to a reception with beverages, drinks, and snacks in the foyer and, weather permitting, on the rooftop terrace located at the Lecture Hall Building, 3rd floor. The reception is included in the conference fee, a separate ticket is not necessary.

Guided Campus Tour

The Countess of Luxburg and her team will lead you through the university's campus where you can find out more about the history of the Goethe University Frankfurt during an entertaining tour on September 20, 11:00 am to 12:00 pm. A ticket needs to be purchased using the registration system.

Conference Dinner with Award Ceremony

The conference dinner will take place at the Eberbach Monastery – a Romanesque monastery church which is nicely located in vineyards above the river Rhine. There will be the option of a guided tour through the monastery with wine tasting. Tickets need to be purchased for both events using the registration system.

There will be bus transportation to Eberbach Monastery. Departure of the buses is on September 21, 2022, 5:00 pm at the parking lot in Fritz Neumann Weg. This parking lot is at the south-west end of Campus Westend and a ten minutes walk away from the seminar building. If you have registered for the conference dinner, you can directly go to the parking lot some minutes befor 5:00 pm. Alternatively, you can go to the meeting point at 4:45 pm in front of the seminar building. Staff members in blue shirts will guide you to the parking lot from there. The bus ride will take 45 to 60 minutes. We will be back at the University at 10:15 PM.

© Stiftung Kloster Eberbach

© Lisa Fakras

Recommended Leisure Activities and Restaurants

You can find a list with some recommendations for leisure activities and restaurants for your stay in Frankfurt and the surrounding area before, during, or after the conference in the list below:

Designation	Link/ URL	Activity	Price level	Food style	Location	How to get there	Details
Frankfurt Sightseeing	https://t1p.de/ikb8w	Excursion	€	n.a.	Frankfurt City center	From "Frankfurt Central station" with all suburban railways to city center. From campus with Subway in the direction of "Südbahnhof".	Take a guided tour of our rebuilt historic city. Book in advance.
Sightseeing by boat on River Main	https://t1p.de/j3903	Excursion	€	n.a.	Frankfurt City center	Boat trips depart from a bridge called "Eiserner Steg", 5 minutes from Subway station "Dom/Römer", with U4 or U5.	Cruise the River Main in Frankfurt for views of the city's skyline. See sights like the pubs of Sachsenhausen and the European Central Bank. Jump off at Gerbermühle and learn about the poet Goethe.
Palmengarten	https://t1p.de/vn5lo	Excursion	€	n.a.	Frankfurt Westend	By walking through the park from Campus Westend to the street "Siesmayerstrasse". Or, from city center with Subway U6 or U7 from station "Westend".	Stroll through tropical greenhouses and traditional gardens at the "Palmengarten", which is celebrating its 150th anniversary this year.
Museums	https://t1p.de/iclkb	Museum	€	n.a.	Franktfurt City center	Most of the 39 museums can be reached from the subway stations "Dom/Römer" or "Schweizer Platz"	Frankfurt is one of Germany's leading museum locations. Famous museums are: Schirn, Städel and MMK
Eppstein, Taunus Mountains	https://t1p.de/hbxb3	Excursion	€	n.a.	Eppstein, 30 km outside Frankfurt	From "Frankfurt Central station" with suburban railway S2 in the direction of "Niedernhausen", station "Eppstein". Half-hour journey to destination. (extra train ticket necessary).	Take a walk around medieval Eppstein in the Taunus mountains: A valley with wild and romantic rocks and well preserved castle ruins amidst a picturesque village.
River Rhine, Loreley Rock	https://t1p.de/1hybg	Excursion	€€	n.a.		From "Frankfurt Central station" with a regional express train, 1.5 h to Boppard. Or by car rental. Boat cruise tickets via the link.	Cruise along the Rhine river between vineyards to the Loreley, where, according to legend, the boatmen were bewitched and drowned by the singing of a blonde beauty.
Oper Frankfurt	https://t1p.de/tuyf	Music	€€	n.a.	City center	Subway U1, U2, U3, U4, U5, or U8 from "Hauptwache".	Classical opera program, concerts and recitals.
Alte Oper Frankfurt	https://t1p.de/m8dqx	Music	€€-€€€	n.a.	Frankfurt Taunusanlage, City center	Subway U6 or U7 from "Hauptwache" to "Taunusanlage". From "Frankfurt Central station" with suburban railway S1, S2, S3, S4, S5, S6, or S8 to "Alte Oper".	Classical opera, ballets, contemporary concerts. At the time of IACAT 2022, the master pianist Igor Levit will perform in Frankfurt. The concert will take place on September 20 in the Alte Oper Frankfurt. Buy tickets online here.
Shopping in Frankfurt	https://t1p.de/hkmnf	Shopping	€-€€€	n.a			Large variety of shops in the city centre at and around the road "Zeil". Exlusive shopping in "Goethestraße". Mall "My Zeil" (near station "Hauptwache") offers a wide range of shops. Smaller and more individual shops in "Bergerstraße" around subway station "Merianplatz".

Designation	Link/ URL	Activity	Price level	Food style	Location	How to get there	Details
Orfeos Erben	https://t1p.de/77oux	Dining	€€	finest regional	Frankfurt Bockenheim, near the fair.	Tram 16 (direction Ginnheim) or 17 (direction Rebstockbad) from Frankfurt Central station to "Varrentrappstrasse".	Restaurant, bar lounge. The chef runs a cinema in the same complex.
Bidlabu	https://bidlabu.de	Dining	€€€	gourmet	City center	Subway U6, U7 to "Alte Oper"	Very good food at reasonable price, relaxed athmosphere, book a table in advance
Villa Merton	https://t1p.de/rad2q	Dining	€€€	gourmet, French, German	Frankfurt Westend	From Campus Westend on foot (1.5 km) or by taxi (€10).	Art Nouveau villa in the diplomatic quarter. Awarded chef de cuisine André Grossfeld (one star by Michelin Guide).
Sushimoto	https://t1p.de/954p5	Dining	€€-€€€	Japanese	City center	Commuter train station "Konstablerwache", three stations with suburban railway (S-Bahn) from "Frankfurt Central station" and a short walk on foot.	Japanese sushi and wagyu beef at the chef's table, that is, Mr. Sakamoto's. Located in "Westin Grand Hotel".
Savanna	https://t1p.de/1jfzf	Dining	€	Eritrean, East Africa	City center	Suburban railway station "Konstablerwache", three stations with suburban railway from "Frankfurt Central Station", one from "Hauptwache" and a short walk on foot.	Eritrean traditional cuisine eaten with your fingers with the delicious injera bread. Best Eritrean address in town, very small location, reservation is a must.
Apfelwein Solzer	Berger Straße 260, 60385 Frankfurt, +49 69 452171	Dining	€	Traditional Frankfurt food	Frankfurt Bornheim	Subway U4, station „Bornheim Mitte", 5 min from there by foot	Traditional apple wine (speciality in Frankfurt area) restaurant, well attended and cozy with good traditional German food and good atmosphere, you may be seated together with other people at large tables
La Trinca	https://t1p.de/4snku	Dining	€€	Spanish, tapas	Frankfurt Sachsenhausen	Subway U1, U2, U3, U8 station „Schweizer Platz"	Nice spanish restaurant with delicious tapas and good selection of wine
Kleinmarkthalle	https://t1p.de/vuqyu	Gourmet shopping	€-€€€	regional, international	City center	Station "Hauptwache" with suburban railway (S-Bahn) from "Frankfurt Central Station". 5 minutes on foot in the direction of the Main river/ "Liebfrauenberg".	Vast range of regional and international delicatessen, wines and spices, and food to go.

For more tourist activities, please visit the website of the tourist information Frankfurt here.

Schedule at a Glance

The conference consists of workshops, keynote lectures, invited symposia, a submitted symposium and thematic paper sessions. Lunch, snacks, cold and hot drinks during the conference and the welcome reception on day 1 are covered by your registration fee.

The Schedule at a glance on the next three pages gives an overview over the IACAT 2022 conference. The sessions are numbered consecutively. The abstracts for the individual presentations within these sessions can be found in the pages following the Schedule at a Glance. The abstracts are ordered by the session numbers.

2022 IACAT Conference (Day 1)

8th Conference of the International Association for Computerized Adaptive Testing
September 20-22, 2022 - Frankfurt am Main, Germany

	Day 1: Tuesday, 20 Sept. 2022				
TIME	Lecture Hall HZ 3	Seminar Room SH 1.106	Seminar Room SH 1.105	Seminar Room SH 1.107	Seminar Room SH 0.106
8:00-10:00		01: Workshop **Computerized Adaptive Testing and Multistage Testing with R (Part 1)** Organizers: Duanli Yan, David Magis, & Alina A. von Davier	02: Workshop **Simulations and CAT (Part 1)** Organizers: Angela Verschoor, Theo Eggen, & Maaike van Groen	03: Workshop **Developing Online Adaptive Tests Using Open-Source Concerto Platform (Part 1)** Organizers: Luning Sun, Joe Watson & Aiden Loe	
10:00-10:30	Coffee Break				
10:30-12:30		01: Workshop **Computerized Adaptive Testing and Multistage Testing with R (Part 2)** Organizers: Duanli Yan, David Magis, & Alina A. von Davier	02: Workshop **Simulations and CAT (Part 2)** Organizers: Angela Verschoor, Theo Eggen, & Maaike van Groen	03: Workshop **Developing Online Adaptive Tests Using Open-Source Concerto Platform (Part 2)** Organizers: Luning Sun, Joe Watson & Aiden Loe	
12:30-1:30	Lunch				
1:30-2:15	**Welcome Ceremony** CIO, Dean, IACAT President, Conference Chair				
2:15-3:15	04: Incoming President Keynote **Anthony Zara: "Where Were All the Psychometricians?"** Chair: Mark Reckase				
3:15-3:45	Coffee and Snacks				
3:45-5:15		05: Invited Symposium **The CAT in the Language Assessments Bag** Chair: Alina A. von Davier	06: Paper Session **Large-Scale Assessments** Chair: Eveline Gebhardt	07: Paper Session **Test Termination** Chair: Alper Sahin	08: Paper Session **CAT for Admission, Selection & Certification** Chair: Bernard Veldkamp
5:30-7:00	**Welcome Reception** *(Rooftop Terrace, Lecture Hall Building)*				

2022 IACAT Conference (Day 2)

8th Conference of the International Association for Computerized Adaptive Testing
September 20-22, 2022 - Frankfurt am Main, Germany

	Day 2: Wednesday, 21 Sept. 2022				
TIME	Lecture Hall HZ 3	Seminar Room SH 1.106	Seminar Room SH 1.105	Seminar Room SH 1.107	Seminar Room SH 0.106
9:00-10:00	09: Keynote **Wim J. van der Linden: "The New Paradigm of Adaptive Testing"** Chair: Andreas Frey				
10:00-10:30	*Coffee Break*				
10:30-12:00		10: Invited Symposium **Computerized adaptive testing (CAT) for the measurement of health outcomes – the Patient-Reported Outcomes Measurement Information System** Chair: Caroline B. Terwee Discussant: Ulf Kroehne	11: Paper Session **Adaptive Measurement of Change** Chair: Theo J. H. M. Eggen	12: Paper Session **Automated Scoring in CAT with AI** Chair: Maaike van Groen	13: Paper Session **New Approaches to CAT** Chair: Samuel Greiff
12:00-1:00	*Lunch*			*IACAT Board of Directors Meeting (Restaurant Sturm & Drang, Lecture Hall Builiding)*	
1:00-2:00	14: Keynote **Ying Cheng: "Cognitive Diagnostic Computerized Adaptive Testing: Recent Developments and Future Directions"** Chair: Anthony Zara				
2:00-2:15	*Break (change building)*				
2:15-3:15		15: Invited Symposium **Adaptive testing in PISA: past, present and future - Part 1** Chairs: Janine Buchholz, Mario Piacentini, & Francesco Avvisati Discussant: Matthias von Davier	16: Symposium **Developing a system that connects learning and adaptive testing for adults learning to read - Part 1** Chair: John Sabatini Discussant: Samuel Greiff	17: Paper Session **Cognitive Diagnosis CAT** Chair: Miguel A. Sorrel	18: Paper Session **CAT applications: Personality Testing 1** Chair: Rodrigo Schames Kreitchmann
3:15-3:45	*Coffee Break*				
3:45-4:45		15: Invited Symposium **Adaptive testing in PISA: past, present and future - Part 2** Chairs: Janine Buchholz, Mario Piacentini, & Francesco Avvisati Discussant: Matthias von Davier	16: Symposium **Developing a system that connects learning and adaptive testing for adults learning to read - Part 2** Chair: John Sabatini Discussant: Samuel Greiff	19: Paper Session **CAT in Educational Contexts** Chair: Nathan Thompson	
5:00-6:00	*Bus transfer to Eberbach Monastery*				
6:00-7:00	*Strolling Wine Tasting (registration necessary) or Individual Tour of Monastery and/or Vineyard*				
7:00-9:30	***Award Ceremony & Conference Dinner***				
9:30-10:15	*Bus transfer to Goethe University*				

2022 IACAT Conference (Day 3)

8th Conference of the International Association for Computerized Adaptive Testing
September 20-22, 2022 - Frankfurt am Main, Germany

	Day 3: Thursday, 22 Sept. 2022				
TIME	Lecture Hall HZ 3	Seminar Room SH 1.106	Seminar Room SH 1.105	Seminar Room SH 1.107	Seminar Room SH 0.106
9:00-10:00	20: Early Career Award Winner Keynote **Miguel A. Sorrel**: **"On the diagnostic power of the items in a pool"** Chair: Alina A. von Davier				
10:00-10:30	*Coffee Break*				
10:30-12:00		21: Invited Symposium **Applications of CAT across multiple fields using the Concerto platform** Chairs: David Stillwell & Luning Sun	22: Paper Session **Test taking experience** Chair: Steven L. Wise	23: Paper Session **Item Calibration** Chair: Angela Verschoor	24: Paper Session **CAT Applications: Ability Testing** Chair: G. Gage Kingsbury
12:00-1:00	*Lunch*				
1:00-2:00	25: Keynote **Bernard P. Veldkamp: "The Double Helix of Adaptive Measurement"** Chair: Theo J. H. M. Eggen				
2:00-2:15	*Break (change building)*				
2:15-3:45		26: Invited Symposium: **Computerized adaptive practicing** Chair: Han L. J. van der Maas	27: Paper Session **Item Selection** Chair: Wim J. van der Linden	28: Paper Session **Usage of Prior Information & Software for CAT Development and Application** Chair: Yigal Attali	29: Paper Session **CAT applications: Personality Testing 2** Chair: Cliff Donath
3:45-4:15	*Coffee Break*				
4:15-5:15		30: Paper Session **Multi-Stage Testing** Chair: Duanli Yan	31: Paper Session **Threats to Validity: Engagement, Position Effects, and DIF** Chair: Muirne Paap	32: Paper Session **Statistical Foundations of CAT** Chair: Peter van Rijn	
5:15-5:30	*Break (change building)*				
5:30-6:00	**Closing Ceremony**				

Abstracts

01: Workshop – Computerized Adaptive Testing and Multistage Testing with R

Duanli Yan, David Magis, & Alina A. von Davier

Abstract: The goal of this workshop is to provide a practical (and brief) overview of the theory on computerized adaptive testing (CAT) and multistage testing (MST), and illustrate the methodologies and applications using R open-source language and several data examples. The implementations rely on the R packages catR and mstR that have been already or are being developed and include some of the newest research algorithms developed by the authors. This workshop will cover several topics: the basics of R, theoretical overview of CAT and MST, CAT and MST designs, assembly methodologies, catR and mstR packages, simulations, and applications. The intended audience for the workshop is undergraduate/graduate students, faculty, researchers, practitioners at testing institutions, and anyone in psychometrics, measurement, education, psychology, and other fields who is interested in computerized adaptive and multistage testing, especially in practical implementations of simulation using R.

02: Workshop – Simulations and CAT

Angela Verschoor, Theo Eggen, & Maaike van Groen

Abstract: This workshop explores the role that simulation studies play in CAT research and test development. Simulations are run by software programs that typically apply a CAT algorithm to a data set of fake (monte carlo) examinees, or real data from past assessments. This provides the researcher control over important algorithms such as items selection methods, exposure controls, and termination criterion, and therefore allows them to imagine and explore experimental designs to investigate aspects of their performance. Simulations are the only method available in CAT development that allow test developers to investigate whether that their CAT results in valid measurements. During the workshop, we will use a Windows software program to have a look at some CAT simulations. Participants are asked to bring their own laptop.

03: Workshop – Developing Online Adaptive Tests Using Open-Source Concerto Platform

Luning Sun, Joe Watson & Aiden Loe

Abstract: Despite their increasing prevalence, only a limited number of test publishers have the capability to develop adaptive tests. The University of Cambridge Psychometrics Centre strives towards making online adaptive testing available to everyone. That is why we've created Concerto: a powerful and user-friendly platform that empowers experts and beginners alike to make better tests, with little to no knowledge of coding experience required. There are minimum set-up costs, no licence fees and no limitations. Concerto harmonises the statistical power of the R programming language, the security of MySQL databases and the flexibility of HTML to deliver advanced online tests. These instruments work in unison, giving users unparalleled freedom and control over the design of their assessments. In-built algorithms for score calculation and report generation ensure a rewarding experience for participants, whatever the context. During this hands-on workshop, participants will learn how to build an online adaptive test using Concerto v5. We will start with an introduction to Concerto, build HTML-based item templates, import item content and parameters and combine them all into a fully-functional online adaptive test. Thus, the target audience for this workshop is for researchers with a basic understanding of Item Response Theory and with a keen interest in developing an online adaptive test for their own research purpose. Participants should bring their own laptops and make sure that the Internet connection is properly configured.

04: Incoming President Keynote - Anthony Zara: Where Were All the Psychometricians?

Chair: *Mark Reckase*

Abstract: The Testing industry was hit particularly hard by the COVID-19 pandemic, particularly high-stakes credentialing testing. The new government health directives caused computerized test centers to close and cease testing for a period of time. This left licensure and certification agencies without an avenue to assess eligible candidates to fulfill their missions to assure public safety through a comprehensive credentialing process. Many agencies reacted by modifying their examination process from in-person test center delivery to remote Online Proctoring. Mostly these changes were made in emergency situations, to keep the agencies relevant and preserve the idea of testing as a valid prerequisite for credentialing. However, most were made without psychometric analysis of comparability of the delivery modalities. This lack of rigorous evidence means that we still do not know whether the test results derived from the two methods are strictly comparable. Even with this validity threat, it seems likely that Online Proctoring will persist as an approved delivery method, even when the pandemic is “over”. The good news is there is a testing method that can help to mitigate many of OP’s known problems. This talk will discuss CAT as a possible solution for validity issues related to test delivery using Online Proctoring.

05: Invited Symposium - The CAT in the language assessments bag

Chair: *Alina A. von Davier*

Abstract: With the growth of digital technology and advances in automated test development tools, ranging from automated item generation to automated scoring, opportunity has come to develop innovative forms of technology-based assessments. This symposium offers an overview of how innovative computer adaptive algorithms, especially when coupled with other advanced technologies, support language tests. The four selected papers cover a bag of CATs with a wide range of specific applications that span the language itself (English and German), the country (Brazil, Germany, USA) and the supportive methodologies and technologies (from automatic item and test development to delivery). The first paper presents a computer adaptive English test from Brazil. It provides a historical perspective that reflects on the changes to the test over time. The second paper provides an overview of the Duolingo English Test automatic item generation (AIG) and CAT algorithm procedures. The third paper describes a German language test for professionals–Goethe Test PRO. The paper illustrates the psychometric considerations for the test. The fourth paper introduces a new paradigm where the AIG and the CAT algorithms are blended together into a dynamic assessment design. This paper will build onto the existing methodologies at Duolingo but integrate them with an Elo rating system for an in-real time difficulty estimation. These studies illustrate the similarities and differences in CAT design across language tests and also contribute to computational psychometric research by blending the computational models behind the automatic algorithm into the more traditional CAT approaches.

Presenters:

Mariana Curi, Elias Silva de Oliveira, & Lohan Rodrigues

The Evolution of the English Proficiency Exam at the University of São Paulo

Abstract: The evolution of the computerized English Proficiency Exam (EPI) for the graduate programs at the Institute of Mathematics and Computer Sciences of University of São Paulo, in Brazil, has been going on since 2013. Until then, the EPI was a paper and pencil test based on the Admissible Probability Measurement (APM) procedure. The transformation to a computer-administered test, followed by the implementation of a computerized adaptive test based on the Samejima graded response model, occured in 2014. Because of some response style issues in the APM design, the format of the questions were transformed to multiple choice in 2017. The option of reporting uncertainty about the correct response, presented in the APM design, was rarely used by the students. As a result, the item format was changed to multiple choice and a plugin developed for Moodle named Adaptive Quizz, was used to implement computerized adaptive testing. The purpose is to enhance the plugin in order to enable several IRT model choices to develop a powerful computing environment for CAT applications on Moodle platform. The flexibility of creating and presenting many kinds of items offered by the platform, as well as the advantages of being free and its large use in education make this project extremely useful in the context of small-scale testing.

Burr Settles, & Geoff LaFlair

Automatic Item Generation and Evaluation

Abstract: A crucial component of computer adaptive testing is trustworthy item difficulty parameters. Traditional approaches of CAT development require extensive item development and piloting in order to establish item difficulty parameters. The Duolingo English test uses an approach that leverages extant open source corpora and natural language processing (NLP) to create items and estimate their difficulties (Settles et al. 2020). We present our methods for item development, difficulty estimation, and evaluation. For item development and difficulty estimation, we extend the passage model from Settles et al. (2020) to employ Bidirectional Encoder Representations from Transformers (BERT) embedding vectors (Devlin et al., 2018) for initial language-based difficulty estimates. We then incorporate scored responses from operational test administrations (n=120,000) by adopting a Linear Logistic Trait Model (LLTM) (Fischer, 1977) in a multi-task learning framework to estimate item difficulty parameters from scored response data and BERT difficulty estimates. Performance of the model is evaluated against model-based metrics (e.g., variance explained (R2) and area under the receiver operating characteristic (AUROC) as well as classical test theory (CTT) metrics (test-retest and split-half reliability). Additionally, we examine the relationship between the model-based difficulty parameters and numerical representations of the linguistic features of the items. These numerical representations were estimated using a multi-dimensional analysis (MDA) approach from the corpus linguistics literature (Biber, 1988). Similar to NLP methods, this method evaluates the language of the items, however, it differs in that it focuses on functionally interpretable co-occurring language features that contribute to the ways in which people use language. As a result, this method provides confirmatory evidence that the NLP algorithms organize items across the difficulty scale along theoretically and empirically appropriate dimensions of language.

Aron Fink, & Katharina Klein

Investigation of the Psychometric Quality of the German Adaptive Language Test Goethe-Test PRO

Abstract: The online German test Goethe-Test PRO: German for Professionals has been carried out worldwide in examination venues of the Goethe-Institut as well as on the premises of corporation partners since April 2017. The test is designed as a computerized adaptive test evaluating listening and reading competence in the workplace on a scale from A1 to C2 of the Common European Framework of Reference for Languages (CEFR). In the first part of our talk, we will briefly introduce the Goethe-Test PRO and its underlying assessment framework. In the second part, we will present the results of a study that investigates the psychometric quality of the Goethe-Test PRO based on the operational test data stemming from N = 5636 test takers from 17 different states. For this purpose, we evaluated the dimensionality, the reliability, the conditional standard error of the ability estimates as well as the degree of adaptivity of the test. In addition, we investigated whether the psychometric quality differs between the countries where the test is conducted. Results indicate good psychometric properties of the Goethe-Test PRO and show that the test provides reliable test results with a reliability of over .9. The study shows similar results regarding the psychometric properties throughout the different countries where the Goethe- Test PRO is offered. The results will be discussed and future steps to improve the Goethe-Test PRO will be mentioned.

Settles, B., LaFlair, G. T., & Hagiwara, M. (2020). Machine-learning driven language assessment. *Transactions of the Association of Computational Linguistics*, 8, 247-263.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding.* Preprint arXiv:1810.04805

Fischer, G. H. (1977). Linear logistic trait models: Theory and application. In H Spada, WF Kempf (eds.), *Structural Models of Thinking and Learning*, pp. 203–225. Huber, Bern.

Yigal Attali & Alina A. von Davier

A Paradigm Shift for CAT Development in a Computational Psychometrics Framework with Language Modeling Affordances

Abstract: Computer-adaptive tests (CAT) have important advantages in educational assessment, promising to shorten tests as well as provide uniformly precise scores for most examinees, by tailoring item difficulty to the ability of examinees. However, the limiting factor in the development of a CAT is that it requires a very large item bank, which often presumes that these items are pretested with human subjects, which is an expensive and time consuming activity.

This paper presents an alternative approach to the development of a CAT assessment, based on a) creating a large item bank using language-model-based automatic item generation techniques, b) estimating preliminary item parameters using natural language processing (NLP) models, c) administering the items in the context of a CAT, and d) a framework for updating item parameters that accounts for the adaptive administration of items. We will illustrate this approach with the Duolingo English Test (DET). The DET is composed of item types that generate a diverse set of responses: from discrete responses, to continuous responses with 0- or 1-inflation, to fully continuous responses. This poses additional psychometric modeling challenges.

In summary, this paper illustrates how a psychometric framework combined with language modeling can support quality assessments for the 21st century.

06: Paper Session - Large-Scale Assessments

Chair: *Eveline Gebhardt*

Wei Buttress & Eveline Gebhardt

National Online Assessment in Australia using adaptive multistage testing design

Abstract: The National Assessment Program – Literacy and Numeracy (NAPLAN) is an online assessment for all students in Years 3, 5, 7 and 9 taken in the 2nd full week in May annually and nationwide in Australia. NAPLAN consists of four test domains, reading, writing, numeracy and conventions of language (spelling, grammar and punctuation) with results provided to schools and parents of each student. NAPLAN online assessment is a tailored computerized multistage test that adapts to student responses, presenting students with questions that may be more or less difficult – resulting in more engaging assessments and more precise results. NAPLAN was first implemented in 2008 as a paper test. Through significant planning, development, research and trialling, NAPLAN transitioned to online test delivery between 2018 and in 2022. In May 2022 nearly 1.3 million students completed 5 million tests over a two-week period including students in very remote locations. In the presentation, we would like to speak about the following aspects of the online NAPLAN:

- Brief history of the NAPLAN assessments
- NAPLAN assessment framework – Curriculum Based
- Test design – multistage branching test design
- Online delivery – Lockdown Browser, Devices, Low and no bandwidth delivery solution
- Accessibility accommodation and enhancements
- Data analysis methodologies and approaches
- Reporting – Student, schools, Jurisdictions and National

Cecilia Marconi, Mario Luzardo, Andrés Peri, & Bruno Fonseca

A Systematic Implementation of a Nation-Wide English Assessment Using a Computerized Adaptive Test

Abstract: This paper presents the design of a National English Computerized Adaptive Test (NECAT) developed and implemented in the Public Education System in Uruguay. Its first edition took place in 2014 and since then, it has been administered annually at national level. The NECAT emerged for developing a measuring instrument tailored for assessment at large-scale in a context of very heterogeneous English learners. At primary school level, three different programs for teaching English as a foreign language (EFL) coexist, reaching all students. One of these programs is Ceibal en Inglés (Ceibal in English); an innovative combination of remote English lessons delivered through videoconference, along with blended learning and collaborative teaching. The latter reaches a very diverse population of children from 4th, 5th and 6th grade of Public Primary Education, with different backgrounds and number of years studying English within the public education system. The development of the program -Ceibal en Inglés- and the implementation of a NECAT, were possible due to the creation of Plan Ceibal in 2007. The latest emerged as a nation-wide policy for guaranteeing digital inclusion and equal opportunities, supporting Uruguayan educational policies through digital technologies. Since those early years, Plan Ceibal has evolved, creating a powerful technological infrastructure for integrating digital technologies in education. This innovative digital ecosystem created the pillars for developing a NECAT, whose results follow the Common European Framework of Reference for Languages (CEFR).

The NECAT was designed to assess learning of all students from 4th to 6th grade of primary education and of students from 1st to 3rd grade of middle-term secondary education. In 2020 and despite the COVID-19 pandemic, its seventh edition reached over 41,000 participants.

The item banks are composed of dichotomous items (multiple choice/3 answers options) and have been calibrated by three-parameter logistic (3PL). The initialization procedure design takes into account the fact that the evaluation is applied annually. Even though it is not a mandatory assessment, the NEAT is addressed to all students.

David Shin

Computerized Adaptive Testing with AI-Scored Items

Abstract: Computerized adaptive testing (CAT) including item level adaptive tests and multi-stage tests (MST) has become more popular due to the availability of modern technology and the potential benefits associated with CAT. While most CAT programs only use items that are instantly machine scorable, some do include artificial intelligence (AI) scored-items to enable a test to measure higher-order skills and allow students to demonstrate complex understanding. In CAT, the selection of next questions relies on student's scores on the previous questions. However, because the response of an AI-scored item may not always be successfully scored by the AI engine immediately, using AI-scored items in CAT implies that a CAT algorithm needs to allow CAT tests to go on when a score from an AI-scored item is not immediately available. This can be done by using the interim theta from the previous question to select next question after the AI item. In this case, the AI-scored items will be scored later, and their scores will be used to compute the final theta. In this study, we refer the AI-scored items that will be scored later as "scored-later" items. In practice, two questions need to be considered regarding the scored-later items. One is the number (or the proportion) of scored-later items that can be inserted. Because the score of these items will be unavailable to updated students' ability to select next questions, the more the scored-later items are inserted, the less adaptive the test may become. Another question is the location where scored-later items are administered in CAT. Since interim thetas in a CAT tend to fluctuate more in the beginning and less at the end of the test, inserting scored-later items in different locations within a CAT test may result in different impact to the final test results. The purpose of this study is to investigate the impact of inserting scored-later items in CAT. The following research questions are investigated:

1. How does the location of scored-later items in a test impact the test precision, item exposure rate, and content on-target rate?
2. How does the number of scored-later items in a test impact the test precision, item exposure rate, and content on-target rate?

The data will be simulated from a Math CAT item pool (828 MC and 47 2-point AI-scored items). The test is an item level adaptive test with 34 items. Two factors will be investigated including (1) the location where scored-later items are inserted (5 conditions), and (2) the number of scored-later items inserted (3 conditions). Totally, there will be 5 x 3 = 15 conditions as listed below.

Factors	Study Conditions
Location of scored-later items	1. within the first 1/3 of the test 2. within the second 1/3 of the test 3. within in the last 1/3 of the test 4. at the end of the test 5. adaptively selected throughout the test
Number of scored-later items inserted (test length = 34 items)	1. 2 items 2. 4 items 3. 6 items

Mark D. Reckase

Setting Performance Standards on Computerized Adaptive Tests: Problems and Solutions

Abstract: Computerized adaptive testing (CAT) has become a common testing methodology for certification/licensure tests and tests of academic achievement. A feature that is shared by these very different types of testing applications is the use of performance standards (cut scores) to facilitate reporting results and making decisions about individuals. The cut scores used to support the uses of the testing programs are typically obtained through a standard setting study that translates a statement of policy that motivates the use of the testing program to a score (the cut score) on the reporting score scale on the test. There are numerous approaches to estimating the cut score on the reporting score scale that is consistent with the stated policy. Most of the approaches were developed for use with fixed length, paper-and-pencil tests and then were modified for use with CATs. Recently some approaches to setting performance standards have been suggested specifically for the unique characteristics of CATs. This paper describes several approaches to setting standards on CATs and also highlights problems that arise from using these approaches. Some solutions to the problems are proposed. Several examples are provided to show the problems and the impact of the proposed solutions.

07: Paper Session - Test Termination

Chair: *Alper Şahin*

Richard Gershon, Saki Amagi, Rina Fox, Aaron Kaat, Micchael Kallen, Benjamin Schalet, & Cindy Nowinski

Alternative Stopping Rules for the NIH Toolbox Emotion Battery

Abstract:

Aims: The current stopping rules for the NIH Toolbox Emotion Battery (NIHTB-EB) CATs are effective for some test takers, but the assessments can be burdensome for high-functioning individuals. Simultaneously, they do not yield adequate reliability for some clinical populations. We evaluated potential stopping rules for CATs, and evaluated which minimized burden while maximizing precision for clinical use.

Methods: We conducted simulations to compare four potential CAT stopping rules to the current rules for 13 NIHTB-EB item banks for simulated general and clinical adult samples. The current rules terminate the test if ≥ 4 items have been administered (minimum), the standard error (SE) of the EAP score estimate is score estimate is < 0.3, or 12 items have been administered (maximum). The potential rules included a SE-change rule, six and eight-item fixed-length CAT rules, and a reduced maximum rule. The SE threshold for interim stopping was reduced to 0.224 for the SE-change and reduced maximum rules. We compared the reliability achieved by each set of rules to the reliability of the current rules, by grouping simulees [Reliability<0.85, 0.85 Reliability<0.90, 0.90 Reliability<0.95, 0.95 Reliability]. We also compared the number of items administered.

Results: Although the SE-change rule increased the proportion of simulations achieving empirical reliability >0.95 (+22.8% general, +29.9% clinical) compared to the current rules, the average percentage of simulations achieving empirical reliability 0.95 (+27.5% general, +36% clinical). Importantly, this did not excessively increase the percentage of simulations achieving reliability <0.85 (+1.6% general, +0.7% clinical). Finally, the reduced maximum rule maintained reliability comparable to the eight-item CAT, while decreasing burden by not always requiring eight items. The mean number of items administered was only 1.13 greater than the current rules for general (7.53 vs 6.40) and 1.83 greater for clinical (7.26 vs 5.43).

Michiel A. J. Luijten, Benjamin D. Schalet, Leo D. Roorda, Lotte Haverman, & Caroline B. Terwee

Reducing administrative burden of CATs in healthcare through advanced stopping rule optimisation

Abstract: To reduce administrative burden of computerized adaptive tests (CAT), we developed a method to determine the optimal standard error reduction (SER) stopping rule, using existing CAT data. We extracted CAT responses (*n* range 3345 – 3397) from pediatric PROMIS Anxiety and Depressive Symptoms item banks collected between April 2020 and November 2020, incl uding estimated trait levels (θ) and standard errors (SE(θ)) for each CAT step. The default stopping rules were a minimum/maximum of 4/12 items administered or a minimum precision of SE(θ) <0.32. We investigated how to optimize an additional SER stopping rule – the difference between the interim SE(θ) and the previous SE(θ) at each CAT step – by imposing increasing SER thresholds (0.01 – 0.20). The following outcome criteria were assessed for each SER threshold using the final θ and final SE(θ) estimates of the CAT; efficiency of the CAT (1- SE(θ)^2/n$_{\text{items}}$), number of items administered (n$_{\text{items}}$), the mean SE(θ) of all respondents (M$_{\text{SE}(\theta)}$) and the mean T-score difference compared to default stopping rules (ˆ†*T*). For optimization of the SER threshold, we looked at the increase in efficiency of subsequent SER values. The default stopping rules resulted in an efficiency of 0.98/1.27, n$_{\text{items}}$=9.02/8.13, and M$_{\text{SE}(\theta)}$= 0.36/0.38 for the Anxiety/Depressive Symptoms item banks. Optimizing the SER stopping rule resulted in a threshold of 0.027 for the Anxiety item bank (efficiency = 1.08, n$_{\text{items}}$ = 5.57, M$_{\text{SE}(\theta)}$ = 4.26, ˆ†*T* = 0.06) and 0.024 for the Depressive Symptoms item bank (efficiency = 1.45, n$_{\text{items}}$ = 4.79, M$_{\text{SE}(\theta)}$ = 4.15, ˆ†*T* = 0.58). For healthy participants with minimum scores, this results in fewest items administered, but a decrease in measurement accuracy and biased T-scores, which may be relevant depending on the goal of assessment. We conclude that this method allows us to determine an optimal SER threshold for different health outcome item banks, however the threshold values will vary depending on the θ distribution of the target population and the IRT model parameters.

Ming Him Tai & David Weiss

Stochastic curtailment in adaptive testing: A new way to end a CAT to improve efficiency

Abstract: In computerized adaptive testing (CAT), one of the most commonly used termination rules is the standard error of measurement (SEM) criterion. However, an examinee, especially those with extreme true θ levels, might not reach the pre-specified SEM level at the maximum test length (hereafter referred to as a low precision case), primarily because of insufficient item bank information at his or her true θ level, particularly in most realistic CAT item banks. In that case, the examinee will take an unnecessarily large number of items without much gain in measurement precision, which increases the burden on the examinee. Inspired by Finkelman's pioneering work on applying stochastic curtailment in sequential mastery testing, we investigated applying it to θ estimation in CAT to shorten test length. In this context, stochastic curtailment was operationalized as the early termination of a CAT procedure if the estimated probability of observing a low precision case is above a user-specified level (for example, 95%). A monte-carlo simulation study was performed to evaluate the performance of the procedure. It was found that for a test with a maximum test length of 40 dichotomous items, the procedure was above 80 percent correct in predicting low precision cases (i.e., a positive predictive value of above 80 percent) for examinees with extreme true θ levels as early as after administering only 10 items; the percentage of correct predictions further increased to nearly 100 percent after 20 items. Our results suggest that stochastic curtailment is a promising approach to significantly shorten test length for examinees whose tests would otherwise be ended only after some arbitrary maximum number of items.

Alper Şahin & Duygu Anil

A Comparison of Test Termination Rules in terms of Ability Estimation Accuracy with Relatively Small Item Banks

Abstract: The accuracy of ability estimation in Computerized Adaptive Testing is based on many different factors. One of these factors is the termination rules used during the estimation process. Especially, if the item bank has a limited number of items, the significance of the termination rule on the accuracy of the ability estimates increases dramatically because the more items there are in an item bank, the better ability estimates are obtained. The aim of this study is to compare the performance of test termination rules in terms of ability estimation accuracy with a relatively small item bank. For this purpose, real data from an English test of 80 multiple choice items which was taken by 965 examinees in paper and pencil format was used. Item parameters were estimated in the three-parameter model using this data. Then, five CAT posthoc simulations were run using the test taker's real response data, the item parameters estimated, the items selected using fisher's highest information as the fixed item selection rule, and five different termination rules (1) when all items are used, (2) when the absolute change in successive theta estimates is less than or equal to 0.001, (3) when the standard error of the theta estimate is less than or equal to 0.200, (4) when the standard error of the theta estimate increases by 0.010 or more and (5) a combination of these four. The classical scores (item correct) of the test takers and their CAT full-bank ability estimates were taken as their pseudo true scores and the correlations between ability estimates obtained from the posthoc simulations and pseudo true scores were calculated within certain classical score intervals (0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69 and 70-80) and latent ability intervals (theta: 3.00 to 2.00, 1.99 to 1.00, 0.99 to 0.00, -0.01 to -0,99, -1.00 to -1.99, -2.00 to -2.99, -3 to -4). Moreover, the average SE of theta estimates for each of these ability intervals was also obtained and compared as well. Findings indicated that the performance of termination rules (3) and (4) were highly satisfactory between the theta 0 and 2 (where the bank had the highest information) and classical scores between 60 and 80. Moreover, when the item banks are relatively small, these termination rules can be used for test-takers whose ability estimates fall within the range of theta where the information is the highest.

08: Paper Session - CAT for Admission, Selection and Certification

Chair: *Bernard Veldkamp*

Tatiana Sango, Precious Mudavanhu & Robert Prince

The National Benchmark Mathematics Test: From Paper Based to Computer Adaptive Testing

Abstract: The National Benchmark Tests (NBTs) provide the South African Higher Education (HE) institutions with additional information to assist in the admission of students in appropriate programs and inform curriculum development. This study presents the current approach behind the criterion-referenced Mathematics (MAT) test design. It outlines the item selection for paper-based and online test delivery models and investigates the necessary requirements and challenges to transition into Computerised Adaptive Test (CAT) mode in the context of high-stakes test delivery.

The content of the MAT test assumes the knowledge in the school leaving National Senior Certificate (NSC) mathematics curriculum and at the same time is aligned to the first year mainstream mathematics needs. Classical test theory (CTT) and the three Parameter Logistic (3-PL) item response theory (IRT) model is used to investigate the items and tests and to score the candidates. The NBT annual testing cycle utilises up to 19 unique MAT test forms which are assembled according to the test framework to ensure the results are comparable and reliable. Each MAT test has 60 multiple-choice items, of which there are 12 anchor items and 8 trial items. The anchor items allow for the equating of different test forms. The trial items are not scored but the anchor items are internal and scored. The items are selected from the item bank according to the blueprint specification including subdomain weighting, the weighting of sections/skills within subdomains, items' cognitive level weighting, as well as the level of difficulty.

The MAT Item bank comprises of more than 1 700 operational items and undergoes bi-annual item development and review. New items created at the item development workshops are critically reviewed for content alignment, cognitive level, language and accessibility, fairness and sensitivity. Item statistics (CTT and IRT) are reviewed and appropriate decisions regarding inclusion of reviewed items are made.

While investigating the transition to adaptive mode of testing, we make certain assumptions and discuss the following five aspects:

- The calibration of the Mathematics Item bank and provision for additional clusters and item level diagnostics.
- The required size of the item bank and methods to ensure comparability of assessments.
- The choice of item selection algorithms to inform the Item Banking specifications.
- The student level information we have currently and the additional information that could be available with CAT.
- The anticipated challenges with using the CAT mode of delivery for our purposes.

We currently deliver the MAT test in both Paper and Online modes, where the candidates are given up to three hours to complete the test. One of the considerations behind using CAT

mode is the ability to reduce the test time while increasing the accuracy of measurement of students' performance level for placement/admission to the most appropriate study program.

Alexandre Jaloto & Ricardo Primi

Can we improve Brazilian High School Exam with CAT?

Abstract: The Brazilian High School Exam (Enem) results are used to enter higher education and obtain scholarships and study financing. The exam comprises an essay and four 45-item tests that measure educational skills in different areas of knowledge: Human Sciences (HS), Languages and Codes (LC), Mathematics (MT), and Natural Sciences (NS). Annually more than four million students take the Enem, whose tests occur in the same two days for all participants. Computerized Adaptive Testing (CAT) can contribute to advances in the exam, such as reducing test size and logistics complexity. Therefore, this work aimed to verify the possibility of reducing the number of items in Enem through a CAT. Our research was divided into two studies. The first study determined the item parameters applied in Enem and equated them to a single scale since the item parameters are not disclosed. The public information about the applications includes participants' answers and their official scores, which are equated throughout the years. We used tests from the 2009 to 2019 editions of Enem. We calibrated items with the mirt package in R from 5,000-participant samples and used the linear transformation method to place the items in the same official Enem metric. Then, we reestimated all participant scores for each application with the expected *a-posteriori* (EAP) method. The correlations between the reestimated and official scores were higher than 0.960, indicating adequate calibration. The second study aimed to simulate CAT analysis using the calibrated items. The bank of items in each area had 765 (HS), 839 (LC), 719 (MT), and 674 (NS) items. A simple random sample of participants from the 2019 edition of Enem was drawn for each area. The sample size was designed to guarantee a mean with a sampling error of 0.03 standard deviation units. By adopting this procedure, we were able to generalize our results to the population of this edition of Enem, which potentially brings our simulation closer to expected situations for future editions with similar characteristics. Then, we simulated responses to each item bank based on the sample participants' scores. Finally, we simulated a CAT with the mirtCAT package, which ended when the error was less than 0.30, or 45 items were applied. On average, the simulations ended with 18.4 (HS), 12.0 (LC), 29.2 (MT), and 22.1 (NS) items. We could reduce Enem to 20 items for the following participant proportions: 71.7% (HS), 94.8% (LC), 39.8% (MT), and 60.4% (NS). This reduction was possible for participants placed in theta ranges of -0.01–2.45 (HS), 0.00–1.91 (LC), 1.15–2.98 (MT), and 0.25–2.45 (NS). Participants with the lowest scores answered 45 items, which indicates the importance of creating easy items to increase information in lower regions of the scales. Future studies should include item exposing control, as Enem is a high-stakes test. Also, it is important to evaluate other stopping rules (e.g., the difference between theta in each iteration) and consider the content proportion in the test.

Burhanettin Ozdemir & Fadi Munshi

Implementing CAT to measure Nursing skills of the candidate: A Post-Hoc Simulation Study

Abstract: The computerized Adaptive Test (CAT) adapts the test to each individual examinee in order to obtain an accurate measurement across the entire ability range. CAT methods combine computer technology with modern measurement theories, such as the item response theory (IRT) model, to increase the efficiency of the exam process. On the other hand, the real data-based post-hoc CAT simulation method allows comparing the paper-pencil (P&P) outcomes of a test to the corresponding alternative CAT designs that utilize different item selection, ability estimation and stopping rules. The purpose of this study is to investigate the feasibility of administering the Nursing licensure exam in CAT format. Additionally, this study aims to determine the most effective CAT design for nursing licensure exam. For this purpose, the data set obtained from Saudi Nursing Licensure Exam (SNLE) administered in 2021 was used to construct the item bank. The Nursing licensure exam is administered to assess the readiness of the nursing graduates for practicing or proceeding to postgraduate training. It is a three-hour exam that consists of 300 dichotomous items. The items were calibrated with the Rasch measurement model using Winsteps-Program to obtain the item parameters. After item calibration, the initial item bank consisted of 2542 items that are related to 4 sections

To determine the best CAT, for SNLE, three different theta estimation (EAP, MLE and BME) methods, two different fisher information based item selection methods and two different Kullback-Leibler based item selection methods, and six termination rules based on two different termination methods (fixed test length, precision) were utilized. In total, 72 different conditions were tested, and the performance of different CAT algorithms were compared with respect to RMSE, bias, the averaged number of administered items, correlation between true-theta and estimated CAT theta and average exposure rate. Results indicate that implication of different theta estimation and item selection methods affected the RMSE, test-length and correlation between the true and estimated thetas as well as exposure rate. Averaged number of administered items was around 100 when precision criterion was set to .20. Aditionally, the correlation between true-theta and estimated theta was around 0.91, However, when precision criterion was set to .15, averaged number of administered items increased to 180 with higher precision and lower RMSE values. Considering the significant increase in the test length, the change in RMSE and precision indicators were negligibly small. On the other hand, utilizing Kullback-Leibler rather than Unweight Fisher’s Information to select items both decrease average number of items administered and RMSE of the test, while increasing the processing time. To conclude, Bayesian ability estimation methods obtained more accurate results with less RMSE and bias. Moreover, using proposed CAT designs with 0.20 precision caused approximately 60 to 65% decrease in the test lengths compared to the paper-pencil version of SNLE. Thus, it is believed that the results of this study will provide guidelines for alternative assessment methods proposed for medical licensure exams.

Araê Cainã, Alexandre Jaloto, Gustavo Henrique Martins, Felipe Dinardi, Thiago Martins Santos, Ricardo Mendes Pereira, & Dario Cecilio-Fernandes

Simulation study for using Computerized Adaptive Testing for a medical residency admission test

Abstract: The first phase of admission to the medical residency at the University of Campinas comprises a knowledge test consisting of 160 multiple-choice questions. Questions are equally divided into five medical knowledge domains. The participants are classified based on their total scores, which corresponds to the sum of correct answers. Alternatively, Computerized Adaptive Testing (CAT) could reduce items while assuring the same level of psychometric properties. This study investigated, through simulation, the use of CAT and the number of items necessary. We were also interested in controlling item exposure. We calibrated and linked the test items from 2017 to 2021 using the mirt package in the R environment. We used the two-parameter logistic model of the Item Response Theory (IRT). We equated the items from the common test takers from 2017 to 2021. We excluded items that did not have any variance (100% correct or incorrect answers) and were eliminated by the test committee. The final bank was composed of 703 items positioned in a single metric. After calculating the IRT score of all students in each edition (true score), we simulated the application of a CAT with the catR package in R. The simulation design included two stopping criteria (maximum of 30 and 40 items) and three item-selection methods (Fisher Maximum Information – MFI, Progressive, and Progressive-Restricted – PR). Under the PR method conditions, the maximum exposure rate of an item was set to 20%. Worth noting that the catR package simulates responses to items based on real theta and item parameters. The theta estimation method on the CAT application (reestimated score) was Expected a-Priori (EAP). Items from each medical knowledge domain had equal proportions in the application. We evaluated each simulation condition based on the correlation between the true and the reestimated scores, the Root Mean Square Error (RMSE), the precision (one minus the square of the RMSE), and the highest exposure rate of an item. Correlations ranged from 0.920 to 0.954; the RMSE from 0.276 to 0.360; the precision from 0.871 to 0.924; the maximum exposure from 0.200 to 1.000. The condition with the highest correlation and highest precision was the 40-item MFI method, but it was the one that had at least one item with 100% exposure. In four conditions, the maximum exposure of an item was 80% or more, which is not desirable in high-stakes tests. The two conditions with the PR method had the lowest exposure with excellent correlation and reliability. The results indicated the possibility of reducing the number of items using CAT without compromising the measurement error and maintaining a high correlation between the true and the reestimated scores. Still, this reduction occurs with low levels of item exposure throughout the applications.

09: Keynote - Wim J. van der Linden: The New Paradigm of Adaptive Testing

Chair: *Andreas Frey*

Abstract: Modern adaptive testing is evolving into much more than just one-item-at-a-time test assembly. In fact, it is quickly becoming the core of a completely new testing paradigm with each of its current cyclical procedures, such as item calibration, fit analysis, cheating detection, item-security monitoring, item-pool maintenance, replaced with real-time continuous processes. I will formally characterize the new paradigm, discuss the necessary tools for its implementation, and show results for a few of these new continuous processes.

10: Invited Symposium - Computerized adaptive testing (CAT) for the measurement of health outcomes – the Patient-Reported Outcomes Measurement Information System

Chair: *Caroline B. Terwee*

Discussant: *Ulf Kroehne*

Abstract: There is increasing interest in healthcare for measuring physical, mental, and social health from the patients‘ perspective in individual patient care to obtain transparent and comparable outcomes for health care evaluations and improvement initiatives. However, health care providers do not yet measure patient-reported health outcomes consistently because of lack of consensus on what to measure, time investment and the excess of questionnaires that differ in content and quality, and have incomparable scores. The Patient-Reported Outcomes Measurement Information System (PROMIS®) was developed by a collaboration between the US National Institute of Health and eight US research institutes to develop one state-of-the-art assessment system to measure patient-reported health with highly accurate, precise and short measures for use across adult and pediatric (patient) populations. A wide range of generic item banks was developed, targeting various constructs, such as pain, physical function, anxiety, depression, fatigue, sleep disturbances, and participation in social roles and activities. Item banks were developed using item response theory (IRT) methods and can be used as standard short forms (e.g. 4-, 6-, 8-items versions), custom short forms (selection of relevant items for a specific context) and computerized adaptive tests (CAT). To make PROMIS widely available and maintain its scientific quality, a number of resources have been established: the PROMIS Health Organization (PHO) was established to maintain and encourage the application of PROMIS. PHO is a growing open membership society with education (e.g. workshops and annual conferences), and on-demand resources. The “HealthMeasures” team (Northwestern University, Chicago) and website is the official information (helpdesk) and distribution center for PROMIS, which also coordinates all translations. The Assessment Center Application Programming Interface (API) was developed to connect to any data collection software application (e.g. REDCap) with the full library of PROMIS measures, CAT software, and standardized item parameters. PROMIS CATs have been built into electronic health record systems, such as Epic, and are available through the PROMIS iPad App. Scoring manuals and interpretations guidelines were developed for research and clinical practice. Linking studies are being performed to convert PROMIS scores to scores of related commonly used questionnaires. PROMIS National Centers have been established in 19 countries. Their role is to coordinate all translation efforts, communicate the value of PROMIS to the scientific and research community, and encourage, facilitate, and support the application of PROMIS in the local country. PROMIS measures have been translated in more than 60 languages. Cross-cultural validation studies are being performed to evaluate content validity, confirm the underlying calibration model, and assess differential item functioning between language versions to test the PROMIS convention to use a single set of IRT item parameters across populations and language versions to express scores on a common scale (T-score metric). The ultimate aim is to develop PROMIS into a gold-standard outcome metric for measuring patient-reported health outcomes in an efficient, precise, and comparable way across the world.

Presenters:

Matthias Rose

Methodology and standards of PROMIS item bank development

Abstract: The development of the Patient-Reported Outcome Measurement Information System (PROMIS) was cross-funded by all National Institutes of Health in the U.S. as one of their key-road initiatives starting in 2004. For the last decade, its implementation in clinical research and practice is facilitated by the Patient-Centered Outcomes Research Institute (PCORI). As of today, PROMIS instruments have been used by more than 40 mio patients.

The main aim of the development of the PROMIS was to facilitate the standardization of patient-reported health status assessment using item-response theory methods. Today, there have been more than 80 item banks developed; following a stepwise approach consented by the PROMIS investigators. All item banks utilizing a two-parameter graded response model. The talk will focus on the development of the physical function item bank (PF) - the mostly used PROMIS item bank - as an example. To develop this bank, 165 instruments have been reviewed, with 1,728 items capturing the PF constructs, focus groups allowed to propose 168 items for empirical testing in >20,000 respondents. In total 40 items had to be excluded due to violations of the unidimensionality assumptions, DIF, residual correlations etc., resulting in an item bank of 124 items. As of today, more than 500 studies evaluated the measurement properties of PROMIS PF tools, and more 3,000 studies have been published using some PROMIS instruments.

Felix Fischer, Sein Schmidt, & Matthias Rose

Comparison of PROMIS Depression CATs based on US calibration with a German calibration

Abstract:

Objective: The Patient-reported outcomes measurement information system (PROMIS) provides large-scale itembanks for over 100 patient-reported outcomes. As PROMIS aims to make PRO measures available worldwide and in different languages, it is unclear how to choose the most appropriate set of item parameters for assessment in a given cultural context. An established, unique set of item parameters facilitates comparability across settings, whereas country-specific item parameters might be more appropriate for measurement in a specific population. The aim of this study is therefore to compare the performance of PROMIS CATs based on 2 different sets of item parameters.

Method: Using data from the Berlin Longterm Observation of Vascular Events Study (BeLOVE), we investigated the performance of a computer-adaptive test based on the PROMIS Emotional Distress Depression itembank in a post-hoc simulation. We compared the performance between two calibrations - a graded response model based on the PROMIS Wave 1 data collected in the US (PROMIS CAT) and a generalized partial credit model based on data collected in the German general as well as clinical populations (GERMAN CAT). We post hoc simulated both CATs (EAP estimation, min SEM = 4 & <= 10) in the sample and compare theta estimates, CAT length and item usage.

Results: 711 respondents suffering from cardiovascular disease or diabetes, were included in this study. The PROMIS Emotion Distress Depression T-Score based on complete response to the full itembank was 46.5 (sd = 9.2), indicating relatively low symptom burden. Overall, the PROMIS CAT resulted in lower theta estimates compared to the GERMAN CAT (mean difference = -0.68) and 95% of the score differences were between -1.34 and -0.07. The correlation was 0.95. The PROMIS CAT presented on average 6.5 items (median = 5), whereas the GERMAN CAT terminated after 7.3 items (median = 8 items). Overall, both CATs used 18 out of 28 available items, with 8 respectively 9 items making up 80% of the responses. Most frequently items used in both CATs were “felt worthless”, “felt depressed” and “felt sad”.

Conclusion: Using different calibrations as basis for a depression CAT leads to differences in item presentation and theta estimation. The large mean difference can be attributed to differences in depression severity in the respective calibration samples. Hence, a challenge for further standardization of PRO assessments is to align different models based on GRM and GPCM on a common scale.

Benjamin Schalet

Adapting PROMIS CAT Stopping Rules to Reduce Patient Burden in Medical Settings

Abstract: The Patient-Reported Outcomes Measurement Information System (PROMIS®) represents a popular system of health outcome instruments, which can be administered via computer adaptive testing (CAT). Developed originally as a research tool, PROMIS instruments are now increasingly deployed in clinical healthcare settings. Although the purpose of PRO measurement in clinical settings varies, they may be classified as (1) screening, (2) monitoring of improvement (or deterioration) over time, or (3) quality of care measurement. A single hospital system may implement PROMIS measures across multiple settings and departments, each with varying interests and priorities. For example, an orthopedics department may be most invested in measuring physical function and pain interference, but less so in depression. A primary care clinic administrator may be required to implement depression screening, and wishes to do this with as few items (and hassle) as possible.

Although PROMIS contains many banks, the 7 PROMIS Profile domains are used most frequently. Currently, CAT algorithms operate individually on each unidimensional bank, which contains items calibrated under the graded response model (GRM). The same CAT item selection algorithm – the Maximum Posterior Weighted Information (MPWI) – is implemented across all banks. CAT stopping rules are also relatively uniform, chosen with research settings in mind and focused on the accurate measurement of clinical populations (1 SD in the direction of worse health). Consequently, the rules were relatively conservative, enforcing a stopping criterion rule of standard error (SE) .91), as well as a minimum of 4 and a maximum of 12 items. But for asymptomatic patients, this SE criterion may never be reached, resulting in the administration of all 12 items. If all 7 PROMIS profile domains are given, this leads to an unacceptably long testing experience for those who least need it. However, some clinical contexts might benefit from more accurate measurement (reliability $>$.95), e.g., for domains directly relevant to important decisions, e.g., ongoing pain management or termination upon depression remission.

In effort to balance brevity and accuracy, we set out to identify new CAT rules that could be used in a variety of clinical health care settings. We conducted CAT simulations in normal and clinical populations of the 7 Profile PROMIS domains, varying the maximum rule, and alternately implementing a predicted SE reduction stopping rule (PSER), a SE reduction rule (SER), as well as a "Best Health" rule. PSER uses the predictive posterior variance to determine the reduction in SE that would result from the administration of additional items. SER checks whether the current SE has been sufficiently reduced to justify the administration of an additional item. The Best Health rule terminates the CAT when 2, 3 or 4 items are consecutively answered in the direction of positive health (e.g., answering "not at all" to the question, "How much did pain interfere with work around the home?"). Simulations showed that PSER and SER rules were especially effective at terminating CATs for healthy simulees, with little meaningful reduction in reliability. Results, future directions, alternative strategies will be presented.

Leo D. Roorda & Caroline B. Terwee

The development of Dutch-Flemish PROMIS® item banks and a CAT platform enabling their use by health professionals en researchers – PROMIS use in the Netherlands and Flanders

Abstract: The Patient-Reported Outcomes Measurement Information System (PROMIS®) has been developed in the United States. It includes a wide range of generic item banks, targeting various constructs, such as pain, physical function, anxiety, depression, fatigue, sleep disturbances, and participation in social roles and activities. Item banks were developed using item response theory methods and can be used as standard short forms (e.g., 4-, 6-, 8-items versions), custom short forms (selection of relevant items for a specific context) and computerized adaptive tests (CAT).

It was decided to translate available PROMIS item bank into the Dutch-Flemish language. Translations were developed using a strict methodology. Once translated, the Dutch-Flemish PROMIS item banks were made available for use in the Netherlands and Flanders as short forms. Once a cross-cultural validation study had been carried out, preferable in a large sample (n~1000) of the general and a patient population, the item banks were considered ready to be used as a CAT. Over 40 PROMIS measures for adults and children have been translated into Dutch-Flemish, and about 20 items banks have been validated and are currently available as CAT. Reference scores of the Dutch general population have been collected and thresholds for mild, moderate, and severe symptoms or (lack of) function have been established, to facilitate interpretation of scores.

A CAT platform was developed in order to make Dutch-Flemish PROMIS CATs available to health professional and researchers in the Netherlands and Flanders. CAT software was developed, using the standard original US item parameters and start- and stopping rules, as per PROMIS convention. The implementation of the CAT platform in these countries, was accompanied by a wide range of challenges, successes and failures. During the presentation, these will be shared with the audience. Lessons learned and critical success factors will also be discussed.

Ulf Kroehne

Discussion

11: Paper Session - Adaptive Measurement of Change

Chair: *Theo J. H. M. Eggen*

Robert Chapman

An Application and Evaluation of Adaptive Measures of Change in the Patient-Reported Outcome Measurement Information System

Abstract:

Introduction: The Patient Reported Outcome Measurement Information System (PROMIS®) was developed as a measurement system of health-related quality of life (HRQoL) which leverages the benefits of item response theory (IRT) and computer adaptive testing (CAT). A PROMIS scientific, clinical and regulatory user community has broadly implemented and worked to validate PROMIS measures across many different conditions, populations and treatment contexts.

While there are expanding applications and growing acceptance of PROMIS IRT and CAT methods, there is a need to explain and compare methods of evaluating score change, particularly in PROMIS CAT. Change is important metric for scientific, regulatory and clinical use, but many users of PROMIS are unaware or do not understand modern statistical methods of detecting change using CAT.

This work conducts a multifaceted evaluation of methods for detecting individual change with PROMIS measures, including simple differences in scores, meaningful/minimally-important change thresholds and modern Adaptive Measure of Change (AMC) for CAT.

Data: Three thousand cases of data are simulated, representing persons who are average (Theta 0, T-score 50), high (Theta 1, T-score 60) and low (Theta -1, T-score 40) on seven PROMIS domains: depression, anxiety, physical function, fatigue, pain interference, sleep disturbance and satisfaction with social roles and activities. Change scores are created, representing small changes (±0.2 Theta, ±2 T-score points), medium changes (±0.5 Theta, ±5 T-score points) and large changes (±0.8 Theta, ±8 T-score points). Person-item responses are simulated using by IRT model-based “most probable responses” and are scored using IRT score estimation methods.

Methods: Multiple methods of detecting individual change are calculated and compared, focusing on comparing simple differences in scores and meaningful/minimally important change thresholds with AMC. AMC is a method of detecting change in CAT measures which uses individual score standard errors to set confidence intervals. The overlap or non-overlap in confidence intervals allow us categorize scores as either likely changed or unchanged. Measures of change are evaluated by root mean square error, mean bias error and effect size, and at simulated ‘True’ levels of the small, medium and large change.

Results: The graphical and table comparison between simple score differences, meaningfully/minimally important change and AMC, show a general but differential advantage for AMC across PROMIS domains (e.g., depression, pain), baseline theta level, magnitude of ‘True’ change and CAT administration (i.e., item selection and test length).

Conclusions: This work provides a detailed explanation and comparison of methods for evaluating change with PROMIS CATs, including established and modern methods. Modern methods for detecting individual change (i.e., AMC) are recommended, with consideration of various HRQOL measurement contexts and facets.

The PROMIS measurement system has made CAT methods more accessible to a broad array of researchers, clinicians and regulators by providing a platform for implementing CAT assessments across a spectrum of HRQOL domains. A greater understanding and acceptance of modern methods for evaluating change with PROMIS CATs will support further applications and a broader user community for CAT.

Allison W. Cooperman & David J. Weiss

Efficacy of adaptive measurement of change under violations of longitudinal measurement invariance

Abstract: Numerous structural equation modeling and item response theory methods have been proposed for the longitudinal measurement of latent traits.[1,2] Whereas the majority of these methods examine changes at the group level, adaptive measurement of change (AMC) is a promising approach for measuring latent trait change at the individual level. In particular, AMC uses computerized adaptive testing (CAT) to identify whether there is significant within-person change for a single examinee on one or more latent traits across testing occasions.[3,4]

Previous simulation research has highlighted the promising statistical properties of various hypothesis tests (e.g., a likelihood ratio test) for AMC. For example, the efficacy of these tests has been examined across multiple testing occasions,[5,6] when measuring multiple latent traits,[6] and in the presence of item parameter estimation error.[7] A common thread among these studies (and other research on the analysis of change) is the assumption that longitudinal measurement invariance (MI) holds across the testing occasions. In other words, the measurement model connecting the items to each latent trait is assumed to be equivalent over time. Only with evidence of longitudinal MI can researchers ensure that observable test score differences are reflective of real changes on the same latent trait, rather than an artifact of measurement error or the influence of other construct-irrelevant factors.[8] If longitudinal MI does not hold, AMC might be less likely to accurately distinguish significant individual change on the trait of interest.

Extending the extant AMC literature, the current study presents a Monte Carlo simulation to evaluate AMC's performance when the assumption of longitudinal MI is violated. Across two testing occasions, AMC is used to measure within-person change on one latent trait. Lack of longitudinal MI is then introduced by modifying the item parameters between testing occasions, a phenomenon called item parameter drift (IPD). AMC's ability to identify change on the trait of interest is examined when manipulating the type of IPD, the percentage of IPD, and the magnitude of latent trait change between testing occasions. AMC's performance then is evaluated using three indices: (a) false positive rates (identifying significant change on the trait of interest when no change is present), (b) true positive rates (identifying significant change when change is indeed present), and (c) change recovery. Results will add to a growing literature on the applicability of AMC for realistic testing scenarios.

References

[1] Little (2013). Longitudinal structural equation modeling.

[2] Wang & Nydick (2020). J Educ Behav Stat.

[3] Kim-Kang & Weiss (2008). Z Psychol.

[4] Weiss & Kingsbury (1984). J Educ Meas.

[5] Phadke (2017). Doctoral dissertation.

[6] Wang et al. (2020). Multivar Behav Res.

[7] Cooperman et al. (2022). Educ Psychol Meas.

[8] Meredith (1993). Psychometrika.

Ming Him Tai & David J. Weiss

Detectability of Profile Patterns with Adaptive Measurement of Individual Change

Abstract: The measurement of individual change is an important topic in psychology and education. Adaptive measurement of change (AMC) is an emerging paradigm to study the detectability of psychometrically significant individual change. To date, research in this paradigm has examined the effects of a number of design factors, with a focus on hypothesis testing methods (Cooperman, Weiss & Wang, in press; Wang et al., 2020; Wang & Weiss, 2017; Finkelman et al., 2010). On finding of this set of studies was that latent trait (theta) change magnitude and pattern have a major effect on its detectability. However, these studies arbitrarily specified several sets of theta patterns. The present study proposed a quantitative framework to systematically design theta patterns. Based on the approach developed by Cronbach and Gleser (1953), this study characterized theta patterns by three parameters: (a) magnitude of change, (b) scatter, and (c) shape. In total there were 431 change patterns examined. This study also included time 1 true theta value (7 levels) and test bank design (3 configurations) as design factors. Altogether there were 9,051 conditions examined.

The dependent variables in the study were true positive rate (TPR; also known as power) and a theta change recovery index (CRI), defined as the root mean squared error (RMSE) of the estimates of change in theta. To obtain precise estimates of theta at all four testing occasions, which is necessary for computing CRI, fixed-length testing was used instead of variable-length testing at all testing occasions. Our preliminary simulations showed that a test length of 25 items is sufficient for the root mean squared error (RMSE) of theta to stabilize.

Throughout the testing administrations, each theta value was estimated with maximum likelihood estimation (MLE). Each simulation condition was replicated 1,000 times (i.e., it consisted of 1,000 simulees). True (error-free) item parameters were used in the simulations because Cooperman et al. (in press) showed that item parameter estimation error has almost negligible effect on AMC performance. Only the likelihood ratio test (LRT) was used to determine omnibus psychometric significance because most other hypothesis testing methods all demonstrated desirable statistical properties in AMC (e.g. Wang et al., 2020; Finkelman et al., 2010). All simulations were conducted using R statistical software (R Core Team, 2021). The catlrt library (Nydick, 2014) was used to estimate theta and the ggplot2 library (Wickham, 2016) was used to create plots.

Key References

Cooperman, A. W., Weiss, D. J., & Wang, C. (2021). Robustness of Adaptive Measurement of Change to Item Parameter Estimation Error (in press)

Cronbach, L. J., & Gleser, G. C. (1953). Assessing similarity between profiles. *Psychological Bulletin*, 50(6), 456–460.

Finkelman, M. D., Weiss, D. J., & Kim-Kang, G. (2010). Item selection and hypothesis testing for the adaptive measurement of change. *Applied Psychological Measurement*, *34*(4), 238–254.

Wang, C., & Weiss, D. J. (2018). Multivariate hypothesis testing methods for evaluating significant individual change. *Applied Psychological Measurement*, *42*(3), 221–239.

Wang, C., Weiss, D. J., & Suen, K. Y. (2020). Hypothesis testing methods for multivariate multi-occasion intra-individual change. *Multivariate Behavioral Research*, 1–17.

Joseph N. DeWeese, Chun Wang, & David J. Weiss

Stochastic Curtailment Stopping Rules for the Adaptive Measurement of Change

Abstract: In many applications of computerized adaptive testing (CAT) it is of interest to identify individuals who have changed across testing occasions. Adaptive measurement of change (AMC) combines psychometric hypothesis testing methods for change at the individual level with CAT to effectively and efficiently detect change in a latent trait (θ) (Weis & Kingsbury, 1984; Finkelman et al., 2010). Previous research on AMC has shown that the likelihood ratio (LR) test has desirable statistical properties in unidimensional, multidimensional, two-occasion, and multi-occasion testing scenarios (Wang et al., 2020).

The LR test can be thought of as a classifier of significant or non-significant change for a given individual and used as a stopping rule, as in classification-based CAT. In a two-occasion testing scenario, the test at the second occasion would continue until psychometrically significant change has been identified or a maximum number of items has been administered. However, testing efficiency could be improved by stopping the test when significant change is sufficiently unlikely to be detected. This idea follows from work by Finkelman (2008), who introduced stochastic curtailment (SC) in CAT for sequential mastery testing. Stochastic curtailment stops a test when the probability of a change in the classification decision is low, given the current data (Finkelman, 2010).

In this study, SC and expected stochastic curtailment (ESC) methods are introduced for multidimensional, two-occasion AMC. Then, a simulation study is conducted to evaluate the performance of SC and ESC as secondary stopping rules when the LR test is used as the primary stopping rule. These methods are compared to the LR test as a solitary stopping rule and to a fixed-length (FL) test with post-hoc application of the LR test to identify change. A three-dimensional graded response model with four response categories is used for all items in ideal and realistic item banks. In all simulation conditions, a fixed-length test is simulated at occasion 1. In a no-change condition, simulees retain their same true θ values at occasion 2. For six change conditions, different magnitudes and patterns of change are added to each simulee's true θ for occasion 2. Then, occasion 2 CATs are simulated using the four different stopping rules (LR, LR+SC, LR+ESC, FL). The evaluation criteria are the true positive rate, false positive rate, true negative rate, false negative rate, average occasion 2 test length, and the proportion of tests stopped by each rule.

References

Finkelman, M. (2008). On Using Stochastic Curtailment to Shorten the SPRT in Sequential Mastery Testing. *Journal of Educational and Behavioral Statistics*, *33*(4), 442–463. https://doi.org/10.3102/1076998607308573

Finkelman, M. D. (2010). Variations on Stochastic Curtailment in Sequential Mastery Testing. *Applied Psychological Measurement*, 34(1), 27–45. https://doi.org/10.1177/0146621609336113

Finkelman, M. D., Weiss, D. J., & Kim-Kang, G. (2010). Item selection and hypothesis testing for the adaptive measurement of change. *Applied Psychological Measurement*, *34*(4), 238-254. https://doi.org/10.1177/0146621609344844

Wang, C., Weiss, D. J., & Suen, K. Y. (2020). Hypothesis Testing Methods for Multivariate Multi-Occasion Intra-Individual Change. Multivariate Behavioral Research. https://doi.org/10.1080/00273171.2020.1730739

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361–375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

12: Paper Session - Automated Scoring in CAT with AI

Chair: *Maaike van Groen*

Angela Verschoor

Open-ended Items and decision trees

Abstract: Immediate scoring is an essential element in computerized adaptive testing, and thus has to be handled by the computer as well. Yet, machine scoring is mainly limited to closed-response items, including hotspot items, drag-and-drop items, and many other beautiful but laborious-to-develop item types. The easiest-to-develop items, such as those for which the answer is a single sentence, while mistakes in spelling and grammar are not to be taken into account, are usually left to paper. Already 30 years ago, there were optimistic research papers announcing that within a few years, machine scoring would be widely used. Since then it has been rather quiet, despite new techniques like Machine Learning.

In this presentation we will take a closer look at the complexity of the problem, and we propose a two-phase method for machine scoring. The first phase detects single keywords in a sentence despite typos and bad spelling, and rewrites the words into their lemmas. Thus, alternative formulations like single / plural or past / present tense will be ignored. The second phase uses a decision tree method to predict correctness of the given response. Decision tree methods use a training set of responses that were scored manually.

Usually, training sets of approximately 1000 responses are claimed to be appropriate. Unfortunately, this is often a serious drawback, especially since correctness of the machine scoring procedure still cannot be guaranteed. We performed a study on the accuracy of the proposed method using responses on approximately 300 different items, taken from tests used in the last year of vocational education. The sizes of the response sets varied from ca 100 to ca 4000. In all cases, random samples from the responses were taken as training sets, while the remaining responses were used to evaluate the accuracy of the machine scores. When the response sets were split in two equal parts, correlations between machine scores and manual scores were between 0.70 and 0.95, while usually, training sets of 100 responses produced correlations between 0.65 and 0.95. At the same time, several response sets were rescored manually by different markers. Correlations between these two human markers were in the same order of magnitude for the vast majority of items. Higher correlations were observed for items, where there are only limited ways to give a correct response, while the “more open” the item was, the lower the correlation was, also when larger training sets were used.

The method we propose is certainly not the final solution to the problem of machine scoring. Especially for high stakes environments, this method is deemed not good enough to be employed. But in spring 2021, experiments have started in low stakes environments. At the moment as linear test, but employment in CAT will probably be the next step.

Daniel Bengs, Ulf Brefeld, Ulf Kroehne, & Fabian Zehner

IRT Models for Automatically Scored Text Responses

Abstract: Test items using open-ended response formats can increase an instrument´s construct validity, but traditionally, their application in educational testing requires human coders to score responses, which entails substantial costs. Moreover, the human-coded scores are not available during on-line testing and cannot be used to inform routing decisions in adaptive designs, such as Computerized Adaptive Testing or Multi-Stage Adaptive Testing. Automatic response coding uses machine learning and natural language processing to alleviate these difficulties by enabling computers to instantly assign scores to text responses. The underlying classification algorithms are trained to reproduce the human-coded scores, but in practice, they do not reach perfect accuracy and thus introduce an additional source of error into the process. The human-coded scores also define the reference frame for item calibration. Hence, the resulting measurement model does not account for classification error and is misspecified when used with automatically coded responses.

We propose to explicitly model classification errors by conditional probabilities of misclassification (in the simplest case, estimated from the classifier's confusion matrix). By incorporating these error probabilities into the IRT model for the human-coded scores, we derive measurement models that allow inferring the latent trait from the automatically coded responses. As a special case within this framework, we recover the four-parameter logistic IRT model as the composite of a response governed by a two-parameter logistic IRT model and a noisy classifier that errs conditionally independently of the latent trait. We study the impact of classifier accuracy on item information with empirical data from 8 items from the PISA 2018 reading domain. Simulating the PISA 2018 Reading Multistage Adaptive Test allows us to evaluate the potential gains afforded by the implementation of automatic response coding for all open-ended items. The results show that, for the studied PISA design, including the information from open-ended items improves the accuracy of interim ability estimates significantly and substantially, while leaving the precision of final ability estimates using human-coded scores unaffected.

Leonidas Bourikas, Stefanos Nikolaou, & Georgios Sideridis

Enhancing scoring of oral responses in CAT based on Convolutional Neural Networks

Abstract: Reading is the foundation of learning that helps us acquire knowledge and information about the world. Given that children learn how to read in the first years of formal education, assessing reading abilities at early educational stages may be critical for the identification of reading-related learning difficulties. One of the best methodologies that has been proposed for measuring reading ability in our contemporary digital world is Computerized Adaptive Testing (CAT), revolutionizing assessment with personalisation and brief yet precise measurement.

Moreover, reading ability is a latent variable measured indirectly and mediated by different important processes, such as text reading comprehension, vocabulary, word decoding and word reading fluency. Some of these processes are assessed using tasks administered in Multiple Choice Question format (MCQ) and scored automatically by a computer system. Some other processes, such as word decoding, require an examiner administering the task and scoring the accuracy of the pronounced words manually. So far, computerized batteries assessing reading ability in Greek have solely focused on evaluating word and pseudoword decoding abilities in MCQ format and not on items calling for word articulation.

Thus, the problem-gap to be addressed here is: “How to automatically and reliably measure decoding ability that doesn’t require the presence of an examiner administering and scoring whether a student uttered real-words or pseudowords correctly or incorrectly?”

To address the gap, we propose a practical application using voice recognition software. The recommended software is based on deep learning techniques. The novelty of this research is the use of voice data to identify the probability of a sound segment to contain a phonetic mistake. This methodology will be used to extend the capabilities of Computerized Reading testing, even further, in comparison with previous decoding MCQ measuring attempts by Greek computerized batteries.

During the presentation, we will go through the architecture of the voice recognition deep learning model and the steps we followed to analyse the voice-related data. Furthermore, we will describe how this application can be used to identify phonetic mistakes, while providing examiners with performance reports with minimal effort.

Importantly, we believe that such an application can be useful in several settings, especially in remote working conditions. Imagine, using this application to screen for reading difficulties online, in cases physical presence is prohibited. Furthermore, such a solution holds the possibility of large-scale student population assessment simultaneously without the need for specialized equipment and staff. The automation of the process will reduce waiting lists for an assessment, which in some countries translates into several months of waiting and consequently free up the workload of educational centres running in-person evaluations. In cases where no access to an assessment centre is possible due to geographic or other restrictions, children could still be assessed using a reliable, user-friendly, and evidence-based tool.

Finally, the proposed solution can be used to automate the scoring of students’ performance, in a way that can be compared with a normative sample. Then it will be useful as a screening tool to timely refer cases to assessment services when needed.

Nathan Thompson

Applications of Machine Learning in Automated Essay Scoring

Abstract: Automated essay scoring (AES) is, like adaptive testing, an important application of artificial intelligence to assessment. This presentation will present the development and validation of a user-friendly tool for AES.

First we will present the problem of AES and how machine learning addresses it. We discuss the high-level algorithm that was selected, the bag-of-words model based on natural language processing (NLP) followed up by machine learning to select features and establish a predictive model.

Next, we validated our algorithm on several data sets. Dependent variables to drive the comparison includes quadratic weighted kappa, correlation of predictive and observed scores, and time required to run. We found that some relatively parimonious models performed nearly as well as extremely complex models, but ran in a small fraction of the time.

Finally, we connected our algorithm to a user-friendly interface that will allow any organization to easily implement this AES algorithm, fitting multiple models and evaluating the result to select the one that best meets their needs. We will discuss the development of that software and how it is configured for general usage.

13: Paper Session - New Approaches to CAT

Chair: *Samuel Greiff*

Johann-Christoph Münscher, Marcus Bürger, & Philipp Yorck Herzberg

Real-time procedural stimulus generation for adaptive testing of attention - The Continuous Matching Task (CMT)

Abstract: The Continuous Matching Task (CMT) is a novel paradigm designed to measure sustained attention and alertness. It is a special type of Continuous Performance Task (CPT) that utilizes truly continuous stimulus material. The tasks offers a range of innovative approaches that leverage computerized adaptive testing. A procedural algorithm is used to generate stimuli in real-time, which which also enables adaptive testing. The task is highly flexible and can be used in either single or dual task configurations that also allow for task mixing. The viability of the CMT was tested and results were compared with similar tasks, i.e., Stroop-Task and Conner's CPT (CCPT), as well as self reports of ADHD in adults in a Multi-Trait-Mult-Method approach in a sample of N=122 participants. Self-reports of task load and measurements of heartrate variability during testing were analysed to infer and compare mental workload during tasks. Employing the dual task CMT with adaptive difficulty resulted in the highest reliability and validity. Results indicate that the CMT is primarily a measure of alertness and processing speed and benefits from adaptive testing. The functionality of the algorithm alongside shortcomings and advantages of real-time stimulus generation as well as adaptive testing are presented and discussed.

Tuo Liu & Sacha Epskamp

Adaptive Form of Network Psychometrics: a simulation study

Abstract: Currently, computerized adaptive testing (CAT) relies heavily on item response theory (IRT). However, when the assumptions of the IRT model are violated, CAT cannot perform well. Especially, most IRT models assume the theoretical realism of the common cause model and local independency, which is not likely to be satisfied in many situations. For example, the p-factor of psychopathology was criticized due to the lack of theoretical background. As a complementary approach, network psychometrics handles the measurement as a stable organization of dynamic components that mutually activate one another, relaxing IRT assumptions. Recently, an adaptive form of network psychometrics was developed, including a network-based item selection algorithm and a final model estimation method using the Ising model. More specifically, this item selection algorithm would administrate the item, which could maximally reduce the entropy to predict the responses of unanswered items given the already answered items. This process would be repeated until all remaining unanswered items can adequately be predicted using an Ising model instead of the IRT model.

Although the adaptive form of network psychometrics theoretically relaxes the assumptions of IRT models and may offer a better alternative in CAT, only rare empirical studies were conducted. Based on this consideration, this study plan to compare the performance of the adaptive form of network psychometrics and CAT by item selection algorithm and model estimation method. First, we compared the item selection algorithm in CAT with the adaptive form of network psychometrics. We used the Kullback-Leibler (KL) divergence criterion in CAT because it is a global information measure compatible with the idea in the adaptive form of network psychometrics. The most used Fisher information (FI) criterion in CAT was used as the baseline. Second, we compared the Ising model and IRT model to estimate the responses of unknown items, no matter which items selection algorithm was used. . Therefore, we conducted one stimulation based on data from previous psychometric research, which has shown that this data largely violated the local independence assumption of IRT, and it is theoretically implausible to have latent variables. This simulation study is based on a factorial design with factor item selection (KL in CAT vs FI in CAT vs Entropy in Network), factor model estimation (IRT vs Ising) and the number of administrated items (from one to all). As the dependent variables, the proportion of misclassification error will be compared among all combinations of factors. As an exploratory approach, the results will provide detailed insights into the complementary application potential of the adaptive form of network psychometrics.

Ray Clifford & Matthew Wilcox

Not every trait is latent: Towards a multi-stage testing approach that rates second-language reading and listening proficiency

Abstract: This paper explicates the theory and research behind a novel approach to language proficiency assessment using a fuzzy logic algorithm to mimic an established method of human-rating.

The American Council on the Teaching of Foreign Languages (ACTFL) uses a robust framework for assessing language proficiency, comprised of five major levels of language proficiency: Novice, Intermediate, Advanced, Superior, and Distinguished, including sublevels at each major level. This framework is derived from, and shares many similarities with, the U.S. Governments' Interagency Language RoundTable (ILR) proficiency standards. For the productive skill of speaking, each major level in this framework has a set of clearly defined criteria that describe what a language learner can do. Highly trained human raters conduct a structured yet flexible Oral Proficiency interview (OPI) that identifies the major level, or "floor" at which the speaker can sustain performance, and also probes the next higher major level, or "ceiling," for linguistic breakdown. In essence, a skilled interviewer applies a human algorithm that tailors the OPI and provides a rating of speaking ability which describes what the speaker can do in real-world situations.

Our research and development explore how this process used in a human tailored OPI can be automated in objectively scored tests of reading and listening proficiency. Whereas other reading and listening tests might apply IRT CAT procedures to a bank of test items, our approach varies significantly in that the theoretical model, test development model, and scoring model (including the branching algorithm) are consistently aligned, as described by Luecht (2003).

Theoretical model. Our test design demands establishing content validity first and foremost; each item must align with established criteria using a 3-factor matrix of Task, Condition, and Accuracy (TCA) with the associated criteria of Author Purpose, Text Type, and Topical Domain.

Test Development Model. Then the test design is aligned with the multi-level Theoretical model.

Psychometric Scoring Model. By definition, these language proficiency tests are criterion-referenced. To maintain alignment with the theoretical construct and test development models, the approaches commonly used to score norm-referenced tests have been replaced with a scoring model that resembles a multistage CAT. Our version differs in that while items are trialed and parameters estimated, the item statistics are used as an added verification that the items meet the qualitative criteria described in the ACTFL proficiency guidelines. Then branching and scoring algorithms apply 'fuzzy' logic during the assessment to mimic the floor and ceiling approach a trained rater would make in an oral interview. Lastly, the final, summative rating is not derived by using multiple cut scores along a continuum of total test scores; rather, it is calculated according to the test-takers ability (or inability) to sustain performance when responding to sets of items that target each functional level.

This presentation will provide both the theoretical underpinnings and evidence for validity for this novel approach to automatic rating of reading and listening.

Bryan Maddox

Extending the analysis of student performance with process data

Abstract: Computer Adaptive Test (CAT) designs including multi-stage adaptive tests, have the advantage of being able to match the difficulty of items with respondent performance, to improve the granularity of student assessment (Rutkowski et al, 2022). They also have wider virtues, such as offering a shorter, more personalised, more inclusive, and more enjoyable test taker experience (Burstein et al. 2021). However, in certain contexts, such as some national school examinations, standardisation rather than differentiation, may be considered as the primary basis of fairness. I.e., the idea that in each person should have the opportunity to demonstrate their ability on the same test (Nisbett and Shaw, 2019). In that way, assessment accuracy and the sense of fairness can pull in different directions.

In this paper we therefore explore whether we might arrive at some of the goals of CAT via other routes. We describe how process data from log files in interactive, technology enhanced items can be used to differentiate and extend the analysis of student performance in the absence of adaptive designs (Salles et al, 2020; Goldhammer et al, 2021). To do this, we present a case study of large-scale, standardised mathematics assessment in French secondary schools. With evidence from an eye tracking and video study conducted in French classrooms, and the analysis of item log data, we show how 'processes models' (Kane and Mislevy, 2017) can be applied to extend 'product' based data on test scores. In conclusion we critically appraise whether such an approach might be considered as a viable alternative to adaptive designs.

References:

Burstein, J. et al. (2021) 'A Theoretical Assessment Ecosystem for a Digital-First Assessment -The Duolingo English Test', Duolingo Research Report DRR-21-04. Available at: englishtest.duolingo.com/research.

Goldhammer, F., Hahnel, C., Kroehne, U., & Zehner, F. (2021). From byproduct to design factor: on validating the interpretation of process indicators based on log data. *Large-scale Assessments in Education*, 9(1), 1-25.

Kane, M., & Mislevy, R (2017). Validating score interpretations based on response processes for the next generation of assessments. In K. Ercikan & J.W. Pellegrino, J.W. *Validation of Score Meaning for the Next Generation of Assessments: The uses of Response Data*. Routledge.

Nisbet, I. & Shaw, S.D. (2019): Fair assessment viewed through the lenses of measurement theory, *Assessment in Education: Principles, Policy & Practice.*

Rutkowski, L., Rutkowski, D., Valdivia, S. D. (2022) Multistage Test Design Considerations in International Large-Scale Assessments of Educational Achievement. International Handbook of Comparative, Large-Scale Studies in Education. Springer.

Salles, F., Dos Santos, R & Keskpaik, S. (2020). 'When didactics meet data science: process data in large-scale mathematics assessment in France'. *Large-Scale Assessments in Education*, 8 (7).

14: Keynote - Ying Cheng: Cognitive Diagnostic Computerized Adaptive Testing: Recent Developments and Future Directions

Chair: *Anthony Zara*

Abstract: This talk will provide a comprehensive and up-to-date overview of cognitive diagnostic computerized adaptive testing (CD-CAT). Compared to the well-known traditional CAT, a key distinction of CD-CAT is that its goal is to obtain the latent mastery profile for each respondent in an efficient manner, typically based on the cognitive diagnostic models (CDMs), also known as the diagnostic classification models (DCMs). In contrast, the goal of the traditional CAT is to reach an accurate latent ability estimate or multiple latent ability estimates quickly, and in some cases make a classification decision on that basis. In this talk, the connections and differences will be discussed between CD-CAT, unidimensional CAT, multidimensional CAT, and classification CAT. Under CD-CAT, this talk will cover both the single-purpose CD-CAT, which focuses on the latent mastery profile itself, and dual-purpose CD-CAT, which intends to estimate both the latent mastery profile and the latent ability simultaneously. In addition, emerging topics such as extension of CD-CAT to incorporate response times, cognitive diagnostic multi-stage testing, and automated test assembly under CDM/DCM will also be explored.

15: Invited Symposium – Adaptive testing in PISA: past, present and future (2 parts)

Chairs: *Janine Buchholz, Mario Piacentini, & Francesco Avvisati*

Discussant: *Matthias von Davier*

Abstract: Starting with the 2018 cycle, the Programme for International Student Assessment (PISA) uses a multi-stage adaptive testing (MSAT) design to assign different test forms that are matched to students' ability. This initial foray into adaptive testing helped PISA address test-fairness concerns (by limiting the share of respondents who are given tests that do not allow them to demonstrate their full proficiency); eliminated the need for country-level adaptations; and achieved some reductions in measurement error, especially for students with exceptionally low or high performance. Specifically, a MSAT design with two branching points and a non-adaptive (random probability) layer was chosen to control exposure of items (for item calibration) and manage non-statistical constraints (coverage of sub-constructs). Only preliminary estimates of item characteristics were available for the adaptive decisions, and all item parameters were re-calibrated after the adaptive administration.

The lessons learned in this first experience have informed the design for PISA 2022. Starting with PISA 2022, multiple domains are being administered in adaptive fashion. Looking beyond 2022, there is a clear potential to improve the designs and associated methodologies to further increase both measurement precision gains and student engagement during the test.

Several papers presented at this symposium will illustrate the challenges that the introduction of adaptive testing in PISA faced and the opportunities that exist to introduce methodological innovations. The first set of presentations review the past and presence of MSAT designs in PISA, while the second set explores ways of introducing future innovations in the context of PISA. The opening presentation by Hyo Jeong Shin will demonstrate the robustness of the adaptive design first implemented in the PISA 2018 reading test. Based on this, Peter van Rijn will review the technical challenges encountered in the design of the PISA 2022 mathematics test and explain how they were addressed. The presentation by Janine Buchholz will examine the potential benefit of adaptive testing in terms of test engagement, which holds particular promise in the context of low-stakes large-scale assessments. Finally, the presentations by Andreas Frey and Hua Hua Chang will explore the potential benefits of introducing greater adaptivity in the design, such as through testlet-based computerised adaptive testing combined with shadow testing (ST), and on-the-fly assembled multistage adaptive testing (OMST). In the discussion, Matthias von Davier will reflect on the five presentations and provide some concluding remarks.

Presenters:

Hyo Jeong Shin, Christoph König, Frederic Robin, Kentaro Yamamoto, & Andreas Frey

Robustness of Multistage Adaptive Testing Designs in Educational Large-Scale Assessments

Abstract: Recent transitions to the computer-based assessments in international large-scale assessments (ILSAs) enabled the introduction of multistage adaptive testing (MST) designs. As one of the most popular ILSAs, the Programme for International Student Assessment (PISA) has started implementing the MST from the 2018 cycle. Consequentially, it is important to evaluate the robustness of the MST designs because the estimated item parameters through MST designs continue to be used for the future cycles. Therefore, this study examined the robustness of the PISA 2018 MST designs impacted by three factors: (1) when the construct is measured through less number of items, (2) when the item-by-country interactions are present, and (3) when the test-takers skip or do not reach the items. We evaluated the item parameter recovery and expected gains in measurement precision through a simulation study with 100 replicates. The simulation study revealed that PISA 2018 MST design is robust to those three factors, showing an acceptable level of parameter recovery and improved measurement precision, about 4-5% compared to the non-adaptive realistic benchmark in which students randomly took the same number of units.

Peter van Rijn, Usama Ali, Hyo Jeong Shin, & Frederic Robin

Stepwise Assembly for Multistage Adaptive Testing: An Application to PISA Mathematics

Abstract: Multistage Adaptive Testing (MSAT) is increasingly used in large-scale survey assessments (LSA) to improve both data collection and measurement. MSAT was first introduced in PISA in 2018 for reading, one of the three major domains assessed. The PISA 2022 MSAT design introduced several changes with respect to the design that was used for the 2018 cycle. The first change is that each item occurs in each stage to achieve the goal of fully balancing item position across all of the test forms. The second change is that a third difficulty level was added in the third stage to allow for further adaptation. The third change was to include linear paths in the design instead of relying on the misrouting of a small proportion of students to ensure large enough sample sizes per item needed for parameter estimation. The fourth change is the use of formal methods for optimal test assembly (van der Linden, 2005).

This presentation will focus on the fourth change and describe the stepwise mixed-integer linear programming assembly approach that the authors used to develop the design for the PISA 2022 Mathematics assessment, with a full specification of all objectives and constraints. Results of a series of simulations conducted to evaluate this MSAT design compared to others, linear and adaptive, will be presented. The designs will be compared in terms of measurement precision, item calibration accuracy and content coverage across the whole range of proficiency. The implications of these findings for adaptive testing designs in the context of PISA and LSA in general will also be discussed.

References:

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: Springer.

Janine Buchholz, Hyo Jeong Shin, & Maria Bolsinova

Test engagement in multistage adaptive testing

Abstract: The core feature of adaptive testing consists in the allocation of test items of appropriate difficulty given a respondent's performance on the test, resulting in the well-known increase in measurement precision. Another, less frequently mentioned effect of the improved match between performance and item difficulty relates to motivational factors over and above a potential reduction of testing time: presenting examinees with test items of appropriate difficulty prevents extreme over- and under-challenging that might cause respondents to disengage from the test. Only scarce evidence exists to show this additional benefit of adaptive testing that holds great promise in the context of large-scale assessments, which are typically low-stakes in nature and for which student engagement is of particular concern.

This study aims to fill this gap by drawing on a specific feature of PISA 2018, i.e. the multi-stage adaptive design with two branching points and a non-adaptive (random probability) layer to control exposure of items (for item calibration) and manage non-statistical constraints (coverage of sub-constructs). At each of two branching points, a randomly selected subset of low- and high performing students each were intentionally misrouted to a difficult and easy booklet, respectively. The misrouting condition, therefore, represents a mismatch between performance and item difficulty. Analyses are based on comparisons between correctly routed and misrouted students on a set of different measures of disengagement such as rapid responding and performance decline. Results point at differential effects of the performance-difficulty mismatch for high- and low-performing students.

Andreas Frey, Christoph König, & Aron Fink

A Highly Adaptive Testing Design for PISA

Abstract: PISA reports results on the population level using plausible values (PVs). The statistical uncertainty of these PV-based estimates has three major sources: sampling error, measurement error, and linking error. The measurement error can be reduced by optimizing which items are presented to the individual students. To this end, from 2018 on the Programme for International Student Assessment (PISA) switched from using a multi-matrix design that specified a number of linear booklets to using multi-stage testing (MST) to assign items to students. The transition to MST led to an increase of 4–7% in the test information compared to the mode of nonadaptive item presentation used before. Constructing an MST design that accounts for all the relevant constraints of PISA and documenting it transparently, however, is labor-intensive and typically contains more restrictions than necessary. One of which is the number of stages. Further gains in terms of test information can be expected by adapting not on only a small number of stages, but on the most fine-grained level possible, while accounting for PISA’s constraints. In order to find out the magnitude of these test information gains, we examined the performance of using testlet-based computerized adaptive testing (CAT) with between- and within-testlet adaptivity combined with shadow testing (ST) for constraint management. The PISA 2018 Reading MST design was examined under PISA-typical conditions (in terms of sample size, omitted responses, and not reached items) with a simulation study. We added research conditions with an individual response probability of .62 instead of .50 to examine the decrements in test information when the transition to CAT is used to foster the test-taking experience by presenting relatively easy items. The MST design was compared to 12 research conditions based on a fully crossed factorial design with the IVs “Item Pool Optimality” (PISA 2018, optimal), “Response Probability (RP)” (.50, .62), and “Ability level” (low, medium, high). In all conditions, the same constraints were accounted for by shadow testing. The main DVs were the relative test efficiency compared to the PISA 2018 MST design, constraint violation, and item exposure. All CAT specifications clearly outperformed the PISA 2018 MST design. Regarding test information, the most restrictive condition using the PISA 2018 item pool and a RP of .62 resulted in a relative test information of 1.27 to 1.28 across the three ability groups and thus a substantial increase in test information of 27% to 28%. When a response probability of .50 is used with the PISA 2018 item pool, the relative test information was 1.28 to 1.35. For an optimal item pool, the relative test information ranged from 2.42 to 3.06 for RP = .62 and from 2.53 to 3.34 for RP = .50. These results underline that it is worthwhile to implement a more fine-grained level of adaptation in PISA, that within-testlet adaptivity is promising for PISA, that presenting relatively easy items will not reduce the test information dramatically, and that systematically developing the PISA item pool further will unlock the full potential of CAT for PISA.

Xiuxiu Tang, Yi Zheng, Tong Wu, Kit-Tai Hau, & Hua-Hua Chang

On-the-fly Multistage Testing Design for PISA Assessment Incorporating Response Time

Abstract: Multistage adaptive testing (MST, a.k.a., multistage testing) has drawn a widespread interest as many large-scale testing/assessment programs began to adopt this test design over the past decades. This study explores the potential use in PISA of a new adaptive test design named "on-the-fly multistage adaptive testing" (OMST; Zheng & Chang, 2015), which combines the merits of computerized adaptive testing (CAT) and MST and offsets their limitations. The main difference between OMST and MST is that modules in OMST are assembled on the fly to match the given examinee's level, while modules in MST are all preassembled before test administration. The tested OMST design also incorporates response time information of the items to improve measurement efficiency. Traditionally, measurement efficiency is solely assessed by the number of administered items required for a given measurement accuracy. However, the amount of time examinees spent to complete the test should also be counted in assessing measurement efficiency. Via simulations mimicking the PISA 2018 reading test settings, including using the real item attributes and replicating the 2018 reading MST design, we compared the performance of the OMST design against the 2018 MST design in terms of (1) measurement accuracy of examinees' latent traits, (2) test time efficiency and stability: mean and standard deviation of test time, (3) test security: the exposure rates of individual items as well as mean and standard deviation of test overlap rates, and (4) constraint violations: the average occurrence rate of the violation of a content constraint.

References:

Zheng, Y., & Chang, H. H. (2015). On-the-fly assembled multistage adaptive testing. Applied Psychological Measurement, 39(2), 104–118.

Matthias von Davier

Discussion

16: Symposium – Developing a system that connects learning and adaptive testing for adults learning to read (2 parts)

Chair: *John Sabatini*

Discussant: *Samuel Greiff*

Abstract: Worldwide, adult reading literacy remains a significant problem. PIAAC estimates that 19.8% of adults read at or below Level One, the lowest proficiency level, and another 34% at Level Two. Levels One and Two span "beginning reader" to "secondary" levels of proficiency. The population encompassed by these estimates is incredibly diverse. Adults in need of further reading literacy development range in lifespan development (age 16 to 80+); racial and ethnic groups; migrant and refugee English language learners; subgroups with known or hidden physical, mental, or learning disabilities; geographical dispersed; and varied in socio-economic status. One cannot rely on valid educational records to document any individual differences that may impact future learning. To further complicate matters, adults may have spent years compensating for their lack of reading proficiency, resulting in a profile of skills with relative strengths and weaknesses, with some compensatory strategies being counterproductive to sustained growth. Further, adults have their own complex lives of work, family, and personal responsibilities – so the time they can devote to learning may be limited, underscoring the importance of efficient, adaptive assessment and instruction in this domain.

This educational context requires tailored instruction informed by rich assessment, not only to determine starting places for an instructional program, but what profile of reading skills/components they present, which instruction to assign, when are they achieving a level of growth in proficiency to move on to different skills, strategies, or levels of instruction over time. This problem context requires our most advanced efforts in combining adaptive assessment with adaptive algorithms to tailor the learning experience to each individual.

In this symposium, we discuss development and merging of an intelligent tutoring system and a computer adaptive testing system. The two systems were initially developed separately and now, through a grant from the US DOE Institute for Education Research, are being brought together. To discuss this work and its implications, we will a) describe an adaptive intelligent tutoring system designed to support reading comprehension gains in adults with low literacy, b) describe a multi-stage assessment system used to measure a specific set of component reading skills to differentiate between struggling and proficient readers, c) seek to establish a technological and logistical framework for linking adaptive assessment and instruction, and d) report empirical efforts to psychometrically align the above systems and frameworks.

Presenters:

Art Graesser, Xiangen Hu, John Sabatini, & John Hollander

AutoTutor for Adult Reading Comprehension: An intelligent tutoring system

Abstract:

Significance: Approximately one in five adults in 33 OECD countries have literacy skills that can be described as “”low levels of proficiency”” [1]. Adult literacy learner populations are characterized by immense diversity in their linguistic, educational, and cognitive backgrounds [2]. As a result, adaptivity in adult-focused educational systems is that much more vital. AutoTutor for Adult Reading Comprehension (AT-ARC) is a web-based intelligent tutoring system that is designed to help adult learners develop effective reading comprehension skills in English through conversational trialogues with computer agents. Lessons span basic reading skills (vocabulary, word parts), comprehension of sentences and texts in different text genres, and rhetorical structures, including digital documents and media. AT-ARC is adaptive on multiple scales. At the item level, conversational agents respond to learner performance to provide guidance and enhance learning opportunities whenever possible. At a lesson level, the system selectively presents texts and item sets at higher and lower difficulties depending on learner performance during universal, medium-difficulty question sets. At a curriculum level, lesson plans can be individualized based on the specific strengths and weaknesses of each learner’s reading comprehension skills.

Methods: To inform the automatic, adaptive decisions, AT-ARC can employ formative, stealth, and summative assessment in various capacities. Data collected from participants in several studies include item accuracy and response times, as well as other features that can be used to continuously model learner characteristics and performance over the course of use. We have conducted several studies using varying analytic approaches to understand what this data can tell us about adult literacy learners.

Results: The data generated from users may be used to understand and improve adults’ reading comprehension skills based on AT-ARC’s theoretical, multilevel framework of comprehension. This presentation will outline the results of several studies using AT-ARC. In one study, learners’ motivation and engagement profiles were determined by data mining their various interactions with the system, which could help maximize learning gains. In another study, AT-ARC was used in conjunction with classroom instruction in adult literacy programs in the US and Canada, resulting in reading comprehension gains. Additional analyses demonstrate the use of AT-ARC data to understand and classify the cognitive and motivational characteristics of learners, which may be used to inform adaptive instruction.

Discussion: In this presentation, we will describe AT-ARC’s approach to adaptive reading comprehension instruction and how this adaptivity is informed by assessment in various capacities. The integration of other methods of adaptive assessment will also be discussed, as well as the generalizability of this type of testing-instruction connection will be discussed.

References:

Fang, Lippert, Cai, Chen, Frijters, Greenberg, & Graesser. (2021). Patterns of adults with low literacy skills interacting with an intelligent tutoring system. *International Journal of Artificial Intelligence in Education,*

Sabatini, O'Reilly, Dreier, Wang. (2019). Cognitive Processing Challenges Associated with Low Literacy in Adults. *The Wiley Handbook of Adult Literacy*.

John Sabatini, Jon Weeks, & Tenaha O'Reilly

A stage-based computer adaptive system combining reading components and scenario based assessments in a longitudinal study

Abstract:

Significance: Skilled reading can be described as the fluent coordination of foundational (sub)lexical and text comprehension skills. These two constructs can be further decomposed into elements which begin as loosely connected strands and become intertwined as reading proficiency improves [1]. In lower ability students, reading component test batteries can provide richer (diagnostic) information because weaknesses in foundational skills can constrain their ability to engage in higher-level text processes. At higher levels of proficiency, performance is more tightly associated with discourse-level comprehension skills and strategies. Using this framework, we consider tests that might optimally target different learners and criteria for routing test-takers to appropriate resources. How do we identify which types and levels of tests to administer to students when priors are unknown? This assessment problem arose from the NCES sponsored 'Middle Grades Longitudinal Study (MGLS)', in which only 30 minutes was allotted for testing reading skills, despite a wide range of reading ability in the population.

Method: Accordingly, we implemented a Multi-Stage Testing (MST) design predicated on the need to administer two different types of reading assessments. Component assessments are ideal for students below expected grade level reading ability [1]. Scenario-based assessments assess complex sets of comprehension skills and are optimized for students who are at or above grade level [2]. To estimate students' reading level, we designed a brief placement test using a compilation of items from three component measures. To identify students who appear to be below grade level expectations and place them into a battery of components, the lowest scorers received foundational tests, and those near the 50th percentile received more traditional comprehension measures. Those who scored at or above grade level were placed into one of two different levels of scenario-based assessments.

Results: Based on prior psychometric studies, we identified a range where students were at the lowest or highest level, and an intermediate range of score cutpoints. We then created a field test design involving random assignment to different, overlapping test blocks for students near cutpoints. This allowed us to gather data to make a more precise decision for the study, and reconfigure test items for boundary blocks to ensure that there were overlapping items in that range to increase precision if students were misplaced. The pilot was successful, both in developing placements of students into four levels (two variants of components; two of SBAs); and to prepare for the longitudinal follow up.

Discussion: While official release of the final report is forthcoming, we will report on the field test results. We will also provide a description of how the logic of this design inspired us to integrate RC/SBAs with our ITS.

References:

Sabatini, Weeks, O'Reilly, Bruce, Steinberg, & Chao. (2019). SARA Reading Components Tests, RISE forms: Technical adequacy and test design. *ETS Research Report Series*.

Sabatini, O’Reilly, Weeks, & Wan. (2020). Engineering a twenty-first century reading comprehension assessment system utilizing scenario-based assessment techniques. *International Journal of Testing*

Xiangen Hu, John Sabatini, Art Graesser, & John Hollander

Designing innovative tasks and test environments: Opportunities and challenges in connecting assessment and instruction

Abstract: Educational assessments are designed with a set of constraints in mind – printing costs, transporting tests, security, test conditions, testing time, student characteristics, objectivity, and costs of scoring. Many of the familiar features of "traditional" test design, administration, scoring, and reporting have taken shape because of such constraints.

Many of these historic constraints no longer apply, have been transformed, or can be relaxed thanks to advances in technology and analytics. These new technologies create the opportunity to make testing less artificial and more face valid by approximating or simulating the situations or contexts in which knowledge, skills, abilities, and dispositions can be tested.

Toolboxes for assessment designers have been dramatically expanded. However, core steps in the assessment design process and validation of evidence remain conceptually sound and necessary to ensure quality test results are produced. In this chapter, we organize our review of TEA principles and technologies, as they align to the elements of test development and validation process.

Motivating the need for change in assessment design, How People Learn I and II reviewed and described the multiple ways that individuals learn in distinct disciplines and domains on a trajectory towards expertise, mastery, or proficiency. The result of successful learning is the ability to flexibly call upon knowledge and skill to identify and solve simple and complex problems in a domain, sometimes as an individual, sometimes collaboratively. Absent this result, we refer to knowledge as inert, useful for answering simple questions on a test, but not much else. Multiple voices in the field of assessment have applied this cognitive or learning science perspective to test design, proposing frameworks and models for transforming assessments from measures of static, inert knowledge into measures with the twin purposes of evaluating an individual's position on a scale of expertise or drawing inferences about the kinds learning or instructional experiences that will likely advance them on this trajectory (Mislevy, 2018, 2019; Mislevy & Haertel, 2007; Pellegrino et al., 2001).

In this paper, we will examine the opportunities and challenges of incorporating innovations in task and item design into systems that connect learning and adaptive testing. Such challenges/opportunities include scoring new types of responses (ranging from simple multiple choice to complex, multi-step visual, or verbal responses in scenario-based environments) or utilizing various forms of process data (e.g., response duration, keystroke or mouse-action log sequences, eye tracking or biometrics.)

References:

Mislevy, R. (2018). Sociocognitive foundations of educational measurement. Routledge.

Mislevy, R. (2019). Advances in Measurement and Cognition. *The Annals of the American Academy of Political and Social Science*, 683(1), 164–182.

Mislevy, R. & Haertel, G. (2007). Implications of evidence-centered design for educational testing. *Educational Measurement Issues and Practice*, 25(4), 6–20.

Pellegrino, J., Naomi, C., & Glaser, R. (Eds.)(2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. National Academy Press.

John Hollander, John Sabatini & Art Graeser

Connecting assessment and learning in the inner loop

Abstract:

Significance: Adult literacy learner populations exhibit substantial variability in their educational, cognitive, and socioeconomic backgrounds [1]. Therefore, when developing and connecting assessments and instructional materials for this population, adaptativity imperative. We seek to address this challenge by analyzing within-subjects data obtained from adults with low literacy who completed a reading component skill assessment battery bookending an instructional program using an intelligent tutoring system called AutoTutor for Adult Reading Comprehension. The Reading Inventory and Scholastic Evaluation (RISE) is a battery of six reading component skills subtests. The conceptual framework for the RISE overlaps significantly with the framework of AutoTutor [1, 2], allowing for a linkage between assessment and instruction. However, in such a system, not all lessons and items optimally serve the same purposes. The objective of this paper is to describe a study in which learner and item characteristics were examined in an attempt to connect adaptive testing and instruction in an adult literacy intervention.

Method: Participants in adult literacy programs in the US and Canada completed the RISE component reading skill assessment battery before and after a hybrid instructional program using the AT-ARC intelligent tutoring system. Data from AT-ARC was used to model learner characteristics. We also matched instructional lessons to theoretically and thematically aligned reading component skills and conducted psychometric analyses on item sets grouped this way.

Results: While the instructional program involving AT-ARC generally supported learning gains, some types of learners made higher or lower gains than others in specific reading skills. To link AT-ARC and RISE lessons, we examined theoretically aligned lessons and subtests and found that reliabilities were generally favorable (É‘ range from .57 to .90), establishing these subtests as formative assessments for lesson planning and placements. We also examined the item difficulty and item-total correlation of each lesson-subtest combined item sets. We developed an item typology by defining items that are *instructive* (best for learning, high correlation with aligned gain scores), *evaluative* (best for assessment, high correlation with with aligned pre-test scores), *motivational* (best for engagement, low difficulty, low discrimination), and *potentially flawed* (poor or negative correlations with aligned pre-test and gain scores). These item types are not mutually exclusive. By removing potentially flawed and motivational items from analyses, the reliability of lesson-subtest hybrid item sets further increased.

Discussion: Taken together, these analyses provide support for the integration of adaptive assessment and instruction: more effective and efficient learning (based on both pre-test scores and continuous modeling of student performance via stealth assessment); more valid post-testing (skill-target items); and consequently, recommendations for more targeted, adaptive pathways through the instructional programs and systems.

References:

Sabatini, O'Reilly, Dreier, Wang. (2019). Cognitive Processing Challenges Associated with Low Literacy in Adults. *The Wiley Handbook of Adult Literacy*.

Graesser, & McNamara. (2011). Computational analyses of multilevel discourse comprehension. *Topics in cognitive science*.

Samuel Greiff

Discussion

17: Paper Session - Cognitive Diagnosis CAT

Chair: *Miguel A. Sorrel*

Miguel A. Sorrel, Pablo Nájera, & Francisco J. Abad

cdcatR: An R Package to Combine Diagnostic Feedback and Computerized Adapting Testing

Abstract: A growing emphasis in education and other areas has been focused on the assessment of attributes of discrete nature (e.g., mastered vs. non-mastered). Cognitive diagnostic models (CDMs) emerged to address the need to assess these attributes. CDMs allow modeling compensatory, non-compensatory, or additive relationships that account for the process of responding to test items. Given the need to provide immediate feedback that facilitates, for example, adapting teaching to the results obtained by students, interest has arisen in the computerized adaptive testing (CAT) application of these models. To this end, some of the methodologies of traditional item response theory have been adapted and other new ones have been developed. A wide range of item selection rules and stopping criteria is now available. In contrast, empirical applications remain still scarce. We developed the cdcatR package in the R environment to provide a platform to compare the different procedures available, to evaluate the functioning of specific item banks, and to facilitate the empirical applications of cognitive diagnosis CAT (CD-CAT). This presentation aims to illustrate the main functions of this package. The main function of the package is cdcat(), which allows to perform a CD-CAT application using different selection rules (GDI, JSD, PWKL, MWPKL, NPS, or random) and stopping criteria (fixed length or fixed precision). The latest version of the package allows to impose content constraints and to define different starting rules. Two other functions, gen.itembank() and gen.data(), are also included to facilitate simulation studies under this topic. These functions and their arguments will be demonstrated using different databases. It is to be expected that in the near future the irruption of Information and Communication Technologies in the classroom settings will make adaptive assessment applications possible with relative ease. This can facilitate dynamics in the context of formative assessment. The availability of this R package allows researchers and practitioners to explore this novel methodology. Future development plans for the package will be discussed.

Pablo Nájera, Francisco J. Abad, Chia-Yi Chiu, & Miguel A. Sorrel

Fixed precision cognitive diagnosis CAT without a calibration sample

Abstract: The growing interest in computerized adaptive testing (CAT) and cognitive diagnosis models (CDMs) has naturally converged on the development of cognitive diagnosis computerized adaptive testing (CD-CAT). CDMs are restricted latent class models that have been regarded as a suitable tool for educational assessment. Namely, they allow classifying examinees according to their mastery or non-mastery of discrete latent variables, often called attributes. In the CDM literature, simulation and applied studies have traditionally focused on large-scale assessments, where both the number of items and sample size are large. CD-CAT helps alleviating the need for long tests by providing accurate classifications in an efficient fashion. Recently, fixed length and fixed precision CD-CAT have been explored in conjunction with several item selection rules. The large sample size requirement, however, has barely been addressed in CD-CAT. This is noticeable, given the useful diagnostic feedback that teachers can extract from CDMs at a classroom level. In this line, the nonparametric classification (NPC) method has been shown to provide more accurate classifications than parametric CDMs under challenging conditions, such as small sample size or bad-quality items. The NPC method has been extended to CD-CAT as the nonparametric item selection (NPS) method. The main advantage of the NPS method is the lack of need for a calibration sample, which makes it an appealing solution for educational settings. However, the NPS method only allows for fixed length CD-CAT applications, since it does not provide posterior probabilities of attribute mastery. Fixed length CAT is suboptimal to the extent that it can result in either low estimation precision or an unnecessary high number of items administered. The aim of the present work is to introduce a new measure of certainty that can be applied within the NPS method to provide a pseudo-posterior probability (PPP) of attribute mastery. The PPP is based on the distribution of the Hamming distances (i.e., the number of discrepancies between an examinee's response pattern and all possible attribute profiles' ideal response patterns) and offers an estimation of nonparametric classification accuracy. This allows for the implementation of fixed precision nonparametric CD-CAT without using a calibration sample. A Monte Carlo simulation study is conducted to explore the performance of the PPP under small sample size conditions in recovering the posterior probabilities and attribute profiles. Item quality, item bank length, and number of attributes are also included as simulation factors. Results show that the PPP fairly approximates the generating posterior probabilities under most conditions, and that it obtains accurate attribute profile classifications. Hence, it can be a suitable and handy solution for CD-CAT applications in educational settings where calibration samples are unavailable. The PPP has been included in the *cdcatR* package to facilitate its usage.

L. Andries van der Ark & Niels Smits

FlexCAT: A flexible CAT for measurement and prediction

Abstract: In child and youth care, testing procedures can be lengthy. For example, administering the National Instrument of the Juvenile Criminal Justice System to a single juvenile delinquent can take up to 8 hours. The long duration takes up precious time from the officers administering the LIJ and affects the quality of responses by the juvenile delinquent, possibly resulting in a biased diagnosis, wrong treatment, or unwanted punishment. The problem can be reduced using computer adaptive testing (CAT). However three issues prevent the use of a traditional CAT: (1) the type of tests and questionnaires we focus on do not allow for the construction of a large item bank, (2) the test data are not (approximately) unidimensional, and (3) the aim of the researchers is not only measurement but also prediction.

We propose FlexCAT to accommodate these three issues. In a first stage, FlexCAT estimates the (discrete) distribution of item-score vectors (denoted p). In a second stage, FlexCAT estimates test scores (y) from pˆ. For the estimation of p, we use a flexible model that can accommodate multidimensionality in the test data and that can pick up higher-order interaction effects. Individual test scores are estimated as E(y|pˆ) using an appropriate model relating y and p. The type of test score is also flexible. If one is interested in measurement, possible test scores include the mean item score and the estimated latent trait value. If one is interested in prediction, a relevant outcome variable (e.g., at risk/not at risk) may be used as a test score. A traditional CAT can be conceived as a FlexCAT that uses the same item response theory model in both stages.

In the presentation we explain FlexCAT for the case that a latent class model is used to estimate p, and where the mean item score is used as a test score. Using a real-data example, we compare the accuracy of FlexCAT and traditional CAT. Finally, we discuss the challenges FlexCAT still faces.

18: Paper Session - CAT applications: Personality Testing 1

Chair: *Rodrigo Schames Kreitchmann*

Jana Dlouhá & Eva Höschlová

4Elements (4El) Personality Inventory: Computerized Adaptive Personality Assessment in the Work Environment

Abstract: The 4Elements personality questionnaire is based on the metaphor of the four elements (Fire, Water, Earth, Air). It consists of 100 items divided evenly into four factors, and it uses a short scale (“yes”, “no”, “don’t know”). Firstly, it was standardized in 2008. The analysis performed so far shows very good psychometric properties of this questionnaire regarding factor structure, reliability, and validity. To improve the assessment’s quality and efficiency, we decided to explore the possibilities of adaptive administration of the questionnaire.

Data from 2011-2018 were used for this purpose (N = 13,298, 58.5% females, average age 36.7, SD = 9.7). Due to the nature of the scale, a polytomous model was considered, but since the results showed that the middle “don’t know” responses behaved more as missing, a unidimensional 2PL model was eventually used.

We performed the Post Hoc analysis using the catR package with the item parameters and the response patterns of the respondents. To ensure content validity, we used content balancing. We compared the item selection methods MFI and bOpt (Urry’s criterion) and used a random selection of items to assess the effectiveness of these methods. While the MFI method worked better for Air and Earth factors, the bOpt method worked better for the factors Water and Fire.

We compared the ML and EAP methods to evaluate the accuracy of the level of ability estimation. In most cases, the EAP method proved to be better. The level of SE(θ)max ‰¤ 0.45 was chosen as the termination rule, which corresponds approximately to the level of reliability r > 0.80, which is the value considered sufficient for personality inventories.

The adaptively administered test achieved the same accuracy as the full-length test using an average of half the items (Pearson’s ðœŒÌ‡ = 0.93). Due to the promising results of the performed Post Hoc simulations, we took steps to create a CAT version 4El. We revised and expanded the item pool and started data collection to calibrate new items. As part of further simulations on a new sample, we will choose a suitable item exposure method. The administration of the fixed test currently runs in a web-based application developed specifically for this purpose. We plan to implement the new adaptive features later this year.

Andrea Giordano, Silvia Testa, Alessandra Solari, & Rosalba Rosato

A provisional Multidimensional Computerized Adaptive Testing version of the MSQOL-54: Individualizing health-related quality of life measures in multiple sclerosis

Abstract:

Background. The Multiple Sclerosis Quality of Life-54 (MSQOL-54) is one of the most used MS-specific health-related quality of life (HRQOL) inventories. Availability of an adaptive short version that immediately processes and scores the items may improve instrument usability and validity. Multidimensional computerized adaptive testing (MCAT) has not previously applied to MSQOL-54 items. Our aim was to develop an MCAT version of the MSQOL-54, and assess its performance.

Methods. Responses from a large international sample of MS patients were assessed. We calibrated items using bifactor item response theory for graded response data model, with 10 group factors and one general HRQOL factor. We used 52 of the 54 items, except for two single scale items. Individual factor scores for the general HRQOL and group factors were estimated via the multidimensional maximum a posteriori estimator. Then, eight simulations were implemented with different termination criteria using a 2X2X2 factorial design. We set standard errors (SE) to 0.40 and 0.32 (corresponding to Cronbach's alpha thresholds of 0.85 and 0.90, respectively) for general HRQOL factor. We also set SE to 0.50 (i.e., Cronbach's alpha of 0.75) and 'no SE threshold' for group factors. In addition to the SE rules, in half of the simulations the MCAT terminated if the change in the general HRQOL factor score (theta) from one item to the next was <0.01. MCAT factor score estimates were evaluated in terms of number of administered items, root mean square difference (RMSD), and correlation.

Results. Our dataset included 3669 MS Italian- and English-speaker patients (mean age 43 years [range 18-87], 74% women, 54% with a mild level of disability). The bifactor model fit the data well. Local dependency was apparent between nine item pairs. By inspecting the item information function within pairs, we removed eight items having the lower information function from the subsequent analysis. Thus, 44 items were used in the simulation studies. Among the eight simulations implemented, two pairs provided the same results, resulting in a total of six simulations. Of those, the simulation with SE set to 0.32 (general factor), and no SE thresholds for group factors provided satisfactory performance. In such case, the mean number of administered items was 26 (range 16-44), representing a 41% reduction in respondent burden; for the general factor, the correlation with the full-length questionnaire was 0.94, and the RMSD was 0.32.

Conclusions. Compared to the original MSQOL-54, the MCAT version required fewer items without loss of precision for the HRQOL general factor, at the same time reducing respondent burden. Further work should be conducted to add/integrate/revise items of the MSQOL-54, in order to make the calibration and MCAT performance efficient also on group factors, so that the MCAT version may be used in clinical practice and research.

Mengyu Zhuang, Chongli Liang, & Danjun Wang

An Application of Computerized Adaptive Testing in Mental Health Assessment

Abstract: Mental health tests are used to assess employees' psychological status in company as more employers wants to help employees to observe their psychological well-being. In Beisen's past practice of our own Mental Risk Assessment which was based on Classical Test Theory (CTT), we found there were several disadvantages: (a) test items in the same dimension cannot be weighted differentially in scoring; (b) it would take a long time to complete several subscales if more dimensions were to be included. (c) besides previous concerns about measurement precision and answering time, it is noteworthy that test items with severe statements description might make employees in normal psychological condition confused and uncomfortable.

In the current study, we develop MR-CAT (Mental Risk – CAT) adopting Computerized Adaptive Testing (CAT) and Item Response Theory (IRT) techniques based on Beisen's original Mental Risk Assessment to improve measurement precision, answering time, and reading experience.

We built the item banks by applying IRT models to the estimation of item parameters and by establishing item selection criterion. In order to achieve differential weighting in scoring between items, MR-CAT applies IRT models to scoring. In item parameter estimation, as a non-cognitive ability test using 4-point Likert scale, MR-CAT is developed using Graded Response Model (Gibbons et al, 2007), and each item has 1 discrimination and 3 difficulty. MR-CAT uses EAP algorithm to estimate theta values. As an assessment with 12 dimensions, each dimension establishes its own item bank. The criterion for selecting items is that the discrimination of each item needs to be between 1 and 3 and could not exceed 5 on Difficulty 3.

We developed the CAT protocol with the item selection principle and test termination principle. MR-CAT integrates maximum Fisher information and randomization for item selection from each dimension's own item bank. The most consistent item is selected based on the severity level of mental risks reflected by the theta value. The test termination principle applies both the fixed length and variable length strategies with a maximum of 6 items for each dimension.

Compared to Beisen's original CTT Mental Risk Assessment, the test length for each dimension was shortened by 3-4 items, and the answering time was reduced from 15 minutes to 7 minutes while the CAT version added two more dimensions. By adopting both CAT and IRT techniques, respondents with normal psychological condition could feel better when they only read the most suitable items.

The CTT reliability (internal consistency) of MR-CAT ranged from 0.793 to 0.934, while the IRT reliability ranged from 0.870 to 0.930. The information means of each dimension ranged from 6.445 to 14.207. As for criterion validity, we compared the scores of the depression dimension of MR-CAT with CES-D-10(r=0.588), anxiety dimension with STAI(r=0.778), mania dimension with HCL-16(r=0.452), delusional hallucinations dimension with CAPE-P15(r=0.866), and 8 personality disorder dimensions with PDQ and PBQ(r=0.440-831). In the next phase of development, items will be continuously added to expand the item bank, providing items according to the level of symptoms more accurately and making the test experience better for employees.

19: Paper Session - CAT in Educational Contexts

Chair: *Nathan Thompson*

Laila Issayeva & Nathan Thompson

Evaluation of Computerized Adaptive Testing for National Progress Monitoring in Kazakhstan

Abstract: An essential component of modern education is a student performance monitoring (SPM) system. The main function of an SPM is to help educators systematically track student performance throughout their school life and apply this data to improve their academic achievements. However, SPM with traditional assessment is time consuming and inefficient, as it requires multiple exam administrations per year, as well as needing a vertical scale for meaningful interpretations. Computerized adaptive testing (CAT) is an ideal solution, since it can be built on one large item bank that can not only deliver an exam multiple times per year without the creation of new forms, but can be vertically-scaled to allow for tracking across years, as well as enabling accurate off-grade assessment. This research project is to evaluate CAT for SPM in Kazakhstan, by calibrating a Mathematics item bank with item response theory, and validating an adaptive version of the test through simulation validity studies. We evaluate item bank requirements, scaling, termination criteria, and the need for exposure or content controls. We find that CAT is a useful approach, given the needs of SPM.

The project began with the collation of a cross-grade item bank, combining assessments from multiple within-grade observation points and across grades. This was calibrated with item response theory and developed into a vertical scale for Mathematics. Additional validation was done on dimensionality with factor analysis and Bejar's method.

The final set of item parameters was used for simulation studies to compare item selection methods, need for item exposure controls, and appropriate termination criteria. Results show that CAT provides a compelling model, because it can be used cross-grade as a vertical scale while producing precise scores even for students that are ahead or behind. It also is useful for SPM because of the capability for multiple administrations per year. We will then discuss issues in application of CAT for SPM in countries like Kazakhstan

Arild Michel Bakken, Bente Rigmor Walgermo, & Per Henning Uppstad

Adaptvurder: a formative adaptive reading test for third grade in Norway

Abstract: Effective reading instruction requires precise assessment of the learner's current skill level (e.g. Vygotsky, 1978; Black & William, 1998). Besides assessment instruments developed locally by teachers, standardized reading tests have an important role to play in this respect, because of their superior validity and reliability. For young learners however, these advantages often come at a great cost. In order to achieve a satisfactory estimate of their proficiency, they may have to sit for what to them is a very long time. In addition, they may be presented with items that are either far too easy or far too difficult for them.

Recent development in adaptive testing has the potential of solving both of these problems (Magis et al., 2017). Some efforts have been made in the last decade towards developing adaptive reading tests for young learners, e.g. in Denmark and Wales. The Danish tests have received criticism, both concerning validity and reliability (Bundsgaard, 2018), and a lot of work remains before adaptive reading tests can replace linear tests in all contexts.

In this paper, I present an ongoing project aiming to develop a formative adaptive reading test of good quality for 3rd grade in Norway. "Quality" here means validity in a broad sense, comprising the suitability of the items, the test design and the use of the test, in relation to both the construct being tested and to the larger societal reasons for engaging in such testing (Messick, 1980; Stobart, 2009). I will present the purpose of the test and the construct as well as the test design: item formats, statistical model (IRT), adaptive design (a combination of item-level CAT and multistage test).

I will present data from a large scale pilot conducted in order to calibrate the items before building the adaptive test. Finally, I will show how results will be communicated to teachers. I will discuss all of these choices in relation to the construct.

If successful, Adaptvurder would be a substantial contribution to the field of educational assessment, greatly reducing the cost of precise assessment for young learners. However, this project relies exclusively on knowledge from the CAT tradition. It does not leverage recent developments in artificial intelligence. I will end the presentation by considering some paths for improving testing by exploiting AI in future projects.

References:

Black, P. & William, D. (1998). "nside the black box: raising standards through classroom assessment. *Phi Delta Kappa.*

Bundsgaard, J. (2018). Pædagogisk brug af test. *Sakprosa*, 10, 2.

Magis, D., von Davier, A. & Yan, D. (2017). *Computerized adaptive and multistage testing with R*. Springer.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, vol. 35, no. 11.

Stobart, G. (2009). Determining validity in national curriculum assessments. *Educational Research*, vol. 51.

Vygotsky, L. (1978). Mind in society: the development of higher psychological processes. Cambridge, Mass.: Harvard University Press.

Vandhana Mehta & Nathan Thompson

Development of multi-stage adaptive Math and ELA assessments during a pandemic

Abstract: ImagineLearning embarked on an initiative to develop fully item-level multi-stage adaptive assessments for Math and English Language Arts (ELA) in January 2020. The goal of these computer adaptive tests was to measure skills in multiple content domains to provide a reliable, valid measure of each student's ability and instructional grade level. Additionally, these tests were also designed to accurately measure growth, whether students are functioning on, below, or above grade level. These assessments designed to be administered several times a year where the results from these assessments will be used to develop a personalized learning path for each student within an adaptive learning platform. The learning path is constructed to address learning needs, i.e., targeting gaps or supporting acceleration for each student.

The COVID-19 pandemic introduced a dramatic overnight shift in everyone's lives and affected every aspect of our lives from our children's education to the way we work. Our children adjusted from in-person to remote learning as we all adjusted from in-office work to remote collaboration. There are unforeseen challenges as well as unique opportunities with this dramatic swing.

There was a distinct advantage with this initiative- an existing, large, secure, fully-calibrated item bank spanning grades K-12 Math and ELA aligned to state standards and initiatives. This avoided the need to conduct a large-scale field test of the item pools for the adaptive assessments. Though, a study to develop a vertical scale needed to be conducted and were faced with prospect of accomplishing this during the pandemic. There were several challenges that we came across including difficulty recruiting a sufficient sample in all grades and a need to accommodate varying modes of test administration- remote vs. in-person.

In the development of these assessments, we experienced these issues, and many testing organizations will also confront during and after the pandemic. As our item parameter estimates were based on pre-pandemic data, these estimates may not be reflective of current student knowledge given unfinished student learning over the past two years. Conversely, the parameter estimates and linking equations that were established from the vertical scaling study perhaps in some ways be more reflective of the current state of student learning than estimates approximated in previous years. Additionally, we developed the necessary software platform to deliver custom-designed multistage adaptive assessments based on a third-party platform but with more sophisticated adaptive logic. Also, we developed our own student interface, onboarding, reports, and integrated everything with our learning environment. After about a year of launching these assessments, we reflect on this journey and implications for assessment development now and in the future.

20: Keynote - Early Career Research Award Winner Miguel A. Sorrel: On the diagnostic power of the items in a pool

Chair: *Alina A. von Davier*

Abstract: In recent years, computerized adaptive testing for diagnostic purposes (CD-CAT) has gained relevance. Under this framework, the item response function is operationalized with a cognitive diagnosis model (CDM). Due to the very narrow definition of attributes in education (e.g., addition, subtraction, simplification), items usually have a complex structure. Despite this, previous research has shown that the available item selection rules may show a preference for administering simpler items. The question arises as to whether simple items are indeed the best way to obtain diagnostic information. As a possible influencing factor in this situation, the item pool calibration protocols and model selection indices available are explored in a Monte Carlo simulation study where several item selection rules are compared in terms of accuracy and item pool usage. The item pool structure (% of items with simple and complex structure), calibration sample size, adaptive test length, starting rule, number and distribution of attributes, and the model estimation and uncertainty consideration methods are manipulated as factors. By relating the calibration sample size to the complexity of the item pool and the calibration protocols available, this research results in a fundamental practical guide on how to approach diagnostic evaluation in low sample size contexts in the most optimal way possible.

21: Invited Symposium – Applications of CAT across multiple fields using the Concerto platform

Chair: *David Stillwell & Luning Sun*

Abstract: The University of Cambridge Psychometrics Centre strives towards making online adaptive testing available to everyone. That is why we've created Concerto: a powerful and user-friendly platform that empowers experts and beginners alike to make better tests, with little to no knowledge of coding experience required. There are minimum set-up costs, no licence fees and no limitations.

Concerto harmonises the statistical power of the R programming language, the security of MySQL databases and the flexibility of HTML to deliver advanced online tests. These instruments work in unison, giving users unparalleled freedom and control over the design of their assessments. In-built algorithms for score calculation and report generation ensure a rewarding experience for participants, whatever the context.

In this symposium, scholars around the world will share their experience of developing online adaptive tests using the Concerto platform. These projects bring forward a number of successful applications of CAT across multiple fields in educational, psychological and clinical assessments.

Presenters:

Conrad J. Harrison, Bao Sheng Loe, Przemysław Lis, & Chris J. Sidey-Gibbons

Transforming healthcare measurement: Computerised adaptive testing and machine learning through Concerto

Abstract: Computerised adaptive testing (CAT) and machine learning are transforming the way assessments are delivered in a range of sectors. These techniques can reduce assessment burden, improve accuracy, increase engagement, and uncover deeper, more actionable insights than traditional assessments. In this presentation, we introduce Concerto as a versatile, secure, and easy-to-use platform for developing advanced assessments that harness the power of CAT and machine learning, and demonstrate how these assessments can be developed and deployed with only minimal programming experience. We explore how Concerto is changing the face of healthcare assessment: from administering shorter, personalised health questionnaires in research and clinical practice, to autonomously analysing open text feedback of doctors' performance.

Eren Can Aybek

Development of CAT Version of Resilience Scale under the Concerto Platform

Abstract: This research aims to investigate the applicability of a 35-item resilience scale as a CAT form. The scale has been developed according to Graded Response Model (GRM). Since the item parameters had already been estimated in the scale development process, this study involved post-hoc CAT simulation and live-CAT application. Data has been collected from two different groups. The first group's data, which consisted of 368 university students, was used for post-hoc simulations, and the live CAT application was applied to 173 university students. Post-hoc CAT simulations were held under the catR package for R. According to the simulation results, the live CAT application was developed under the conditions as follows: IRT model as GRM, termination rule as .20 standard error, MFI as the item selection, and MAP as the ability estimation method. Simulation study showed that 7.32 mean number of items were sufficient to terminate the test below .20 standard error, and the correlation between CAT estimation and the full item estimation was observed as .93. After the post-hoc simulation results were obtained, a live CAT application was developed with Concerto Platform v.5.0.20 installed on Amazon Web Services t2.micro instances (2GB RAM, 1 CPU, 8GB SSD) as suggested on the Concerto Platform's website. The CAT application includes a welcome page, a form page that is used to collect demographic information, the CAT form, results, and feedback pages. Results page shows the estimated theta level, standard error of measurement, number of administered items, and a plot that displays examines' location under the normal distribution curve. Examines also received some questions to get some feedback about their CAT application experience. Live CAT results were consistent with the simulation results. Mean theta level of the participants was obtained as -.20 while mean standard error was .19. The mean number of administered items was 6.20. The participants reported that they did not encounter any problem during the application, 86.1% of the participants stated that they would prefer CAT application over paper-pencil test forms. Moreover, only 8.5% of participants were not satisfied with the results page. Some participants suggested more information and graphical explanation for the results, and only a small number of them did not satisfy with the number of administered items. They indicated that they would prefer to take a test with more items, and they questioned the results' validity. On the server-side; the maximum concurrent examines on the application was 31, and considering the capacity of the server, this number is impressive. In conclusion, based on the initial findings, it is safe to conclude that the resilience scale was successfully adapted to CAT form. Concerto Platform provided comprehensive tools for CAT development and made the development process as convenient as possible. Future studies could focus on the comparison between full-length and CAT estimations, and investigate how many items would be sufficient for satisfying examinees on the validity of estimates.

Ecosse L. Lamoureux

Optimizing the patients' perspectives in research, clinical trials, and real-world clinical care using Item Banks and CATs in Ophthalmology

Abstract: Recognizing the psychometric and practical limitations of 'paper-pencil' questionnaires, my group's research in Ophthalmology over the last ten years has focused on item banking, operationalized with CAT methods. Our objective is to comprehensively assess the quality-of-life impact of eye diseases, and associated vision impairment, and the effectiveness of treatment modalities from the patients' perspectives in clinical trials, as well as real-world clinical care. We have developed one vision-specific (the Impact of Vision Impairment [IVI-CAT]) and two eye-disease specific CATs (Diabetic Retinopathy [RetCAT] and Glaucoma [GlauCAT]), using Concerto as our online adaptive testing platform. We are currently developing CATs for two other ocular pathologies (Age-Related Macular Degeneration [MacCAT] and Myopia [MyoCAT]), and Diabetes [DiabCAT]. Our CATs contain between 3-12 QoL domains and 15-303 items. Our CATs are currently being implemented in eye hospitals in the USA and Singapore, respectively. In this talk, I will share my group's development and validation journey of our existing CATs; our implementation experience in the clinical world, including challenges and lessons learned with integrating CATs in the health care sector; and our future plans, including building capacity for widespread home-based CAT administration to adapt to the new COVID-19 world.

Bao Sheng Loe, Przemysław Lis, & Vesselin Popov

The importance of project management in the development of CAT

Abstract: The Concerto platform was developed to make better computerised tests for both experts and beginners across different fields. However, more discussions are needed about the collaboration between the team creating the CAT and relevant stakeholders that make the project successful. This talk provides an opportunity to showcase the use of Concerto within an educational setting. I will share the University of Cambridge Psychometrics Centre's experience of collaborating with the Northern Irish Exam Board to develop a suite of computer-adaptive numeracy and literacy assessments. This project that has started off as a small-scale CAT pilot has now been used to assess thousands of students per year. This case study provides insights on several interesting aspects of CAT development, including how technological innovation and project management need to go hand-in-hand. It also highlights the importance of working with a variety of stakeholders in the process of CAT development and introduces critical Concerto features that contributed to the success of this project.

22: Paper Session - Test taking experience

Chair: *Steven L. Wise*

Hanif Akhtar, Silfiasari, Boglárka Vekety, Balázs Klein & Kristof Kovacs

The Effect of Computerized Adaptive Testing on Motivation and Anxiety: A Systematic Review and Meta-Analysis

Abstract: Although the *psychometric* aspects of computerized adaptive testing (CAT) have been extensively studied, there is very limited research on the *psychological* effect that CAT has on test-takers. Regardless, it has been frequently claimed that CAT is more motivating and induces less anxiety than traditional fixed-item tests (FIT), because in CAT the presented items are matched to test-takers' ability. However, there seems to be no clear-cut evidence to support this claim. The purpose of this systematic review and meta-analysis, performed under the PRISMA protocol, was to gain a comprehensive understanding of the effects of CAT on motivation and anxiety in comparison to FIT. Seven databases (PsycINFO, PubMed, Web of Science, Scopus, Google Scholar, Proquest, and EbscoHost) were examined. Articles were eligible if they contained an empirical study containing a comparison of motivation and/or anxiety between CAT and FIT. Two reviewers screened and extracted relevant data pertaining to the study. Eleven articles were included in the analysis; seven articles dealt with anxiety only, two articles dealt with motivation only, and two articles dealt with both anxiety and motivation. In the reviewed studies, CAT was applied in educational, military, and corporate settings. Meta-analytical results showed no overall effect of test type when comparing CAT with FIT (k = 11, g+ = 0.06, p = .28). The overall effect was significantly heterogeneous, with a high proportion of observed variance ($I^{2\,=}$ 84%) reflecting differences in effect size. When analysis was performed for each psychological aspect, there was a non-significant effect of testing type on anxiety (k = 9, g+ = 0.09, p = .23) as well as on motivation (k = 4, g+ = 0.03, p = .75). However, in studies where easier CAT (i.e., a CAT targeted at higher success rates) was compared to FIT, the effect size was small but significant (k = 2, g+ = .22, p < .001). In the eleven studies included, they used different CAT administration procedures and outcome measures, which might explain the differences in the results. The result of our review is in contrast with the claims articulated in early work on CAT. Certain modifications in CAT administration might have positive psychological effects on test-takers.

Serkan Arikan & Eren Can Aybek

Investigating the Effects of Selecting initial items according to Test Anxiety on Measurement Precision: A Post-hoc Simulation with PISA data

Abstract:

Introduction Test anxiety level of students is known to affect student performances, especially in high stake exams. In measuring a construct, it is desired to eliminate or minimize the effects of the confounding variables, such as test anxiety. In the current study, it is aimed to investigate the effect of selecting the first set of items according to students' anxiety level in a post-hoc CAT study. It is hypothesized that if anxious students start with easier items, their ability level would be estimated with higher precision in CAT. The research question of the study is; keeping test length fixed in a post-hoc CAT simulation study, what is the effect of selecting the first set of items according to students' test anxiety level on ability estimation precision?

Method

Variables In the current study, all the variables were obtained from PISA 2012 dataset. The variables were mathematics achievement test scores estimated by PISA, the test anxiety scores of students, and students' responses for one of the booklets of PISA 2012 mathematics test. As the effect of providing easier items to anxious students will be evaluated, the standard error of the measurement of test scores per students is another variable.

Procedures In the PISA 2012 dataset, booklet 1 was selected. In booklet 1, there were total of 24 items and this booklet was administered to 27253 students. Half of the students will be randomly assigned to a control group or an experimental group. These students will be divided into three groups according to their anxiety level: high ($z_{anxiety} \geq 1$), moderate ($-1 < z_{anxiety} < 1$), and low ($z_{anxiety} \leq 1$). All groups will be simulated to take CAT version of the PISA test. For experimental group, the difficulty level of first set of items will be decided according to their anxiety level. Highly-anxious students will start with easy items (b=-1), moderately-anxious students will start with average difficulty items (b=0) and low-anxious students will start with difficult items (b=1). All students in the control group will start with average difficulty items. All students will take total of 12 items as the test length is kept fixed. The mean SEM for student groups in experimental and control groups will be compared. To see the effect of item length, the same procedure will be repeated with 10 and 14 items. Post-hoc simulations will be conducted with catR package of R.

Expected Results As an outcome, it is expected to identify the effects of using students' anxiety level in deciding the first set of items on ability estimation precision. If this approach is found to be related to a less measurement error, it might be possible to control effects of some confounding variables, such as anxiety level of students, when measuring a construct. This study is planned to be the post-hoc simulation part of a larger live CAT study. The same procedures will be used to test the same research question in a live CAT as a future study.

Michelle Lennon-Maslin & Claudia Quaiser-Pohl

Closing the gap: Reducing gender bias in STEM through the development of new approaches to assessment of spatial abilities

Abstract: Despite globalisation, gendered career choices persist worldwide (Makarova et al., 2019). The World Economic Forum estimates that currently only one third of female students pursue higher education or research careers in STEM fields (WEF, 2017, p.17). A sobering quote from their 2020 annual report states that: "None of us will see gender parity in our lifetimes, and likely nor will many of our children" (Global Gender Gap Report 2020, 2022) Spatial ability and its components are cognitive processes crucial to success in the STEM arena (Buckley et al. 2019; Newcombe, 2017). One component, Mental Rotation (MR), has been studied extensively in psychology and education due to significant gender differences found in the ability to rotate mental representations of two and three dimensional objects (Neuburger et al., 2012). Although less evidence of cognitive disparity has been found in more recent research, MR tasks consistently yield large and reliable differences in performance favouring males with no significant reduction (Wraga et al., 2006). MR is traditionally assessed using psychometric instruments such as the Mental Rotation Test (MRT; Vandenberg & Kuse, 1978), often referred to as paper-and-pencil tests. During computerized MR testing, participants respond to each item or series of stimuli on a computer or touch screen device (Monahan et al., 2008). Stimuli usually take the form of pictorial representations of an object, such as a cube or a letter, an animal or a toy (Rahe et al., 2021). Current research on assessment of spatial ability focuses on issues, which may influence the participant's performance, such the characteristics of the tests, e.g. gender-stereotypical stimuli (Ruthsatz et al., 2017) or the chronological order of items (Eggen & Verschoor, 2006). The latter is the subject of this study, in which the authors intend to develop a computer adaptive test of mental rotation (CAT-MR) in order to address test anxiety and gender differences. During this method of testing, a computer algorithm automatically presents items to participants and selects the next item based on their previous response; hence the test adapts to their performance (Hogan, 2013). This approach has been found to elicit less anxiety which in turn can negatively impact cognitive performance and contribute to gender differences (Fritts & Marszalek, 2010). Female primary school students for example achieved better results, reported higher motivation, and more positive subjective test experience after CAT (Martin & Lazendic, 2018). Furthermore, CAT offers teachers convenience and flexibility, faster, more accurate scoring and reporting, and potentially shorter tests (Chuesathuchon, 2008). This study is part of the EU-financed (Horizon 2020) research network SellSTEM (Spatially Enhanced Learning linked to STEM), which aims to raise spatial ability in children (girls in particular) across Europe, so that they are better prepared for the cognitive demands of STEM education. Ultimately, the goal of the working group is to promote more successful STEM learning, trigger migration in larger numbers towards STEM careers, consequently generating a more gender-balanced ratio in the field (SellSTEM MSCA ITN, 2021).

Paulius Satkus & Kyung Han

Considering Test-taking Experience for Item Selection in CAT

Abstract: Adaptive tests provide numerous benefits over linear tests by administering items that are optimal for each examinee (Chang, 2015). From psychometrics perspective, optimal items are items that provide the greatest amount of statistical information for estimating examinee's ability. Statistical information is modeled as a function of examinee's true ability and the item's parameters (e.g., difficulty) by adopting an item-response theory (IRT) model. Regardless of the chosen IRT model, an item, for which item difficulty is matched to the examinee's ability, results in the greatest amount of information. Typically, items to which examinees have .5 probability of correctly responding are items that provide the most amount of information (under 1PL or 2PL IRT models).

The psychometrically optimal items that are administered in adaptive tests may not be considered optimal from the examinees' perspective. Specifically, if examinees correctly respond to only 50% of the items on the test, they may feel as if the test is too difficult, which could lead to negative test experience. Studies report that examinees are more anxious when completing adaptive tests compared to linear tests (Martin & Lazendic, 2018; Ortner & Caspers, 2011), although some examinees also report positive emotions (Ortner et al., 2013).

One approach to reduce the difficulty of adaptive tests is to administer items that have higher than .5 probability of success. Researchers have found that modifying the item selection algorithms to select easier (success probability = .6 or .7) items led to negligible loss in measurement precision (Eggen & Verschoor, 2006; Widiatmo & Sullivan, 2014). Administering even easier items (success probability > .7) is not recommended for the entire test. However, the use of these items (termed "motivator" items) for 50% of the test can result in comparable measurement precision and higher examinee self-confidence than selecting optimal items (Hausler & Sommer, 2008).

The purpose of the current study is two-fold. First, we surveyed examinees completing a high-stake adaptive test about their perspective on test difficulty. We found that examinees deemed 50% as too difficult. At least 40% of examinees responded they would feel frustrated, discouraged, anxious, and mad about themselves if they were asked to complete a difficult test. On the other hand, a third of examinees completing a test that is too easy for them (defined as a test where they correctly respond to 80-90% of items) reported that they would be comfortable, hopeful, satisfied, encouraged, and excited.

Given the survey results, the second purpose of the study is to experiment with several adaptive test designs, in which test difficulty would be altered. Specifically, we plan to explore the effects of modifying item selection algorithms for choosing easier items and/or including "motivator" items on the test for the recovery of examinee ability estimates. Additionally, we plan to vary the expected success probability of administered items (e.g., 50%-90%), the number of "motivator" items, and the placement (e.g., at the start of the test). In the full paper, practical guidelines/insights for CAT developers will be provided with a comprehensive discussion over the study results.

23: Paper Session - Item Calibration

Chair: *Angela Verschoor*

Jumoke Oladele

Item Bank Development and Validation for Computer Adaptive Testing Mental Well-being Scale

Abstract: Reports by WHO show that relatively few people around the world have access to quality mental health services. In low and middle-income countries, more than 75% of people with mental, neurological and substance use disorders receive no treatment for their condition at all. This is compounded by stigma, discrimination, punitive legislation and human rights abuses that are still widespread. This report shows that close to 1 billion people are living with a mental disorder, 3 million people die every year from the harmful use of alcohol and one person dies every 40 seconds by suicide. And now, billions of people around the world have been affected by the COVID-19 pandemic, which is having a further impact on people's mental health. The purpose of the research was to develop an optimal item bank for assessing well-being while detailing all the procedures, including planning, content analysis, and test blueprint going by Bloom's taxonomy of educational objectives. The developed item bank was pilot tested using FastTest and subjected to psychometric analysis using the Graded Response Model (GRM). The target population would be university undergraduates in South Africa and Nigeria. The sample would be drawn purposively with a focus on undergraduates who are engaged in virtual learning. The instrument for the study would be a Likert scaled questionnaire with items based on indices of mental well-being. The scale would be face and content validated by medical, sociology and educational psychological experts while a trial test would be carried out to determine the reliability of the test items. The scale parameters would be analysed using Xcalibre 4.2 programme to ensure an optimal mental well-being scale while the procedures would be carefully outlined also ensure replicability within other programmes. The developed scale would be deployed adaptively to support the national COVID-19 recovery plan for sub-Saharan African countries with a dart of mental health services. This study is also germane to the attainment of SDG goal 3 of ensuring healthy lives and promoting well-being for all.

Stéphanie Berger, Charles C. Driver, Laura A. Helbling, Stella Bollmann, & Martin J. Tomasik

Vertical scaling of an item bank for computer-adaptive formative assessment

Abstract: The formative assessment system Mindsteps is an online item bank, which can be used by teachers and students across Switzerland for data-based decision making, and informs them on students' current performance and learning gains. The item bank has been developed in accordance with a competence-based curriculum covering eleven competence domains (e.g., "listening comprehension" in English or "form and space" in mathematics) and seven grades in school (i.e., grade 3 in primary school until grade 9 in secondary school). As a consequence, the tens of thousands of items in the current item bank are referenced to the hierarchically structured competency levels stated in the curriculum. Besides this theoretical anchoring of the items in the curriculum, item difficulty parameters have been estimated based on the Rasch model on a vertical IRT scale.

The reference to the curriculum in combination with the empirical item difficulty support teachers and students in creating assessments according to their current questions and needs. Depending on the selected use case, the items are either assembled to linear or adaptive assessments. In the past years, more than 100,000 students from Switzerland have been using the system, sometimes irregularly but sometimes also intensively. However, given the large item pool, the seven target school grades, and the different options to assemble assessments, the number of observations per item are sparse and unbalanced and item calibration is challenging. At the same time, the increasing amount of collected response data promises to improve the estimation of item difficulty parameters and to provide information on the item quality.

Against this backdrop, we present the results of different calibration approaches implemented in a newly developed R package capable of handling large amounts of sparse response data. Subsamples of data collected in different use cases (i.e., linear versus adaptive assessments) are compared in order to validate and improve the vertical scale. We will discuss the implication of the different data subsets on grade-to-grade growth in students' ability as well as patterns of results across different competence domains and school grades.

Paul Barney & Paul Jaquith

Pre-Calibration of CAT Items Using Expert Comparative Judgment

Abstract: For many institutions interested in implementing Computer Adaptive Testing, one of the most significant challenges is the need to pre-calibrate all items before using them in a scored administration. However, if new item calibrations can be estimated with sufficient accuracy, then new items can be used on a CAT without pre-testing, served to candidates using these provisional calibrations, and subsequently scored using the more accurate calibrations derived from IRT analysis of the actual results.

This presentation will demonstrate that by using a modern application of Thurstone's Law of Comparative Judgment, subject matter experts are able to estimate item calibrations with sufficient precision to accurately serve these items on a CAT. The procedure, which is currently used in the United Arab Emirates' national educational assessments, allows rapid deployment and calibration of new items on Computer Adaptive Tests without needing to use pre-testing and without sacrificing scoring accuracy.

During the presentation we will detail the technique used, show how comparative judgment works in general, and demonstrate one popular resource that can be used to experiment with comparative judgment. Based on our practical use of these techniques in national assessment, we will summarize our learned best practices for maximizing the efficiency and accuracy of the resulting provisional calibrations and their optimal use in CAT exams. Both empirical data from national administrations and simulations will be used to demonstrate that candidate scores from an administration using comparative-judgment-estimated item calibrations are equivalent to those using traditionally calibrated items.

Tobias Deribo, Ulf Kroehne, & Frank Goldhammer

Studying the Impact of Rapid Guessing on Item Bank Calibration – A Monte Carlo Simulation Study

Abstract: Behind the administration of *Computerized Adaptive Tests* (CAT) lies the assumption that the item parameters obtained through calibration have been precisely estimated (Veldkamp & Verschoor, 2019). This assumption can be at risk if unaccounted estimation error is part of the calibration process. One possible source of error are responses obtained under rapid guessing behavior (R_{RG}, Wise, 2017), as they appear to be unrelated to test-takers ability. We therefore conducted a simulation study to quantify the effect of untreated R_{RG} on item parameter estimation in the calibration process and compare the efficacy of different treatments of R_{RG} to recover said parameters.

For this we simulated a medium (N = 1000) calibration sample and a medium (I = 250) item bank. Items from the bank were clustered to 25-item sets, combined to 50-item booklets in a *Youden square design* and spiraled out to the calibration sample (Frey et al., 2009). Appearance of regular responses or R_{RG} on an item was assumed to depend on the interplay of a test-takers latent rapid guessing propensity and an itemwise rapid guessing difficulty parameter. As indicated by prior research a negative correlation between ability and rapid guessing propensity (Deribo et al., 2021), a negative correlation between rapid guessing difficulty and item difficulty (Goldhammer et al., 2017) and an decrease in rapid guessing difficulty through later item position (Linder et al. 2019) were taken into account. The mean proportions of introduced R_{RG} for the simulated conditions was 5.59% or 10.19%, which seems in line with prior findings (e.g., Goldhammer et al., 2017). Furthermore, response time based methods have been suggested to identify R_{RG} in large item banks (Wise & Ma, 2012). For these methods the proportion of identifiable R_{RG} appears to be dependent on how clearly response time distributions for regular responses and R_{RG} (Wise, 2017) can be distinguished. To take this dependence into account we assumed either 50% or 100% of R_{RG} to be identified. Finally, for calibration identified R_{RG} have been attended through multiple possible treatments. These encompass treating R_{RG} as not-administered (e.g., Rios et al., 2016), as incorrect (e.g., Wright, 2019) or with a model based on Mislevy and Wu (MW Model, 1996). For each condition 100 simulations were conducted. *Bias* values indicate a systematic overestimation of item difficulty for easier items and an underestimation for harder items when R_{RG} are untreated. This effect seemingly becomes more pronounced at the tails of the difficulty distribution and in part with the proportion of (unidentified) R_{RG}.

Bias values showed a maximum underestimation of -1.03 (.02) and overestimation of 1.73 (.01) of item difficulties for untreated R_{RG}, possibly leading to biased trait estimations in the operative phase. Furthermore, treating identified R_{RG} as not-administered showed the lowest *Mean Absolute Error* off all treatments. Implying it to be the method of choice under the simulated conditions.

The simulation will be extended to encompass smaller and larger item pools and sample sizes, as well as the estimation of latent ability in a subsequent, simulated operative CAT phase in the final presentation.

24: Paper Session - CAT Applications: Ability Testing

Chair: *G. Gage Kingsbury*

Nathan Thompson, Raylene Paludneviciene, Wanda Riddle, & Fernando Austria Corrales

Development of an Adaptive Test for American Sign Language

Abstract: Gallaudet University is the only university in the world where students live and learn using American Sign Language (ASL) and English. This provides a unique situation for university admissions and placement, leading Gallaudet to drive the development of an computerized adaptive test (CAT) for ASL. This exam was designed from the outset to be entirely video-based, with recordings of speakers using ASL and asking the examinees to select the correct response from a list of other videos, to provide much higher fidelity than image-based ASL assessment.

This presentation will describe the process of developing, publishing, and validating a fully adaptive assessment of ASL. After an initial bank of 40 items was planned and recorded, it was piloted to a representative pilot sample of 327 students. The results from this sample were calibrated with item response theory (IRT), comparing both the Rasch and three-parameter models. The Rasch results were then used for simulation studies to determine appropriate adaptive testing specifications. It was found that CAT, as expected, can substantially reduce the test length with negligible reduction in precision.

Both linear and CAT scores were compared to professor classifications of student proficiency level. It was found that there were some range restriction issues, both with the item bank and pilot sample, leading the team to plan future enhancement of the item bank. A pilot version of the actual CAT was administered to a small group of real students, with a survey to obtain feedback on the new experience. It is hoped that this assessment will be used more widely to increase educational access for both native and non-native ASL speakers.

Anna-Lena Jobmann

Development of a computer-adaptive test of figural matrices (CAT-FM) for personnel selection

Abstract: The paper presents the development of a computer-adaptive test of figural matrices (CAT-FM) for personnel selection. Figural matrices were constructed using an automated item generator (AIG) based on six different rules with three varying element groups. A total of N = 7838 applicants for German for middle, upper and highest grades of public service are used for calibration of 210 generated figural matrices in a high-stake situation. Items were presented in 11 blocks of 15 to 30 items each. Sample sizes for calibrations of item blocks differed due to organizational reasons. Items are analyzed and stepwise reduced with regard to internal consistency, item-total correlations, fit to a one-dimensional model, item fit (RMSD) for the 2pl model as well as differential item functioning. The final item bank consists of 190 items with appropriate 2pl model fit. Evidence of convergent validity is provided by high latent correlations to numerical and verbal reasoning. Simulations are used to compare different stop-criteria for practical use of CAT-FM: With approximately 9 items a reliability of Rel = .874 and with 22 items a reliability of Rel = .938 was reached. The item bank allows to measure reasoning ability with low standard error (SE <= .316) especially for ability levels between -1 and 1. Based on the results of simulations recommendations for practical use of the CAT-FM are presented and discussed.

Hanan M AlGhamdi, Ioannis Tsaousis, & Georgios Sideridis

Evaluating a Computerized Adaptive Testing Version of a Cognitive Ability Test Using a Simulation Study

Abstract: This study aimed to evaluate the efficacy and psychometric quality of a CAT system measuring general cognitive ability using simulation. To achieve the study's objectives, a statistical protocol put forth by Han (2018a) was implemented. First, the psychometric characteristics of the item pool were examined. An item pool with strong psychometric characteristics is a prerequisite for an effective and robust CAT system. The results from the CFA and the Latent Unidimensionality Analysis revealed that the GCAT item pool was unidimensional representing a general cognitive ability. Next, the psychometric quality of each item was examined using the 3PL IRT model. Item parameters, item fit, and item local dependency were evaluated. The final item pool consisted of 165 items. Monte-Carlo (MC) simulation was run by first generating response vectors for 10,000 participants from a standard normal. True ability scores were estimated from those response vectors. Next, the GCAT's 165-item bank was applied to 10,000 participants and estimated theta (θ) was computed for each participant using maximum likelihood. The Maximum Fisher Information (MFI) and the Maximum Likelihood Estimation with Fences (MLEF) algorithms were utilized as score estimation and item selection methods, respectively. For item exposure control, we selected the Fade Away (FAM) method, for which there is convincing empirical evidence that it is very effective in controlling both item exposure and item overlap compared to other popular control methods (Han, 2018b; Ozturk & Dogan, 2015). Finally, the minimum standard error criterion was used as the termination criterion, representing acceptable international standard (i.e., 0.33).

The present simulation revealed that the required number of items needed to achieve and maintain a .33 standard error of measurement was 14.96 (S.D. = 12.1). This finding suggests that with the CAT version of the GCAT, the item administration is reduced dramatically by 81.25% (with the full-length version of the test, 80 items should always be administered). Moreover, the level of precision in estimating the participant's ability score was extremely high, as demonstrated by the CBIAS, the CMAE, and the CRMSE. A point that deserves attention is that at very low and very high θ, the error of measurement was higher than at areas closer to the grand mean of theta (i.e.,zero). Moreover, more items are needed to be administered at the lower and the higher end of the θ ability range, probably because there are not enough discriminating items at both areas of this ability range.

Finally, the exposure rate of most of the items in the item pool was very low (< 10%).The CAT system showed a preference for items with high discrimination parameters; indeed, highly exposed items had high discrimination values. This finding suggests that the method used for controlling item overexposure (i.e., Fade Away –FAM) was very effective. This study has several limitations. First, the size of the item pool was moderate in length, a larger item pool is oftentimes recommended. Additionally, highly discriminating items are necessary, especially at the two ends of the θ area.

25: Keynote - Bernard P. Veldkamp: The Double Helix of Adaptive Measurement

Chair: *Theo J. H. M. Eggen*

Abstract: When we think about adaptive measurement, we generally think about adapting the difficulty of the items to the level of the respondent, in other words, about CAT. In the past twenty years, CAT has become more and more popular in the fields of psychological, health and educational measurement. One of the main reasons why CAT became so popular lies in the reduction in test length without any loss in measurement precision. CAT has made testing much more efficient. In most applications, CAT relies on IRT. Unfortunately, this might be quite restrictive, because of the underlying assumptions of the different kinds of IRT models that can be applied. The question arises whether CAT fully benefits from all the less structured data that is currently available and whether it is ready for the age of big data. In many applications, (big) data coming from multiple sources is used for measurement. Besides responses to test items, underlying traits could be measured using, for example, physiological data, process data, logfile data, video data and/or combinations of them. The process of combining data from all these sources is also referred to as adaptive measurement. Within this context, adaptivity not only refers to adapting to various data sources, but also to adapting the measurement to individual differences in data availability. For some respondents, data might be missing, incomplete or not usable because of data reliability and data quality issues. To handle these kind of challenges, AI based algorithms have been applied successfully (see, for example, Dolmans et al., 2021). In this keynote, the focus is on combining both adaptive measurement paradigms. What are the benefits, the limitations, the opportunities and the costs? Initial attempts have been made by combining information about response times and item responses in one hierarchical framework. One step further was to apply a Bayes framework for the combination of various sources of information. The ultimate challenge though, will be to integrate both CAT and AI in one algorithm to fully optimize adaptive testing and to create a double helix of adaptive measurement.

26: Invited Symposium - Computerized adaptive practicing

Chair: *Han L. J. van der Maas*

Abstract: Computerized adaptive practicing (CAP) is a variant of Computerized adaptive testing (CAT) combining the goals of formative and summative measurement, i.e., practicing and testing. Both are essential in education. It is well known that learning skills such as arithmetic requires intensive practice adapted to the level of ability of the individual (cf. zone of proximal development, deliberate practice). It is also evident that adaptive practicing requires precise assessments of ability, the goal of adaptive measurement.

In the last 15 years we developed an algorithm for CAP and applied this technology in a popular online educational system used by 2000 Dutch primary schools, in which we collect about two million item responses per day in about 50 games concerning arithmetic, intelligence, and language (Dutch and English).

The algorithm is based on the Elo rating system developed for chess competitions, but incorporates response time in scoring responses to items. Both items and person parameters are estimated on the fly, such that pre-testing the 60.000 items in the item bank is no longer required.

In this symposium a) we explain the educational and psychological concepts underlying this approach and introduce the Elo estimation algorithm , b) describe how and why this algorithm has been optimized in 12 year of Math Garden practice , c) explain what role AB testing plays in this optimization and how the data can be utilized to provide learning analytics beyond the basic IRT estimates of ability, d) discuss limitations of the Elo algorithm and provide insights in trackers of ability in a developmental (learning) context, and e) propose a new algorithm for computerized adaptive practicing that allows for unbiased statistical testing of educational and developmental hypotheses.

Presenters:

Han L. J van der Maas

An introduction to Computerized adaptive practicing

Abstract: The development of a Computerized adaptive practicing system was motivated by basic research questions in developmental psychology, which required reliable high frequent measurements of large numbers of subjects. As a solution to the practical problems for this type of research, we thought of a way to make use of the fact that primary school children practice exercises in arithmetic and language on a daily basis. This led to Math Garden, an online computerized adaptive monitoring and practice system, based on an extended Elo algorithm, with Rasch measurement scale properties. By using an explicit scoring rule for the trade off between speed and accuracy, we can make use of response times in scoring. This is important since we present children with relatively easy items to keep them motivated.

The win-win-win of our approach is that children playfully exercise at their own level of ability, teachers are assisted in realizing adaptive education and freed from the task of checking exercises, and researchers get access to a very rich database, utilized in dozens of scientific publications in the last 12 years. In this talk I will discuss the scientific ideas behind Math Garden and introduce the extended Elo algorithm that we applied initially. I will give some insight in the practical use of our system and our plans for the future. I end with an overview of the scientific studies performed using Math Garden.

Alexander Savi, Gunter Maris, Han van der Maas, Maria Bolsinova, & Benjamin Deonovic

From adaptive practicing to one-to-one tutoring

Abstract: The formative assessment established by computer adaptive practice systems creates important opportunities. First and foremost, it has the potential to deliver the most pertinent promise of online learning: one-to-one tutoring. However, while matching student abilities and item difficulties might be necessary, it is not sufficient. In this talk, I will discuss two distinct efforts that bring one-to-one tutoring a step closer. First, large-scale computer adaptive practice systems enable A/B tests: massive online field experiments. The scale of such experiments enables the estimation of heterogeneous treatment effects, such as for meaningful subgroups that may benefit from distinct interventions. Second, cognitive diagnostics can map students' inabilities by tracing their errors. Such error tracing models may for instance enable personalized instructions.

Abe Hofman, Matthieu Brinkhuis, & Nick te Broeke

Optimizations for Learning

Abstract: With the extended Elo algorithm in place, the CAL systems of Prowise Learn have grown rapidly in both the number of players and number of games (and items). This resulted in new opportunities and challenges. In this talk, I will highlight some of the adaptations that aim to optimize learning. First, with data coming in, we should monitor the working of the system. I will present some cases where we found unexpected behavior and will discuss what we learned from maintaining the system. Second, with more and more games being launched, how can we steer students to self-select the most optimal games without explicit guidance from teachers? I will present the results of an experiment (A/B test) aimed to investigate the effects of a new way of selecting which games require more practice. Third, I will present some preliminary results on quitting behavior. Can we understand why students stop playing in our system by analyzing the collected log data?

Maria Bolsinova, Gunter Maris, Matthieu Brinkhuis, Abe Hofman, & Han van der Maas

Urnings: a new rating system for computer adaptive practicing

Abstract: A popular method to track the development of student ability and item difficulty in computer adaptive systems is the Elo Rating System (ERS). The ERS allows for tracking the ability of the learners and the difficulty of the items by updating the ratings of the learners and the items after every response. However, the system does not provide a measure of uncertainty (standard errors) about the ratings, which makes it impossible to evaluate the reliability of measurement and to make statistical inferences based on the ratings (e.g., to determine whether abilities have grown or difficulties have changed). Furthermore, adaptivity of item selection in computer adaptive practicing systems may lead to systematic bias in the ratings with the variance of the ratings artificially inflated. To solve these issues we introduce a new urn-based rating system called Urnings, where every person and item is represented by an urn filled with a combination of green and red balls. Urns are updated after every response, such that the proportions of green balls represent person ability or item difficulty. The main advantages of this approach is that the standard errors and reliability of the urn-based ratings is known, and that the adaptive item selection can be explicitly accounted for such that no artificial inflation of the ratings would occur. We highlight features of the Urnings rating system and compare it to the ERS in a simulation study and in an empirical data example from a large-scale computer adaptive practicing application.

27: Paper Session - Item Selection

Chair: *Wim J. van der Linden*

Angela Verschoor & Sebasitaan de Klerk

Linear-on-the-fly testing: flexible and secure high stakes testing

Abstract: The main strategy for society to cope with the Covid pandemic has been a rapid digitization. Examination is no exception to this phenomenon. Flexibilization and security may well become even more important in the near future. Linear-on-the-fly-testing (LOFT) is an option with advantages, especially for examination: Candidates will all take a unique exam form without the technical complications of computerized adaptive testing (CAT). A further advantage of LOFT over CAT is, that it is possible to work with a single fixed cut off score instead of hard-to-explain ability estimates.

The easiest, and most popular, implementation of LOFT is a random selection of items according to the test blueprint. However, it is almost impossible to control the difficulty, and thus the cut off score, for those exams. For a test form of 40 dichotomous items, the standard deviation of the expected test score may be in the order of magnitude of 1, making it unacceptable to fix the cut off score to a single value.

In this presentation, we compare two automated test assembly (ATA) models with these random selections. Not only the difficulty of the exam forms will be controlled to about any given margin, also item exposure can easily be controlled, thus making it possible to find an optimal compromise between exposure control and reliability. When restrictions regarding exposure control are strict, ATA models generally outperform random selection in the sense that the expected score is stable within the given margin, the item exposure can be set even more evenly than in the random case, while reliability is comparable. With more relaxed exposure restrictions, reliability is higher and expected score is still within the admissible range.

Rodrigo Schames Kreitchmann, Miguel A. Sorrel, & Francisco J. Abad

Investigating block pool design in the context of pairwise forced-choice computerized adaptive testing

Abstract:

Purpose: The multidimensional forced-choice (MFC) response format has been frequently claimed to mitigate the effect of acquiescence and social desirability biases. A major drawback for the MFC format, however, is that it may provide ipsative sum scores. Recently, it has become clear that the ipsativity of the scores is not a consequence of the response format itself and can be solved by properly accounting for the underlying response process of comparative judgements. Recent item response theory (IRT) models have been developed to address MFC format, enabling the estimation of non-ipsative scores. However, it has also become clear that the assembly of MFC questionnaires can greatly affect the precision of the IRT scores. Specifically, it has been shown that these questionnaires are especially prone to hold some degree of empirical underidentification of score estimates. Such empirical underidentification depends on the quotients between the items (i.e., statements) scale parameters in a block (i.e., pair of statements) and, further, on the variance of such quotients through the questionnaire. The assessment of positively correlated traits is especially susceptible to this underidentification. In applied settings, computerized adaptive testing (CAT) can be crucial to retrieve precise trait estimators with feasible questionnaire lengths. However, due to the empirical underidentification, the MFC block pool composition can greatly affect the accuracy of the adaptive assessment. In this sense, this study investigated the effects of the block pool design over the precision of trait estimates in FC adaptive assessments. Subsequently, it aims to evaluate the effect of using proper information-based item selection criteria in each pool condition.

Method: Starting from a simulated five-dimensional 480-item pool, this study evaluated the effect of block pool design by comparing a) a 240-block pool formed by randomly pairing items, and b) an optimal 240-block pool formed with a genetic algorithm maximizing the average posterior marginal reliability. For each pool condition, three block selection criteria (T-optimality, A-optimality, and D-optimality) were considered for the CAT implementation. Two CAT lengths were investigated (30 and 60), and trait correlations were either generated as zero or those (positive) found in the NEO Personality Inventory-Revised validation study. The trait score recovery in each condition was compared in terms of squared correlations between true and estimated trait scores and mean trait correlation bias.

Results: As expected, the optimized block pool offered consistently better trait score recoveries than the randomly paired block bank. Such differences were greater for the condition with positively correlated traits. Within each bank condition, the A and D-optimality criteria offered better trait estimates than the T-optimality criterion. Finally, the characteristics of the selected pairs are presented and recommendations on block bank construction are made.

G. Gage Kingsbury

Using allowable item subsets with common stimuli in adaptive tests

Abstract:

When adaptive testing (Weiss, 1973) was first being implemented, stand-alone multiple-choice questions were the predominant item style. Since then there has been a rapid expansion of item styles. Many item styles associate multiple test questions to a single stimulus (common-stimulus items). Depending on the particular item type, these test questions may be related to one another in a variety of ways.

These item relationships go far beyond the common constraints that are seen in test blueprints. This study describes some of the challenges that occur when trying to use common-stimulus items within an adaptive test. It then introduces allowable item subsets as a tool to meet some of these challenges. Finally, it details procedures for using allowable subsets within several operational adaptive testing settings to describe how they might be used to improve the quality of the adaptive tests.

Challenges in using common-stimulus items: One of the challenges associated with the use of common-stimulus items is the potential for statistical dependencies to occur among the common items. This may lead to inaccurate trait level estimates, growth estimates, and classification outcomes.

Another challenge is that agencies using adaptive tests commonly have ongoing item development for operational stimuli. This may change the dependencies that exist among the items associated with a particular stimulus as items are added or deleted.

Different approaches to item and stimulus selection are also available today. Some programs use adaptive stimulus selection, but administer items associated with the stimulus non-adaptively. Other programs administer items adaptively once a particular stimulus is chosen. Rules used for selection and administration of items may be suboptimal, so they present challenges to test design.

So, some challenges associated with using common-stimulus items in an adaptive test include:

- Cross-item information
- Statistical dependencies
- Ongoing item development
- Adaptive stimulus selection

Adaptive item selection

Allowable item subsets: One way of addressing the challenges described above is through identification and utilization of allowable item subsets.

Briefly defined, an allowable item subset is a group of items that can be administered together with a particular stimulus.

So, if items A, B, C, and D are available to be administered with a particular stimulus, but items A and B have cross-item information, we might create one allowable subset with items A, C, and D and another allowable subset with items B, C, and D. Only items within one allowable subset would be allowed to be administered to a particular test taker.

The example above could be used to eliminate cross-item information by careful selection of the subsets. Allowable item subsets can also be used to address the other challenges mentioned above. To demonstrate this, four examples seen in operational adaptive tests are described.

Within each example, a description of the current testing practice is presented, and then an alternative practice is presented using allowable item subsets. The alternative model details how the use of allowable item subsets addresses each of the challenges revealed in current practice.

Niels Smits

The apriori algorithm as an engine for computerized adaptive assessment

Abstract: Both in medical research and clinical practice, questionnaire-based assessments are increasingly used to obtain information about patients. Such information is often used to make clinical predictions, classifying respondents into categories, such as ‘at risk’ and ‘not at risk’ for pathology. An important consideration in designing questionnaires is to minimize respondent burden, and computerized assessment provides an opportunity to make test administration efficient. For questionnaires meeting certain measurement properties, adaptive testing algorithms have been developed which not only allow for early stopping, but also for the dynamic selection of items to minimize testing time. By contrast, questionnaires not meeting such standards, such as those assessing symptomatology, do not allow for this methodology, and therefore alternative methods are needed. In the current study it is illustrated how the apriori algorithm, commonly used in market basket analysis, may be utilized as a method for computerized adaptive testing.

28: Paper Session - Usage of Prior Information & Software for CAT Development and Application

Chair: *Yigal Attali*

Safir Yousfi

The CAT algorithm of Psychological Service of the German Federal Employment Agency

Abstract: The Psychological Service of the German Federal Employment Agency has seven computerized adaptive tests with more the 100.000 test takers per year. Until now, the CAT algorithm relies on selecting the item with the maximal Fisher information for the MLE. However, the recently implemented new algorithm allows for MAP and WLE ablity estimation and Sympson-Hetter item exposure control. The method of ability estimation used for item selection may differ from the ability estimates reported to the test users. We intend to take advantage of this feature by offering CAT online-pretests to the test-takers applied via the Internet based of a subset of item pool with low to medium item discriminations. The results of the online-test will be used as prior information for MAP estimation used for item selection for the on-site validation CAT based on an item pool with high discrimination. The results reported to the test users will rely exclusive on WLE estimates of the on-site CAT. This procedure is meant to reduce the testing time for on-site testing which might have positive effects on the interaction process of test takers and test users.

Dries Debeer & Benjamin Becker

Automated assembly of MST modules, a demonstartion of the eatATA R-package

Abstract: For the development of multi-stage testing (MST) administrations, items have to be selected and combined into multiple item modules. For every stage of the MST, one or more modules should be assembled out of the pool of available items, so that both content-related as well as statistical specifications are met. For instance, often the information function of a module should meet a pre-specified target information function, and this target can differ over modules and stages. At the same time, a module should contain a fixed number of items from each item type as well as a fixed number of items related to specific content topics or domains. The resulting combinatorial assembly problems are typically complex, making human assembly inefficient and far from optimal. Yet these assembly problems can be efficiently addressed by viewing them as (and translating them into) mixed integer linear programming problems. Several software tools exist (both commercial and open source) with high-end algorithms for solving mixed integer linear programming model. Although these tools can be used to perform assembly tasks automatically, they are generally not build for typical test assembly problems. Consequently, these software tools are often unknown or unavailable to practitioners who are confronted with test assembly problems.

In this paper we present the eatATA R-package, which allows using several mixed-integer linear programming solvers specifically from the perspective of automated test assembly problems. We will describe the general functionality and demonstrate how the workflow targeted at R-users. Using an example from MST, we will demonstrate how eatATA can be used to assemble modules for and MST administration. In addition, we will illustrate how in an easy and convenient manner eatATA can translate complex assembly problems into mixed integer linear programming models, so that multiple MST modules can be assembled simultaneously.

Nathan Thompson

Assess.ai: Rapid CAT Development with No Code

Abstract: Two major hurdles in the greater adoption of computerized adaptive testing (CAT) across disciplines are the investment required in a software platform and the psychometric expertise necessary to both design and run it. This presentation will discuss an innovative platform that allows assessment organizations to easily build online assessments, run IRT calibrations, and publish both item-level adaptive and multistage tests, without having to worry about code, nodes, or any other technical issues. This allows the experts to focus on the content and the psychometrics, which are the two core aspects of validity.

We will begin by discussing the reasons for the platform, based on feedback from users. We will then discuss the design and development of the platform from a technological perspective. Finally, we will provide a discussion of how the platform can be used to quickly develop and publish adaptive tests; if existing pilot data is available, the entire process can be done in hours. Primarily, this includes the IRT calibration software as well as the interface to design the CAT parameters, from termination criterion to item selection to custom multistage routing rules. As a follow-up, we will present case studies of some projects which have successfully used the platform to develop adaptive tests.

Of course, even if the software provides all the necessary functionality out of the box, the practitioner needs to perform the legwork needed to develop a CAT that meets the organizations needs and has sufficient supporting validity documentation. We will provide a developmental model designed to help organizations accomplish these two things, especially if they are new to CAT.

Niek Frans, Johan Braeken, Bernard P. Veldkamp, & Muirne C. S. Paap

Empirical Priors in Computerized Adaptive Testing: Risks and Rewards

Abstract: Given that the trait level of a participant is unknown before any items have been administered, it is common practice in CAT initialization to assume an average trait level as the starting estimate of each participant. However, other sources of information about the participant are frequently available and may be used to obtain a more accurate starting estimate. While concerns about the fairness of using so-called ‘collateral’ information as an empirical prior in test results have suppressed its application in summative educational tests, these concerns are vastly different in the context of health care measurement. In addition, by generally using unbiased prior information and excessively large simulated item banks with highly informative dichotomous items, studies on the use of prior information in CAT have thus far mainly focused on the benefits of empirical prior information under ideal circumstances. While these studies have shown that the benefits of using empirical information as a prior in CATs can be substantial, the potential risks when this information is biased or when more realistic but less than ideal item banks are used have been neglected. In this simulation study, we explore the benefits and potential risks involved, when using empirical priors in realistic educational and clinical CAT scenarios.

The potential impact of prior information on fixed precision CAT outcomes and efficiency was studied in two scenarios. The first scenario uses larger item banks with dichotomous items that are common to educational settings. The second scenario utilizes a small item bank with highly informative polytomous items that are commonly found in clinical settings. Bias and precision of the prior information, as well as item bank size, were systematically varied in the first scenario to explore its impact on CAT length and estimation bias. The second scenario was set up using empirical data to illustrate use of prior information in an applied clinical setting. All outcomes were compared to a baseline standard normal prior that is commonly used in current practice.

The results of these simulations show that empirical priors can substantially increase the CAT efficiency in both applied settings. However, there are two major risks involved with important implications for this otherwise positive outcome. First, contrary to expectations, an unbiased starting estimate did, under certain circumstances, lead to substantially longer tests compared to a single constant starting point for all participants. This surprising finding could be explained as a direct result of the item bank properties which sometimes had few informative items on the prior location. Second, although a highly precise prior vastly reduced test length, this will often result in CAT termination before the trait estimate had a chance to recover from a biased starting point. Given the promising results of using an empirical prior in small polytomous item banks, and the applicability of prior information in diagnostic tests, we expect that the application of empirical priors may yield substantial rewards in clinical CATs. However, this study shows the importance of investigating potentially disadvantageous interactions between the prior and item bank properties.

29: Paper Session - CAT applications: Personality Testing 2

Chair: *Cliff Donath*

Gustavo Henrique Martins, Alexandre Jaloto, Felipe Dinardi, Araê Cainã Zani de Souza, & Rodolfo Augusto Matteo Ambiel

Would it be possible to optimize the assessment of a short measure of interests by areas of Psychology?

Abstract: In the Brazilian context, favorable validity evidence was published for an instrument that assesses the interests of Psychology students and psychologists regarding the areas of activity of the profession, named Scale of Interests in Areas of Psychology (EIAPsi). The short version of the EIAPsi has five items for evaluating each of the 11 factors of the instrument, totaling 55 items. Although the number of items per factor of the EIAPsi is already reduced, would it be possible to reduce this number of items answered by the subject? In this direction, we can use the test application format known as Computerized Adaptive Testing (CAT). However, the CAT literature usually suggests the elaboration and calibration of large banks of items so that the test is efficient in this application format. Therefore, the objective of this study was to test the efficiency of the application in the CAT format of EIAPsi. To achieve this objective, we surveyed in two stages. Step 1 aimed to identify the best stopping criterion for the EIAPsi in CAT format. The sample used in Step 1 consisted of 2026 Brazilian adults, 1433 psychology students and 593 were psychologists. All participants responded to EIAPsi in a linear format. Through post-hoc simulations, we compared the number of items answered in different stopping criteria based on the Standard Error of Measurement (SEM). The item selection method was maximum Fisher information and the estimation of scores we calculated using the Expected a Posteriori (EAP) method. We identified that when adopting an SEM < 0.40, we obtained precision (i.e., SEM average of linear EIAPsi = 0.32) and scores (i.e., r ≥ 0.95) similar to the linear version, requiring an average response of 28.22 items in total. Step 2 sought to test validity evidence for the EIAPsi in the CAT format. We considered the algorithm selected in Step 1. 206 Brazilian Psychology students participated. Data collection took place online through an application developed in the R environment with the “mirtCAT” package. In this sample, the application of the EIAPsi in the CAT format required a total average of 36.5 items answered per participant, representing a 34% reduction in the average number of items in the instrument. Mean SEM by factor ranged from 0.33 (Sport) to 0.42 (Clinical). We also tested the relationship between EIAPsi factors and professional interests in the RIASEC model, evaluated by the 18REST instrument. The correlations were primarily positive and consistent with what was hypothesized, for example, the correlations between Teaching and Research and the Investigative type; Social and the Social type; Organizational and Entrepreneurial and Conventional types. The results made it possible to identify that the EIAPsi assessment could be optimized by adopting the CAT format. More broadly, the results suggest that in some low-stakes contexts (e.g., career context), it is possible to promote a reduction in the number of items when using the CAT format for application, even when the item bank is small (e.g., five items per factor).

Jeremiah McMillan & Nathan Carter

Computerized Adaptive Testing for Ideal Point Personality Assessment: A Comparison of Test Characteristics

Abstract: The use of computerized adaptive testing (CAT) for personality assessment is gaining popularity in employee selection contexts for boosting efficiency, reducing costs, and improving examinee reactions. Extant research has primarily established that CAT provides utility over static personality testing when the response model is monotonic (i.e., higher standing on the trait results in participant endorsement of a higher response option; e.g., Forbey & Ben-Porath, 2007; Hol, Vorst, & Mellenbergh, 2008; Makransky, Mortensen, & Glas, 2013; Reise & Henson, 2000; Simms & Clark, 2005). Given the recent emergence of the use of ideal point item response theory (IRT) models for personality testing —which assume that higher response probability is inversely related to an individual's distance from the item —it is important that research examine whether these models support effective CAT, and the test characteristics that may play a role. This is because ideal point models require more individual response data than monotonic models to accurately estimate theta, potentially hindering the utility of CAT. Additionally, holding test length constant, ideal point models may require different considerations around balancing number of points of adaptivity with the stability of theta-hat estimates used for adaptation, compared with monotonic response models. The present study used real-data simulations to examine the performance of different CAT conditions using a pool of Likert-type conscientiousness items calibrated under the generalized graded unfolding model (GGUM; Roberts, Donoghue, & Laughlin, 2000) on a sample of 1,724 Amazon Mechanical Turk workers. General measurement accuracy (bias, RMSE, and theta-theta-hat correlation) and the accuracy of dichotomous employee selection decisions based upon theta estimates were examined while manipulating the cut-score adopted for employee selection, total test length, number of pre-adaptive items presented (i.e., an initial testlet), and the use of a sequential versus multistage testing design. Results indicate that adaptive tests outperform ideal point static tests on general measures of accuracy but not on employee selection decision accuracy. The most critical test characteristic for successful adaptive testing is the presence of an initial testlet. Implications for testing theory, CAT design considerations, and future research directions are discussed. In particular, plans for a full simulation study to replicate findings under carefully controlled testing conditions are proposed.

30: Paper Session - Multi-Stage Testing

Chair: *Duanli Yan*

Vandhana Mehta & Nathan Thompson

Test Design and Monte Carlo Simulations to develop a fully item-level multi-stage adaptive assessments for Math and English Language Arts in an event-driven architecture

Abstract: Multi-stage adaptive testing offers many advantages over traditional assessment approaches. Personalized assessment improves the student experience and the resulting information, especially when integrated in a continuous loop with adaptive instructional systems. The development of such an assessment is not trivial and becomes more challenging when designing a vertical scale across all grades. This session provides a framework for the development of such assessments, describes our implementation, and discusses lessons learned that can help organizations improve their own assessments.

In January 2020, we embarked on an initiative to develop fully item-level multi-stage adaptive assessments for Math and English Language Arts (ELA). The assessments were designed to measure skills in multiple content domains to provide a reliable, valid measure of each student's ability and instructional grade level regardless of their grade level. They were also designed to be administered several times a year to monitor student progress during the school year. Students' results from these assessments will be used to develop a personalized learning path within an adaptive learning platform. This learning path will be adjusted according to students' overall and domain results.

Both the Math and ELA assessments cover the skills and knowledge defined by standards from the Common Core Framework and from several key states (such as NY, TX, CA, IL, and FL). Psychometricians and test developers engaged in many discussions to develop the domains and the standards associated with them. Initial design discussions are critical.

Monte Carlo Simulations have been conducted throughout the development of these CATs. Various aspects were examined, such as Bayesian variance and boundaries for grade levels, as these simulations were conducted. Data was collected on test length, SEM, and correlations between estimated and true student ability to examine the impact of multiple administrations. Similar data points were also collected at the domain level.

Currently, about one year of student data in an evented format has been collected from these CATs. The evented data has been converted and stored into Snowflake for analysis purposes. This allows us to examine the data collected from the CAT in many ways that were not possible before. For example, we can compare data from our MC simulations to actual student data collected at the end of each stage of the test, investigate student ability at the domain-level from each of their administrations and even examine time spent on responding to each item or to the entire test. We created data visualizations that summarize the performance of students across and within a grade level for a particular subject so that we can constantly monitor student performance as they complete their journey with us.

After a year of launching these assessments, we reflect on our decisions with respect to test design and item selection, and implications for future assessment development. In summary, we attempted to make smarter assessments that provide quality information on student knowledge, in fewer items, to directly impact their learning and save instructional time. We want to empower our students and teachers that learning can happen anytime!

Hanan M AlGhamdi & Dimiter M Dimitrov

Piloting of Multistage Testing under D-scoring Method

Abstract: The National Center for Assessment (NCA) in Saudi Arabia is transitioning from linear forms of Computer Based Testing (CBT) to Adaptive Multistage Testing (MST) in a variety of large-scale assessments developed and administered by the NCA. The MST is under current piloting for the General Aptitude Test (GAT) which is taken by high school graduates who apply to universities in Saudi Arabia. In this context, the MST scoring and routing of examinees across stages of the selected design (e.g., 1-3-3) is based on the *D*-scoring method of measurement (DSM) which is largely adopted by the NCA (Dimitrov, 2016, 2018).

Under the classical version of the DSM, the *D*-score is based on the person's response vector of (1/0) scores on the test items and their expected difficulties ("”*deltas*""). The *D*-scores range from 0 to 1 and form an interval scale. They are often multiplied by 100 to range from 0 to 100 and show what percent of the ability required for total success on the test is demonstrated by the examinee.

This paper aims to evaluate the adequacy of the MST design and routing rules to assess the precision and quality of the MST-GAT under the DSM. Previously, the efficiency of MST under DSM has been examined in a simulation study (Han, Dimitrov & Al Saud, 2018) and real-data studies at the NCA (e.g., Dimitrov, AlGhamdi, & Alqataee, 2019), with focus was on the performance of modules, routing rules, and paths of examinees' scores compared to their counterparts under linear CBT.

This study examines the (1-3-3) MST design for GAT, with one module at the first stage and three modules at the second and third stages. The modules differ in level of item difficulty (easy, medium, difficult) specified by the range and average difficulty (*delta*) for each module. The routing of examinees across stages is based on their performance at all previous stages. Specifically, as the examinees' *D*-scores and expected item difficulties (*deltas*) are on the same scale [0 to 1], an examinee with a given *D*-score is routed to the next-stage module with the closest "”distance"" between its average difficulty and the person's *D*-score.

The study results showed that the routing distribution of examinees is concentrated in the main paths: easy-easy, medium-medium, and difficult-difficult. This attests to the stability of the examinee's performance and appropriateness of routing rules. The results also revealed a logical trend that the averages of examinees' *D*-scores on paths increase upwards with the increase of their difficulty, thus supporting the adequacy of items assembly and module specifications. The results also revealed high coloration between the examinees' *D*-scores obtained under the MST and previous Classical Linear Test (CLT) of GAT. It was also found that, on average, the differences between the scores of MST-GAT and those on previous attempts on CLT-GAT were relatively small, with a slight increase of the difference (in favor of MST scores) for examinees with the highest performance on CLT-GAT.

Cagla Alpayar & Deha Dogan

Comparison of the Performance of Hybrid and Traditional Multi-Stage Tests

Abstract: Adaptive tests may be classified in terms of adaptivity level into item-level and module-level (Han, 2020). In large-scale assessments, only item-level or only module-level adaptive tests are commonly conducted. However, tests combining two adaptive test types are not common. Designs with both adaptation units applied consecutively in the same test are called hybrid. Hybrid tests may provide better measurement precision because they contain more adaptation points thanks to the item-level adaptive section. The module-level adaptive section prevents the low-accurate estimations when especially provisional estimations are not close to respondents' true ability level. In the existing hybrid designs, the CAT is applied within the modules without any module-level adaptation (Ma, 2015) or module-level adaptive sections are applied at just the beginning of the test (Tay, 2015; Wang, Lin, Chang, & Douglas, 2016). Therefore at the initial stage, when limited information is available about the individual's ability level, estimations can be made based on the respondent's performance in a module rather than in a single item. In the following item-level adaptive section, more adapted items are applied depending on the estimations made based on the performance in the MST section and provisional estimations closer to the respondent's true ability level. In this study, both approaches in the literature will be applied together in hybrid designs. Taking into account the contribution of MST to measurement precision at the beginning of the test, a medium-difficulty module will be applied before CAT section. It will be examined how two hybrid designs (**Hybrid_1:** *Medium Module + CAT + MST* and **Hybrid_2**: *Medium Module + MST + CAT*) and *traditional MST* differ from each other in terms of measurement precision.

- **Traditional MST**: 1-3-3 MST design is applied.
- **Hybrid_1**: Respondents answer the medium difficulty module first. Then they complete the CAT section depending on their previous performance. The CAT section is used as the routing stage of the MST. Therefore, the first stage of the 1-3-3 pattern will be completed in the CAT section, they will be directed to one of the easy-medium-hard modules adaptive to their provisional ability estimation. At the last stage, the process is repeated and the test is terminated.
- **Hybrid_2:** Respondents complete 1-3-3 MST with a medium difficulty routing stage. Then, depending on their performance in this section, they are routed to the CAT section.

The test length will be fixed at 36. Each section will contain 12 items. In the MST section, the modules will be pre-assembled. Since a simulation pattern will be designed for a single content, content balancing methods will not be used. Criterian-based routing will be used as the routing rule. In the CAT section, there is no need for a starting rule as the respondents complete a module before the CAT section. As item selection method, the highest expected posterior weighted-information method will be used. The test designs will be compared based on the measurement precision (RMSE, bias, conditional bias) according to different ability levels and item pool sizes. 100 replication will be used for each condition.

31: Paper Session - Threats to Validity: Engagement, Position Effects, and DIF

Chair: *Muirne Paap*

Steven L. Wise & G. Gage Kingsbury

Diminished Performance During CAT Test Events: Identification, Implications, and Amelioration

Abstract: Achievement tests are designed to ascertain what a student knows and can do, by administering a set of items during a test event, with the student's achievement level estimated based on the collective set of item responses. This process implicitly assumes, however, full engagement and effort throughout a test event. Otherwise, the resulting test score may underestimate the student's true achievement level.

While this assumption is usually not investigated with operational tests, there are reasons to believe that it may not always be met. First, research has found evidence of item position effects, in which an item's difficulty tends to increase across item positions. Second, in low-stakes testing, the prevalence of rapid guessing —which indicates disengagement — increases with item position. Third, in timed tests speededness tends to affect responses later in a test event, suggesting that a student's responses to those items may be most adversely affected by increased time pressure.

Our study investigated the stability of test performance on NWEA's Map Growth assessment in Math (50 items; N = 60,000) and Reading (40 items; N = 60,000). MAP Growth is a widely used, interim CAT administered to U.S. students in grades K-12. It is a low-stakes assessment that is considered unspeeded because it is typically administered without time limits.

Our initial analyses revealed an interesting pattern in our CAT data. For each test event, maximum-likelihood estimates (MLEs) were computed for both the first half of the items and the second half. In each subject area, mean test performance decreased .60 standard errors during the second half of the test, indicating diminished performance — which is contrary to the assumption of stable performance. We then estimated mean demonstrated achievement at each item position, finding that achievement level initially rose during early items and then showed a steady decline, which reached a full standard error by the last position. The score decline persisted even when rapid guesses were excluded from scoring. These results indicate that mean test performance varied considerably during test events, with the performance decline during the second half being especially concerning. This effect could be due to multiple factors, including increased boredom, fatigue, anxiety, or perceived speededness (i.e., feeling time pressure despite the absence of a time limit).

Collectively, our findings suggest increasing disengagement (i.e., factors that decrease the probability of passing items). This implies that some students perform at a maximum level only during a portion of their test event. Arguably, scores based only on this portion might yield more valid scores than that based on all item responses, because they would be less distorted by disengagement.

Our study will investigate this phenomenon and study the use of scores based on rolling item subsets to identify the prevalence of students who showed meaningful performance decline. Additionally, we will assess the degree to which validity is improved if we exclude portions of

test events that are consistent with diminished performance. Finally, we will suggest testing practices that should reduce the prevalence of diminished performance during MAP Growth test events.

Mirshod Ermamatov, Davronbek Alimov, Akmal Sulaymonov, Akmal Baratov, Abdulaziz Sattiev, & Tatyana Kasimova

Invariance of the difficulty of test items on mathematics test

Abstract: The paper analyzes, the group invariance of the difficulties of the test items, using the data from mathematics test used for the admission to 5th grade of the specialized schools with the in-depth study on Computer Science and Information Technology in the Republic of Uzbekistan. The testing results from the ten different regions are explored separately and totally, putting together results from all the groups. Comparison of these two cases shows invariance of the difficulties of test items, proved in item response theories, with respect to the groups within measurement errors.

32: Paper Session - Statistical Foundations of CAT

Chair: *Peter van Rijn*

Eren Can Aybek

Using the relation between Classical Test Theory and Item Response Theory in Computerized Adaptive Testing

Abstract: The relation between Classical Test Theory (CTT) item statistics and Item Response Theory (IRT) item parameters is well known. Especially item difficulty statistics p, and item difficulty parameter b can be transformed to each other by using the area under the normal distribution curve. The present study aims to go beyond the relationship between CTT and IRT in the manner of item parameters and to find an answer to how practical is the CTT to IRT transformation using this relationship for Computerized Adaptive Test (CAT) applications? In this regard, 36 different data sets created with catR package were studied. For each data set, item difficulty parameters and transformed item difficulty parameters were calculated and the correlation coefficients between these parameters were analysed. Then, CAT simulations were performed using these parameters. It is seen that the transformed-b parameter (b*) has a higher bias and RMSE value than the b parameter in CAT simulation. However, it was found that bias and RMSE values in the simulations using b* also decreased, especially when the sample size was 250 and above. On the other hand, while the correlation coefficients between the estimates were found to be around .85, the correlation coefficients between the ability levels estimated by CAT and the ability levels estimated from all items were found to be around .90 when both the b and b*parameters were used. In both cases, the simulation terminated with less than 10 items.

All these findings reveal the potential of b* parameter IRT-based studies. The simulation results are affected more by the sample size than by the item pool size (except item pool size was 10). Although the findings show that the b*parameter is not as effective as the b parameter, the similarity of CAT simulation results is promising. Especially due to COVID-19 pandemic, the practicality of measurement and assessment processes in distance education has become even more important. In this process, tailored test solutions such as CAT are beyond being available to educators who are not particularly familiar with IRT.

In this context, it is expected that CAT applications can be made by easily converting parameters from CTT to IRT with the proposed transformation. However, the data used in the research were produced in accordance with the IRT assumptions with the catR package. Investigating the performance of the b* parameter where IRT assumptions are not met, as well as applying real data-based post-hoc CAT simulations will provide a deeper understanding to see how effective the transformation is. In addition, the transformation applied in the research assumed that student ability is normally distributed. Further studies are required to be conducted on how violating this assumption may affect the b* parameter and the results of the analysis.

Denis Federiakin

Effects of the basic properties of the numerical integration on the Expected-a-Posteriori estimation of persons' ability for the marginalized unidimensional dichotomous logistic Rasch model

Abstract: Many studies investigated the Expected-a-Posteriori (EAP) estimation of persons' abilities and compared it to other Bayesian and frequentist estimators. Overall, EAP is one of the most popular estimators of persons' abilities in the paradigm of the Item Response Theory, among regular Maximum Likelihood and the Warm's Weighted Maximum Likelihood procedures. This happens due to its capabilities of estimating extreme response profiles, non-iterative variants, and computational simplicity. Moreover, the Posterior Standard Deviation (PSD) of the EAP estimate can be treated as its standard error, overall leading to the computational effectiveness of EAP. It is also well established that the bias of EAP tends to underestimate the abilities of "strong" persons and overestimate the ability of "weak" persons, leading to the inward bias and general underestimation of variance across persons. A number of studies try to introduce some correction for the EAP in order to compensate for this bias. Significantly fewer studies, however, investigate how EAP behaves under different technical conditions. Almost all studies utilize simple Gauss-Hermite quadrature for numerical integration, apparently, due to its simplicity and solid theoretical foundation. This study compares the precision, bias, and overall behavior of EAP estimates of persons' ability under various numerical integration rules using extensive simulations. First, we compare the precision of EAP characteristics under numerical integration by Gauss-Hermite quadrature, rectangular integration, trapezoidal integration, and Simpson's quadratic and cubic rules. Despite analytically established theoretical ideas about the bias in those integration techniques, the exact amount of bias they introduce in the EAP estimates has not been studied yet. We compare different integration techniques using the amount and direction of bias they introduce into the estimates under different test lengths and distributions of item parameters relative to the distribution of person parameters. Second, we study these integration techniques in terms of the amount and direction of bias they introduce under various numbers of intervals in the grid of integration. The literature contains some general rules of thumb for the number of Gauss-Hermite quadrature nodes (like the square root of the test length for the relatively long tests). Despite it, the nature of the relations between different integration techniques and the number of intervals in the grid of integration has not been studied yet. We also study the computations complexity and the computational time of these integration techniques to produce some reasonable recommendations for the number of intervals in the grid of integration depending on the integration technique.

Conference Sponsors

Platinum Sponsors

Gold Sponsors

Silver Sponsors

Graduate
Management
Admission
Council™

CiTO

Northwestern Medicine®
Feinberg School of Medicine

Bronze Sponsors

PABST

NCME national council on measurement in education

ASC
FOR GOOD MEASURE

duolingo
english test

optimum
assessment