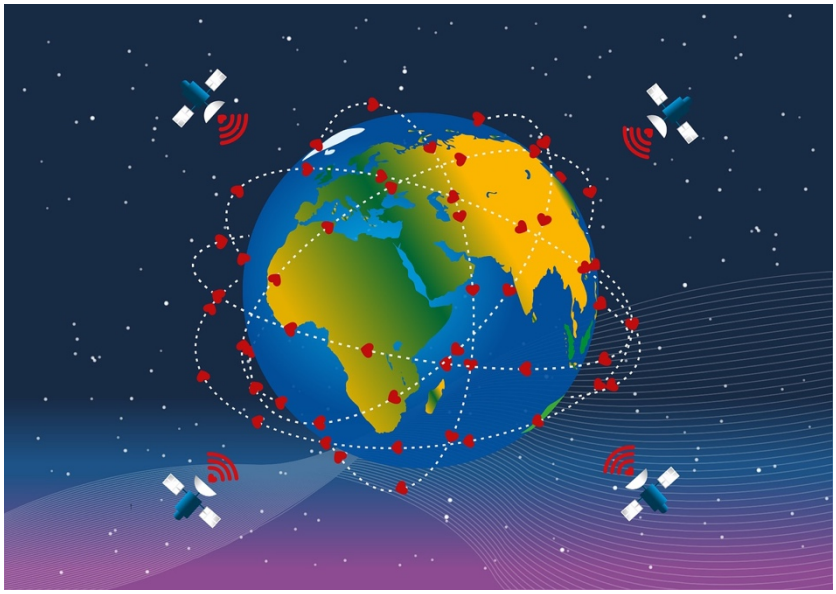
Testing International

Volume 46, December 2021

Editor: Nicky Hayes

Testing International is a publication of the International Test Commission

Contents



ITC President's Message

Aletta Odendaal

President,
International Test Commission

I trust this issue of *Testing International* finds you all well. The academic year is winding down in South Africa, and we are approaching our summer and holiday season. Work has continued amid the continuous interruptions and volatility caused by the Covid pandemic. As I write this report, we are facing a new wave of infections globally, but if the last 12 months have taught us anything, it is to truly appreciate the preciousness of life.

Helping to halt the spread of COVID-19 should be everyone's main concern, and in the fight against the pandemic the most powerful tool we have is vaccination. We have seen vaccination rollouts globally, although a vaccination inequality in developing countries remains a concern for everyone. Those sounding the alarm have frequently repeated the same mantra: "No one is safe until everyone is safe."

That mantra will also impact and guide activities of the ITC in the year to come. The transition to new ways of conducting our business, and in particular relying more on technology to connect with our members, has allowed us to develop new initiatives, such as our very successful virtual Colloquium in July and the way that we have conducted our executive and council meetings online. It not only saves costs; it provides us with the opportunity for more regular contact, and gives the opportunity for more members of the ITC to get actively involved in ITC matters by means of involvement on ITC committees. On that note, too, I'd like to urge any members and friends who feel they would like to get involved in any of the ITC committees to get in touch. Either contact the chairs of the committees directly, or email Ananda van Tonder at ananda.vantonder@intestcom.org.

This newsletter carries a report on the successful hosting of the 12th ITC Conference as a Colloquium on Tests and Testing from 9-12 July. We regard the Colloquium as a great success: not only financially but also the quality of the content that was uniformly praised by delegates, who also welcomed the ability to watch presentations at any time and to have ongoing access to them for the next 6 months. In all, 255 people registered for the Colloquium, and provided us valuable feedback regarding their experiences that we will be using in our formal conference planning going forward. Given our experiences with lockdown and the current covid conditions, we are likely to explore the application of a hybrid format going forward to future events. The Colloquium also illustrates how our experiences of lockdown and continuous isolation have helped us to make the transition towards a new normal, where technology is seamlessly integrated with our daily activities.

Latest news from the Executive and Council of the ITC is that the move of the ITC from US-based organisation operating in US dollars to a UK-based one operating in Sterling has now been successfully completed, with our assets now residing in the UK. Making this transition has been a mammoth task and I would truly like to thank Dave Bartram, our Treasurer, as well as Ananda van Tonder, April Zenisky and Kurt Geisinger for their ongoing commitment, guidance, and support in this process. The transition has established a firm grounding for the sound financial governance of all ITC matters.

I am furthermore pleased to announce that the ITC Council has given formal approval to our support towards the organization of the July 2022 IPCP working congress. The International Declaration of Core Competences in Professional Psychology (IPCP) seeks to

identify a set of internationally recognized and endorsed competences that can serve as the foundation for a coherent global professional identity for psychologists, which will also allow for a possible international recognition system for professional preparation systems, program accreditation, professional credentials, and the regulation of professional competence and conduct. Both Dave Bartram and Dragos Iliescu have been active members of the work group since its adoption in 2013 and will represent the ITC at the two day working congress planned in July 2022.

The work of the ITC is done through our committees, and I want to use this opportunity to thank Steve Stark, University of South Florida, for his excellent work as Editor for the International Journal of Testing that during his term as Editor from July 2014 – September 2021 greatly enhanced the stature of IJT. His contribution was significant with a special issue on *Equity and Fairness in Testing and Assessment in School Admissions* as well as an issue on the *Use of Technology for Assessment in Organisational, Psychological or Educational Research and Applications*. . I then also want to forward a warm and official welcome Chris Nye, of Michigan State University, as the incoming editor of the International Journal of Testing (effective from 1 October 2021). We are looking forward to the future growth of the journal under your leadership.

The Membership, Involvement and Marketing Committee continues its sterling work. They are particularly working to develop our social media and external communication strategy, and have given more detail about this later in this newsletter. I would also like to urge members to be on the look-out for the announcement of the release of our new guidelines on Technology-Based Assessments, in collaboration with the Association of Test Publishers (ATP) under the lead of Steve Sireci and John Weiner. The Guidelines provide information about the key issues to consider when designing and delivering tests using digital platforms, and guidance to test developers, test administrators, and test users on how to best ensure fair and valid assessment in a digital environment. Their goal is to promote best practices in test development, administration, and scoring to facilitate fair and valid measurement of the psychological and educational characteristics targeted by contemporary assessments.

Conferences are a key focus of ITC activities and under the auspice of the Conference Committee an expression of interest to host ITC conference 2024 and 2026 was circulated to members and friends. Council is currently reviewing the applications received and the venue for the 2024 conference will be announced at the 2022 conference in South Africa. South Africa is often viewed as the gateway into Africa, and hosting our conference there will contribute to further strengthen ITC's vision of building capacity in measurement and psychometric expertise and encouraging best practice in assessment. We regard this as especially important in developing countries, which face unique challenges in the domains of applied psychological assessment, development, and practice.

I invite you to read more about the 13th ITC Conference in this newsletter. It is aligned to summer in the southern hemisphere, taking place from 12-15 December 2022. Exploring the theme: "*Advancing diversity, equity, and inclusion: Opportunities and challenges towards culturally responsive assessments*", the 2022 conference will highlight cultural diversity and the rapid advancements in the field of technology-based assessment. It is my sincere hope that life will have become sufficiently normalised by December 2022 for us to be able to host the 13th ITC conference as a face-to-face event, but should circumstances surrounding the pandemic be unchanged we will explore alternative formats of hosting the conference, using the valuable lessons we have learnt from our virtual Colloquium.

The dispersion of our membership around the world is

our strength, and my aim as President is to enhance our international visibility and reputation still further. Our focus is on continuous growth and development, and we plan to continue expanding ITC activity and relevance where we do not yet have a footprint or presence. To help us continue to extend the reach of

the ITC, I encourage all members to share any ideas and concerns that you may experience in your context directly with me. You can contact me at President@intestcom.org.

Finally, I wish everyone a healthy, safe and happy 2022, and would like to end with this African proverb: "*If you want to go fast you travel alone, if you want to go far you travel together*".

Let us work together to further our mission of promoting effective testing and assessment policies and to the proper development, evaluation and use of educational and psychological instruments worldwide.

Aletta Odendaal.
President, International Test Commission

Please remember to let us know if you change your email address. Contact our Office Manager and all-round organizing angel: Ananda van Tonder (ananda.vantonder@googlemail.com).

Editor's Message

Nicky Hayes

Editor "Testing International"

Welcome to the latest issue of Testing International! In this issue, we are pleased to be able to make the first announcement about the planned Conference of 2022, which looks to be very exciting, so we are all hoping that the world health situation will allow us to go ahead with it as planned.

Our other items feature reports from the Conference and Pubications committees, the welcome and introduction of our new journal editor, Chris Nye, and of course our regular report from the Membership, Involvement and Marketing Committee, which includes the invaluable survey of testing-relevant events around the globe produced by Peter Macqueen. In addition, Dave Bartram provides us with some fascinating feedback on this summer's highly successful Colloquium.

We conclude this issue with a detailed and not always complimentary discussion of the latest version of Raven's Progressive Matrices2, by no less a person than John Raven. I suspect it will give all of us plenty of food for thought.

Happy Reading!

Nicky Hayes

SAVE THIS DATE!

The 13th Conference of the International Test Commission
12-15 December 2022

Hosted by the Department of Industrial Psychology, Stellenbosch University, South Africa

We are delighted to announce that the ITC 2022 conference will be hosted in South Africa from 12-15 December at Stellenbosch University. We invite you to participate in this historical event of hosting the first ITC conference on the African continent, so please save this date!

The date was specifically chosen to be aligned with summer in the Southern hemisphere. Stellenbosch is a university town in South Africa's Western Cape province. It is nestled between vineyards and secluded by magnificent mountain ranges. The town's oak-shaded streets, lined with cafés, boutiques and art galleries, are bordered with two world heritage nature reserves, Jonkershoek and Simonsberg. Stellenbosch is uniquely positioned for hosting events, as it is situated just 40 minutes from Cape Town International Airport and 50 minutes from Cape Town, in the scenically beautiful and peaceful Cape Winelands.

The conference promises to be an exceptional professional, cultural, and scientific experience in a unique environment which is renowned for its friendliness, exciting and diverse cultural experiences, excellent wining and dining opportunities and moderate Mediterranean climate.

Exploring the theme: *Advancing diversity, equity, and inclusion: Opportunities and challenges towards culturally responsive assessments*, the 2022 conference will highlight cultural diversity and the rapid advancements in the field of technology-based assessment. The programme will be organised according to the following subthemes:

1. Test development, adaption, and translation: advancing culturally responsive assessments
2. Innovation and advances in psychometric theory
3. Next generation technology-enhanced assessment, privacy, and test security
4. Global differences in equity approaches, policy, and solutions
5. Best practices in testing and assessment: preserving human rights in the era of big data assessment

As we face a new wave of infections, we remain optimistic that life will sufficiently normalise with further worldwide vaccination rollouts to allow us to host the 2022 event in-person. Having the 2022 conference as an in-person-event will give us an ideal opportunity to see old friends again, and make new

ones. The conference is also the place to share research, to network, and to develop international partnerships. The Local Organising Committee will, however, continue to monitor the covid situation and and engage with the ITC Council, in case it becomes necessary to host the event as a hybrid conference.

Please keep a look-out for the second conference announcement. Keynotes, workshops and the call for papers will be announced when we launch the conference website 2022 on the ITC website. We plan that the stimulating scientific programme will also be complemented by a packed social programme, including the opening ceremony, the gala dinner, and magnificent sightseeing tours. Regular updates on the conference will be communicated in the Newsletter, through email communication with our membership and friends, and also on the conference website.

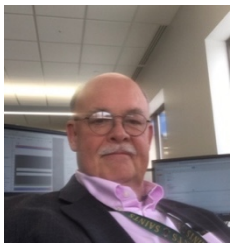
Stellenbosch is an inspiring place to meet, and we invite you to view the following video:
https://www.youtube.com/watch?v=CeYTDY0dyGI

We look forward to welcoming you to South Africa in December 2022!

The Local Organising Committee is:
Prof Aletta Odendaal,
Prof Deon de Bruin,
Prof Gina Görgens,
Dr Michele Visser
Doctoral Candidate Francois van der Bank.

Cape Town

Conference Committee Report



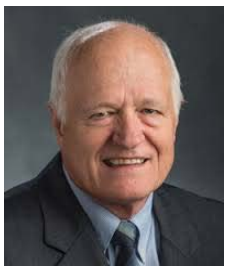
Jon S. Twing, Ph.D.
Conference Committee Chair

With Paula Elosua's move to ITC Secretary General, I have been asked to step in as the Conference Committee Chair and I wanted to say thank you to both Paula and to Aletta Odendaal, ITC President, for this wonderful opportunity to help guide the ITC and our members in developing fun, interesting, and meaningful conferences in the future. I am also grateful to both Paula and Aletta for the wonderful council they have provided during my transition from elected member to Chair and look forward to my continued work with them as active Council members in the future.

While I am new in the job, the work of planning future conferences continues, and the great news is that we have received solicitations from three different countries to host ITC conferences in 2024 and 2026. I am working to schedule conference committee meetings such that we can evaluate, inquire, and clarify such proposals and will be back in our next Newsletter to tell you the great news if we reach a decision before then. Planning international conferences is no easy task! The committee must evaluate many facets of such organization from the cost, distance, and local support as well as ensuring the conference will support the mission, goals and values of the ITC. As such, and given the great expense and impact, we will make no hasty decision so our members can rest assured that when we bring a recommended venue to the ITC Council, it will be well vetted and the best plan possible.

Jon S. Twing PhD
Pearson Assessments & Qualifications, US

The ITC Publications Committee



Neal Schmitt, Chair

Our ITC journal *The Internal Journal of Testing* has now completed its editorial transition from Steve Stark to Chris Nye, and there is a report from Chris in the current edition of this Newsletter. The book series now has nine volumes either published or in process. We would still like to acquire a volume on personality, and there may also be opportunity for one on clinical assessment, and possibly also one on licensing or certification. Our list of existing volumes is as follows:

Iliescu, D. Adapting Tests in Linguistic and Cultural Situations

Scott, J. C., Bartram, D., & Reynolds, D. H. Next generation technology-enhanced assessment.

Schmidt, W. H., Houang, R. T., Cogan, L. S., & Solorio, M. L. Schooling across the globe.

Wells, C. Assessing measurement invariance for applied research.

Laher, S.et al. International histories of psychological assessment. (To be available end 2021)

Stark, S. E., Wiernik, B., & Bornovalova, M. Introduction to Measurement and Decision Making. (currrently being reconfigured in discussion with publisher).

Woo, S., Tay, L., & Behrend, T. Technology and measurement: Research and practice. (In process, several chapters completed, commitments from all authors have been made and chapters are being written).

Scherbaum, C., Goldstein, H., et al. Cognitive abilities and the modern world of work. (Currently in the writing stage).

Arthur, W. et al. Proposal under development on the nature and measurement of adverse impact resulting from testing and assessment.

Any suggestions as to authors or books are welcome: please send them to me, Neal Schmitt, in the first instance, at mailto:schmitt@msu.edu.

International Journal of Testing

Christopher D. Nye
Michigan State University
Editor: International Journal of Testing

News and Updates

The *International Journal of Testing (IJT)* is dedicated to the advancement of theory, research, and practice in the areas of testing and assessment in psychology, education, counseling, organizational behavior, human resource management, and related disciplines. IJT publishes original articles addressing theoretical issues, methodological approaches, and empirical research, as well as integrative and interdisciplinary reviews of testing-related topics and reports of current testing practices. All papers are peer-reviewed and are of interest to an international audience.

In October of this year, I took over as Editor of IJT. I had served as an Associate Editor of the journal for nearly 3 years and am looking forward to my new role. I want to thank **Dr. Stephen Stark** for all of the work that he did as Editor of IJT over the past several years. Under his leadership, the journal has continued to grow and increase in both popularity and importance. The journal now receives around 125 submissions and publishes 20 articles per year. Dr. Stark has played a significant role in this growth and I am grateful to be taking over the editorial role with the journal in such a good position. I am also grateful that **Dr. Elias Mpofu** will be continuing as an Associate Editor. His insights and expertise will continue to make important contributions to the journal.

The other current Associate Editor, **Dr. Sang E. Woo**, will be leaving the journal at the end of the year. I would like to thank Dr. Woo for her service to the journal. To fill her role on the editorial team, we will be looking for a new Associate Editor to expand the

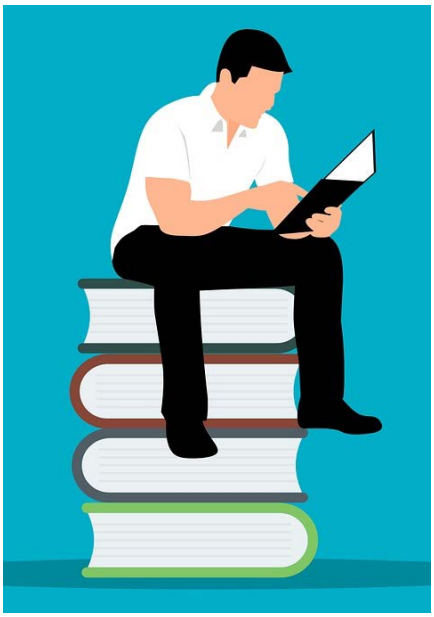
expertise as well as the geographical and cultural diversity of our editorial team.

Nominations for this position are welcome. To be selected, an individual must have an accomplished publication record and expertise in measurement / testing methodology and practice.

As the editorial team moves forward, we hope to continue to increase the visibility and impact of the journal. Over the next year, we will be publishing two special issues on highly relevant and timely topics. The first will be on **Equity and Fairness in Testing and Assessment in School Admissions** and the second will be a special issue on **New Technology in Testing**. We are also continuing to accept papers for our new **Test Validity and Adaptations** section. The papers for this new section are generally brief (approximately 2,000 words) and will go through an abbreviated review process.

Finally, I want to **thank everyone who has supported IJT** over the past year and who will continue to support the journal as we move forward. This includes our tremendously helpful editorial board, associate editors, ad hoc reviewers, and, importantly, authors. The journal would not be where it is without all of you and I look forward to working with all of you over the next couple of years.

Chris Nye



Membership, Involvement and Marketing (MIM) Committee

Peter Macqueen, Chair

No doubt all readers are keen to see the end of 2021, with the prospect of a brighter 2022. COVID-19 has produced disparate effects around the globe but ultimately it has adversely affected many individuals, families, organisations and societies. Let us hope that the situation keeps improving, and we have the opportunity to meet in Cape Town in December next year.

I would like to make a special mention of Ananda van Tonder. Ananda has been dealing with health challenges during 2021, and COVID-19 has not helped. This has made it very difficult for Ananda to follow through on a range of membership matters, and she has done a sterling job in light of the difficulties.

This is another reason why I would ask all members to review their contact details, or that of their organisation; and to consider the state of their 2021 ITC subscriptions.

A few weeks ago Ananda did contact a number of you. It is recognised that organisations often pay their ITC fees in the second half of the calendar year, but with COVID-19 some of these important administrative undertakings can be overlooked, I believe.

A reminder as well that our member services have increased in recent times, with the advent of the **ITC Learning Centre. (**see https://www.itc-learning.net/pages/home) Furthermore, the increased use of social media is likely to provide an added benefit for ITC members. This is discussed in the section below.

MIM Committee Activities:

We had a very fruitful Zoom meeting recently, with committee members from around the globe.
Two key areas that were discussed during this meeting included:

1. Membership structure of ITC. This may need to be modernised in light of our overall mission, and nearly 50 years after our formation. This will be an ongoing matter and something the ITC Council will likely consider thoughtfully during 2022.

2. Social Media. This is an important issue, and is not unrelated to the above element. There is certainly an expectation from many people that online, up to date information exchange is an important part of modern professional activity. An enhanced social media presence should also assist in increasing engagement with members, including student members.

   It was resolved that three members of the MIM Committee will take this further, and with great expedition a report was presented at the Council meeting on 17 November. The three members involved in this important activity are Pia Zeinoun (The Netherlands, and formerly of Lebanon), Sabrena Arosh (Malaysia), and Justin August (South Africa).

As readers would be aware, there are various matters (strategic and practical) that do need to be considered: for both of these key issues.

Forthcoming Events

As previously, I have endeavoured to provide a range of possibilities in regard to testing and assessment events being held globally.

You are of course recommended to check websites carefully for the latest information.

There does appear to be a trend for conferences to move to a hybrid mode (rather than just virtual), but even so, certain major events have been postponed such as ICAP 2022 (Beijing) which is now scheduled for July 2023 (Beijing).

I am sure you have your own thoughts about the relative benefits of physical, virtual and hybrid conferences, but apart from the potential impact of COVID-19 I would not be surprised if issues of environmental sustainability are considered increasingly when organising international conferences. And of course, some of us still may be subject to significant travel or quarantine restrictions, despite being fully vaccinated. **So do check which components may be just virtual or in person, and the window of opportunity to access material following the conference.**

The following represent some of the events of which we are aware @ 15 November 2021

CAUTION: DETAILS MAY CHANGE!

AFRICA:

42nd Annual Conference of Assessment Centre Study Group of South Africa (ACSG)
7-14 March 2022
https://acsg.co.za/conference-information

The ITC Conference

Stellenbosch, SOUTH AFRICA
12th - 15th December 2022

ASIA:

Pacific Rim Objective Measurement Symposium (PROMS)
4-6 December 2021 Virtual
via Nanjing Normal University, China
The planned 2020 conference was cancelled due to COVID. The details of the 2021 Conference have just been released. You may not receive this TI newsletter in time – but the information is provided for future reference. Rasch scaling is a central theme of PROMS.
https://tinyurl.com/rbt8jnme

30th International Congress of Applied Psychology (ICAP)
24-28 July 2023 Beijing CHINA
http://www.icap2022.com/

EUROPE:

20th European Association of Work and Organizational Psychology (EAWOP) Conference, with BPS DOP
11 – 14 Jan 2022 Glasgow SCOTLAND
https://eawop2022.org/

Theme: Interventions at Work – Integrating Science and Practice

17th European Congress of Psychology
5-8 July 2022 Ljubljana SLOVENIA
https://www.ecp2022.eu/
Theme: Psychology as the Hub Science: Opportunities & Responsibilities

E-ATP (Europe Association of Test Publishers)
10 - 12 October 2022 London UK
https://www.testpublishers.org/european-atp-conference

NORTH AMERICA:
62nd International Military Testing Association (IMTA) Conference
Raleigh NC
7-11 March 2022 postponed from 2020 and 2021
http://www.imta.info/Conference/Conference_Home.aspx

Association of Test Publishers (ATP) Conference
Innovations in Testing In Person & Virtual
20-23 March 2022 Orlando FLORIDA
http://www.innovationsintesting.org/

37th SIOP Conference
Seattle WASHINGTON STATE
28-30 April 2022
Virtual & In Person (but not all content)
https://www.siop.org/Annual-Conference

OCEANIA:
There are no specific Testing events scheduled although broader conferences should offer sessions related to testing and assessment.

14th APS Industrial & Organisational Psychology Conference
GOLD COAST AUSTRALIA
7-9 July 2022
A physical conference at this stage (but likely to have virtual content)
Theme: IOP at the forefront: Leading transformative and global change
https://www.psychology.org.au/APS-IOP-Conf/2022

SOUTH AMERICA:

39th Inter-American Congress of Psychology (SIP)
TBA: The last was July 2021: see below
https://38cip.sipsych.org/

If you are aware of any forthcoming conferences in 2022 / 2023
(virtual, physical or hybrid), please let us know!

PLEASE SEND YOUR SUGGESTIONS to:
secretary@intestcom.org

Ananda van Tonder (ITC Office Manager) or Paula Elosua (ITC Secretary-General) will direct your email for action.

The ITC Colloquium

9th to 12th July 2021

(and thereafter to 31st December 2021)

REPORT

Dave Bartram

Despite being something of a risky venture in comparison with the tried-and-tested format of the ITC biennial conferences, the Colloquium was a great success. Financially, it was successful in helping us offset the losses made in 2020 from the cancellation of the Luxembourg conference. Other aspects of it were more nuanced in terms of people's reactions. This article is abstracted from the full report written for ITC Council. It will focus on lessons learned and feedback on aspects of the Colloquium which were new to us, due to its virtual format.

Finance

We had budgeted for 300 people to attend the Colloquium. In the event we had 238 people registered. This included 59 who were complimentary registrants, and the rest who had paid Early Bird rates either at the general rate or as students in Income Categories A, B or C. The budget had been set assuming 240 would be paying and 60 would be complimentary. In the event we achieved 179 paying delegates and 59 complimentary.

We managed to underspend on the budget (for which expenditure was set at just over £30,000) by around £13,000. This was due to most of the work being done by us rather than by paying others to do it. We no longer had a separate Local Organizing Committee (LOC) to do everything. We had settled up with the University of Luxembourg on this issue and had split the costs of the losses with them. Unfortunately, Ananda van Tonder, our office manager, had to undergo surgery and was unwell for a long time, which left us under more pressure than we had planned for!

However, financially, the Colloquium was a great success and enabled us to recoup all the losses made in 2020.

The platform

The feedback we obtained on the content and the overall experience was varied. The quality of the content was uniformly praised by delegates, who also welcomed the ability to watch presentations at any time and to have ongoing access to them for 6 months. There was less consensus on the merits of the platform. The use of Zoom was liked very much, and people enjoyed the live interactions in the 'breakout rooms'. There was less agreement on the ease of navigation around the system. The main criticism was that there were too many supporting videos and articles to read. The arrangement of information in the downloads could also have been better designed in some cases. What people really wanted were single A-4 'cheat sheets 'showing how to carry out the key procedures.

There were also criticisms of the length and complexity of the induction process. We were aware of the problem of there being too many steps: Registration on the ITC web site; a delay of a few days; and then invoicing, with payment through Stripe; log in to the Learning Centre and then enrolment on the Colloquium bundle. Each step introduced potential areas for confusion and required a lot of our time to manage.

People would also have liked to get access to the Learning Centre site sooner. We had given a full 24 hours of access prior to the Opening Session. We had intended this to be a week. However, our ability to set things running was compromised by a few presenters submitting materials very late. We had issues with sponsors as well as delegates on this. We have learnt a lot from this event: we know what to do next time (if there is a next time) and what not to do.

In essence the virtual event is highly cost-efficient for delegates as it costs only the registration fee – no travel or accommodation to pay for and registration fees were reduced. These fees were just 50% of a standard face-to-face sessions fee. The ITC could consider running a similar virtual event occasionally or on alternate years to our traditional biennial face-to-face conferences.

There were also two by-products emerging from this event that may prove useful to those involved in future conference organisation. The first is a more formal set of conditions governing the relationship between the ITC and a Local Organising Committee. The second is a standard agreement for use with sponsoring organisations.

The Colloquium Assistants

The ITC Colloquium was staffed during the 4 conference days around the clock by a staff of 8 volunteers: masters and doctoral students from the University of Bucharest, Romania. The team was selected, trained, and supervised by Dr. Andrei Șerban Zanfirescu, also from the University of Bucharest, Romania. They operated primarily from their base in the Zoom Forum. Access to the Zoom Forum was via the Zoom Portal in the Learning Centre

Separately, Mrs. Adelina Neagu (an assistant of Prof. Iliescu), also contributed with consistent administrative help before and during the Colloquium. The volunteers staffed the Zoom sessions, i.e., the main lobby and the various breakout rooms and had the following responsibilities:

- address technical issues (microphone and sound tests)
- help audience navigate the zoom lobby and breakout rooms
- record sessions
- update audience on event-related changes

One of the Colloquium Assistants (CA) was assigned to each session and helped ensure everything runs smoothly. We had three sessions designated as Workshops, three panel discussions and a number of events like the opening and closing ceremonies all ZOOM-based and all different from each other in format. We also had 50 Sponsored Breaks (15-minute sessions) which were held in one of 6 meeting rooms (a meeting room is simply a Zoom Breakout room).

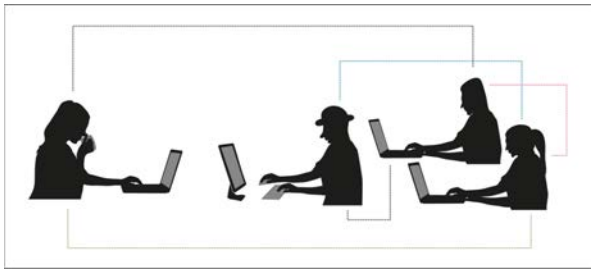
One CA managed each Zoom Breakout room. They needed to check with the presenters that the information we held was correct. The CA needed to facilitate moves from the Learning Centre (e.g. from Room A) to the Zoom Forum and its breakout rooms (e.g. to BR 01). We did recommend that session presenters provided a PDF copy of their slide set if applicable. They were also invited to provide PDFs of one or two papers people might want to read in relation to their topic.

The CAs dealt with recording the sessions and preserving the recorded version for the ARCHIVE. They managed the flow of people. They also helped with the scheduling of activities and the close of the sessions.

The actual Colloquium shows an average level of 10.27 visits per delegate (nb: n=255 for number of delegates) over the four days, giving an average per day of over 2.5 visits. Monday was a half day only, so if one adjusts for that the average visits per day comes to just under 3 (2.93)

Records of presentations

The virtual format of the event provides access to new measures of how successful the event was. As an



example, if we look at the Fremer-Foster keynote presentation we can see what sort of data is obtained (session KEY-A07). We looked at the data from the beginning to the end of July.

- The session was scheduled for playback on Sunday 11th July from 13:30 to 14:15 in Room A.

- There was a Zoom feedback session scheduled for 15:00 to 15:15 the same day in Breakout room 1.

We found that the video was played 41 times in total from the time it was loaded to 12:15 UTC+1 on 21/07/2021. We can tell whether one person was playing the same video repeatedly or if it was different people. The 41 plays were made by 25 people. One person played the video 5 times (person #18). [Note that individual identities are confidential as individuals did not give their permission explicitly for any further use of this data. 'Anonymous 'means that we do not know the ID of the person.]

The data show clearly that people were quite content to play the video at different times from before the event started to well after it had finished. This pattern is very typical. Looking at July 11th, we see people playing the video from 08:53 to 18:32 on Sunday (in relation to UTC+1 times).

The Learn Worlds software provides some other useful data on the video presentations. For each video it provide an average engagement score (e.g.39% for the Fremer-Foster keynote), based on a total of 11.47 hours of viewing divided by the total number of plays x the video length.

It also provides a plot of number of viewers by elapsed time and other more detailed outputs. Analyses were carried out of the viewing rates for each video. In general, there was not much difference between videos but the "in Memoriam" sessions had the highest engagement scores.

Feedback from Delegates

A questionnaire was developed and distributed to all delegates on 20th July. Submissions closed on 1 August 2021. 69 people (27% of the total) returned completed questionnaires. (The results are presented in Appendices to the full Conference Report.) Detailed comments were provided by 35 of these people (13.7% of the total). For the most part scores on the statements were on a six-point scale, with '5 'and '6 ' indicating 'good 'to 'excellent 'ratings.

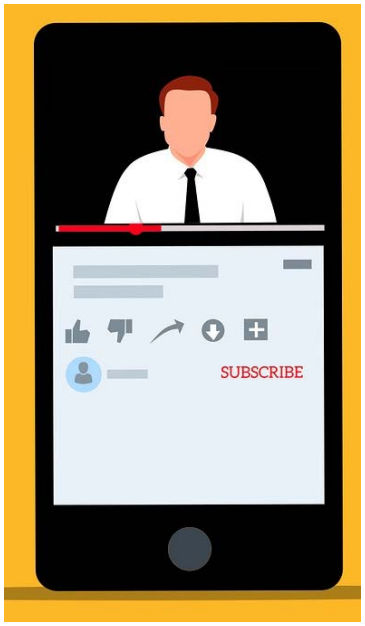
Most of the modal scores were sixes or fives, and the average scores were around 5. Typically, the problems tended to arise from the ratings that were to do with usability of the technology. For example, the quality of the current content was rated quite high while how easy it was to move around the learning centre system was rated low (This last question got the lowest average rating of 3.03).

So, in general the ratings were all satisfactory to good, but the main criticisms were reserved for those that were to do with the ease of using the technology. The extended comments that some people gave were also along the same lines. They were generally very complimentary about the content and the presentations but less so about the technology involved in using the system. It is very difficult to know how much that is an issue of the actual system that we had developed as opposed to virtual conferences in general. However, there were clearly problems with the particular system that we had which may have been peculiar to its design.

Afterwards

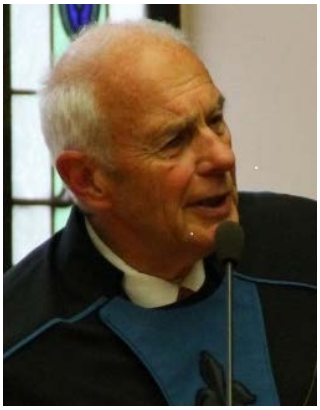
After the Colloquium was over, people still had access to all the materials. The use was tracked up to 5th August. People were informed that they would have access until the end of December 2021. Delegates accessed the materials on 406 occasions in the 25 days following the Colloquium: an average of 16.24 times per day across all delegates. We are continuing to track use and will report on the analysis of delegate behaviour in a future paper.
We do not have any active plans to do another virtual conference, but who knows what external factors might arise to push us to do so. We believe the Colloquium was successful and obtained generally positive reviews of it. The technical features of the platform need improvement, but the main content of the conference was impressive and well-appreciated.

Dave Bartram
Chair of the Local Organising Committee.



Food for Thought

Commentary on the *Raven's 2 Progressive Matrices* Tests and Manual *(Pearson, 2018)*



John Raven

I feel obliged to write some kind of commentary on Pearson's *Raven's 2 Progressive Matrices* tests and *Manual* in part because we had earlier developed tests (the ***Parallel*** *Coloured* and *Standard Progressive Matrices* and the *Standard Progressive Matrices* ***Plus***[i]**)** to meet the needs[ii] the authors advance to justify the production of the new tests, and in part because I feel somewhat embarrassed to find my name, and that of *Progressive Matrices,* attached to tests many of the items of which bear little resemblance to those in the *Raven Progressive Matrices.*

While I have little doubt that the new tests will serve the purposes for which the *Progressive Matrices* are most commonly used, it is less clear that they will be as well suited to the purposes for which the tests were originally developed or, indeed, to some specialised uses to which the tests are currently put.

It cannot be too strongly emphasised that J.C.Raven's tests (which include the *Vocabulary* scales) were developed as theoretically-based tools for use in research and not primarily as tools for practical applications such as in personnel selection. As such, particularly in the context of the cross-cultural and historical data that have accumulated, they have proved of inestimable value. I am much less sure about the ethics of many of their practical applications[iii].

It is not at all clear that the *Raven's 2* tests will be able to contribute in the same way to the advance of scientific understanding.

First, a brief description of Raven's 2.

The *Raven's 2* suite of tests depends on a pool of 329 items all of which have been assessed for conformity to the "Rasch" measurement model[iv] and assigned difficulty levels determined through the application of Item Response Theory. From this pool it is possible to extract thousands of individual constellations of items which differ from each other in specific content but are nevertheless statistically equivalent. As a result, it is not possible for any one individual to have memorised the answers or to copy from his or her neighbour.

The on-line program routinely constellates a set of 60 items arranged in five Sets, roughly equivalent to the *Standard Progressive Matrices.* However, users can request a test consisting of three Sets of easier items (total 36 items) (for children and the less able) and four Sets (48 items) of more difficult items for adults. Each of these extracted sets of items is virtually unique so, although no two people take "the same" test, each set of items is equivalent in terms of difficulty and other properties as judged in Item Response Theory terms.

There is also a published, printed, set of 60 items drawn from the full set of 329 items and arranged, as in the Classic *Standard Progressive Matrices* test, in five sets of 12 items which increase in difficulty within each Set and then revert to easier items at the beginning of the next Set.

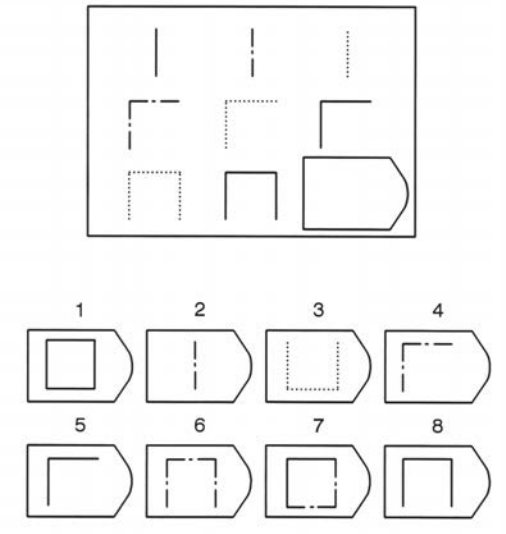
Testing can be arranged such that only Sets A, B, and C are administered to younger children or B, C, D and E to older children and adults. Time limits are set for the testing sessions, although most people finish well within those limits. In other words these are not timed tests. Although the computerised administration incorporates a "discontinue" rule whereby testing stops if the respondent gives 6 incorrect responses in a row, the administrations are not "adaptive" in the sense that such programs present additional items around the point at which the respondent begins to fail and, in this way, generate a more accurate score.

The Items

Raven's *Progressive Matrices* tests[v] consist of a series of 2x2 or 3x3 dimensional matrices, or patterns, the cells of which display non-verbal figures showing progressive change following the same logic in two dimensions (vertically and horizontally). The bottom right hand cell is left empty and the person taking the test is asked to choose from a number of options that which is required to logically complete the pattern. An example, not from any of the tests, is shown in Figure 1.

Many of the items included in the *Raven's 2 Progressive Matrices* (and incidentally, also in the "Matrix Reasoning" subtest of the *Wechsler Adult Intelligence Test*), do not follow this format but consist of a one-dimensional, linear series of figures such as would be obtained by extracting the bottom row of the stimulus matrix shown Figure 1 (although the series presented consist of six rather than three cells).

Fig 1 An Illustrative Progressive Matrices Item



This suggests that the authors have not fully grasped what the term "matrix" is intended to imply. Although it has a number of uses, the term typically refers to an array of figures, symbols, organs, or tissues bearing some meaningful relationship to each other. When the terms in the matrix bear some mathematical relationship to each other, in mathematics, the array, or matrix, is referred to as a *determinant*.

It is not clear that all the items of the Raven's 2 have this property. These items can thus be said to be *indeterminate*.

In the Raven *Progressive Matrices* tests, the patterns in the cells of the array display the same progressive change in two dimensions and thus uniquely determine the nature of the piece required to complete the overall pattern, or matrix. The nature of the piece required to complete the matrix, or pattern is determined by the features of the matrix presented, and can be generated without reference to the range of options presented[vi].

A "solution" which fails to complete the progression in *both* rows and columns using the same logic is unsatisfactory. This has the great advantage that whoever is taking the test can, by checking in two (or three) directions, be *certain* that he or she has the correct answer. But, to set this *commentary* in context, the matrix display can itself be used to illustrate what the test is intended to measure[vii].

In technical terms the *Raven Progressive Matrices* tests were designed to measure *eductive* ability, a term which Spearman[viii] introduced to refer to one of the two major abilities constituting ***g.*** The other component being *reproductive* ability[ix]. The first major component, *eductive* (not educative) ability consists of the ability to educe, or draw, meaning out of, apparent confusion, draw logical conclusions from the insights gained, and test those conclusions.

Unfortunately, the phrase "make and test logical inferences" already leads us into a trap because it seems to imply verbalisation which is not necessarily the case. In much of life – in the booming, buzzing, confusion in which we live – the elements of the multi-dimensional matrix around us are not pre-formulated and have to be educed, or abstracted, from the whole. A recursive process which involves paying attention to the whole to discern, and conceptualise (beware the verbalisation overtones of the term) the parts, and then using that conceptualisation of the parts to reconceptualise the whole is needed.

This is essentially the process of perception itself. Contrary to what was once taught in elementary psychology courses the visual field is not projected onto the retina in something like a photographic image and from there transmitted to the brain. The retina consists of an overlapping array of different types of nerve endings with a major gap at the point of entry of the optic nerve and sparse provision at the periphery. The brain has to make sense of the complex and distorted information it receives. It fills in gaps so that we see what is not there and does not see things that are there.

It is misleading to call variance in the ability to do these things "cognitive ability". Yet it is this very process that the *Raven Progressive Matrices* sets out to measure.

When most readers of this article look at the easier items in the RPM, the whole process takes place automatically, unconsciously. Many deride the items as "merely perceptual". But, as the item analysis[x] shows, this is anything but the case for the less able individuals. When they come to more difficult items many more able individuals[xi] begin to ask themselves (often in non-verbalised form) "What is going on here? If I am right, I should pay attention to this and this and *this* should happen. Whoops. No. That does not work. Let's try again. Now I am attracted by *this* idea". And so on. Certain details of the pattern come to attract and demand attention. Feelings of excitement, elation, frustration, and satisfaction wax and wane.

This is not the kind of process which comes to mind when one hears the term "cognitive ability" Yet it is the process which occurs as people set about "intuitively" organising their lives to achieve their often unclear goals. They say to

themselves "There are many things going on here, but for this purpose I need to concentrate on these and these things … and change my perceptions if things don't work out as I expect". The more things they consider, the more systemic their thinking. And it is the quality and comprehensiveness of their perceptions and inferences which determines their success.

All of these processes are required to engage with the classic *Progressive Matrices* items but are much less likely to be called upon in relation to items which consist merely of linear series. Unfortunately, everyday observation tells us that many of those who (for example) make vital policy decisions have *not* considered the whole, taken account of multiple "variables", or checked their inferences from a range of perspectives[xii].

The internal properties of the Raven's 2 test.

Range of item difficulties

One of the reasons we set about developing the *Standard Progressive Matrices* ***Plus*** was to introduce more difficult items to compensate for what has become known as the "Flynn effect", i.e. for a previously unsuspected, inter-generational secular increase in performance on measures which relied on, or indexed, eductive ability[xiii], i.e. the ability to make meaning out of confusion (which, as we have seen, Spearman had identified as one component of ***g*****).**

From the point of view of evaluating the *Raven's 2* test it is important to note how it came about that we were among the first to demonstrate this increase. It stemmed, at least in part, from the fact that the same (unchanged) test had been widely used in the same way in many different countries[xiv] over a long period of time. Note that this calls into question the wisdom of making fundamental changes to the test.

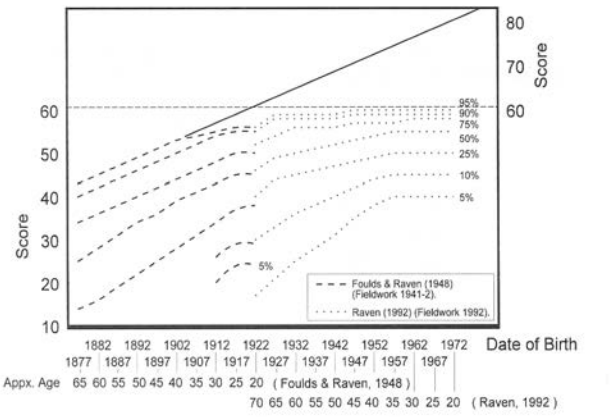
More specifically, we were able to compare the results of a 1979 UK-wide norming study with data collected some 40 years earlier. We published our data in tabular form and focussed on the differential increase at different levels of ability.

We did not express the overall increase in terms of the percentage of test variance that the increase represented. We would, in any case, have been reluctant to do so because, as will be seen from Figure 2 below, the scores are not Gaussianly (often misleadingly called "normally") distributed. This renders the application of the statistical procedures traditionally used to summarise data – i.e. means and standard deviations – inappropriate.

As it happened, Thorndike had documented a similar effect with the Stanford-Binet test a few years earlier[xv], and Flynn[xvi] had become seriously concerned about the implications of the increases.

Undeterred by statistical niceties, Flynn converted our data to means and standard deviations. The results were striking. As reported by Flynn, scores were increasing at about 1 Standard Deviation per generation. Some fifty percent of our grandparents would have been consigned to special education classes had their scores been judged by reference to the more recent norms.

Figure 2 Classic Standard Progressive Matrices



The graphs in Figure 2 show the rise in *Standard Progressive Matrices* (SPM) scores across the best part of a century. The X axis shows both the age of the respondents at the time of testing and the years in which they were born. The lines to the left come from data collected between 1941 and 1942 (as reported in 1948) and those to the right from the 1979 study.

As the graph shows, the main result of the increase has been to undermine the SPM's ability to discriminate among more able adolescents and adults. (Since the test has 60 items there is a marked ceiling effect.) As shown by the solid line to the right (which shows a projection of the increase in scores for the 95th percentile) a dozen or so additional, more difficult, items would be required to restore the test's ability to discriminate among high-scoring individuals.

However, generating additional difficult items turned out to be more easily said than done and, as reported in our Manual[xvii], we had great difficulty - indeed found it impossible - to generate items that were more difficult than the most difficult items of the *Advanced Progressive Matrices*.

So the question we have for the *Raven's 2* is whether its authors have succeeded not only in this task but also in making provision for further increases in the future.

So far as I can make out from Table A1 in their Manual, the last two or three items in the paper version of their test are indeed extremely difficult.

Few of those in the standardisation sample (and they are actually the 25 year olds) have abilities, expressed in IRT terms, of above 650. This corresponds to a raw score of 45 on the "adult" selection of items included in the paper version of the test. There does therefore seem to some room for a further increase in scores as per the "Flynn effect".

It is important to note that, expressed in IRT ability terms, there is a big jump in the abilities required to progress from raw scores of 45 to 46, 46-47 and 47-48, i.e. in the difficulty levels of the three most difficult items. One implication of this is that, if scores continue to increase, it will, once again, be impossible to discriminate between more able individuals.

I will return to this in a moment. But first let us note that the normative data published in the *Raven's 2 Manual* actually suggest that such an increase is likely – because the top scores in the norm table decline from age 25 onward!

This may sound like an odd statement. And so it would be but for the fact that what Flynn showed (and what Figure 2 above shows) is that the cross-sectional data that had previously been interpreted as indicating a decline in scores with advancing age in reality represented mainly an increase with date of birth. This is of profound significance because what it reveals is that *there had always been* massive amounts of evidence for the "Flynn effect"! It just had not been recognised for what it was.

Very many authors had reported from one-off, cross-sectional[xviii], norming studies that scores on many tests – especially those which rely heavily on eductive (meaning-making) ability [in contrast with reproductive ability (ability to store and retrieve information)] – declined with age. If graphed and the X axis renamed "date of birth" instead of "age" (as has been done on the X axis on the Graph shown in Figure 2 above) the graphs reveal an unmistakable increase with the date of birth of the respondents.

When the data the available for many tests are re-plotted in this way it quickly becomes apparent which scores have been increasing over the years and which have not. A brief summary of the outcome of doing this is that, by and large, eductive (meaning making) ability has been increasing while reproductive ability (knowledge) has not.

In other words, while our ability to make sense of the world has increased, it seems that we cannot register and retrieve any more information than we could a century ago. We can run faster and think better but do not retain more in our heads. Why not? This, not the "Flynn effect", is the real puzzle to be addressed by psychologists. But will the *Raven's 2* test be able to register the increase which the authors' own data suggest is likely to occur?

Yes … And No.

At this point it is useful to go on a slight digression before we consider the terminology in the *Raven's 2* Manual.

Fluid and Crystallised Intelligence

It is important to caution against the substitution of the words "fluid" and "crystallised" "intelligence" for the terms eductive and reproductive ability. Although Horn, along with Cattell, was responsible for introducing the fluid/crystallised terminology, Horn[xix] later joined Spearman in emphasising that reproductive ability is not a crystallised form of "fluid" ability. The abilities differ at birth, have different genetic origins, are influence by different things in the environment, and predict different things in life. They simply work closely together (see endnote[xx] [xxi]). And, as to the word "intelligence", Spearman had long ago shown that this was such a slippery concept, being used by different people to refer to different things and by one person to refer to different things at different points the same discussion[xxii] that it was best to avoid the use of the word at all costs.

The terminology in the Raven's 2 Manual

With this behind us, let us consider the terminology in the *Raven's 2 Manual*.

The very first sentence in that *Manual* immediately leads into a quagmire. It says: *"The Raven's Progressive Matrices 2 … is a nonverbal assessment of general cognitive ability".*

Unfortunately, the term "cognitive ability" is itself used, measured, and understood in many very different ways. This leads to endless disputes and misunderstandings between researchers, educational administrators, school psychologists, and in courts of law. The problem could be partially corrected by inserting the words *one component of* before "cognitive ability".

The authors then proceed to rescue themselves in the remainder of the paragraph. But the gain is short-lived. A few sentences later they are talking about "mental ability" as if this were some kind of unitary "thing". They even write about "a full range of cognitive ability". They then fully exonerate themselves by introducing material which looks as if it was copied out of the main Raven, Court and Raven Manual for the Raven Progressive Matrices tests[xxiii]. However the intrusion of a paragraph referring to "the contemporary theory of intelligence" suggests that they have not fully understood and accepted what is being said.

As one progresses through the Raven's 2 Manual, the authors drift back into the *general cognitive ability* terminology,

implying that the Raven's 2 test provides an index of this. The nature of the authors' thoughtways emerges clearly on page 43 where, under the chapter heading *Interpretive Considerations,* they repeat: "*The Raven's 2 is a nonverbal assessment of general cognitive ability …*"

The question of terminology (and associated thoughtways) is no academic matter. It acquires profound significance as one enters the endless quagmire of contradictory legislation and cut-off points relating to access to different types of educational provision and other administrative procedures relating to access to, and compulsion to engage in (or be excused compliance with) social provisions … including such things as whether or not one is "too dumb to die"[xxiv] [xxv].

Such terms as "IQ" and "Cognitive Ability" are used and understood by researchers, legislators, school administrators, school psychologists, lawyers, judges, and jurors in very different ways. There are serious questions about how substitutable the results of one test are for another. It behoves the writers of test Manuals to be as clear as they can about what their tests do, and do not, measure.

J.C.Raven claimed only that the *Progressive Matrices* measures "the ability to perceive and think clearly" … technically *eductive* ability[xxvi]. This is only one component of ***g*** – which is itself only one kind of "intelligence", concerned with only one domain of "ability". In contrast, the term "cognitive ability" is deeply contested and sloppily used … for example educational credentials (which mainly measure *reproductive* ability) are often used as a sufficient index of it[xxvii].

Although they are not alone in this, the authors' failure to choose and use their words carefully is, to say the least, alarming … particularly because the *Raven Progressive Matrices* tests are used to take decisions which strongly influence the lives and livelihoods of millions of people's and thus the operation of whole societies and the future of the planet[xxviii].

In the past, the RPM has intruded directly into the lives of billions of people worldwide and had a dramatic influence on their lives and careers. Via its use in the allocation of personnel in the military systems of many countries it has probably even had a major influence on the outcomes of wars[xxix].

Language free and culture free

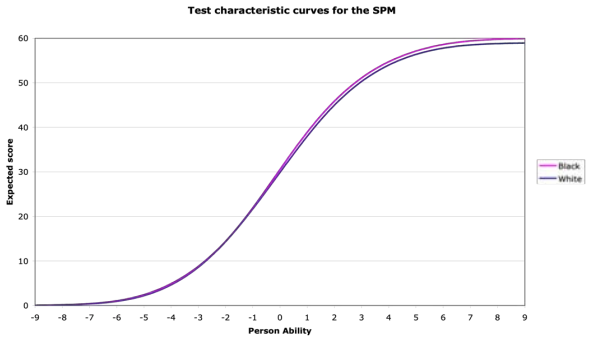
The notion that the tests are *"nonverbal"* has been widely disputed by researchers, some of whom argue that language is required to solve the problems. Yet the fact that it is, in some sense, language free is readily apparent. Unfortunately, this has led many to believe that it is "*culture free*". This again turns out to be an ambiguous term carrying with it many varying assumptions and overtones. The briefest dip into research shows that there are huge differences between the scores obtained by different cultural groups[xxx]. The ramifications and implications of this are far from obvious.

The *Raven's 2* manual offers only overall US norms, not broken down by cultural group. The implication seems to be that these can be used, without further ado, to do such things as assign pupils to educational programmes or to select personnel. But think about this. Readers of the *Raven's 2* Manual are referred to the *Research and References* sections of the Raven, Raven, & Court Manual for evidence on the validity of the test[xxxi]. If one visits this material, what quickly emerges is that, regardless of the language-free nature of the test, different cultural groups within the United States of America (never mind elsewhere) achieve very different scores.

Stability of Internal properties.

Yet the test works - scales - in much the same way in most of the groups that have been studied. We checked that, however different their mean scores, the sequence of item difficulties was much the same in many cultural groups, but the point is neatly illustrated in the graph in figure 3, generated by Nicola Taylor[xxxii] from data derived from applicants for jobs in the mines of South Africa[xxxiii]. But, before looking at it, note that there are huge differences in the mean scores of miners from different backgrounds.

Figure 3



Predictive Validity

Not only does the SPM, to all intents and purposes, have the same internal properties within different cultural groups, it also has similar predictive validity to external criteria within the groups. Figures 4 and 5 show the regression lines for Anglos and Hispanics for the RPM against a couple of scores on the California Achievement Test in Douglas, Arizona around 1990[xxxiv]. Figure 4 shows the regressions of CAT Math scores on the Coloured Progressive Matrices among first-grade students. Figure 5 shows the regressions of CAT Reading scores on the Standard Progressive Matrices among fourth-grade students. The graphs are essentially parallel but operate at different levels of ability, and similar graphs are

available for other subject areas (Redrawn from Hoffman, 1990).

Figure 4 First grade students

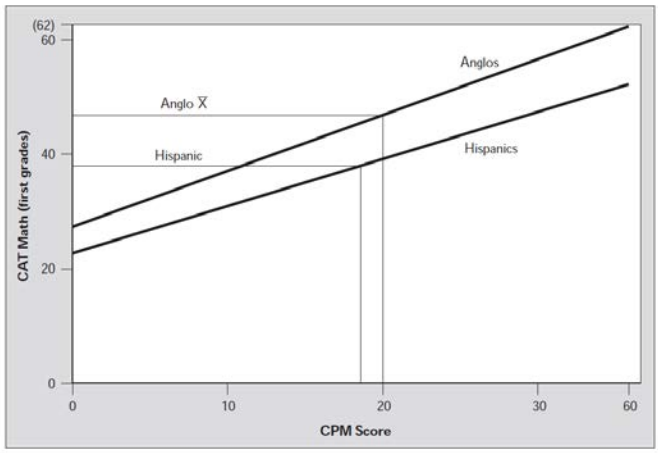
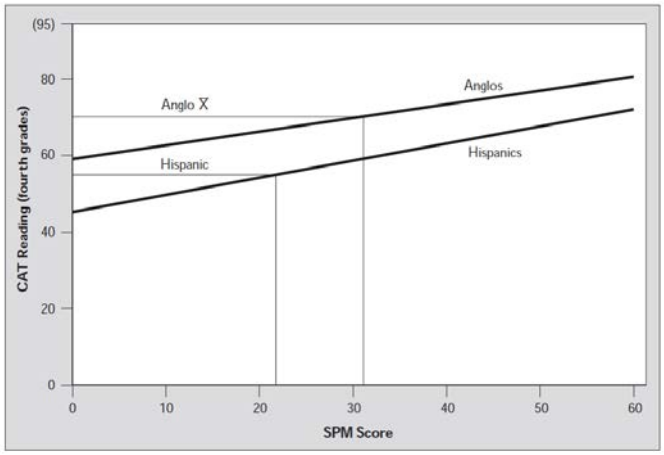


Figure 5 Fourth grade students



In other words, despite the fact that the test is not "culture free" in the sense in which that term is commonly understood, the test works, and works in the same way, for different cultural groups and also has similar predictive validity to external criteria within those groups.

Some implications

So far, so good. But what happens when the overall norms are used, for example, to assign pupils from different cultural backgrounds to mandated variance in educational provision (such as programmes for those deemed in need of "special" or "gifted" education) or to select people for jobs? Well, obviously, a disproportionate number of the highest scoring group are going to be selected for "gifted" programmes and vice versa for allocation to "special education". *And that decision will indeed be a valid decision when evaluated against performance criteria*.

Hmmm**.**

There appears to be a danger, in this way, of perpetuating the very social differences which may have contributed to the differential scores in the first place[xxxv]. In short, one has to challenge the widespread unverbalised assumption that, once one has a "language free" test, one has solved the problem – that that is the end of the matter. It clearly is not … and the issue merits explicit discussion[xxxvi].

The Test Characteristic Curve and the meaning of "difference" or change scores.

I turn now to another question about the internal properties of the test.

It has to do with the regularity of the increase in item difficulty from item to item and thus the relative meaning of any given *increase* in raw score at different points in the scale. For example: Is a raw score increase from say 8 to 11 in any way the same as an increase from 48 to 51?

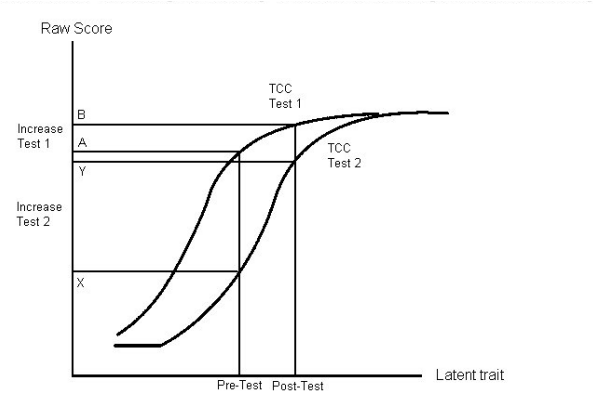
It has taken many years for the significance of this question to become fully apparent although its neglect has resulted in endless misleading conclusions being drawn from research,such as in the assessment of the relative impact of remedial or educational enrichment programmes on the more or less able. It has resulted in the unjustifiable advocacy of a number of derivative (difference) scores, such as "Learning potential" or "sensitivity to stress", where a given difference is assumed to have meaning and predictive validity - that is, correlation with other variables - regardless of the point on the scale at which the difference occurred. It also shows up in legal procedures relating to claims about such things as the effects of accidents or injuries on mental abilities, where so much turns on the magnitude of the decline in comparisons of scores before and after the injury.

Although J.C.Raven and others involved in the development of the tests mentioned trying to make sure that the items were "as far as possible" equally spaced in difficulty, no great attention was paid to doing this and there was no recognition whatsoever of the implications of not doing so. In short, the problems involved in the measurement of change were simply not recognised.

We were fortunate in having the matter forcefully drawn to our attention by a student of G.H.Fisher (Joerg Prieler) who happened to get involved in our work via the Schuhfried Company in Vienna. Prieler (2008) illustrated the problem by means of the graphs shown in figures 6 and 7.

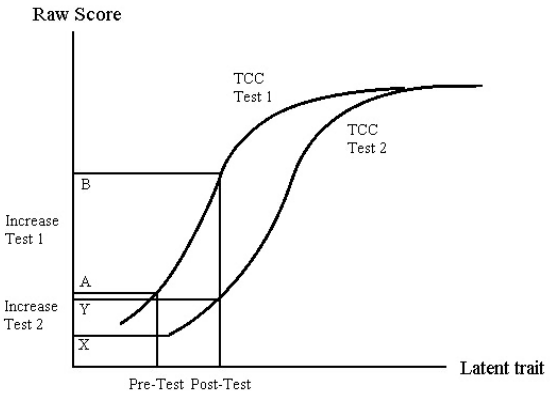
Figure 6 illustrates the problem in connection with trying to evaluate the impact of a managerial development programme on high ability personnel and Figure 7 on low ability personnel. If we employ a test having the Test Characteristic Curve shown on the left in Figure 6, the mean scores of the high ability group increase from A at the pretest (i.e. before training) to B at posttest (i.e. after training). This is a relatively small increase. But if we use the more difficult test shown on the right, the same increase in score on the latent trait of the high ability group shows up as a *huge* increase in raw score, moving from X to Y.

Figure 6 Illustration of changes in raw scores on "easy" and "difficult" IRT-Based tests of cognitive ability for identical changes in latent ability *High* ability group only



As can be seen from Figure 7, exactly the opposite effect occurs at the other end of the scale. The apparent increase in score from pretest to posttest is huge on Test 1 and trivial on Test 2.

Figure 7 Illustration of changes in raw scores on "easy" and "difficult" IRT-Based tests of cognitive ability for identical changes in latent ability *Low* ability group only



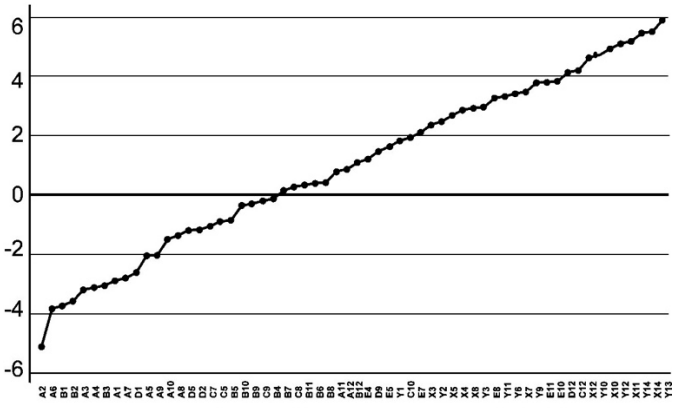
Putting the two cases together, it is obvious that, if a researcher employs Test 1 to assess the impact of a training (or other) programme, the relative gains of the low ability group are huge while those of the high ability group are trivial. On the other hand, if the researcher employs Test 2, exactly the opposite findings emerge.

The general, and vitally important, conclusion which emerges from these examples is that the apparent magnitude of any real increase in latent ability arising from a developmental experience, accident, or natural change over time depends (a) the general difficulty level of the test relative to the ability tested and (b) the distribution of the item parameters relative to the interval on the latent trait where change occurs. This makes it virtually impossible, without employing techniques which are described in the *2004 update* of the Y2K *Standard Progressive Matrices* section of the Raven Court & Raven Manual and in Prieler and Raven[xxxvii] , to make any meaningful statement about the *relative* magnitude of gains or losses of high, medium, and low ability groups. It also renders the use of most difference scores ("Learning Potential", "Sensitivity to Stress") untenable.

Except …. that there remains the intriguing possibility of using a test with a *linear* Test Characteristic Curve.

The possibility of developing such a test – which is to say a test with an equal interval scale – has widely been regarded as a pipe dream in psychology. Nevertheless, the procedures employed in the development of the SPM+[xxxviii] resulted in the test having the graph of item difficulties shown in figure 8.

Figure 8 Standard Progressive Matrices *Plus* Item Difficulties (in Logits)

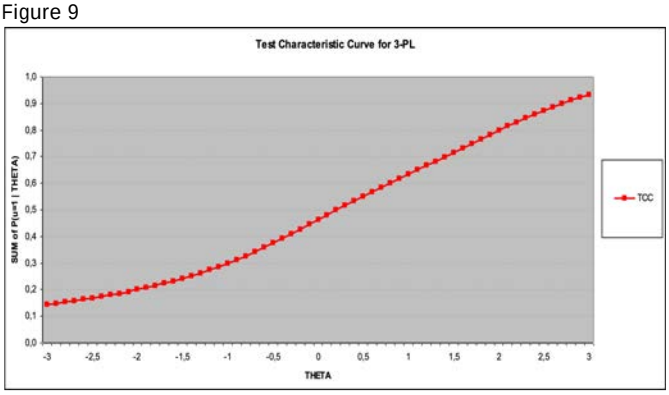


Final 60-item test from 1996 item-equating study, arranged in order of difficulty.
(Reproduced from: Raven, J., Raven, J. C., & Court, J. H., 2000.)

A cross validation of the test's properties in the course of a Romanian standardisation yielded the following Test Characteristic Curve[xxxix] (Figure 9).

In short, the SPM+, almost uniquely among psychological tests, offers its users something approaching an interval scale. The question now is whether the *Raven's 2* does the same.

Figure 9



According to Table A1 in the *Raven's 2* Manual, there are large increases in the IRT ability indices needed to achieve a one point raw score gain in the tails of the distribution yet only a small increase in ability is needed to achieve a one raw score gain in the middle of the scale. So, no, the *Raven's 2* test does not yield this property and so is not suitable for making easily interpretable statements about the relative effects of interventions or accidents.

More on discrimination in the tails of the distribution: The test information function.

Most practical *uses* of psychological tests are in the tails of the distributions … where their ability to discriminate is poor[xl]. In other words, most tests provide poor information in the domains in which it is most needed.

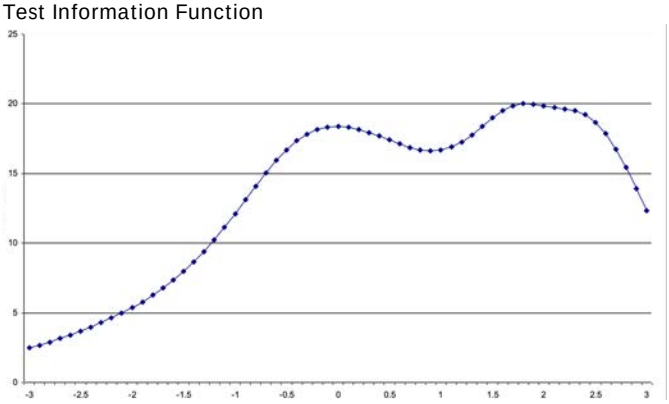
Worse, as mentioned earlier, those who use tests to make discriminations in these areas get tangled up in an impenetrable morass of legislation drawn up by a variety of administrators and legislators with varying backgrounds and objectives. They have varying levels of familiarity with the problems associated with cut off points especially in relation to eligibility for, or compulsory consignment to, "services". Small score differences in the tests used or scores in the area around cut off points (specified in different ways) can have major implications.

The huge jumps in differences in ability required to secure a one raw point gain at both ends of the *Raven's 2* scale have already been noted. As discussed by Hambleton et al (1991) the ideal shape of this graph would be rectilinear; one would get as much information from scores at all points in the distribution.

Ironically, as can be seen from the graph of the test information function for the SPM+ (figure 10), the SPM+ test goes some way toward correcting this.

Unfortunately, a test having a more appropriate overall Test Information Function would not necessarily yield such a function *within* e.g. age groups …. which is where such discriminations are typically made.

Figure 10 Standard Progressive Matrices Plus Romanian Standardisation



Conclusion

As we have seen, the *Raven Progressive Matrices* tests have, over the years, contributed enormously to the accumulation of research insights and played a huge role in the administration of education systems, military systems, organisations, and societies. What are the chances of the *Raven's 2* test contributing in that way? Who is to know? But one thing is certain. *Raven's 2* scores are not convertible to, and thus integratable with, Classic RPM scores. In contrast, not only are the *Standard Progressive Matrices* ***Plus*** scores directly convertible to scores on the classic tests, many of the items directly parallel items in those tests.

Given that the *Raven's 2* test does not possess many of the desirable properties evinced by the *SPM+* one cannot help wondering why the staff at The Psychological Corporation/Harcourt/Pearson have chosen to go this route.

John Raven

References

Flynn, J. R. (1984). IQ gains and the Binet decrements. *Journal of Educational Measurement, 21*, 283-290.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*, 29-51

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*, 171-191.

Flynn, J. R. (1989). Raven's and measuring intelligence: The tests cannot save themselves. *Psychological Test Bulletin, 2(2)*, November, 58-61. Hawthorn, Victoria: ACER.

Flynn, J. R. (1999). Searching for justice: The discovery of IQ gains over time. *American Psychologist, 54(1)*, 5-20.

Flynn, J. R. (2000). *How to Defend Humane Ideals*. Nebraska: University of Nebraska Press

Flynn, J. (2008). Excerpts from how to defend humane ideals. In J. Raven & J. Raven (Eds.) *Uses and Abuses of Intelligence* (see below) Chapter 25, pp. 556-567). http://eyeonsociety.co.uk/resources/UAIChapter25.pdf

Garfinkel, R., & Thorndike, R. L. (1976). Binet item difficulty: Then and now. *Child Development, 47*, 959-965.

Hambleton, R. K. (1989). Constructing tests with item response models: A discussion of methods and two problems. *Bulletin of the International Test Commission, No.28/29*, 96-106. Strasbourg: ITC

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Press.

Horn, J. L. (1994). Theory of fluid and crystallized intelligence. In R. J. Sternberg (Ed.), *Encyclopedia of Human Intelligence*, pp. 443-451. New York: Macmillan.

McKinzey, R. K. (2008). Too dumb to die: Mental retardation meets the death penalty. In J. Raven & J. Raven (Eds.) *(See Below)* Chapter 24, pp. 543-555. http://eyeonsociety.co.uk/resources/UAIChapter24.pdf

Prieler, J. & Raven, J. (2008). Problems in the measurement of change (with particular reference to individual change [gain] scores) and their potential solution using IRT. In J. Raven & J. Raven (Eds.) *Uses and Abuses of Intelligence* (see below) Chapter 7, pp. 173-210. Also available at: http://eyeonsociety.co.uk/resources/UAIChapter7.pdf also https://www.researchgate.net/publication/340900090_Problems_in_the_Measurement_of_Change_with_Particular_Reference_to_Individual_Change_Gain_Scores_and_their_Potential_Solution_Using_IRT

Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology, 41*, 1-48. But see also https://www.researchgate.net/publication/255565943_Change_and_Stability_in_RPM_Scores_Over_Culture_and_Time_The_Story_at_the_Turn_of_the_Century

Raven, J. (2008). General introduction and overview: The Raven *Progressive Matrices* Tests: Their theoretical basis and measurement model. In J. Raven & J. Raven (Eds.) *Uses and Abuses of Intelligence* (see below) Chapter 1, pp. 17-68. Also available at http://eyeonsociety.co.uk/resources/UAIChapter1.pdf also at https://www.researchgate.net/publication/255605513_The_Raven_Progressive_Matrices_Tests_Their_Theoretical_Basis_and_Measurement_Model

Raven, J. (2008). Intelligence, engineered invisibility, and the destruction of life on earth. In J. Raven & J. Raven (Eds.) *Uses and Abuses of Intelligence* (see below) Chapter 19, pp. 431-471. Also available at http://www.eyeonsociety.co.uk/resources/UAIChapter19.pdf also at https://www.researchgate.net/publication/350661300_INTELLIGENCE_ENGINEERED_INVISIBILITY_AND_THE_DESTRUCTION_OF_LIFE_ON_EARTH

Raven, J. (2019). The pervasive and pernicious effects of neglecting systems thinking (especially when combined with a disposition toward fascism). http://eyeonsociety.co.uk/resources/Unwillingness-to-engage-in-systems-thinking.pdf also at https://www.researchgate.net/publication/339325826_The_Pervasive_and_Pernicious_Effects_of_Neglecting_Systems_Thinking_especially_when_combined_with_a_disposition_toward_fascism

Raven, J. (2020). 'Closing the gap': Problems with its philosophy and research – A keynote address prepared for BPS Education Section Conference, September 2019 *The Psychology of Education Review, Vol. 44, No. 3, Special Issue, 2020 pages 2 – 40.* ISSN: 1463-9807. https://shop.bps.org.uk/the-psychology-of-education-review-vol-44-no-3-special-issue-2020-0 also available at: http://eyeonsociety.co.uk/resources/CAT-2376.pdf also at https://www.researchgate.net/publication/337925783_Fundamental_problems_in_and_with_policy-relevant_research_illustrated_from_research_relating_to_Closing_the_Gap

Raven, J., Prieler, J. & Benesch, M. (2008). Using the Romanian data to replicate the IRT-based Item Analysis of the SPM+: Striking achievements, pitfalls, and lessons. In J. Raven & J. Raven (Eds.) *Uses and Abuses of Intelligence.* (see below) Chapter 5, pp. 127-159. Also available at: http://eyeonsociety.co.uk/resources/UAIChapter5.pdf also at https://www.researchgate.net/publication/340899446_Using_the_Romanian_Data_to_Replicate_the_IRT-_Based_Item_Analysis_of_the_SPM_Striking_Achievements_Pitfalls_and_Lessons_Chapter_5_in_Raven_J_Raven_J_ed_Uses_and_abuses_of_intelligence_published_by_Roy

Raven, J., & Raven, J. (Eds.). (2008). *Uses and Abuses of Intelligence: Studies Advancing Spearman and Raven's Quest for Non-Arbitrary Metrics.* Unionville, New York: Royal Fireworks Press; Edinburgh, Scotland: Competency Motivation Project; Budapest, Hungary: EDGE 2000; Cluj Napoca, Romania: Romanian Psychological Testing Services SRL. Also available at: http://eyeonsociety.co.uk/resources/Uses-and-Abuses-of-Intelligence.pdf and http://eyeonsociety.co.uk/resources/fulllist.html#uses_and_abuses

Raven, J., Raven, J. C., & Court, J. H. (1998, revised and up-dated 2003). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 1: General Overview*. San Antonio, TX: Harcourt Assessment.

Raven, J., Raven, J. C., & Court, J. H. (2000, updated 2004). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3: The Standard Progressive Matrices, Including the Parallel and Plus Versions.* San Antonio, TX: Harcourt Assessment.

Raven, J., Raven, J. C., & Court, J. H. (2000, updated 2004). *Manual for Raven's Progressive Matrices and Vocabulary Scales.* Sections 1 to 7 with three Research Appendices. San Antonio, TX: Harcourt Assessment.

Raven, J., Rust, J., & Squire, A. (2008). *Manual: Coloured Progressive Matrices and Crichton Vocabulary Scale.* London, England: NCS Pearson, Inc.

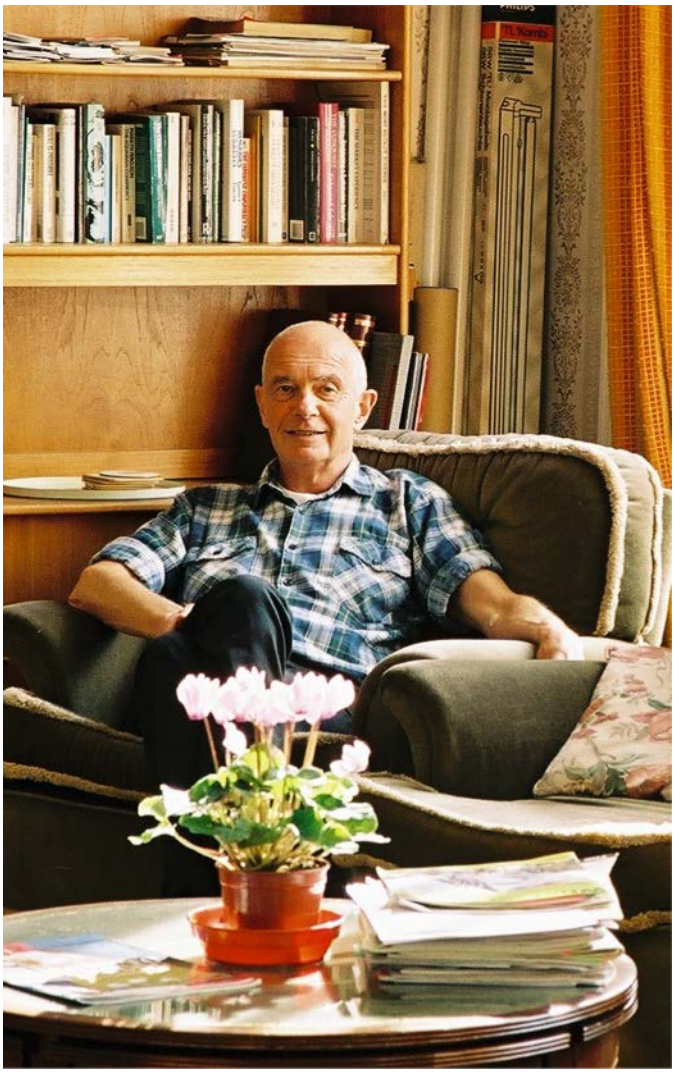
Raven, J., Rust, J., & Squire, A. (2008). *Manual: Standard Progressive Matrices - Plus Version - and Mill Hill Vocabulary Scale.* London, England: NCS Pearson, Inc.

Spearman, C. (1927). *The Abilities of Man: Their Nature and Measurement*. London, England: MacMillan.

Taylor, N. (2008). Raven's *Standard and Advanced Progressive Matrices* among adults in South Africa. In J. Raven & J. Raven (Eds.) *Uses and Abuses of Intelligence* (see above) Chapter 15, pp. 371-391. Also available at: http://eyeonsociety.co.uk/resources/UAIChapter15.pdf

Thorndike, R. L. (1975). *Mr. Binet's Test 70 Years Later*. Presidential Address to the American Educational Research Association

Wicherts, J. (2007) *Group differences in intelligence test performance*. University of Amsterdam.https://www.researchgate.net/publication/34419948_Group_differences_in_intelligence_test_performance



Your Newsletter Needs You!

We are always looking for new articles and ideas for Testing International.

This is our forum for professional discussions, information-sharing and opinions. So if you have anything you'd like to discuss, report or just raise as food for thought, we'd be delighted to hear from you!

Please send your contribution or idea to:

newsletter@intestcom.org

Many thanks!

Nicky Hayes, Editor

Endnotes

[i] Ironically, these tests and associated Manuals (generated by the author and colleagues) were, and are, published by Pearson US (see Raven, J., Raven, J. C., & Court, J. H. (2000, updated 2004) [Section 3] and the ***SPM+*** had been standardised by a team at Pearson (UK) and published in the UK (Raven, Rust & Squire, 2008).

[ii] These needs revolve around such things as the lack of security of the tests arising from such things as the availability of copies on the internet and the problems posed by the intergenerational increase in scores that has become known as the "Flynn effect". As discussed below, the latter generates misleading information when old norms are used as reference data against which to view the scores of people tested more recently.

[iii] Some examples of these abuses are discussed in Raven (2008) *Intelligence, engineered invisibility, and the destruction of life on earth.* (See *References* section for full citation.)

[iv] We are not told which application of this model was applied. As will be seen from the discussion in endnote x, item-equivalence according to a 3-parameter model including both difficulty and shape of item characteristic curve would be extremely demanding.

[v] There are, in fact, multiple versions of the tests now known as *Classic* Standard, Coloured, and Advanced versions, *Parallel* versions of the Coloured and Standard SPM, and the *Standard Progressive Matrices* ***Plus*** version.

[vi] It seems that more able respondents characteristically mentally "solve" the problem before looking at the options while less able are more likely to resort to looking at the options in their quest for a solution.

[vii] See Raven, J. (2008). *General introduction and overview* or the *General* section of the Manual for more details.

[viii] Spearman (1927)

[ix] Spearman distinguished the two as follows "*To understand the respective natures of eduction and reproduction – in their trenchant contrast, in their ubiquitous co-operation and in their genetic inter-linkage – to do this would appear to be for the psychology of individual abilities the very beginning of wisdom*".

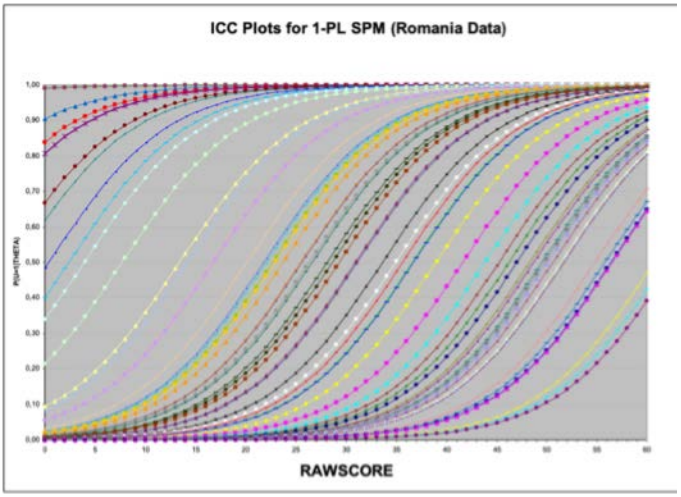
[x] Although somewhat out of place in this *commentary,* it might be of interest to introduce the evidence needed to justify this statement using material drawn from Raven, Prieler and Benesch (2008).

The graphs in the Figures below are the Item Characteristic Curves for all items in the SPM+.

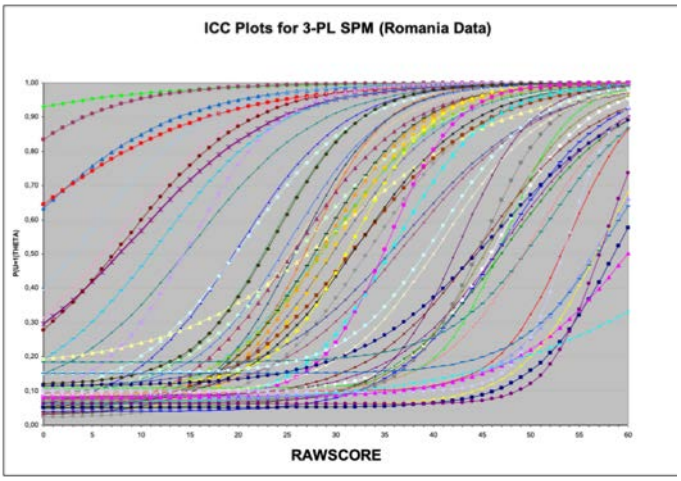
Each individual graph plots the proportion of respondents with each total score who get the item right. Thus we see that while many low ability people fail to get the easier items right, 100% of more able people do so. On the other hand, while most low ability people fail to get the most

difficult items right (those that do so do so as a result of randomly selecting the correct answer from one of the options available) more of the more able do so.

The program used to generate the graphs plotted in the first figure – known as a 1 parameter IRT model – has smoothed the raw data rather heavily. The 3-parameter plot shown in the next Figure creates a more realistic impression.



Although few of those who use the off-the-shelf statistical packages appear to understand it, the mathematical indices generated by these packages indicate how closely the set of items in a test conform to an ideal 1-parameter IRT model. The graphs derived from a 3 parameter plot show more deviance from this ideal.



What these Figures show is that, at least to a considerable extent, the abilities required to solve each more difficult item build on, and extend, those required to solve the easier items. No "new" abilities are required and there are no transformations or "metamorphoses" in the abilities required to solve the more difficult items. What is more, the exact same abilities are required to solve the easiest items. The psychological processes required to generate "mere perception" are the same as those required to engage in complex systemic thinking.

[xi] Many, to the amazement of most researchers, continue simply to see even the most difficult items as patterns immediately implying the figure required to complete them.

[xii] The failure to engage in systems (systemic) thinking lies at the heart, not only of many inappropriate policy decisions, but also at the heart of many of the erroneous conclusions drawn from "scientific" studies Raven (2019, 2020).

[xiii] Interestingly, this increase in some psychological test scores parallels a similar increase in other biological characteristics like height and athletic ability. Thus the real puzzle for psychologists is, not the "Flynn effect" but why other psychological test scores have *not* been increasing.

[xiv] The significance of the cross-cultural data became even more apparent when it came to generating an *explanation* of the "Flynn effect". It emerged that, at any point in time, the norms for countries with a tradition of literacy were remarkably similar – this despite the fact that they had very different income levels, diets, and access to education and television. It followed that the variables usually put forward to explain the increase over time actually had very little demonstrable effect.

[xv] Thorndike (1975)

[xvi] Flynn (1984)

[xvii] Raven, J., Raven, J. C., & Court, J. H. (2000, updated 2004). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3: The SPM.*

[xviii] i.e. studies in which a cross section of age groups had been tested at a single time point.

[xix] Horn (1994).

[xx] Note that, except via the inclusion of "marker variables" they do not emerge as separate *factors* in factor analyses.

[xxi] J.C.Raven attached considerable importance to *discrepancies* between estimates of ability derived from *Progressive Matrices* (eductive ability) and *Vocabulary* (reproductive ability) test scores. Given that the *Raven's 2* is not accompanied by a parallel measure of reproductive ability, it will not be possible for users to generate such diagnostic information.

[xxii] Flynn (1989 and 1999) found himself ensnared in this mess when he found himself asking whether the increase he had documented "really" represented an increases in "intelligence".

[xxiii] Raven, J., Raven, J. C., & Court, J. H. (2000, updated 2004). *Manual for Raven's Progressive Matrices and Vocabulary Scales.* Sections 1 to 7.

[xxiv] McKinzey, R. K. (2008).

[xxv] Serious disputes often arise around small differences in scores around legislatively prescribed cut off points. As the authors note when discussing the concept of "age equivalents", a small difference in raw score can make a big difference to the mental age assigned. Reversing this observation, a few months difference in age can make a big difference to the standard score assigned and thus to the educational programme to which a child is assigned. This is because there are big differences in item difficulties, particularly in the tails of the distribution. The test is simply not suitable for mechanical application in such situations. Interestingly, in discussing Case Study No 3 the authors find it necessary to recommend broadening the range of assessment instruments used.

[xxvi] Interestingly enough even this is never made explicit. It is only implied by an introductory quote, which now appears in endnote ix but which initially appeared simply as a stand-alone paragraph at the front of J.C.Raven's *Guide* to the use of the Progressive Matrices.

[xxvii] Raven (2020).

[xxviii] See Raven, J. (2008). *Intelligence, engineered invisibility* … (Raven & Raven 2008 chapter 19) for a fuller discussion.

[xxix] The ethics of this process has been discussed by Flynn in Flynn (2000) … which is summarised in a chapter in our *Uses and Abuses of Intelligence* (Flynn, 2008).

[xxx] Some of these are reported in various sections of the main Raven, Court and Raven Manual and more are available in in Raven & Raven (ed) 2008.

[xxxi] This is more than a little odd since this section was first superseded by a computerised version and then by the digitisation of the entire Raven archive by The Psychological Corporation, San Antonio.

[xxxii] Taylor, N. (2008).

[xxxiii] Prospective miners still come from very different tribes who speak different languages and often do not understand each other, never mind English.

[xxxiv] The Hispanic/White difference has probably declined since then, but this does not affect the point being made.

[xxxv] John Rust has drawn my attention to the way in which non-verbal algorithms (Artificial Intelligence [nb use of the term "intelligence"]) are being used to cyclically evolve discriminations (unverbalised constructs) which are then recursively accorded differential treatment on an undiscoverable basis to enhance those discriminations and without regard to their wider consequences.

[xxxvi] An important discussion of this and several other issues raised in this review has been provided by Wicherts (2007).

[xxxvii] Prieler and Raven (2008).

[xxxviii] See Raven, J., Raven, J. C., & Court, J. H. (2000, updated 2004). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3*.

[xxxix] Raven, Prieler, and Benasch (2008). On the face of it, this provides a solution to our problem. Unfortunately, there can be no guarantee that, just because the overall distribution is as illustrated, the distributions for different ability groups will be similar. I leave the task of checking on this to others.

[xl] Raven, Prieler and Benasch (2008) but see important discussion of Test Information Function in Hambleton (1991).