Personality and Individual Differences 103 (2016) 128–134

Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid

Electrophysiology, chronometrics, and cross-cultural psychometrics at the Biosignal Lab: Why it began, what we learned, and why it ended

Paul Barrett *

Cognadev Ltd., New Zealand
University of Auckland, New Zealand

ARTICLE INFO

Article history:
Received 17 November 2015
Accepted 6 April 2016

Keywords:
Hans Eysenck
Intelligence
Chronometrics
Evoked potentials
Personality
Measurement

ABSTRACT

This article is about why and how Hans Eysenck began investigating the brain evoked potential and sensory nerve conduction correlates of psychometric IQ, his investigations into the timed/speeded performance (chronometrics) results being published by Art Jensen, and the initial results being reported by Doug Vickers and Ted Nettelbeck on the relation of Inspection Time to psychometric IQ. In the midst of this experimental work, Hans and Sybil Eysenck were also engaged in the final stages of their multi-country investigation into the universals of human personality and temperament, using the Eysenck Personality Questionnaire (EPQ). Although many regarded both the experimental and cross-cultural work as 'mildly amusing but not his best work', we did actually learn a great deal from these efforts. (1) None of the previously reported evoked potential parameter correlations with psychometric IQ replicated with any substantive consistency. Neither did the sensory nerve conduction parameters. (2) The expected relationship between Reaction Time parameters and psychometric IQ was replicable, but theoretically and scientifically deficient. Why? Because we observed many individuals with low IQ possessing reaction times faster than high IQ individuals, even when retested on a second occasion. (3) The expected relationship between Inspection Time parameters and psychometric IQ was replicable, but again, theoretically and scientifically deficient. Too many cases were observed showing the opposite of the expected effect (i.e. low IQ individuals possessed much shorter Inspection Times than high IQ individuals). (4) The four personality constructs assessed with the EPQ showed 'good enough' replicability within datasets acquired across 35 different countries, although the Psychoticism scale was the weakest in terms of factor recovery.

1. Introduction

My very personal 'what exactly was it like working with Hans Eysenck' article was published in 2001, in a special issue of *Personality and Individuals Differences* devoted to publishing Hans's bibliography, his impact among those who knew him, and his contribution to psychology. This article is just as informal, but now explaining *why* he devoted nearly 11 years of his life to investigating the biological and later chronometric correlates of the psychometric assessments of abilities/IQ and why/how it ended.

When I first began work with Hans back in 1983. I was employed to:

1) Help Sybil (and Hans) continue their ongoing work examining the cross-cultural comparability of their Eysenck Personality Questionnaire (EPQ).
2) Enable Hans to investigate, replicate, and perhaps extend the Reaction Time parameter x IQ results published by Lally and Nettelbeck (1977) and Jensen (1982) with correlations varying between −.40 to as high as −.74.
3) Replicate and extend the Inspection Time results which were showing correlations of up to −.92 with IQ (Lally & Nettelbeck, 1977), and −.78 (Brand, 1981).
4) Replicate Alan and Elaine Hendrickson's work showing that IQ test scores were correlated up to −.83 with parameters computed from auditory evoked potentials.

Given the magnitudes of effect sizes in points #2–4, it is pretty obvious why Hans wanted to explore the phenomena further. Furthermore, each possessed rudimentary theory as to what might be causal for the phenomenon, and Hans always valued theory. So, the stage was set for a multiple-approach experimental push on each phenomenon, with Sybil continuing with her primary responsibility for the cross-cultural work.

2. The cross-cultural psychometrics project

As described more comprehensively in S.B.G. Eysenck (1983), the purpose of this project was to establish the universality of the scales comprising the EPQ, confirm item loadings on each scale-factor, generate a suitable score-key for the scales in that country, and form scale score comparisons between a UK reference sample and the country

* 19 Carlton Road, Pukekohe, Auckland 2120, New Zealand.
E-mail address: paul@cognadev.com.

http://dx.doi.org/10.1016/j.paid.2016.04.038
0191-8869/

data. This was achieved by back translations of items, acquiring samples of approximately 500 males and females in a particular country sample, factor analysing the data, computing configural factor comparisons, and generating score keys, with the 'in-common with the UK' items scored and compared with the scores generated from the same items in the UK reference sample data. So, it required a fair amount of sophisticated statistical and computing work; overseen originally by Owen White with myself stepping into that role as Owen gradually retired from his methodologist position. This work was ongoing throughout 1983–1998, culminating in a summary article (Barrett, Petrides, Eysenck, & Eysenck, 1998), and finally the release of all data to the wider community by S.B.G. Eysenck and Barrett (2013).

3. Reaction time (RT) and intelligence

This was the initial area in which Hans began his work into investigating the relationships between simple, choice, and complex reaction time (the Odd-Man-Out task) and psychometric IQ. We contracted an electronics company to build us the same stimulus–response boxes as Jensen was using/recommending; enabling 0–3-bit RT acquisition, a separation of motor and decision-times, and a more complex "Odd-Man-Out" task invented by Hans and a PhD student, Warwick Frearson (Frearson & Eysenck, 1986). I wrote the software programmes which controlled the experiments and boxes via a linked computer. In addition, Hans was also working on a cross-cultural project, which required myself and the electronics company to design and fabricate an entirely self-contained computer system (incorporating screen, custom keyboard, computer, and custom-automation software) to which a 'Jensen' box could be attached, and a complete sequence of RT experiments undertaken by an investigator, just by making a few menu selections on-screen.

These 'experiments in a box' were literally shipped out in crates to an investigator in another country; the investigator would test a few hundred children, send back the data discs, and ship the boxes back to us when the experiment was complete. Given no cost-effective or practically-useful portable laptop computers were available at the time, we had to source the smallest desktop available and incorporate the screen, a new limited-key keyboard, main processor unit with disc drives, Jensen box interface unit, and smoothed/externally-fused power supply filters, all into a monolithic experiment unit which could be transported from site to site by the local investigator. Looking back now on what Hans set out to do and how we made it happen, I do not think he or I realised just how 'adventurous' was this project. The shipping, customs-paperwork, and transport arrangements also required a new set of skills to be learned by the research assistant working for Hans! But, these boxes actually 'did the business' as designed; it was actually very exciting at the time, especially when the data discs would begin arriving back at the Lab. I cannot remember why, but at some point during this exercise, we handed all the data and boxes over to Richard Lynn, who continued with the experiment and eventually wrote up all the results in several publications!

Probably the landmark publication of our own reaction time work was the article reporting two major studies in the Lab (Barrett, Eysenck, & Lucking, 1986). This publication showed that contrary to expectations, not everyone fitted Hick's Law[1] (Hans included!) — and this was a replicable phenomenon. Furthermore, we did not attain the large correlations reported by others, although still observing the negative correlations between reaction times, their variability, and IQ test scores (between about −.25 and −.45, increasing in negative magnitude with increasing information content of the stimulus). So, this area was looking promising and we began to expand the scope of experiments incorporating the new 'Inspection Time' procedure alongside reaction time acquisition.

4. Inspection time (IT) and intelligence

IT is simply an assessment of the time required by an observer to accurately discriminate between two stimuli. In most operationalisations of the task, the two stimuli were Light Emitting Diode bars formed from multiple segments placed side by side, which could be illuminated in such a way that more segments on one bar would be illuminated compared to the other, so providing a stimulus which varied the length of illuminated bar on the left or right-hand-side. The LED bar display could be energised in three standard ways, showing a longer red bar on the right side, the left side, and a mask that energised all LED bars that were not illuminated as part of the stimulus. So the visual impression is of a longer (or shorter bar) then both bars extended to the same length. The assessment protocol varies the display times, finding the minimum time the lines must be displayed in order for an observer to discriminate the length-difference reliably; this time is defined as their Inspection Time. Typical experiment instructions were:

> "This task will be measuring how little time you need in order to accurately discriminate between one short and one long bar of light. The long bar will be randomly varied between the left and right positions. Whichever side you see the long bar on, press the button which matches that position (left or right). This is not a reaction time task-you have as much time as you like in which to respond. You can make your response whenever you like."

Again, in conjunction with our electronics company, we built our own Inspection Time stimulus presentation apparatus, incorporating stimulus display and timing, which was connected to a computer. I again wrote the software which conducted the IT assessment; the algorithm and a summary of our experiment results were summarised in Barrett, Petrides and Eysenck (1998). As usual, what we found was that those huge initial correlations being reported were virtually double or treble what we could observe in many experiments. A result similar to that reported in a meta-analysis of Inspection experiment results conducted by Kranzler and Jensen (1989), and finally in a huge meta-analysis of over 50,000 cases of data (Sheppard & Vernon, 2008).

When one looked very closely at those initial experiments reporting the very large correlations, it was clear that the very wide range of IQs coupled with low samples sizes were contributing to the magnitude of observed correlations. For example in one study (Nettelbeck & Lally, 1976) the 10 individuals comprising the sample possessed IQs varying from 47 to 119. And, many of these early studies possessed sample sizes rarely exceeding 20 individuals (Table 1, Brand & Deary, 1982). As Irwin (1984, p. 63) noted, others had also commented on the manner in which IT had been assessed (or in some cases extrapolated). From his own experimental work in this area, he concluded:

> "Published high correlations between inspection time and intelligence should be viewed with skepticism because they are often based on small samples of excessively large ranges of intelligence. Many of the published values of inspection time that provide the basis for these correlations must be unreliable because they involve large extrapolations or interpolations, and because they are derived from a flat portion of the psychometric function where small changes in the percentage correct are accompanied by large changes in stimulus duration. The evidence of this study, which attempted to avoid these defects, suggests a small negative relation between inspection time and intelligence."

[1] The time taken to make a response is a linear function of the number of bits of information in a stimulus presentation. An "Information bit" is the base 2 logarithm of the number of elements within a stimulus presentation. For example, when only element can change, that corresponds to 0-bit information. When one of two elements can change, that corresponds to 1-bit of information requiring to be processed, When one of four elements can change, then 2-bits of information are required to be processed.

5. The electrophysiological correlates of IQ

During the late 1980s, Hans decided we would finally tackle the biological correlates of intelligence, attempting a careful exact replication of the initial Hendrickson studies. Hendrickson and Hendrickson (1980, 1982) had derived two measures which could be computed from an Averaged Evoked Potential (AEP) recorded using conventional scalp electrodes which acquired voltage changes emanating in several brain regions associated with the brain's response to an auditory stimulus. A complexity measure (otherwise known as the 'string measure') was assessed by computing the contour perimeter of the AEP waveform; the larger this value, the higher an individual's IQ. The second measure, the variance, was computed by taking the average variability of each sample point on an AEP over a number of epochs. The greater the variance, the lower an individual's IQ. Essentially they proposed that the neural transmission characteristics of high-IQ individuals was such that fewer propagation-transmission errors are made than was the case with low-IQ individuals. Consequently, within the AEP, high-IQ individuals will tend to have more complex AEPs; the individual component traces being less variable from trial-to-trial, thus preserving more of the detail of the single evoked potential response. In contrast, the low-IQ individuals would produce a more varied evoked response from trial-to-trial, yielding (when averaged) a smoother, less complex AEP. Thus, the high-IQ AEP trace should yield a longer string measure than that seen in a low-IQ individual's AEP, and the variability measure should yield a lower value than that observed for low-IQ individuals' AEPs.

The empirical evidence regarding the complexity and/or variance of the AEP was drawn from two studies. The first, reported by Blinkhorn and Hendrickson (1982), correlated the complexity (string) measure with performance on Raven's Advanced Progressive Matrices and a variety of verbal ability tests. The correlation, after corrections for unreliability and restriction of range was .84 ($n = 33$). This value was reasonably close to one obtained by Hendrickson and Hendrickson (1980) in an analysis of some previously published data of Ertl and Schafer's (1969) for which they obtained a correlation of .77 between WISC-IQs and the string scores from the AEPs. In the second study (Hendrickson & Hendrickson, 1982), a sample of 219 schoolchildren (121 boys, 98 girls) completed the WAIS to assess IQ, comprising Performance, Verbal, and Full-Scale IQ. In addition to the complexity and variance measures defined before, D.E. Hendrickson defined a new composite measure: the variance score minus the string score. The stimulus presentation, data acquisition, epoch length, EEG derivation, and montage were the same as those reported in the first study, effectively the paradigm methodology. The correlations among the WAIS-IQ and string, variance, and composite AEP measures were .72, −.72, and −.83, respectively.

In order to replicate exactly that which the Hendricksons claimed was essential to their paradigm and observing its results, we needed state-of-the-art recoding equipment that would allow complete flexibility for programming, graphical display, digital signal analysis, digital filter design and construction, and high simultaneous sampling speeds across 32 channels for sustained-duration epochs (sometimes up to 15 min). We also needed recording equipment accurate to ±1 microvolt within a low electromagnetic-noise (shielded) environment, and signal generation/monitoring equipment of a high specification. These days it is trivial to buy such systems that will fit on a desktop. But, back in 1988/9, there were only basic equipment setups for the more routine psychophysiology experiments. So, Hans obtained a huge amount of money from the tobacco company, Philip Morris International,[2] which enabled us to purchase what was then a state-of-the-art signal-processing minicomputer (the Masscomp) that ran Unix as its operating system, a decent Fortran compiler for analysis work, and enough signal acquisition and timing hardware to run the kind of experiments we had in mind. It also came with substantive storage and a handy set of subroutines which allowed me to programme computer-controlled experiments which would also link with other stimulus devices. And we could run the cross-cultural psychometrics off the same machine. But, the heat output of this behemoth was huge. So, Hans eventually moved the entire Lab down to the Bethlem Royal Hospital into a suite of rooms that enabled us to have a dedicated machine room with its own air-conditioning (the Masscomp was permanently "on" – and modem-enabled so that I could access it and its data/compilers from home), several testing rooms where we could undertake chronometrics and group psychometrics assessments, a large signal monitoring/experiment control room, a dedicated 'biological signal acquisition' room in which experiment participants would be connected up and take part in the experiments, and an office for Hans and Sybil, as well as for the Lab team and myself.

Hans had style. There was no hand-wringing, penny-pinching, or consideration of half-measures. He wanted to do powerful scientific research on something which was potentially field-changing if it held up. And he found the money to get the job done.

So, off to work we went. Our first major replication attempt was published in great detail in Barrett and Eysenck (1992). But during this time we also carried out two more attempts and extended replications, incorporating both reaction time and inspection time as ancillary tasks, and looking at the cross-relationships between the various tasks. Barrett and Eysenck (1994) summarised all our results, and their implications. I have quoted the abstract from this article below because it gives you the flavour of where we finished up after such intense research:

> "Two-hundred subjects provided data from within three separate studies that attempted to replicate correlations between averaged evoked potential (AEP) indices and psychometric IQ. In addition, AEP zero-cross analysis was undertaken as a specific test of a proposition made within the Weiss quantum theory of intelligence. Measures of AEP variability, mean individual epoch amplitude, and P180 component latencies were found to correlate negatively with IQ at around −0.50 across the three studies. However, the consistency and size of relationship in the results was found to be a function of selecting subjects whose AEP P180 component amplitude was greater than some specified, sample dependent, target value. The zero-cross analysis, contrary to predictions, yielded no correlations with IQ. Robinson's cerebral arousability theory was noted as a possible explanatory framework for the results. In addition, it was noted that if Robinson is correct in his assertion of the complex analogue nature of the evoked response, conventional AEP analysis is no longer relevant."

Our work resulted in two main conclusions:

1. The Hendrickson work was essentially non-replicable. A presentation downloadable from http://www.pbarrett.net/presentations/string.pdf showcases 10 positive replications, and 10 negative replications. Equally non-replicable were the neural efficiency hypothesis-parameters and their relation to psychometric IQ from Ertl and Schafer (1969), the proposed continuous EEG spectral frequency/coherence relationships, and basically almost every derived parameter proposed by a raft of investigators over the preceding years. How is it that we failed to replicate virtually every finding published in this area? My opinion is that many researchers were carefully selecting the cases they used to form their final analysis samples; a questionable research practise now known to be more common than originally thought among many social science investigators (Fanelli, 2009; John, Loewenstein, & Prelec, 2012). This may not have been deliberate but simply the result of not applying objectively implemented criteria for data exclusions, resulting in 'expectancy-bias' creeping into their more

[2] Philip Morris were interested in the potential relationships between cigarette smoking behaviours, preferences, personality, and brain activity. We committed to looking at the relationship between these variables (I must admit with little real interest, but that is what they wanted to see!). However, the really important research for us was that associated with replicating the results concerning psychometric IQ and biological/chronometric correlates.

personal, 'visually-guided' selections. As noted in our own publications, we always reported the computer-implemented algorithms for data exclusions, and never relied upon visual inspection of records as a basis for retention or rejection. We also carefully reported the conditions under which we could seem to attain a replication, and destroy it, solely on the basis of record-selection or other computational-algorithm criteria.

2. David Robinson's stellar work on the constituent continuous synchronous waveform components of the AEP was quite possibly correct (Robinson, 1999, 2000). With carefully designed digital filters on the Masscomp, I was able to show that entire sequences of a hundred evoked potentials (constituting the epoch records to be averaged) could, indeed, be separated out as comprising three discrete synchronous continuous waveforms, perturbed by the auditory stimulus but returning to their basal synchronous frequency. But I was never able to finish these computational investigations as I was carrying out this work within days of the Lab closing and myself entering unemployment. To this day, I cannot understand the intellectual indifference of so many who dismissed David Robinson's theory and work in this area. My own impression when I spoke with his detractors was that they were technically naïve when it came to matters of digital and analogue signal engineering, large-scale custom data-processing, and the kind of experimental design work required to investigate his claims and empirical phenomena. Their indifference merely reflected their inability to undertake replication rather than any coherent line of argument.

But before you think this was all we investigated, I must mention the work we carried out as an extension to this EEG work, and alongside it, in two extra experiments. Hans was fascinated by the work coming from Philip Vernon, T. Edward Reed, and Arthur Jensen on peripheral and cerebral visual pathway nerve transmission time/nerve conduction velocity and its relation to psychometric IQ (Vernon & Mori, 1992; Reed & Jensen, 1991, 1992). I remember sitting down with Hans one day and speculating that perhaps whatever is part-causal for what we are observing within the brain is actually a feature of the entire nervous system. What if the 'noise' we attribute to differentiation of neural efficiency is a fundamental property of all nervous transmission systems, that is, even a sensory nerve like the median nerve at the wrist? So, we set out to measure nerve conduction velocity and variability at the wrist, with electrical pulse stimulation at various currents starting at 1 mA on the third finger. The entire experiment was automated, with stimulus delivery at exactly 34 °C using a computer controlled heat-lamp placed near the wrist/hand, and four skin-surface thermistor temperature sensors. The two experiments revealed no replicable correlation between nerve conduction velocity and IQ, but potentially a replicable correlation between conduction variability and IQ scores. However, given the multiple-fibre composition of the median nerve, with groups of fibres conducting at different speeds relative to the applied stimulus, this was always going to be a tricky experiment even when we fixed the impulse stimulus at 2 mA above each individual participant's personal sensation threshold. The experiment setup, results, and in-depth discussion can be found in Barrett and Eysenck (1993).

In late 1993, the Lab closed its doors; we had exhausted our research funds and the entire line of 'correlates' research that had sought to generate causal theory of replicable and substantive-effect phenomena. Hans had done what he set out to do: we had carefully examined the biological and chronometric correlates work and found the evidence bases far weaker than originally proclaimed, so weak that any coherent causal theory in this area was rendered untenable. Every potentially field-changing phenomenon with huge initial effect sizes was eventually shown to be no more than the usual small-to-moderate effects found in the area of individual differences research, with replication studies behaving as Francis (2013) would later explain why. There was quite literally nothing more to investigate given *how* we were conducting our investigations. That is, we had been constantly engaged in a process of 'phenomena-detection'; looking for phenomenal effects of such outstanding magnitude that coherent explanatory theory might be generated to explain their occurrence. The actual position we arrived at in 1993 was that there were no such phenomena for which any coherent explanatory theory might be generated.

As argued in seven slides of reaction/inspection time/nerve-conduction AEP experiment data drawn from the several studies undertaken at the Biosignal Lab, (http://www.pbarrett.net/presentations/chrono_bio_IQ.pdf), all that can be claimed is that for some individuals, speeded performance on timed tasks does seem to be associated with their magnitude of psychometric IQ. The same kind of 'caveat-laden' result as finding that about 20% of individuals do not fit Hicks Law. As Hans noted, it was only when we publicly presented this replicable fact, others in German laboratories who had been investigating Jensen's work told us they had observed the same phenomenon, but had not published their data thinking it was due to faulty experimentation or participants not properly attending to the experiment itself. And remember, Hans himself did not perform according to Hick's Law — and that was replicated several times.

Over a 10 year programme of research – not a few one-off studies but a real programme of research – Hans sought to generate causal theory and establish the replicable detectability of certain phenomena associated with human intelligence and psychometric IQ scores. That was an impressive feat alone; the funding required to run a lab, its equipment and staff, was not trivial. But, the more rigorous we were with our experimentation and controls, the more tenuous were the results we found. By the end of 9 years or so, we were both losing interest because we had discovered that the results we could generate were only observable by selecting out particular cases of data. We could do this via objective computer algorithms based upon particular exclusion-criteria, but there was no substantive theory-evident reasons why such criteria might be required. In the end it was clear that we were dealing with 'statistical' phenomenal occurrences with no explanation for those many cases who did not behave as expected.

For those who may doubt Hans Eysenck's integrity as a scientist, note that at no time did Hans (or myself) seek to engage in any questionable research practise that might have produced the kind of results that previously published findings had led us to expect. Hans let the data speak, and accepted that the optimism he had shown along with the speculative theories he and others had proposed in his 1982 *Model for Intelligence* landmark book, had effectively dissipated to a few general correlational relationships, with no causal theory as to why the relationships were only observable in the aggregate with small-to-moderate magnitudes (Ferguson, 2009). That is why he ceased research in this area; scientific integrity was paramount for him. There was literally no point in continuing with a programme of research which had revealed such clear, albeit disappointing results.

Ultimately, Hans was a scientist; he wanted to investigate causal theory, not just present statistical relationships among variables. For myself, I had stumbled across many features of the methodologies such as nerve conduction neurophysiology and measurement issues, the generation of adaptive test/stimulus-presentation algorithms which were driven by psychological rather than statistical/psychophysical concerns, and the work from David Robinson on the complex synchronous waveform componentry of an evoked brain potential. But, while each might have been worth continuing specialised study, they were an aside to the research programme. We had done our level-best to investigate what looked to be powerful phenomena, and after 9 years ended up with a few correlations and some broad, generic statements about phenomenal occurrences in some individuals but not others, with no more understanding of *why* those phenomena occurred than when we had begun our investigations.

At the same time, the cross-cultural psychometric project wound down, because so many countries had now been investigated. This was another 25 year programme of consistent, dedicated research. At the time and while it was ongoing, it looked routine. But when we look back on what Sybil and Hans set out to achieve, the discipline, rigour, and expertise from Owen White which was applied to kick-start

and sustain the programme, it is seriously impressive. I was just the man who helped shepherd it to its conclusion; the 'heavy lifting' was all Sybil, Hans, and Owen (and their invaluable research assistant, Jackie Marshall[3]).

It was in 1997, the year Hans passed away, that the reasons for our failure to discover causal explanatory theory for consistent, replicable, powerful phenomena were explained. Not by any individual difference psychologist, experimentalists, or methodologists, but by two philosophers of science and measurement: Joel Michell and Michael Maraun. In 1997/1998 they published their respective landmark articles: Michell on the definition of quantitative measurement within science, and Maraun on the meaning of making a claim about 'a measurement' of some attribute (Michell, 1997; Maraun, 1998). These two publications are what triggered the 1998 book by the late Paul Kline, on the 'New Psychometrics', where he explained why he also had stepped away from statistical test theory (Kline, 1998).

So what was it that Michell said, that caused such an intellectual upheaval in both myself and Paul Kline? Look at the abstract to Michell's, (1997) article:

> "It is argued that establishing quantitative science involves two research tasks: the scientific one of showing that the relevant attribute is quantitative; and the instrumental one of constructing procedures for numerically estimating magnitudes. In proposing quantitative theories and claiming to measure the attributes involved, psychologists are logically committed to both tasks. However, they have adopted their own, special, definition of measurement, one that deflects attention away from the scientific task. It is argued that this is not accidental. From Fechner onwards, the dominant tradition in quantitative psychology ignored this task. Stevens' definition rationalized this neglect. The widespread acceptance of this definition within psychology made this neglect systemic, with the consequence that the implications of contemporary research in measurement theory for undertaking the scientific task are not appreciated. It is argued further that when the ideological support structures of a science sustain serious blind spots like this, then that science is in the grip of some kind of thought disorder."

In this article, Michell set out the constituent properties of quantitative measurement, defining that which constitutes a 'quantity', and restating the axioms which define, classes, orders, and quantities. Let me set out the definitions of key-terms provided in Michell (1999):

Additivity: a relation between levels of a quantitative attribute. For any two distinct levels of a quantitative attribute, a third always exists such that the greater of the two is the sum of the third and the less.

Quantity: an attribute possessing ordinal and additive structure. For example, length is a quantity because lengths are ordered according to their magnitude, and each specific length is constituted additively of other specific lengths.

Magnitude: a specific level of a quantitative attribute (or quantity). For example, each specific length that an object might have is a magnitude of the attribute, length.

Measurement: the discovery or estimation of the ratio of a magnitude of a quantity to a unit of the same quantity.

Unit: a specific magnitude of a quantity relative to which measurements are made.

Basically, a quantitative variable requires that a standard unit is defined, against which other magnitudes may be compared in the form of a ratio. Now think 'IQ', 'Extraversion' or any psychological-attribute latent variable of your choosing from a SEM or IRT analysis (e.g. job satisfaction, trait emotional intelligence, verbal reasoning). By assuming such variables are quantitatively structured, psychologists commit to three fundamental assumptions:

- magnitudes of the variable vary additively (linearly);
- there is a named, standard unit against which other magnitudes may be expressed as a ratio of that unit;
- the variable is no different in the qualities defining its magnitude-variation from that of the seven Système International Base Units within physics (such as length, time, electrical current, thermodynamic temperature; http://physics.nist.gov/cuu/Units/units.html).

So, trait emotional intelligence or even human intelligence is assumed to vary in exactly the same way as electrical current or mass; as a continuous, real-valued, additive-unit, quantitative variable. It is a preposterous assumption, always left untested, and invariably made on the back of a claim: "ordered classes are simply approximations of quantitative variation, so the assumption is reasonable." As Michell (2009, 2012a, 2012b) demonstrated, this is not the case at all except under very special constraints. I should also add Michell also explained why classical and modern psychometrics is a 'pathology of science/pathological science' (Michell, 2000, 2004, 2008). I am afraid to say, most psychologists know nothing of these publications and expositions, and those who do mostly ignore them.

However, in 1998 Michael Maraun published his landmark paper on the role of meaning, measurement, and validity. Let me quote a couple of passages which again, brought me and my thinking/work as a psychological scientist to a complete standstill:

> "Measurement practice in psychology misdiagnoses the nature of measurement, since it is uniformly formulated under the assumption that measurement claims are justified in large part through empirical case-building [aka construct validity]…. The problem is that in construct validation theory, knowing about something is confused with an understanding of the meaning of the concept that denotes that something…. This is mistaken. One may know more or less about **it**, build a correct or incorrect case about **it**, articulate to a greater or lesser extent the laws into which **it** enters, discover much, or very little about **it**. However, these activities all presuppose rules for the application of the concept that denotes **it** (e.g. intelligence, dominance). Furthermore, one must be prepared to cite these standards as justification for the claim that these empirical facts are about **it**." (pp. 436–438)

He also noted the distinction between two kinds of concepts employed in the natural and social sciences, technical and common-or-garden concepts:

> "A technical concept is a concept defined by a specialized or expert community, and employed within a narrow, technical field of application. A common-or-garden concept, on the other hand, is a concept with a common employment in everyday life…. Common-or-garden concepts are taught, learned and understood by the person on the street, and have meanings that are manifest in broad, normative linguistic practices… The physical sciences rest on a bedrock of technical concepts, for example mass, force, gravity, hydrogen, ganglia and neutrino. Importantly, whether a technical concept plays a role in measurement is up to the inventor of the concept, since the

[3] The late Jackie Marshall was a remarkable woman, a truly magnificent administrator, a punch-tape, punch-card, and eventually key-to-disc data processor who worked for many years with Sybil Eysenck providing all those essential background services which enabled the day-to-day mechanics of the cross-cultural project. When I joined the Lab in 1983, it was clear Jackie was far more capable of other kinds of work than just that which she was doing for Sybil. So, with some additional assistance, she evolved into the senior experiment administrator for Hans, undertaking EEG and other bioelectric electrode placements, conducting all the Biosignal Lab experiments, the chronometrics, as well as the psychometric group testing of IQ and Personality. She maintained our data and experiment-respondent databases, continued working closely with Sybil, and was in the end, absolutely critical to the day-to-day running of the Lab itself. Experiment automation can only get you so far; without a person of Jackie Marshall's calibre, dedication, responsibility, and competence, I doubt whether we could have achieved what we did in the time available to us.

meaning of the concept is stated by the inventor. Mach's concept of mass, for example, is measurable by definition". (p. 453–454)

And:

"… the claim that common-or-garden psychological concepts are not measurable follows from the simple observation that common-or-garden psychological concepts as they stand are not embedded in normative practices of measurement. This is a simple observation, because it rests on whether there exist normative, rule-governed techniques for taking measurements of dominance, intelligence, creativity, tension, and so on. The normativity of rules mean that, if they exist, they are public, surveyable standards of correctness for behaviour. Rules of measurement, if they existed for common-or-garden concepts, would be manifest in descriptions of the correct employments of these concepts (see, e.g., Ter Hark, 1990). But they do not exist. There is no public, normative status at all to assertions like 'Tomorrow we are going to measure little Tommy's dominance'. What does this mean? In contrast to the teaching of the use of concepts such as weight and height, the teaching of the use of concepts such as dominance and intelligence does not involve the teaching of rules for measuring. There is no common language standard of correctness for a claim like 'I measured Sue's leadership this morning'. In other words, there is no public, standardly taught notion of what it is to be correct in making such an assertion; instead, it sounds merely curious." (p. 455)

The challenge for me after reading these articles was answering the double-barrelled question: 'what exactly is the consequence of understanding this new knowledge, and if I am to be working within a non-quantitative science, how should I proceed to investigate psychological phenomena in future?' It took me almost 10 years to come to terms with the questions and their answers. Quite simply: the consequence was the rejection of latent variable psychology, Item Response Theory and Rasch measurement, all classical and modern psychometrics, and all aggregate statistical methodologies as the means to develop causal scientific explanations of psychological phenomena.

Nowadays, my primary focus is on measurement, reliability, and validity as might be addressed within any other science, quantitative or non-quantitative; devoid of all statistical test-theory/psychometrics dogma. Likewise I separate out 'scientific' research from the 'pragmatic'. That is, for pragmatic purposes, I will use any methodology which helps me arrive at a useful outcome, for prediction, classification, or any other purpose. But my 'toolbox' now extends into machine learning, custom computational algorithmics, and whatever else can be utilised to form a useful cross-validated prediction or classification model. For scientific research, unless working specifically at phenomenon detection, the measurement procedures for any attribute are those constrained by the theory which dictates how an attribute must vary, and what is causal for that variation. Which is where James Grice's (2011) Observational Oriented Modelling is so appropriate; it is a methodology designed to test causal explanatory theory where attribute/phenomenal relations are not assumed as quantitative.

Unlike many psychologists, methodologists, and psychometricians who simply ignore the deep implications of the above knowledge, I think Hans would have found these explanatory expositions from Michell and Maraun absolutely riveting and intellectually challenging. As he told me many times, he really wanted to be a physicist, had not the Nazi Party intervened, forcing him to leave Germany. For me, looking back, it was unfortunate that I spent my time with him as someone steeped in conventional psychometrics, test theory, and statistical analysis. But maybe that kind of routine 'status-quo' methodology was required at that particular point in time in order to pursue the kinds of experiments we set out to conduct. But who knows what might have resulted from a possible new research programme targeted directly at the investigation of individual differences phenomena; dedicated to explanation rather than correlation, and grounded in a coherent approach to measurement and meaning rather than that reliant upon so many incoherent assumptions and propositions which form the basis of the current aggregate statistical (Freedman & Berk, 2003) modelling and individual-differences psychometrics research. Hans was a pioneer and rebel at heart — and this new knowledge demands a pioneering response and significant field leadership, exactly what he possessed and demonstrated. But, it all came too late. Both my mentors, Paul Kline and Hans Eysenck died within a year or so of each other.

So, there we have it, a brief summary of why Hans set up the Lab, what stimulated his interest, what he was looking for, and why the research ended. Although many colleagues thought little of this research, especially when it ended, these are the same people who enthusiastically promote the small/moderate effect sizes in their own research without realising the pointlessness of any investigative science based upon such aggregate statistical trivia. The tragedy is that no-one in individual differences research is now engaged in the kind of causal explanatory-theory investigation that Hans always set out to conduct. We have lots of correlational workups, descriptive meta-analyses, descriptive latent variable models, and a host of longitudinal and applied statistical/epidemiological findings promoted as "cognitive epidemiology" etc., but these are no more than variations of phenomena detection; of general social interest but hardly the basis of an investigative science which seeks to explain the causes of phenomena, accurately. An opinion shared by Denny Borsboom (2013) appearing in a recent article for the Centre For Open Science:

"My field – psychology – unfortunately does not afford much of a lazy life. We don't have theories that can offer predictions sufficiently precise to intervene in the world with appreciable certainty. That's why there exists no such thing as a psychological engineer. And that's why there are fields of theoretical physics, theoretical biology, and even theoretical economics, while there is no parallel field of theoretical psychology. It is a sad but, in my view, inescapable conclusion: we don't have much in the way of scientific theory in psychology."

Hans Eysenck was dedicated to exploring explanatory causal theory for phenomena, as a scientist; that is what differentiated his systematic, disciplined programme of investigative work in this area from those who were, and remain satisfied with simple phenomenon detection and speculative description.

6. A more personal perspective of Hans Eysenck

Working with Hans (and Sybil) on a daily basis for the best part of 11 years was an adventure. There really is no other word for it. Many have commented on the extraordinary impact Hans had on other researchers and practitioners, not only could he inspire, enthuse, and motivate others, he could also provide leadership with what looked like effortless ease. He could also inflame others to the point of apoplexy by his impassive style of argument. In fact, arguing with Hans was like arguing with a logic machine with a large working memory and huge long-term storage capacity. But how he loved debate and challenge! I kept my head down for the first couple of years with Hans, but as usual, my own love of debate and challenge got the better of me and so I would begin to slowly probe some of Hans's work and thinking. Usually I was soon shown the error of my thinking rather smartly – but I was able to learn. And this again was such a marvellous feature of Hans, regardless of where he was in his own work, he would always take time out to explain a concept or piece of evidence for you. Often, he would bring me photocopies of the articles he had been mentioning. I know from speaking with many others who also interacted with Hans that this characteristic of offering help and advice was global.

I think the one thing that always stood out with Hans, almost irrespective of which domain of psychology within which he was working

at the time, was that he could always see the big picture. That is, his grasp of theory was such that many diverse results in different domains were construed as part of some integrated picture in his own representational system. Thus, he was able to suggest propositions and provide insights that might sometimes have eluded others because of his global theoretical view. This, allied to his fundamental principles of what constituted scientific investigation, laid the foundation for the enormous impact and stimulus that he provided to so many for such a long time.

Although generally calm and assured, one source of irritation to Hans was being kept waiting (usually by me) for tennis! I would often get dragged out by Glenn Wilson or Hans — right in the middle of some work. I remember sometimes wondering (whilst turning the air blue) whether I should give up any pretence at work and just settle for 'tennis practise-partner' as my occupation! However, truth be told — we were as bad as one another about our love for the game. When I went to the second ISSID conference in Toronto – I was "requested" to bring my tennis stuff – and there we were, mid-conference, playing in almost 90% humidity in a temperature of 90 F. The trouble was, I was as worn out as Hans — and he was nearing 70, almost twice my age at the time!

Maybe some of the above surprises some readers. But I worked closely with Hans on a day-to-day basis for many years. There were times when we had major arguments and disagreements — and I once or twice gave notice to quit. However, his good grace, his wonderful sense of humour, and the growing realisation of how to manage my own more explosive energies ensured that we would continue our working relationship!

References

Barrett, P. T., & Eysenck, H. J. (1992). Brain evoked potentials and intelligence: The Hendrickson paradigm. *Intelligence*, *16*, 361–381.

Barrett, P. T., & Eysenck, H. J. (1993). Sensory nerve conduction and intelligence: A replication. *Personality and Individual Differences*, *15*, 249–260.

Barrett, P. T., & Eysenck, H. J. (1994). The relationship between evoked potential component amplitude, latency, contour length, variability, zero-crossings, and psychometric intelligence. *Personality and Individual Differences*, *16*, 3–32.

Barrett, P. T., Petrides, K. V., & Eysenck, H. J. (1998a). Estimating inspection time: Response probabilities, the BRAT IT algorithm, and IQ correlations. *Personality and Individual Differences*, *24*, 405–419.

Barrett, P. T., Petrides, K. V., Eysenck, S. B. G., & Eysenck, H. J. (1998b). The Eysenck Personality Questionnaire: An examination of the factorial similarity of P, E, N, and L across 34 countries. *Personality and Individual Differences*, *25*, 805–819.

Barrett, P., Eysenck, H. J., & Lucking, S. (1986). Reaction time and intelligence: A replicated study. *Intelligence*, *10*, 9–40.

Blinkhorn, S. F., & Hendrickson, D. E. (1982). Averaged evoked responses and psychometric intelligence. *Nature*, *295*, 596–597.

Borsboom, D. (2013). Theoretical amnesia. *Centre for Open Science*. Retrieved from: http://centerforopenscience.github.io/osc/2013/11/20/theoretical-amnesia/

Brand, C. R. (1981). General intelligence and mental speed: Their relationship and development. In M. P. Friedman, J. P. Das, & N. O'Connor (Eds.), *Intelligence and learning* (pp. 589–593). New York: Plenum Press.

Brand, C. R., & Deary, I. J. (1982). Intelligence and 'inspection time'. In H. J. Eysenck (Ed.), *A model for intelligence* (pp. 133–148). New York: Springer.

Ertl, J., & Schafer, E. (1969). Brain response correlates of psychometric intelligence. *Nature*, *223*, 421–422.

Eysenck, S. B. G. (1983). One approach to cross-cultural studies of personality. *Australian Journal of Psychology*, *35*, 381–391.

Eysenck, S. B. G., & Barrett, P. T. (2013). Re-introduction to cross-cultural studies of the EPQ. *Personality and Individual Differences*, *54*, 485–489.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PloS One*, *4*(5), e5738.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*, 532–538.

Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology*, *57*, 153–169.

Frearson, W., & Eysenck, H. J. (1986). Intelligence, reaction time (RT) and a new 'odd-man-out' RT paradigm. *Personality and Individual Differences*, *7*, 807–817.

Freedman, D. A., & Berk, R. A. (2003). Statistical assumptions as empirical commitments. In T. G. Blomberg, & S. Cohen (Eds.), *Law, punishment, and social control: Essays in honor of Sheldon Messinger* (pp. 235–254) (2nd ed.). Berlin: Aldine de Gruyter.

Grice, J. (2011). *Observation oriented modeling: Analysis of cause in the behavioral sciences*. New York: Academic Press.

Hendrickson, D. E., & Hendrickson, A. E. (1980). The biological basis of individual differences in intelligence. *Personality and Individual Differences*, *1*, 3–33.

Hendrickson, D. E., & Hendrickson, A. E. (1982). The biological basis of intelligence, part II: Measurement. In H. J. Eysenck (Ed.), *A model for intelligence* (pp. 197–228). New York: Springer.

Irwin, R. J. (1984). Inspection time and its relation to intelligence. *Intelligence*, *8*, 47–65.

Jensen, A. R. (1982). Reaction time and psychometric g. In H. J. Eysenck (Ed.), *A model for intelligence* (pp. 93–132). New York: Springer.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532.

Kline, P. (1998). *The new psychometrics: Science, psychology, and measurement*. London: Routledge.

Kranzler, J. H., & Jensen, A. R. (1989). Inspection time and intelligence: A meta-analysis. *Intelligence*, *13*, 329–347.

Lally, M., & Nettelbeck, T. (1977). Intelligence, reaction time, and inspection time. *American Journal of Mental Deficiency*, *82*, 273–381.

Maraun, M. D. (1998). Measurement as a normative practice: Implications of Wittgenstein's philosophy for measurement in psychology. *Theory and Psychology*, *8*, 435–461.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*, 355–383.

Michell, J. (1999). *Measurement in psychology: Critical history of a methodological concept*. Cambridge, UK: Cambridge University Press.

Michell, J. (2000). Normal science, pathological science, and psychometrics. *Theory and Psychology*, *10*, 639–667.

Michell, J. (2004). Item response models, pathological science, and the shape of error. *Theory and Psychology*, *14*, 121–129.

Michell, J. (2008). Is psychometrics pathological science? *Measurement: Interdisciplinary Research & Perspective*, *6*, 7–24.

Michell, J. (2009). The psychometricians' fallacy: Too clever by half? *British Journal of Mathematical and Statistical Psychology*, *62*, 41–55.

Michell, J. (2012a). "The constantly recurring argument": Inferring quantity from order. *Theory and Psychology*, *22*, 255–271.

Michell, J. (2012b). Alfred Binet and the concept of heterogeneous orders. *Frontiers in Quantitative Psychology and Measurement*, *3*, 1–8.

Nettelbeck, T., & Lally, M. (1976). Inspection time and measured intelligence. *British Journal of Psychology*, *67*, 17–22.

Reed, T. E., & Jensen, A. R. (1991). Arm nerve conduction velocity (NCV), brain NCV, reaction time, and intelligence. *Intelligence*, *15*, 33–47.

Reed, T. E., & Jensen, A. R. (1992). Conduction velocity in a brain nerve pathway of normal adults correlates with intelligence level. *Intelligence*, *16*, 259–272.

Robinson, D. L. (1999). The technical, neurological and psychological significance of 'alpha', 'delta' and 'theta' waves confounded in EEG evoked potentials: A study of peak latencies. *Clinical Neurophysiology*, *110*, 1427–1434.

Robinson, D. L. (2000). The technical, neurological, and psychological significance of 'alpha', 'delta' and 'theta' waves confounded in EEG evoked potentials: A study of peak amplitudes. *Personality and Individual Differences*, *28*, 673–693.

Sheppard, L. D., & Vernon, P. A. (2008). Intelligence and speed of information processing: A review of 50 years of research. *Personality and Individual Differences*, *44*, 535–551.

Ter Hark, M. (1990). Beyond the inner and the outer: Wittgenstein's philosophy of psychology. *Dordrecht: Kluwer Academic*.

Vernon, P. A., & Mori, M. (1992). Intelligence, reaction times, and peripheral nerve conduction velocity. *Intelligence*, *16*, 273–288.