SOUTH AFRICAN
Journal of Science

Prince
Edward Islands:
South Africa's
Subantarctic
laboratory

*Lab on a disc
for point-of-care
diagnostics*

Economic mineral
potential of the
Western Cape

*Responsible
conduct of
research in
South Africa*

volume 110
number 1/2

JANUARY/FEBRUARY 2014

ISSN: 0038-2353

ASSAf
ACADEMY OF SCIENCE OF SOUTH AFRICA

JANUARY/FEBRUARY

ISSN: 0038-2353 (Print) | ISSN: 1996-7489 (Online)

SOUTH AFRICAN Journal of Science

volume 110
number 1/2

Leader

News & Views

Book Review

Commentary

Review Article

Research Article

Published by
the Academy of
Science of South Africa
(www.assaf.org.za) with financial
assistance from the Department of
Science & Technology.

Design and layout
SUN MeDIA Bloemfontein
T: 051 4300 459
E: sunmedia@sunbloem.co.za

Correspondence and enquiries
sajs@assaf.org.za

Disclaimer
The publisher and editors accept no responsibility for statements made by the authors.

Submissions
Submissions should be made at
http://mc.manuscriptcentral.com/sajs

Subscriptions
Subscription information can be
found at
www.sajs.co.za

The statesman of education, science and technology

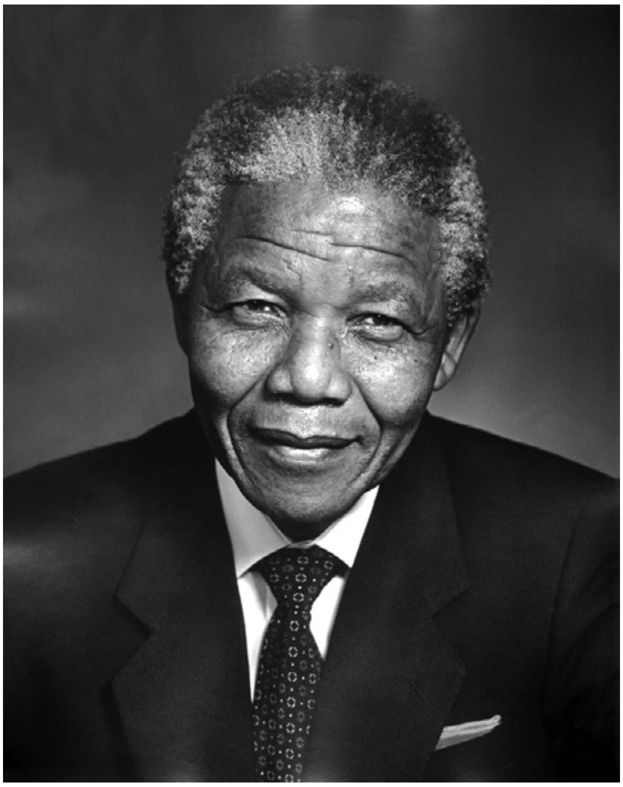


Photo: Festival Karsh (CC)

(http://creativecommons.org/licenses/by-nc-nd/2.0/)

Nelson Mandela, 1990.

> *What has become of our rationality, our ability to think? We have used our reason to make great advances in science and technology, though often using those for warfare and plunder. We have placed people on the moon and in space; we have split the atom and transplanted organs; we are cloning life and manipulating nature. Yet we have failed to sit down as rational beings and eliminate conflict, war and consequent suffering of innocent millions, mostly women, children and the aged.*
>
> *Nelson Mandela: Address on receiving the International Gandhi Peace Prize, 2000*

Almost every South African who works in the fields of education, science or technology (EST) knows the famous Nelson Mandela statement that 'Education is the great engine of personal development'[1]. But the words that the late president included in his acceptance speech on receiving the International Gandhi Peace Prize in 2000 take that idea into a much wider realm. Education is, of course, the basis for personal growth; yet, when coupled with its application in the realms of the broad world of all the sciences and of technology, it becomes the key to social and economic growth and justice. The *South African Journal of Science* pays tribute, here, to a statesman and leader who recognised that relationship and who strove, through his insights and leadership, to change the way in which all three fields could and would operate in a new democracy faced with vast challenges of social and economic justice.

Nelson Rolihlohlo Mandela's own words on education and science are enlightening. About his first experience in a 'proper hospital' after injuring his heel while on Robben Island, aggravating an earlier injury, he wrote[2]:

> *I found the trip* [to the hospital] *instructive ... because in that hospital I sensed a thawing in the relationship between black and white. The doctor and nurses had treated me in a natural way as though they had been dealing with blacks on a basis of equality all their lives. This was something new to me, and an encouraging sign. It reaffirmed my long-held belief that education is the enemy of prejudice. These were men and women of science, and science had no room for racism.*

In the *New Scientist* of 06 December 2013, Calestous Juma, Professor in the Kennedy School of Government at Harvard, points out[3]:

> *For much of the world Nelson Mandela was the icon of the age of modern liberation that started with Mahatma Gandhi and reached its height with South Africa's first democratic elections in 1994. What is less well known is that the struggle for political freedom was closely associated with the desire to develop scientific and technological capacity.*

During the years of apartheid – and before – educational opportunities for Black South Africans were severely limited. With the exception of missionary schools (and the University of Fort Hare, which grew out of a mission school) and, later, ethnically based universities and the University of Natal Medical School, opportunities for education at all levels were limited to very small numbers for anyone who was not White.

Transformation and national development had, as a consequence, to include major reformation in EST as a necessity – not just for 'the things in themselves', but in order to provide the knowledge and the applications – the broader engines of development – that would help to drive economic growth and the many changes (still) needed. As Cloete et al.[4] explain:

> [...T]*here is increasing evidence that high levels of education in general, and of higher education in particular, are essential for the design and productive use of new technologies, while they also provide the foundations for a nation's innovative capacity, and contribute more than any other social institution to the development of civil society.*

As President of South Africa, Nelson Mandela clearly grasped these relationships – both those between education and economic development and those among science, technology and the general development of society. So it is not surprising that Simon Connell, a member of the

HOW TO CITE: Butler-Adam J. The statesman of education, science and technology. S Afr J Sci. 2014;110(1/2), Art. #a0052, 2 pages. http://dx.doi.org/10.1590/sajs.2014/a0052

Academy of Science of South Africa, and Professor of Physics in the University of Johannesburg, wrote[5]:

> *Apart from creating a general political climate favourable to research, Mandela was known to intervene directly when science became politicised. An example was the contentious politics surrounding health issues in South Africa. He appealed, 'In the face of the grave threat posed by HIV/AIDS, we have to rise above our differences and combine our efforts to save our people. History will judge us harshly if we fail to do so now, and right now.'*

Mr Mandela did not simply offer lip service to the role of EST but made it possible for the needed changes to happen and, when necessary, intervened in the contestation between science and politics on the side of science.

His commitment is also evident in the number of educational and research organisations to which he agreed to give his name – and, in some cases, his personal support and encouragement. Motivated by his insightfulness, Mandela lent his name to the creation of a new generation of African Institutes of Science and Technology. Two have already been established – in Tanzania and Nigeria – and a third is planned in Burkina Faso. His name is also that of a university (the Nelson Mandela Metropolitan University), a medical school (the Nelson R Mandela School of Medicine at the University of KwaZulu-Natal) and a pupil support centre in Uitenhage (the Nelson Mandela Bay Science and Technology Centre). Not to be overlooked is Mandela's patronage of the Academy itself; at the Academy's launch in 1996, Mandela made the observation, core to his view of EST, that the formation of the Academy 'is not an isolated act but part of the building of a new society which freedom has made possible'[6].

In the same speech, Mandela went on to say

> *South Africa's need for rapid expansion of its scientific and technological skills is immense. It is inhibited by the disastrous restriction which apartheid imposed on the level of scientific and technological education: and by an image of science tarnished in the eyes of the majority by associations with the past. On your shoulders rest the challenge of giving science a face that inspires our youth to seek out science, engineering and technology.*[6]

We would be neglectful were we to overlook the challenging problems that still beset the primary and secondary education subsystems in South Africa. While not exonerating universities, the Council on Higher Education's report *A proposal for undergraduate curriculum reform in South Africa* makes it clear that 'it is widely accepted that student preparation [for higher education] is the dominant cause of the poor performance patterns in higher education' and 'that if higher education is to rely on improvement in schooling to deal with the systemic faults affecting it, there needs to be a rigorous assessment of the prospects of sufficient improvement being achieved within that sector.' This is an area in which there is still a long, long 'walk to success' to be completed.

Setting aside the current conditions of the primary and secondary subsystems, which require a return to the late president's understanding of the importance of science over politics in dire situations of need, his contributions to EST as key to South Africa's success have been monumental. His fundamental understanding that education and the sciences are the foundations of a non-racial society, his commitment to promoting this position (directly when necessary) and his insights into the relationships between EST and both economic growth and social justice as the basis of a successful democracy are the legacy he has left us.

Perhaps it is fitting then to end this tribute to a remarkable human being and champion, a true statesman of EST, with the words of another member of the Academy, Professor Salim Abdool Karim, President of the South African Medical Research Council: 'We will humbly try to continue following in [Nelson Mandela's] footsteps in the enduring quest to make our world a better place for all'[7].

References

1. Department of Basic Education. Nelson Mandela quotes [document on the Internet]. c2014 [cited 2014 Jan 10]. Available from: http://www.education.gov.za/LinkClick.aspx?fileticket=8d6cPhef%2FL8%3D&tabid=656&mid=1849

2. Mandela N. Long walk to freedom. Boston: Back Bay Books; 1995.

3. Juna C. Mandela's unsung legacy of science in Africa [Opinion]. New Sci [serial on the Internet]. 2013 Dec 06 [cited 2014 Jan 10]. Available from: http://www.newscientist.com/article/dn24712-mandelas-unsung-legacy-of-science-in-africa.html#.UtjJ1hCSwrX

4. Cloete N, Bailey T, Pillay P, Bunting I, Maassen P. Universities and economic development in Africa. Cape Town: Centre for Higher Education Transformation; 2011.

5. Connell S. What Mandela meant for science in South Africa [homepage on the Internet]. c2013 [cited 2014 Jan 10]. Available from: http://www.uj.ac.za/EN/Newsroom/News/Pages/What-Mandela-meant-for-Science-in-South-Africa aspx.

6. Mandela N. Address by President Nelson Mandela on the inauguration of the Academy of Science of South Africa [speech on the Internet]. c1996 [cited 2014 Jan 10]. Available from: http://www.mandela.gov.za/mandela_speeches/1996/960322_science.htm

7. Makoni M, Grange H. Mandela 'aided scientific renaissance in South Africa'. SciDev.Net [online]. 2013 Dec 19 [cited 2014 Jan 10]. Available from: http://www.scidev.net/global/education/news/mandela-aided-scientific-renaissance-in-south-africa-1.html

Sentinels to climate change. The need for monitoring at South Africa's Subantarctic laboratory

AUTHORS:
Isabelle J. Ansorge[1]
Jonathan V. Durgadoo[2]
Anne M. Treasure[1]

AFFILIATIONS:
[1]Department of Oceanography, Marine Research Institute (Ma-Re), University of Cape Town, Cape Town, South Africa
[2]GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany

CORRESPONDENCE TO:
Isabelle Ansorge

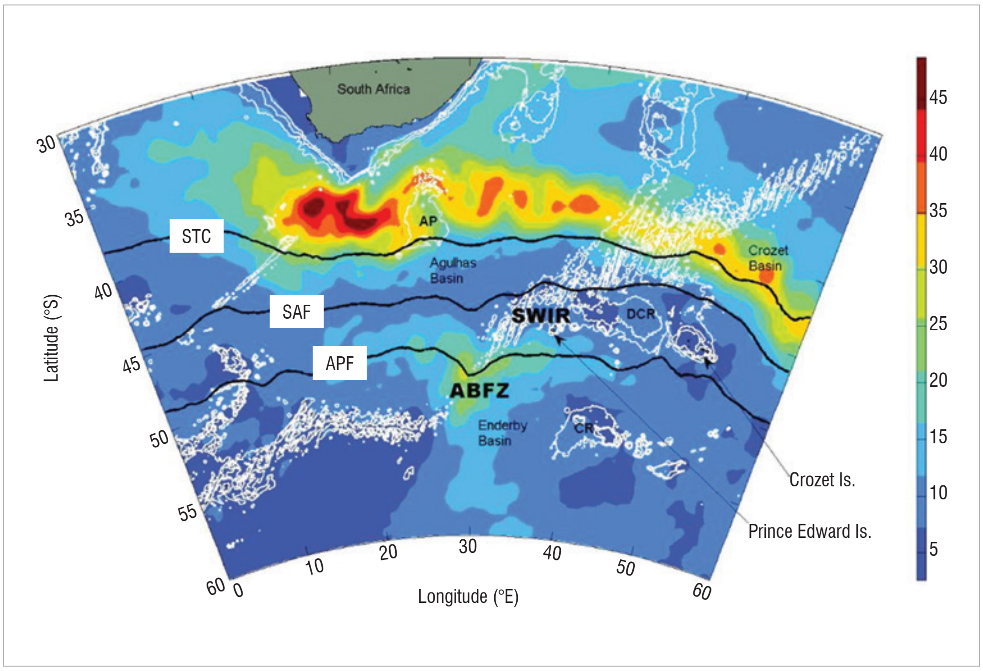
EMAIL:
Isabelle.Ansorge@uct.ac.za

POSTAL ADDRESS:
Department of Oceanography, Marine Research Institute (Ma-Re), University of Cape Town, Private Bag, Rondebosch 7701, South Africa

KEYWORDS:
Prince Edward Islands; Subantarctic Front; Antarctic Circumpolar Current; climate change; oceanography

HOW TO CITE:
Ansorge IJ, Durgadoo JV, Treasure AM. Sentinels to climate change. The need for monitoring at South Africa's Subantarctic laboratory. S Afr J Sci. 2014;110(1/2), Art. #a0044, 4 pages. http://dx.doi.org/10.1590/sajs.2014/a0044

Subantarctic islands form ideal sentinels to climate change. These islands support terrestrial and marine ecosystems that are relatively simple and extremely sensitive to perturbations.[1] They provide an ideal natural laboratory for studying how ecosystems respond to a changing climate in the Southern Ocean. Initial studies[1] and subsequent analyses[2,3] at the Prince Edward Islands in the Indian sector of the Southern Ocean have all shown that there has been a climatological rise of >1 °C in the sea surface temperature since 1949. Mirrored to this rise is a decrease in rain, an increase in extreme events, an increase in winds from the warmer sector in the northwest, and annual sunshine hours have risen by 3.3 h since the 1950s.[1,2,4] It has been proposed[5] that climate changes reported at the Prince Edward Islands correspond in time to a southward shift of the Antarctic Circumpolar Current (ACC) and in particular its frontal systems the Subantarctic Front (SAF) and the Antarctic Polar Front (APF), between which the islands lie (Figure 1). The Prince Edward Islands, like many other oceanic islands within the Southern Ocean, are seasonally characterised by vast populations of marine organisms and a diversity and abundance of seabirds that use the islands as breeding grounds.[6,7] It is estimated that the islands support over 5 million breeding pairs of top predators including seabirds, penguins and seals during the peak breeding season.[7] The energy necessary to sustain these top predators is derived from the close interaction between the oceanic environment and the islands themselves. Changes in the intensity and geographic position within these frontal systems are likely to coincide with dramatic changes in the distribution of species and total productivity within the Southern Ocean. Long-term research in both the offshore and near-shore environments of the Prince Edward Islands is critical if the mechanisms needed to sustain high concentrations of marine life in a changing environment are to be better understood and conserved. If ecosystems are pushed beyond certain thresholds or tipping points, there is a high risk of dramatic biodiversity loss and accompanying degradation of a broad range of ecosystem services.[8] Furthermore, a real threat brought on by a warmer climate will be the increase in ease in which pristine Subantarctic islands such as the Prince Edward Islands can be invaded by alien species.[1]



Source: Figure courtesy of Durgadoo et al.[10]

Figure 1: The root mean square in sea level anomaly (cm) calculated from a 13-year record of altimetry products. Two broad regions of variability are observed in the African sector of the Southern Ocean: the Agulhas regime (comprising the Agulhas Current, the Agulhas Retroflection, the Agulhas Return Current and the Subtropical Convergence) and the South-West Indian Ridge (SWIR), a separate region to the south. Bathymetric contours (-3000 m, -2000 m, -1000 m) are overlain in white. The average positions of the three major fronts associated with the Antarctic Circumpolar Current are also shown: the Subtropical Convergence (STC), the Subantarctic Front (SAF) and the Antarctic Polar Front (APF). The Andrew Bain Fracture Zone (ABFZ) is shown as the core of eddy variability. The Del Cano Rise (DCR) lies to the east of the Prince Edward Islands.

Recent observations by the International Programme on the State of the Ocean[9] have highlighted concerns that the cumulative impacts on the ocean are greater than previously thought and the need to effectively manage fragile ecosystems is now critical. It is likely that global climate change may in the future become associated with the disruption of the 'life support system of the Prince Edward Islands'[6] and further investigations integrating offshore dynamics with the near-shore circulation are required to understand better what impact these changes will have on the system as a whole, and at what rate. On 9 April 2013, the Minister of Water and Environmental Affairs, Edna

Molewa, declared the Prince Edward Islands a marine protected area (MPA). The new MPA is intended to contribute to the protection of unique species, habitats and ecosystem processes. It will provide a scientific blueprint that can advise future management in understanding better the impact climate change is having on the Southern Ocean. We highlight the importance of long-term oceanographic research at the Prince Edward Islands if the mechanisms needed to sustain total productivity in a changing environment are to be better understood and the objectives of this newly constituted MPA are to be realised.

Offshore environment: Ocean corridors

Hydrographic data collected during the recent South African National Antarctic Programme (SANAP)-funded Marion Island Offshore Study (MIOS), Southern Ocean Climate and Biodiversity (SOCAB) and Dynamics of Marion Island's Impacts on the Marine Ecosystem (DEIMEC) programmes[10] have shown that the ecosystem of the Prince Edward Islands benefits substantially from their location immediately downstream of an eddy corridor (Figure 1). Past investigations[11,12] have shown that the advection of eddies between the Subantarctic and Antarctic produce meridional heat and salt fluxes that balance much of the ocean–atmosphere exchange across the ACC. If the assumption is correct that a part of this flux comes about through mesoscale eddies, then their quantification at regions where mesoscale turbulence is high may be crucial. Recent studies have shown that since the 1950s, the ACC has warmed with possible shifts in frontal features between 50 km and 70 km.[13,14] The impact this migration may have on the Prince Edward Islands ecosystem over the next century is indeed complex. We speculate that wind stress can induce changes in the intensity and frequency of the Southern Ocean eddy field. In short, an increase in the poleward eddy flux can be attributed to the strengthening westerly wind belt, which in turn has been shown to contribute to the warming of the Southern Ocean.[13,15] Closer examination of Gille's data set suggests these trends occur directly south of each band of high eddy kinetic energy – the South-West Indian Ridge being one such example (Figure 1). These eddies move into the inter-island region and have been shown[16] to carry with them organisms from their region of origin, suggesting that the near-shore waters of the islands have a more diverse spectrum of biota than one would normally expect from a Subantarctic island. Changes in the intensity and geographic position within these frontal systems are likely to coincide with dramatic changes in the distribution of species and total productivity within the Southern Ocean. The impact that a long-term shift in the SAF may have on the near-shore dynamics provokes an interesting discussion.

Near-shore environment: A through-flow system?

The Prince Edward Islands are seasonally characterised by dense populations of top predators[6]; subsequently, any changes in the frontal dynamics, either in the vicinity of these islands or further afield, may have strong implications on their near-shore foraging behaviour. In addition to the eddy corridor there is growing evidence that the geographical position of the SAF in the proximity of the Prince Edward Islands plays an important role in defining the short-term local macro- and mesoscale oceanographic conditions.[17,18] Previous investigations have shown that when the SAF lies far north of the islands, the interaction between the ACC and the island group results in water retention over the inter-island region which encourages algal blooms.[19] During these periods, signature Antarctic species dominate and suggest that water masses typical of modified Antarctic surface water prevail. In contrast, when the SAF meanders southwards in closer proximity to the islands, advection forces prevail and a flow-through system associated with a long-term decline in the frequency of phytoplankton blooms is observed.[20,21] Recent decline in stable isotope carbon values of the bottom-dwelling shrimp *Nauticaris marionis* indirectly postulates a decrease in the occurrence of bloom conditions in the inter-island region between the 1980s and 2000s.[5,22] Mirrored to these findings, there have been increases in the population sizes of offshore feeding land-based predators, such as wandering and grey-headed albatrosses[23] and Subantarctic and Antarctic fur seals[24], which may be a result of a decrease in foraging distances from the Prince Edward Islands. In direct contrast, it appears that the populations of inshore predators feeding on *Nauticaris marionis* – such as Gentoo and Rockhopper penguins and Crozet shags[25-28] – have decreased. The decline in the abundances of predators that rely heavily on prey inhabiting the inshore system of the islands suggests an indirect effect of climate change on prey availability.[5] Climate models suggest a restructuring of the ACC under changing westerly winds, such that its associated frontal systems display an average southward shift.[14]

Ocean monitoring as a tool for marine protected areas

Separating short-term natural variability through eddy shedding from the long-term southward migration of the ACC remains a key challenge. In-situ long-term oceanographic measurements remain sparse in time and space because of ongoing logistical and cost constraints of marine research in the Southern Ocean.[29] The need to effectively monitor the long-term change in the position and velocity associated with the ACC in relation to the island group through dedicated research programmes is now critical. On 9 April 2013, the Minister of Water and Environmental Affairs, Edna Molewa, declared the Prince Edward Islands an MPA. This new Prince Edward Islands MPA, the first South African offshore MPA, will serve to significantly contribute to global initiatives towards protection of offshore and deep ocean areas. Minister Molewa added[30]:

> *The new MPA is intended, among other things, to contribute to the protection of unique species, habitats and ecosystem processes. It will also provide scientific reference points that can inform the future management of the area and to be able to understand better the impacts of climate change on the whole Southern Ocean. It will also contribute to integrated and ecologically sustainable management of marine resources of the area.*

It has been shown[31] that the success of this MPA will be dependent on the ability to nest the region within a broad management framework encompassing the island's exclusive economic zone and its broader region. A key outcome of this policy is that the Prince Edward Islands are reliant on a combination of ecosystem processes that occur within the inter-island region and further afield at the South-West Indian Ridge. It is unclear whether this change is expected to continue and at what rate. However, it may have already promoted long-term changes in land-based top predator populations on the islands, possibly through food-chain modifications.[22] A key recommendation by the MPA committee[31] to the Ministry is the need to continue studies that

> *detect climate related changes that may impact the boundaries of the proposed MPA, sea surface temperatures and shifts in the position of the major frontal systems ... and their biotic responses ... to the north and south of the islands.*

This recommendation further mirrors the key focus of the Department of Science and Technology 2018 Grand Challenge which calls for an improved understanding of how the Southern Ocean, and in particular its Subantarctic islands, will respond to climate change through changes in ecosystem function and structure. National long-term environmental monitoring is already supported by government initiatives such as SANAP, the South African Environmental Observation Network and the South African Weather Service. Other organisations also contribute to environmental monitoring but, because such activities need to be long term, 'monitoring and data management may be threatened by the vagaries of funding cycles and changes in research priorities'[31].

The way forward: Future foci

The need to integrate near-shore and offshore environments through a series of core objectives is recognised. Our understanding of the eastward passage of eddies extending from the South-West Indian

Ridge (Figure 1) as well as the implications of a southward migration of the SAF[14] has improved through the use of high-resolution ocean models and direct observations through both international (i.e. ARGO, Global Drifter Programme, Aviso) and national (SANAP) programmes. However, interactions of these processes – for instance the impact a change in position of the SAF has on the frequency and intensity of eddy generation at the South-West Indian Ridge – remain unknown. Given recent findings[5] which show that the southward migration of the SAF is likely to be associated with a long-term decline in phytoplankton blooms, addressing the importance of this biological–physical link is critical. If the SAF is indeed shifting southwards, what impact will this shift have on the inter-island ecosystem in the future? Past investigations have shown that the ability for water to be retained between the islands is dependent on the position of the SAF. Thus, the need to monitor the position of the front in close proximity to the Prince Edward Islands is critical in order to separate the impacts of short-term variability through eddy shedding from the long-term southward migration in response to global climate change. In order to achieve this, a long-term mooring array spanning the inter-island region directed at measuring the position of the SAF, through temperature and velocity profiles, needs to be deployed. Using a full-depth thermistor chain, temperature profiles will highlight both seasonal- and frontal-driven changes in the inter-island region, while velocity profiles from mounted acoustic doppler currrent profilers will provide an insight into the dominant circulation between the islands. It is expected that should the SAF be shifting southwards as inferred by coupled ocean–atmospheric models,[14] temperature and velocity profiles between the islands will increase over the next decade.

The unique biological diversity of Subantarctic islands provides an ideal environment in which methods can be employed to enhance our ability to predict the impact a changing world has on a fragile ecosystem. The recently launched Prince Edward Island MPA will serve to protect its marine environment by offering South Africa's Subantarctic community a glimpse into the current status quo. It will provide a measure on how ecosystems are currently functioning, which will become increasingly valuable as climate change gradually disorders them. Only when these results become available and are linked with similar long-term biological and biogeochemical observations, will some understanding of the direct role that the SAF has on the near-shore island dynamics and the vulnerability of its ecosystem be fully gained.

Acknowledgements

The Department of Science and Technology, Department of Environmental Affairs – Antarctic Logistics and Support division and Oceans and Coasts as well as the South African National Antarctic Programme are acknowledged for financial and logistical support. We thank the various captains, officers and crew of the research and supply vessel SA *Agulhas* and more recently the SA *Agulhas II* for their endless support and outstanding professionalism over the past decade. The collegial collaboration with biological oceanographers from the Southern Ocean Group at Rhodes University is much appreciated.

References

1. Smith VR. Climate change in the sub-Antarctic: An illustration from Marion Island. Climatic Change. 2002;52:345–357. http://dx.doi.org/10.1023/A:1013718617277
2. Mélice J-L, Lutjeharms JRE, Rouault M, Ansorge IJ. Sea-surface temperatures at the sub-Antarctic islands Marion and Gough during the past 50 years. S Afr J Sci. 2003;99:363–366.
3. Ansorge IJ, Durgadoo JV, Pakhomov EA. Dynamics of physical and biological systems of the Prince Edward Islands in a changing climate. Papers and Proceedings of the Tasmanian Royal Society. 2009;143:15–18.
4. Le Roux PC, McGeoch MA. Changes in climate extremes, variability and signature on sub-Antarctic Marion Island. Climatic Change. 2008;86:309–329. http://dx.doi.org/10.1007/s10584-007-9259-y
5. Allen EL, Froneman PW, Durgadoo JV, McQuaid CD, Ansorge IJ, Richoux NB. Critical indirect effects of climate change on sub-Antarctic ecosystem functioning. Ecol Evol. 2013;3(9):2994–3004. http://dx.doi.org/10.1002/ece3.678
6. Bergstrom D, Chown SL. Life at the front: History, ecology and change on southern ocean islands. Trends Ecol Evol. 1999;14:472–477. http://dx.doi.org/10.1016/S0169-5347(99)01688-2
7. Ryan PG, Bester MN. Pelagic predators. In: Chown SL, Froneman PW, editors. The Prince Edward Islands: Land-sea interactions in a changing ecosystem. Stellenbosch: Sun Press; 2008. p. 121–164.
8. ICSU. Foresight analysis – 2031. International science in 2031 – exploratory scenarios. Paris: International Council for Science; 2011.
9. Rogers AD, Laffoley D d'A International Earth system expert workshop on ocean stresses and impacts: Summary report. Oxford: IPSO; 2011.
10. Durgadoo JV, Ansorge IJ, Lutjeharms JRE. Oceanographic observations of eddies impacting the Prince Edward Islands, South Africa. Antarct Sci. 2010;22(3):211–219. http://dx.doi.org/10.1017/S0954102010000088
11. Gordon AL, Taylor HW. Heat and salt balance within the cold waters of the world ocean, in numerical models of ocean circulation. Washington DC: National Academy of Science; 1975. p. 54–56.
12. De Szoeke R, Levine M. The advective flux of heat by mean geostrophic motions in the Southern Ocean. Deep-Sea Res. 1981;28:1057–1085. http://dx.doi.org/10.1016/0198-0149(81)90048-0
13. Gille ST. Warming of the Southern Ocean since the 1950s. Science. 2002;295:1275–1277. http://dx.doi.org/10.1126/science.1065863
14. Downes SM, Budnick AS, Sarmiento JL, Farneti R. Impacts of wind stress on the Antarctic Circumpolar Current fronts and associated subduction. Geophys Res Lett. 2011;38:L11605. http://dx.doi.org/10.1029/2011GL047668
15. Meredith M, Hogg A. Circumpolar response of the Southern Ocean eddy activity to changes in the Southern Annular Mode. Geophys Res Lett. 2006;33:L16608. http://dx.doi.org/10.1029/2006GL026499
16. Bernard ATF, Ansorge IJ, Froneman PW, Bernard KS, Lutjeharms JRE. Entrainment of Antarctic euphausiids into the sub-Antarctic by a cold eddy. Deep-Sea Res Pt I. 2007;54(10):1841–1851. http://dx.doi.org/10.1016/j.dsr.2007.06.007
17. Pakhomov EA, Froneman PW, Ansorge IJ, Lutjeharms JRE. Temporal variability in the physico-biological environment of the Prince Edward Islands (Southern Ocean). J Marine Syst. 2000;26:75–95. http://dx.doi.org/10.1016/S0924-7963(00)00041-5
18. Ansorge IJ, Lutjeharms JRE. The hydrography and dynamics of the ocean environment of the Prince Edward Islands (Southern Ocean). J Marine Syst. 2002;37:107–127. http://dx.doi.org/10.1016/S0924-7963(02)00198-7
19. Perissinotto R, Duncombe-Rae CM. Occurrence of anti-cyclonic eddies on the Prince Edward plateau (Southern Ocean): Effects on phytoplankton productivity and biomass. Deep-Sea Res Pt I. 1990;37:777–793.
20. Pakhomov EA, Froneman PW. Composition and spatial variability of macroplankton and micronekton within the Antarctic Polar Frontal Zone on the Indian Ocean during austral autumn 1997. Polar Biol. 2000;23:410–419. http://dx.doi.org/10.1007/s003000050462
21. Pakhomov EA, Chown SL. Appendix VI: Freshwater invertebrates of the Prince Edward Islands. In: Chown SL, Froneman PW, editors. The Prince Edward Islands: Land-sea interactions in a changing ecosystem. Stellenbosch: Sun Press; 2008. p. 398–399..
22. Pakhomov EA, McClelland JW, Bernard KS, Kaehler S, Montoya JP. Spatial and temporal shifts in stable isotope values of the bottom-dwelling shrimp *Nauticaris marionis* at the sub-Antarctic archipelago. Mar Biol. 2004;144:317–325. http://dx.doi.org/10.1007/s00227-003-1196-3
23. Ryan PG, Jones MGW, Dyer BM, Upfold L, Crawford RJM. Recent population estimates and trends in numbers of albatrosses and giant petrels breeding at the sub-Antarctic Prince Edwards Islands. Afr J Mar Sci. 2009;31:409–417. http://dx.doi.org/10.2989/AJMS.2009.31.3.13.1001
24. Hofmeyr GJG, Bester MN, Makhado AB, Pistorius PA. Population changes in Subantarctic and Antarctic fur seals at Marion Island. S Afr J Wildl Res. 2006;36:55–68.
25. Williams AJ, Siegfried WR, Burger AE, Berruti A. The Prince Edward Islands: A sanctuary for seabirds in the Southern Ocean. Biol Conserv. 1979;15:59–71. http://dx.doi.org/10.1016/0006-3207(79)90015-6

26. Crawford RJM, Cooper J, Dyer BM, Wolfaardt AC, Tshingana D, Spencer K, et al. Population, breeding, diet and conservation of the Crozet shag *Phalacrocorax [atriceps] melanogenis* at Marion Island, 1994/95-2002/03. Afr J Mar Sci. 2003;25:537–547.

27. Crawford RJM, Ryan PG, Dyer BM, Upfold L. Recent trends in numbers of Crozet shags breeding at the Prince Edward Islands. Afr J Mar Sci. 2009;31:427–430. http://dx.doi.org/10.2989/18142320309504043

28. Crawford RJM, Whittington PA, Upfold L, Ryan PG, Petersen SL, Dyer BM, et al. Recent trends in numbers of four species of penguins at the Prince Edward Islands. Afr J Mar Sci. 2009;31:419–426. http://dx.doi.org/10.2989/AJMS.2009.31.3.15.1003

29. Treasure AM, Moloney CL, Bester MN, McQuaid CD, Findlay KP, Best PB, et al. South African research in the Southern Ocean: New opportunities but serious challenges. S Afr J Sci. 2013;109(3/4), Art. #a009, 4 pages. http://dx.doi.org/10.1590/sajs.2013/a009

30. Minister Edna Molewa declares Prince Edward Islands a Marine Protected Area [statement by the Department of Environmental Affairs]. 2013 Apr 13 [cited 2013 Nov 12]. Available from: http://www.polity.org.za/

31. Lombard AT, Reyers B, Schonegevel LY, Cooper J, Smith-Adao IB, Nel DC, et al. Conserving pattern and process in the Southern Ocean: Designing a marine protected area for the Prince Edward Islands. Antarct Sci. 2007;19:39–54. http://dx.doi.org/10.1017/S0954102007000077

The distribution of the economic mineral resource potential in the Western Cape Province

AUTHORS:
Luncedo Ngcofe[1]
Doug I.Cole[1]

AFFILIATION:
[1]Council for Geoscience, Cape Town, South Africa

CORRESPONDENCE TO:
Luncedo Ngcofe

EMAIL:
lngcofe@geoscience.org.za

POSTAL ADDRESS:
Council for Geoscience, PO Box 572, Bellville 7535, South Africa

KEYWORDS:
mineral resources; industrial mining; deposits; working mines; former mines

HOW TO CITE:
Ngcofe L, Cole DI. The distribution of the economic mineral resource potential in the Western Cape Province. S Afr J Sci. 2014;110(1/2), Art. #a0045, 4 pages. http://dx.doi.org/10.1590/sajs.2014/a0045

South Africa is blessed with vast mineral resources. The formal exploitation of mineral resources in South Africa started with copper mining in Namaqualand (near Springbok in the Northern Cape Province) in the mid-1800s. The discovery of diamonds near Kimberley in 1867 and gold on the Witwatersrand in 1886 led to the significant growth of the mining industry in South Africa. In 1995, it was reported that approximately 57 different minerals were sourced from 816 mines and quarries,[1] with the most important mining provinces being the North West (as a large producer of platinum group metals and gold); Gauteng (gold); Mpumalanga (coal) and the Free State (gold). Although the Western Cape is classified as being the least productive in terms of mineral resources,[1] it has a large potential for the exploitation of industrial minerals. These often neglected resources comprise a highly diverse group of vitally important minerals that are used in a variety of applications ranging from everyday products to highly sophisticated materials. Here we describe economic or potentially economic minerals occurring in the Western Cape in terms of their geological setting.

Geology

The occurrence of a mineral resource is always determined by the geological setting of a region. In the Western Cape Province, the oldest rocks are gneisses and granites of the Mokolian Namaqua-Natal Metamorphic Province (~1100 million years old) exposed north of Vredendal. These rocks are overlain by the Gariep Supergroup rocks, which are approximately 650 million years old, and similar-aged rocks of the Malmesbury Group. The Kaaimans and Cango Groups occur in the southwestern and southern parts of the province, respectively (Figure 1). The Malmesbury and Kaaimans Groups are intruded by the 550–510-million-year-old Cape Granite Suite. The slightly younger Vanrhynsdorp Group occurs in the northwestern part and the Klipheuwel Group in the southwestern part of the province. The rocks of the Table Mountain, Bokkeveld and Witteberg Groups of the Cape Supergroup follow unconformably upon the older rocks described above. The younger Dwyka, Ecca and Beaufort Groups of the Karoo Supergroup were deposited from 300 to 255 million years ago in the northeastern part of the province (Figure 1). Rocks of the Karoo Supergroup and older strata were tectonically deformed during the Cape Orogeny, which finished about 215 million years ago and was followed by uplift and intrusion of a vast network of dykes and sills of the Karoo Dolerite Suite into the Karoo rocks some 180 million years ago. Fluvial sandstone and gravel overlaid by lacustrine clay of the ~145 million year old Uitenhage Group occupy small fault-bounded basins between Worcester and Plettenberg Bay. The youngest geological sequence is the Cenozoic sediments, which consist of fluvial, marine and predominantly windblown sandy deposits. They are assigned to the Sandveld Group on the western coastal plain and the Bredasdorp Group on the southern coastal plain.

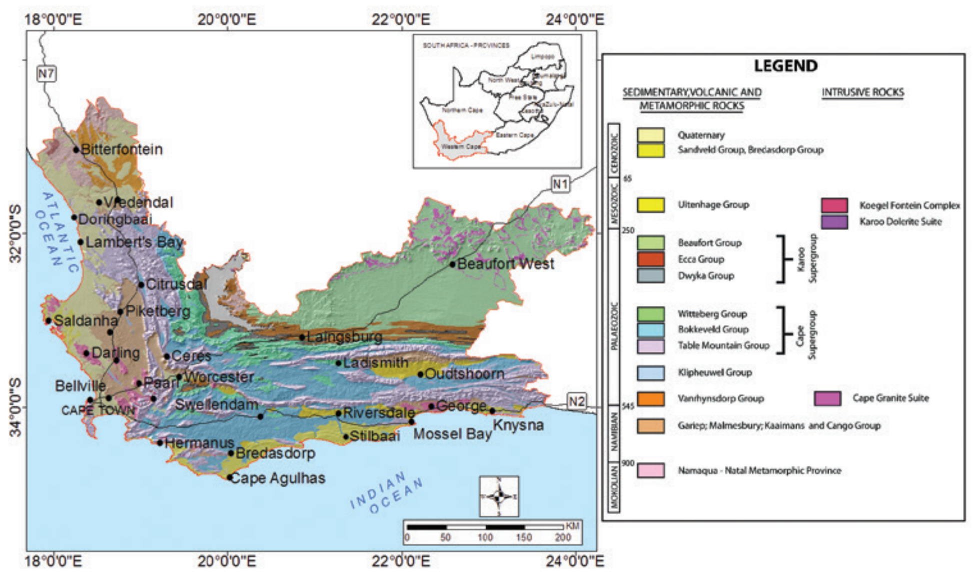


Figure 1: Geological map of the Western Cape Province.

Mineral occurrences

A total of 26 mineral commodities with economic or potential economic variability for exploitation have been delineated in the Western Cape Province.[2] In approximate order of economic importance and present status these comprise: stone aggregate, brick clay, building sand, limestone, dolomite, diamonds, heavy minerals, gypsum, uranium, bentonite, dimension stone (granite, sandstone, marble), rare earth elements, silica sand, plastic clay, salt, phosphate, gravel, kaolin, industrial sand, shale gas, tungsten, mineral pigment, lignite and manganese. Only the first 11 minerals are discussed here. Figure 2 shows the distribution of the discussed eleven minerals in the Western Cape Province. The information about each mineral was derived from the South African Minerals Database (SAMINDABA), metallogenic maps of the region and through field identification and verification processes. Our findings will be used to update the SAMINDABA database.

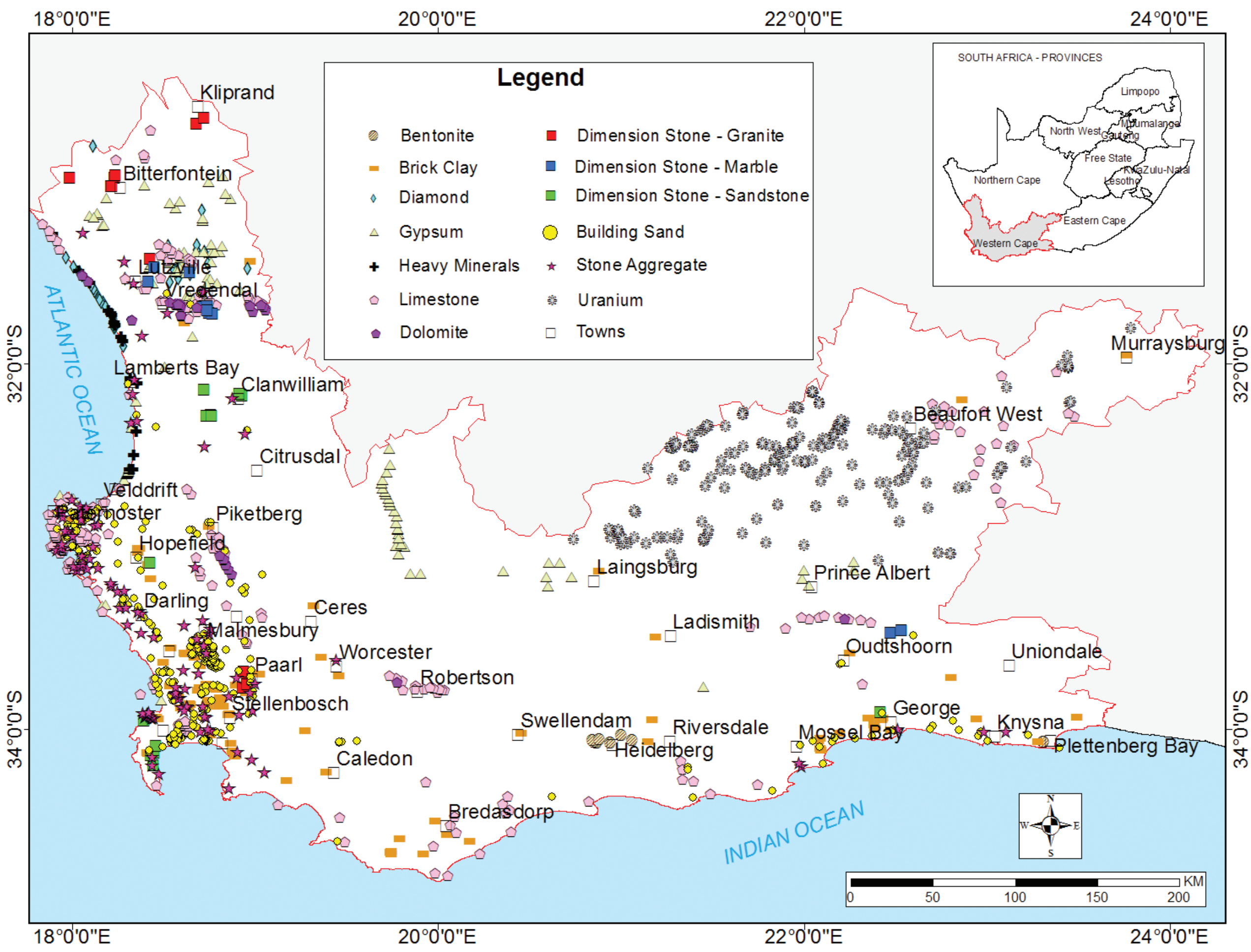


Figure 2: Map of the mineral potential of the Western Cape Province.

Stone aggregate

Stone aggregate is a crushed rock generally coarser than 6.7 mm, largely used for road, concrete and other pavement construction.[3] Quartz-rich rocks produce a good-quality aggregate. In the Western Cape, these rocks are sourced from the quartzitic sandstone of the Piekenierskloof, Peninsula and Skurweberg Formations of the Table Mountain Group,[3,4] granite of the Pan-African Cape Granite Suite and hornfels of the Malmesbury Group.

Brick clay

The short supply of wood and natural building stone led to the use of brick clay as an alternative building material. Today, brick clay is the most predominant building material. Clays suitable for brick-making material must contain the minerals kaolinite, quartz and illite.[5,6] Van Strijp[7] concurs, reporting that kaolinite has a good sintering effect while quartz acts as a stabiliser and illite provides plasticity. In the Western Cape, clays meeting these requirements comprise residual clay of the Malmesbury Group shale of Cape Town, Ceres and Hopefield; residual clay of the Gifberg Group (Gariep Supergroup) schist near Klawer; residual clay of the Saasveld Formation (Kaaimans Group) schist and phyllite in the George area[8,9] and residual clay of the Kirkwood Formation (Uitenhage Group) mudstone near Oudtshoorn and Swellendam.[2]

Building sand

Sand mining is normally short to medium term in duration, creating relatively few job opportunities. However, it contributes significantly to the local and regional economy. Building sand is commonly used for the manufacture of plaster, mortar and concrete.[9] Building sand is distributed over most of the province but is generally absent in areas underlain by Karoo Supergoup sedimentary rocks and dolerite between Laingsburg, Beaufort West and Murraysburg. In the Greater Cape Town area, plaster and mortar sands are obtained from hairpin parabolic dunes of the Holocene Witzand Formation in the Philippi and Macassar regions. Concrete sand is obtained from hillwash deposits in the zones southwest of Malmesbury and northwest of Darling.[2] In the Saldanha and Vredenburg urban area, all three types of building sand are sourced from colluvial and hillwash sands. In the southern part of the province east of Port Beaufort (Figure 2), enormous resources of mortar and plaster grade building sand are derived from coast-parallel dunes, up to 54 m high, of the Strandveld and Wankoe Formations (Bredasdorp Group).[10] Concrete grade sand in the same area is exploited from hillwash, colluvial and alluvial sands.

Limestone/dolomite

Limestone is used for several applications, including cement manufacture, metallurgical flux, paper coating, water purification, steel production and as a neutraliser of acid soils.[2,11,12] Pure limestone is composed entirely of calcium carbonate; when it contains variable amounts of magnesium, it is called dolomite.[11] Limestone and dolomite are discussed together because they form in similar geological environments and are often associated in the field. To a certain extent they have similar uses, although limestone is the more valuable material. Limestone occurs in the western and southern portion of the province and can be divided into high-grade and low-grade categories. The high-grade limestone occurs in the Cango Group north of Oudtshoorn, in the Malmesbury Group in the southwestern part of the province and in the Gifberg and Vanrhynsdorp Groups in the vicinity of Vanrhynsdorp and Vredendal. Low-grade limestone occurs in the De Hoopvlei, Wankoe and Waenhuiskrans Formations of the Bredasdorp Group between Stanford and Mossel Bay and in the Langebaan and Velddrif Formations of the Strandveld Group between Cape Town and Velddrif. Dolomite occurs in the Malmesbury Group near Robertson and in the area southeast of Piketberg and in the Vanrhynsdrop Group in the region of Vredendal and Vanrhynsdorp.

Diamonds

Diamonds are only present in the northwestern part of the province and are hosted by both alluvial and marine placers. Alluvial placers occur in fluvial deposits of the middle Pliocene Quagga's Kop Formation in the Knersvlakte and in fluvial terrace gravels in the Olifants and Hol River valleys.[2] Marine placers occur along the coast northwest of Donkin's Bay in gravels that overlie Pliocene to Pleistocene wave-cut terraces and in gravel within recent offshore gulleys and potholes.[13] The diamonds were derived from an inland kimberlite source via a late Cretaceous 'Karoo River' that connected the diamondiferous kimberlite region of central South Africa to the Knersvlakte and mouth of the Olifants River.[14] The diamonds of alluvium placers are of low grade and no longer exploited.[15] Present mining of the marine placers is concentrated on the diamondiferous gravel that fills gullies, joints and potholes in bedrock on the seafloor.[2,15,16]

Heavy minerals

Heavy minerals provide a source of both titanium, which is hosted in the minerals ilmenite and rutile, and zirconium, in the mineral zircon. Other heavy minerals such as magnetite, monazite, garnet, tourmaline and hematite are commonly present but are not processed. In the Western Cape Province, heavy mineral bearing sands are present along the west coast, north of Dwarskersbos within Quaternary beach and aeolian placers.[17] A total of 16 placers have been discovered but only 2 are economically viable, the others being small or of low grade.[17-19] The largest is Namakwa Sands hosted by aeolian sand some 52 km northwest of Lutzville (Figure 2).

Gypsum

Gypsum occurs in the northwestern part of the province. It is hosted by clay and occurs in the form of selenite crystal aggregate, powder, alabaster and satin spar veins.[2] The gypsum-bearing clay is derived from underlying bedrock of the Namaqua-Natal Metamorphic Province, Gariep Supergoup, Vanrhynsdorp Group and Whitehill Formation. The clay has an average thickness of 1 m but can attain thicknesses of up to 4 m. In the region northwest of Vanrhynsdorp, approximately 4.5 million tons of gypsum has been mined since the 1930s, leaving remaining resources of 13 million tons.[20] Given the rate of production of gypsum, the province will have an adequate supply for the next 100 years. The applications of gypsum are largely in the building industry, as cement retarder, in plaster and ceiling boards, and as a soil conditioner in the agriculture industry.

Uranium

The Western Cape Province has deposits of uranium with economic potential hosted by sandstone of the Beaufort Group between Laingsburg, Beaufort West and Murraysburg, which have never been exploited. During the period between 1976 and 1979 when the price of uranium reached USD110/lb U_3O_8, extensive exploration occurred in the Laingsburg and Beaufort West regions, but the exploration stopped when the price dropped.[2] Currently, the price ranges between USD40 and USD55/lb U_3O_8 and it is presumed that should the price rise to sustainable levels – above USD80/lb U_3O_8 – exploration of this resource might occur again.[15]

Bentonite

Bentonite is a clay-rich mineral consisting mainly of montmorillonite. It is used for a variety of applications, for example, as a wine and fruit juice decolourant, as a binder in foundry sands and as a soil sealant.[5] It occurs in five rift-related basins of the Late Jurassic to Early Cretaceous age in the Cape Fold Belt between Robertson and Plettenberg Bay:

1. Robertson and Swellendam basin – bentonite occurs in the Kirkwood Formation but no prospecting has been carried out.[2]
2. Heidelberg–Riversdale basin – bentonite in the Kirkwood Formation is divided into two units: an upper multicoloured, pale greyish yellow and reddish sandstone and mudstone unit and a lower zone comprising olive-green and greenish grey mudstone and sandstone beds.[5] The lower zone is up to 10 m thick and hosts several bentonite horizons.
3. Mossel Bay basin – bentonite has been prospected with insignificant results.
4. Plettenberg Bay basin – bentonite occurs in lenticular- to pod-like bodies that vary in thickness from 0.15 m to 2.5 m and were mined between 1993 and 2004.[5]
5. Oudtshoorn basin – bentonite occurs in the Uitenhage Group but exploration has not been undertaken.[2]

Dimension stone

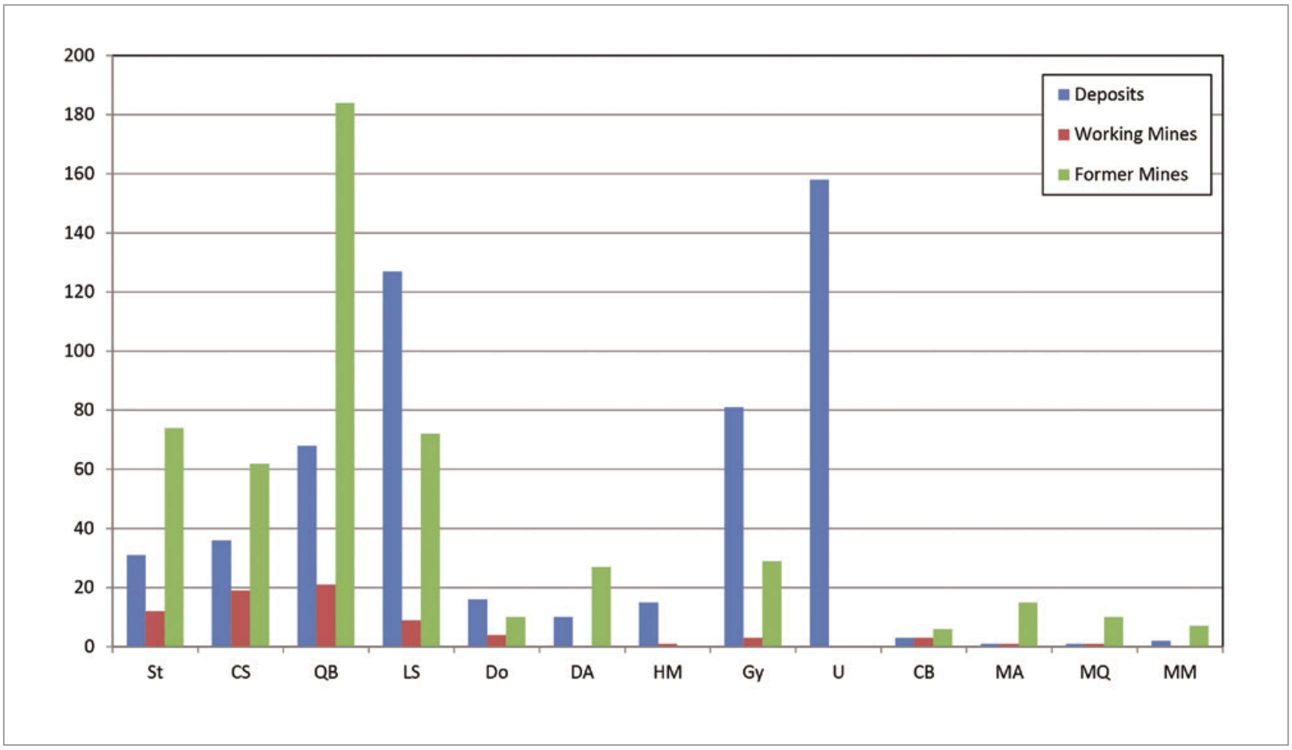
Dimension stone is a collective term for various natural stones used for structural or decorative purposes in construction and monument applications. There are three types of dimension stone that occur in the Western Cape Province: granite, sandstone and marble. These stones are used for building construction, tiles, cladding and memorial art (tombstone). There are two types of granites mined in the province: a light grey, medium-grained granite (known as Paarl Grey), which occurs on the eastern side of the Paarl Pluton, and a light green, charnokitic granite (known as Green Granite) which is located in the northwest part of the province around Bitterfontein.[15] Although no longer popular, sandstone was formerly exploited from the Peninsula Formation (Table Mountain Group) near Cape Town, Piketberg and Vanrhynsdorp, from the Robberg Formation at Mossel Bay and from the Kirkwood Formation at Oudtshoorn for local building stone.[2,21] The marble in the province is a result of the metamorphism of limestone during the Pan African Orogeny.[13] The marble resources are found in the Widouw Formation (Gariep Supergroup) in the northwestern part of the province near Vanrhynsdorp and in the Cango Group in the De Rust area.

Conclusion

The status of the selected resources in term of deposits, working mines and former mines is shown in Figure 3. There are still a number of deposits to be exploited. The extent of former mines does not necessarily relate to the depletion of resources, but is driven rather by market value and demand. The resources of building sand in the Greater Cape Town area are limited and new resources from other regions or crushed sand from the Table Mountain Group sandstone east of this area will have to be utilised. The other construction materials – stone aggregate, brick clay and limestone (for use in cement manufacture) – are in abundant supply. There are approximately 20 000 tons of uranium in the identified resource category, but the uranium price must rise to higher sustainable levels before exploitation becomes viable. The current updated distribution of mineral resources in the Western Cape Province will be vital to town and rural planners, as well as to conservation organisations such as CapeNature, for the delineation of nature reserves so as to avoid sterilisation of economic important mineral commodities.

Acknowledgements

We thank Coenie De Beer, Ishaam Davids and Ponani Mthembi for their valued contributions to this paper.



St, stone aggregate; CS, brick clay; QB, building sand; LS, limestone; Do, dolomite; DA, diamonds; HM, heavy minerals; Gy, gypsum; U, uranium; CB, bentonite; MA, granite (dimension stone); MQ, sandstone (dimension stone); MM, marble (dimension stone).

Figure 3: The status of potentially economic minerals of the Western Cape Province.

References

1. Minerals Bureau. An overview of the South African mineral industry. In: Wilson MGC, Anhaeusser CR, editors. The mineral resources of South Africa: Handbook. Pretoria: Council for Geoscience; 1998. p. 5–10.
2. Cole DI, Ngcofe L, Halenyane K. Mineral commodities in the Western Cape Province, South Africa. Report number 2013-0165. Pretoria: Council for Geoscience; 2013.
3. Roux PL. Aggregates. In: Wilson MGC, Anhaeusser CR, editors. The mineral resources of South Africa: Handbook. Pretoria: Council for Geoscience; 1998. p. 40–45.
4. Cole DI. The metallogeny of the Cape Town area. Explanation and metallogenic map of sheet 3318 (scale 1:250 000). Pretoria: Council for Geoscience; 2003.
5. Horn GFJ, Strydom JH. Clay. In: Wilson MGC, Anhaeusser CR, editors. The mineral resources of South Africa: Handbook. Pretoria: Council for Geoscience; 1998. p. 106–135.
6. Heckroodt RO. Clays and clay materials in South Africa. J S Afr Inst Min Metall. 1991;91:343–363.
7. Van Strijp LT. Brickmaking materials. In: Wilson MGC, Anhaeusser CR, editors. The mineral resources of South Africa: Handbook. Pretoria: Council for Geoscience; 1998. p. 85–89.
8. Roberts DL, Viljoen JHA, Macey P, Nhleko L, Cole DI, Chevallier L, et al. The geology of George and environs. Explanation and geological map of sheets 3322CD and 3422AB (1:50 000). Pretoria: Council for Geoscience; 2008.
9. Cole DI, Viljoen JHA. Building sand potential of the Greater Cape Town area. Council for Geoscience Bulletin. 2001;129:31.
10. Roberts DL, Botha GA, Maud RR, Pether J. Coastal Cenozoic deposits. In: Johnson MR, Anhaeusser CR, Thomas RJ, editors. The geology of South Africa. Pretoria: Council for Geoscience; 2006. p. 605–628.
11. Martini JEJ, Wilson MGC. Limestone and dolomite. In: Wilson MGC, Anhaeusser CR, editors. The mineral resources of South Africa: Handbook. Pretoria: Council for Geoscience; 1998. p. 433–440.
12. Martini JEJ. Limestone and dolomite resources of the Republics of South Africa, Bophuthatswana, Ciskei, Transkei and Venda. Geological Survey 9. Pretoria: Department of Energy and Mineral Affairs; 1987.
13. De Beer CH, Gresse PG, Theron JN, Almond JE. The geology of the Calvinia area. Explanation of sheet 3118 (Calvinia). Pretoria: Council for Geoscience; 2002. p. 92.
14. De Wit MCJ. Post-Gondwana drainage and the development of diamond placers in western South Africa. Econ Geol. 1999;94:721–740. http://dx.doi.org/10.2113/gsecongeo.94.5.721
15. Cole DI. The metallogeny of the Calvinia area. Explanation and metallogenic map of sheet 3118 (scale 1:250 000). Pretoria: Council for Geoscience; 2013. p. 77.
16. Lynn MD, Wipplinger PE, Wilson MGC. Diamonds. In: Wilson MGC, Anhaeusser CR, editors. The mineral resources of South Africa: Handbook. Pretoria: Council for Geoscience; 1998. p. 232–258.
17. Cilliers LM. The geology of the Graauwduinen heavy mineral sand deposit, west coast of South Africa [MSc thesis]. Stellenbosch: Stellenbosch University; 1995.
18. Palmer G. The discovery and delineation of the heavy-mineral sand ore bodies at Graauwduinen, Namaqualand, Republic of South Africa. Explor Min Geol. 1994;3:399–405.
19. Wipplinger PE. Titanium. In: Wilson MGC, Anhaeusser CR, editors. The mineral resources of South Africa: Handbook. Pretoria: Council for Geoscience; 1998. p. 621–632.
20. Cole DI. Revised report on the mineral potential of the proposed Knersvlakte Biosphere Reserve. Report number 2012-0228. Pretoria: Council for Geoscience; 2012.
21. Wybergh W. The building stones of the Union of South Africa. Memoir, Geol Surv S Afr. 1932;29:244.

Using SNPs to find my roots

AUTHOR:
Brenda Wingfield[1]

AFFILIATION:
[1]Faculty of Natural and Agricultural Sciences, University of Pretoria, Pretoria, South Africa

CORRESPONDENCE TO:
Brenda Wingfield

EMAIL:
brenda.wingfield@fabi.up.ac.za

POSTAL ADDRESS:
Faculty of Natural and
Agricultural Sciences, University
of Pretoria, Pretoria 0002,
South Africa

KEYWORDS:
genetics; ancestry; genome; single nucleotide polymorphism; lineage

HOW TO CITE:
Wingfield B. Using SNPs to find my roots. S Afr J Sci. 2014;110(1/2), Art. #a0046, 1 page. http://dx.doi.org/10.1590/sajs.2014/a0046

South Africa has been a true democracy for almost 20 years. There have been many positive developments in this time; some (including those relating to science) seem almost miraculous when one considers for instance some of the consequences of the Arab Spring. Despite our two decades of freedom from apartheid, we sadly remain a substantially racially divided nation. As a scientist with a strong interest in genetics, my own racial heritage has always been of interest to me. Although sequencing one's own genome remains inordinately expensive for the average person, determining one's single nucleotide polymorphism (SNP) genotype has become relatively inexpensive. And it is against this background that I now have confirmation of who my own ancestors were.

It is almost fashionable in South Africa to be able to claim a southern African ancestor. My DNA analysis shows that while I can claim to be 1% African, there is nothing to suggest that any of my ancestors ever lived in southern Africa. This finding must clearly be viewed against the backdrop of the fact that humans originated in Africa. My ancestors obviously left Africa a very long time ago! I can, however, report that I share 2.9% of my genome with Neanderthals, for whatever that is worth. My current theory – and I am sticking to it – is that Neanderthals were blonde!

I grew up with the knowledge that my great-great-grandfather, 23 generations back, was Robert the Bruce. For those not raised on European history, he was King of Scots in 1306. Much time has passed since he was king and I do not have any claims on the Scottish throne nor is there any evidence of any family treasure to which I might have rights. However, my father was very proudly Scottish. My mother can trace her English ancestors back to the Tudors and her mother was Irish. It is thus not at all surprising that my DNA profile is 50–70% English or Celtic and that overall I am 98% European with a very small Northern African and Asian influence.

My mitochondrial genome is also one of the European lineages. In his popular book *The Seven Daughters of Eve*, Bryan Sykes[1] named the originator of my mtDNA haplogroup Jasmine. I obviously come from a very long line of European women and I have passed my mitochondrial genome on to my children.

The story regarding my 'roots' is exactly what I expected. But it was interesting (somewhat reassuring) to have it confirmed and to understand some of the other aspects of my gene pool. It was also fascinating to learn how much we know already as a result of Genome Wide Association Genetics. Based on my SNP genotype, I am predicted to be blonde and to have blue eyes. The fact that this reflects my phenotype exactly is both reassuring and also shows how much one can tell from just a million SNPs.

Testing for genetic diseases based on SNP linkage has become a component of the commercially available tests that the medical profession is already using. It is thus important that we come to understand more about genetics and our own genomes. Yet, there are also some complex consequences to having access to all these data. For example, in the South African context, the availability of these data may force a re-evaluation of our various racial classifications. In parts of the USA, the 'one drop rule' has been applied to race classification – any mixed race ancestry determines one's race. I would suggest that having one's genomic data available vividly illustrates how pointless these various classification systems really are. We are all probably a bit of a mixture – some, such as myself, a bit more 'European', while others are finding unexpected 'roots' to celebrate.

The poet Langston Hughes[2] wrote in his 1940 memoir

> *You see, unfortunately, I am not black. There are lots of different kinds of blood in our family. But here in the United States, the word 'Negro' is used to mean anyone who has any Negro blood at all in his veins. In Africa, the word is more pure. It means all Negro, therefore black. I am brown.*

A colleague of mine who has a mixed-race son once told me that, in her homeland, her child is 'neither White nor Black but rather just beautiful'. The data of our genomes will ultimately be available to us all. And hopefully they will emphasise the fact that the human race is something truly beautiful to behold.

References

1. Sykes B. The seven daughters of Eve. London: Bantam Press; 2001.
2. Hughes L. The big sea, an autobiography. New York: Knopf; 1940.

Entomological evidence of misdemeanour

BOOK TITLE:
Insect evidence: Basic collection procedures at the death scene

AUTHOR:
Melvin R. Bishop

ISBN:
9781436358132 (softcover)

PUBLISHER:
Xlibris Corporation, USA, USD19.99

PUBLISHED:
2008

REVIEW TITLE:
Entomological evidence of misdemeanour

REVIEWER:
Martin H. Villet

EMAIL:
m.villet@ru.ac.za

AFFILIATION:
Southern African Forensic Entomology Research Laboratory, Department of Zoology & Entomology, Rhodes University, Grahamstown, South Africa

POSTAL ADDRESS:
Department of Zoology & Entomology, Rhodes University, PO Box 94, Grahamstown 6140, South Africa

HOW TO CITE:
Villet MH. Entomological evidence of misdemeanour. S Afr J Sci. 2014;110(1/2), Art. #a0050, 1 page. http://dx.doi.org/10.1590/sajs.2014/a0050

The face of publishing is indisputably changing. As a print-on-demand publication, *Insect Evidence* is interesting for reasons besides its content. The Internet has freed us of many constraints when it comes to publishing – a situation that is exemplified by the popular fanfiction and academic open-access movements. Fanfic has an effective quality-assurance mechanism, because poor writing receives no encouragement from its reading community, while academic publishing has a two-stage system in which peer reviewers and editors perform an initial gatekeeping role, and the citation system is a secondary mechanism that functions more like that of fanfic. The print-on-demand publishing service industry allows authors to be their own publishers, thus bringing the fanfic model to academic publication.

Under this print-on-demand publication model, the publisher supplies information about a client's opus on their website and on the sites of large online book suppliers, and when a copy is ordered online, they reproduce and deliver it. The author receives a royalty and retains control of the design and rights to their work, which can be ordered at any time, from anywhere, more or less in perpetuity. Some online book suppliers allow you fanfic-like means to submit voluntary reviews on each item, and provide links to cognate works based on keywords and customers' buying patterns. The missing ingredients are the editing services and pre-publication review of traditional academic publishing, but it is not clear if that matters. Various authors[1-3] have raised alarm about the reliability of the peer-review and editorial processes of academic publishing, and there is a voluminous literature about what citations might actually indicate about academic merit.[4]

Bishop is qualified in his topic by both experience and training. He graduated from the Virginia Forensic Science Academy and has done research on nocturnal oviposition of blowflies under artificial lighting at the Department of Entomology, University of Nebraska. He is now a forensic consultant who retired with 23 years of service as a Master Police Officer and Forensic Technician with the Charlottesville VA Police Department. He has assisted with forensic entomology in more than 70 criminal and civil cases between 1995 and 2003, although reportedly rarely travels to the scenes. He taught forensic entomology to graduate and professional audiences in Virginia and Florida, which is reflected in the approach of this book.

This book is effectively a concise, illustrated training manual, along the lines of Catts & Haskell's[5] longer, spiral-bound classic. It contains chapters on the nature of forensic entomology and of death scenes; general etiquette, protocols and equipment for approaching and sampling outdoor death scenes for insect evidence and environmental conditions; additional considerations for indoor and mortuary cases; the general life cycle of blow flies; and layman's accounts of the flies and beetles that one may expect to collect in North America. The presentation is process oriented and includes mnemonics for the stages of the protocols. The science is informal and monochrome photographs (some reproduced in colour at http://www.forensic-topics.com/) clarify explanations. As a training document, this book has a very clear target audience which it addresses with the right tone and level of detail. It is available in three formats – hardcover, softcover and e-book. My well-printed and robustly bound softcover arrived directly from the publisher by post unscathed and promptly. Prices vary by format and some suppliers undercut the publisher.

This book is ideal for its purpose, and yet it is not cited in the rapidly growing literature about forensic entomology. To be fair, the subject is specialised, but even so, there have been citation opportunities. I discovered this book about 5 years after it was published, on a tangentially related Internet search. Although it is marketed by, among others, Amazon and Barnes & Nobel, the book did not appear within the first 10 pages of hits generated by their search engines. Clearly, one of the biggest challenges of publishing academic works privately lies in advertising.

If the possibilities of print-on-demand publication inspire you, there is a caveat: do your homework. While *Insect Evidence* is a wholesome example of print-on-demand, there are counterexamples with a different reputation. Some publishers using print-on-demand take advantage of permissive Creative Commons and ShareAlike licences to sell anthologies of articles from *Wikipedia*, *Wikia* and even the *South African Journal of Science* as books through well-known online book suppliers. The cover of one book on forensic DNA analysis (selling for US$56) announces boldly "High quality content by Wikipedia articles!". Authors publishing privately on demand should weigh up the advantages of professional editing, peer review and marketing.

References

1. Epstein WM. Confirmational response bias among social work journals. Sci Technol Hum Val. 1990;15:9–38. http://dx.doi.org/10.1177/016224399001500102
2. Sokal AD. A physicist experiments with cultural studies. Lingua Franca. 1996;6:62–64.
3. Bohannon J. Who's afraid of peer review? Science. 2013;342(6154):60–65. http://dx.doi.org/10.1126/science.342.6154.60
4. Wang D, Song C, Barabási A-L. Quantifying long-term scientific impact. Science. 2013;342:127–132. http://dx.doi.org/10.1126/science.1237825
5. Catts EP, Haskell NH. Entomology and death: A procedural guide. Clemson, SC: Joyce's Print Shop; 1990.

Academic freedom as a keyword

BOOK TITLE:
Academic freedom in a democratic South Africa: Essays and interviews on higher education and the humanities

AUTHOR:
John Higgins

ISBN:
9781868147519

PUBLISHER:
Wits University Press, Johannesburg, R290

PUBLISHED:
2013

REVIEW TITLE:
Academic freedom as a keyword

REVIEWER:
Lis Lange

EMAIL:
LangeML@ufs.ac.za

AFFILIATION:
Directorate for Institutional Research and Academic Planning, University of the Free State, Bloemfontein, South Africa

POSTAL ADDRESS:
Directorate for Institutional Research and Academic Planning, University of the Free State, PO Box 339, Bloemfontein 9300, South Africa

HOW TO CITE:
Lange L. Academic freedom as a keyword. S Afr J Sci. 2014;110(1/2), Art. #a0051, 2 pages. http://dx.doi.org/10.1590/sajs.2014/a0051

John Higgins's *Academic Freedom* is a worthy example of what Said refers to in the book as the fugal and contrapuntal approach to literacy (p. 219). The author's attention to the simultaneity of voices is not unlike Glen Gould playing Bach and produces a similar effect: a will to listen more and to engage further.

Academic Freedom is organised in two parts. Part One consists of five essays published by the author in different formats between 1987 and 2011. Two of them focus directly on academic freedom (Chapters 1 and 2), while the other three focus on different aspects of the position of the humanities in higher education policy and practice in the last 30 years. Part Two consists of interviews with Terry Eagleton, Edward Said and Jakes Gerwel, all of whom have directly contributed to the debate about the humanities and the role of universities and intellectuals in society. Finally, the book closes with a conclusion that, instead of summarising already stated arguments, presents the issues at stake against the current policy and political context in South Africa. A foreword by John Coetzee in the form of a letter to the author adds flair to a text that is fundamentally dialogical.

Like all good writing Higgins's *Academic Freedom* is susceptible to several different readings. I would like to propose four: an historical reading, a political reading, a pedagogical reading and an intellectual reading.

In the historical reading Higgins deals with academic freedom as the construction of a problem from the debates about De Klerk's Regulations of 1987 to the launch of the Charter for the Humanities (2011) and the amendments to the *Higher Education Act* introduced by the Minister of Higher Education and Training in 2012. The local chronology might be misleading. Far from concentrating on South African higher education, Higgins explores the changing relationship between higher education and the state and higher education and society as part of the global expansion and influence of neo-liberalism. From this perspective Higgins navigates through fundamental higher education texts contraposing and combining the detailed analysis of policy and policy reports with specialised higher education literature. The main concern here is the seemingly easy transition from a conception of higher education 'as a sub-set of the political system to a sub-set of the economic system' (p. 156) and the instrumentalist conceptualisation of higher education which arises from this. In doing this, Higgins does not stop at the door of governments, states and markets, but goes inside higher education institutions, exploring how the overriding instrumental motive manifests itself in the tensions between universities' management and academics. The continuum in this historical reading of Higgins's work is the impact that these changes have had on the humanities disciplines in terms of self-definition as well as for their position, valorisation and funding within universities. Higgins's argument is far from simplistic, antiquated or romantic. He dissects the contradictions and tensions of 21st century higher education, particularly in the context of post-apartheid South Africa, at the same time that he makes the case for the humanities as eminently necessary disciplines for the 21st century informatics society.

A political reading of *Academic Freedom* brings us back to Said's Reith Lectures (1973) on the representations of the intellectual and his/her role in 'disturbing the easy flow and circulation of received ideas' (p. 2). Higgins's minute dissection of the notion of 'occasioned' writing lays out the difficulties of a type of writing which is developed academically but that is stated for and in public. Each one of the essays comes from a combination of what the American Association of University Professors calls in its definition of academic freedom 'extramural utterances' and rigorous academic writing; thus newspaper articles and public lectures are woven through Higgins's text. The point of this writing is to take a position, which Higgins does in every chapter. The political reading of *Academic Freedom* is confirmed in the final counterpoint between the arguments contained in the five essays and the three interviews with academics who are also intellectuals and whose participation in public debate and political life has been exemplary. The notion that the humanities disciplines are necessary for democratic life because they educate people in the art of dissecting discourse, constructing arguments and asking questions about the questions, argued throughout the book, speaks, in my view, to the fundamental bastion of the political in higher education: the conceptualisation of education itself as a political act in an agoric sense and it serves to show in practice the social value of critical literacy.

Academic Freedom is also susceptible to a pedagogical reading. This reading is particularly true of the five essays but can also be detected in the conversation about education that takes place in all three interviews. From Chapter 1 onwards, in subsequent levels of complexity and explicitness, Higgins proposes a humanities pedagogy. He usefully argues that critical literacy is the name of the project 'standing behind or within' the common ground of all humanities disciplines (p. 83), a notion that some local academics still find difficult to grasp. From the specific pedagogical perspective, Higgins not only analyses text and discourse, he shows how to do it. In chapter after chapter, Higgins deploys the three dimensions of textuality, theory and history that constitute critical literacy as exemplars of what he means. The analysis of institutional culture as a keyword, the observations about the difference between statement and address in policy documents (and in any argument for that matter) and the introduction of the NAIL (narrative, analysis, interpretation and language) disciplines are all teaching moments, as is his minute analysis of the debates on academic freedom in Chapters 1 and 2. Inevitably, the pedagogical moment is also a political moment as Higgins insists on the importance of understanding the uneven distribution of cultural capital among university students and the possibility of looking at universities' 'institutional cultures' from the point of view of the 'pedagogic culture and its forms of transmission' (p. 127). Once again the pedagogical reading is sustained in the interviews as all three interviewees use their experience as university teachers to reflect on the difficulties of teaching critical literacy to the millennial generation.

My last reading of Higgins is intellectual in the sense of the author's intellectual perspective in the analysis of the problem at hand. The analysis presented in *Academic Freedom* is unambiguously and unashamedly materialistic. Higgins's own deployment of the theoretical, the historical and the textual is grounded in the material connection between culture and society. Raymond Williams's oeuvre, on which Higgins has worked extensively, becomes a primary frame of reference for the analysis of the humanities in the context of academic freedom. The exploration of institutional culture as a keyword (Chapter 4) is probably the best, but by no means the only, example of this. The notion of keyword as 'an item of contested vocabulary in a conflicted and disputed social process' (p. 103) sets the scene for an exploration of the manner in which words and concepts change meaning under the pressure of social and political change. Both text and footnotes rely on Marx and historical materialism as still useful lenses to interpret reality; thus we move from *The German Ideology* to the *Critique of Hegel's Theory of the State* and the *Theses on Feuerbach* as part of the same argument. The reader should not be misled to think that what Higgins's text offers is a simplistic Marxist analysis of a social problem. This materialist approach engages with Bourdieu, Derrida and Foucault as much as with more recent postcolonial theory. As in the other proposed readings of *Academic Freedom*, in the intellectual reading the five essays and the interviews operate in a fugue structure and Eagleton, Said and Gerwel, from different personal experiences, multiply the reflection on the value and need of thinking with Marx against Marx, to reappropriate a phrase coined by Benhabib apropos of Arendt. Of particular interest for South African reflection is the dialogue with Gerwel about race and class in local politics in a post-socialist world.

It is possible and necessary to argue with the author from a variety of points of view in order to explore further the implications of some of the notions and definitions proposed, to examine again the role of whiteness and given epistemologies in institutional culture, to introduce the possibility that the author's conceptualisations are very marked by his South African institutional context, and to explore further the permanent tension in the definition of the university and its role in society under different historical moments. Yet this does not detract from the quality of the argument or the importance of the book. As with all good arguments *Academic Freedom* is bound to induce debate and provoke the possibility of public deliberation. At a time when there is not sufficient public debate in the country as to the suitability of our interpretive schemes to understand and act in our society, Higgins's work is a good incentive to reflect on the extent to which academia can hold on to the double responsibility of interpreting the world and helping to change it, and on the political conditions under which this might be possible.

A new look at demographic transformation: Comments on Govinder et al. (2013)

AUTHOR:
Nico Cloete[1]

AFFILIATION:
[1]Centre for Higher Education Transformation, Cape Town, South Africa

CORRESPONDENCE TO:
Nico Cloete

EMAIL:
ncloete@chet.org.za

POSTAL ADDRESS:
Centre for Higher Education Transformation, PO Box 18094, Wynberg 7824, South Africa

KEYWORDS:
higher education; demographics; equity indices; massification; knowledge production

HOW TO CITE:
Cloete N. A new look at demographic transformation: Comments on Govinder et al. (2013). S Afr J Sci. 2014;110(1/2), Art. #a0048, 4 pages. http://dx.doi.org/10.1590/sajs.2014/a0048

I noted when I read the draft of Govinder et al.'s[1] paper on equity indices that it equated equity with transformation, and delinked equity from development and performance. This draft version of the paper fell into the trap of a prevailing South African condition: using transformation as a code word for race. Furthermore, the formula used in the paper produced a result in which several of the most equitable institutions were those being run by a government-appointed administrator. By this, the authors implied their promotion of high equity, yet also regarded the existence of dysfunctional institutions as a given in their proposed model for the South African university system.

The paper on equity indices, published in the *South African Journal of Science*[1], certainly responds to both these criticisms. Firstly, equity is used mostly in reference to the formula as described in the paper,[1] although the focus of equity is racial, being mainly African. Secondly, a serious attempt is made to reconcile the well-known Harold Wolpe tension between equity and development, as described by Cloete et al.[2] While I will argue that the attempt is not entirely successful, the approach of developing empirical indicators to reflect the equity–development duality of transformation is to be lauded as it is a step towards developing South African indicator-based performance clustering systems. My time spent at the Shanghai Jiao Tong University's Centre for World Class Universities during early November 2013 has made it even clearer to me that, while for the foreseeable future the Jiao Tong type of methodology will continue to make a considerable contribution to debate and controversy, it will not assist much in strengthening universities in Africa.

Govinder et al.[1] are correct when they assert that equity-weighted research output goes beyond the Centre for Higher Education Transformation (CHET) clusters,[3] which were based mainly on performance and efficiency in knowledge production. The more recent CHET clustering[3] in 2010 has been expanded to include factors such as staff qualifications, undergraduate-to-masters graduation rates and high-level knowledge production (doctorates and research publications). This latest CHET clustering has shown that, in addition to those usually in the top group of higher education institutions (such as UCT, Stellenbosch University and Wits University), some 'on-the-move' institutions, such as UKZN, North-West University and the University of the Western Cape, have moved into the top group.[3]

Govinder et al.[1] are also correct in pointing out that some of their results do not square up with the CHET differentiated clusters[3]. For example, their high rating for Unisa – in terms of both the graduation Equity Index and the weighted research output – is completely contradictory to the performance of Unisa in the South African system as shown by CHET. Similarly, their low ranking for Rhodes University is contradictory to the CHET finding that Rhodes is one of the three most efficient knowledge producers in terms of weighted publication per staff member. It appears that by not using staff:research ratios, the Equity Index formula has skewed results in favour of larger institutions.

Stellenbosch University, amongst others, can be used to illustrate the difficulty of finding a measure that adequately combines equity and development. Stellenbosch comes last in the equity indices for students and staff, and ninth in the equity-weighted research output.[1] However, CHET has shown that Stellenbosch has the highest undergraduate and doctoral throughput in the South African system. For 3-year degrees, 68% of students graduate after 6 years at Stellenbosch. UCT is second with 64%, while the national average is 40%.[3] At the doctoral level, Stellenbosch's graduation rate after 7 years is 71%. Here, Wits shows in second place at 69%, while the national average is 46%.[4] Proportionally, Stellenbosch also produces the most female doctorates[4]; however, gender does not feature in the racially orientated Equity Index. The African Doctoral Academy at Stellenbosch has 60 students from five sub-Saharan African countries,[5] but black Africans from countries other than South Africa also do not count for equity on the Equity Index.

The role of Africans from countries other than South Africa in academia is becoming a sensitive issue, and it has some resonance with demonstrations of township competition between local and foreign traders. In the Govinder et al.[1] report, it seems rather disingenuous to exclude the Africans from countries other than South Africa when calculating student demographic ratios, but to include them when counting publications, especially because at certain institutions – such as UKZN, Fort Hare and North-West – publications by black Africans are substantially by Africans from countries other than South Africa. In recent presentations, CHET has highlighted the fact that in 2010, for the first time in history, there were more black African than white doctoral students enrolled in South African higher education.[3] Instead of expressions of delight at this emerging trend, one usually hears the murmured lament that 'The majority are foreigners'. We seem to be reaching a rather indefensible position in which we count black Africans who are not from South Africa only when it suits us!

I leave further comment on the statistical and the demographic methodologies of the Equity Index to other contributors to this issue of the *South African Journal of Science*, and focus in the rest of this response on the educational and political implications.

In terms of the nature of higher education, there appears to be a fundamental flaw in the Equity Index assumption that the university should be a mirror of national demographics. The university is a specific institution in society that is supposed to lead rather than reflect society. A forthcoming book by Castells and Himanen[6], in discussing Amartya Sen's *Development as Freedom* (1999) and John Rawls' *A Theory of Justice* (1971), highlights the argument that, while all citizens are equal before the law and are all entitled to dignity, this is not the case in terms of capability, particularly if capability is understood as performance rather than potential.

In almost all countries, educational performance – capability – is skewed because of historical contestations and struggles, with socio-economic class showing as a worldwide distorter of representivity. In the long term, it is part of the South African universities' developmental role to play a part in redressing these distortions, within the broader context of debate and policies on affirmative action. Nonetheless, it is generally accepted that this is a long-term and secondary task. The first task of universities is to enrol and educate the most educationally capable – those with the highest educational attainment – in order to contribute to development. The first question that must be asked is thus whether the universities are reflecting educational attainment.

In their conclusions, Govinder et al.[1] ask whether the reasons behind the slow progress in transformation of higher education are passive resistance, denial, the abuse of autonomy or an abhorrence of accountability. The assumption that the lack of transformation is simply the result of a bad attitude is a common South African form of accusatory politics. This kind of thinking assumes that there is a university-ready pool of applicants reflective of racial demographics and that they are not admitted to top-performing institutions because of prejudice and a bad attitude. In reality, in certain areas such as doctoral enrolments, overall enrolments grew by 149% between 1996 and 2011; however, the enrolment of black African students exploded over this same period by 795%.[4] This growth is not slow change: no other country in the world has been identified to have had such a dramatic change in equity. There is also ample evidence, such as that provided by Wickham[7], that the system is already admitting candidates who are not educationally prepared for university study. And, while the whole university system must accept blame and take more responsibility for poor school performance, also implicated in this failure are the national education system and the government.

Another assumption underpinning the arguments of the authors is that the slow progress is as a result of a lack of institutional compliance. Not once is the question raised as to the role of the national Department of Higher Education and Training and its contribution to the problem.

As research director for the National Commission on Higher Education (NCHE) over the period 1994–1996, I was part of the ongoing equity-development debates, both within the Commission, and between the Commission and the then Department of Education. It is widely accepted that the NCHE was essentially an equity commission; development, knowledge production and differentiation were raised but did not feature in the final report.[8]

Furthermore, although equity was dominant in the report, there was no unanimity about how to redress it. One redress suggestion was to award a disadvantage subsidy from the government block grant for each black student enrolled. This process would serve as an incentive for historically advantaged universities to enrol more black students and offset some lost tuition fees. Furthermore, for the historically disadvantaged universities whose enrolments were almost 100% black, a disadvantage subsidy would have been a redress bonus. The group supporting this recommendation in the NCHE even made financial projections based on different scenarios; and it seemed a simple-to-implement and affordable redress mechanism. However, another group, led by the historically black university vice-chancellors in the NCHE, wanted institutional – rather than individual – redress. The incentive group abandoned their proposal when it became clear that the Department of Education leadership, headed by a minister who was also a former historically black university vice-chancellor, was also not supporting their position. Of course, the Education Minister never expected that the Ministry of Finance would turn a deaf ear to the institutional redress pleas. Apparently Treasury rejected the institutional redress proposals owing to, amongst others, a combination of the 1996 currency crisis[9] and a lack of confidence in the institutional absorptive capacity of the historically black universities.

The second redress argument was that of the massification of the post-secondary system. The NCHE was heavily influenced by Peter Scott's book, *The Meanings of Massification*[10], which appeared in 1995, soon after the NCHE began operating. Scott asserted that in the evolving knowledge society, massification of higher education was inevitable and was already happening in most advanced countries. Even the United Kingdom, with its elitist system, had by then taken a decision to massify: it increased participation from under 15% in the late 1980s to over 40% in 2002, simultaneously reducing the cost per student.[11]

In essence, a massified and differentiated system requires a dramatic increase in higher education participation, while also accommodating top-end research universities. The knowledge economy/society needs much larger numbers of post-school educated citizens, both for skills and for democratic citizenship. Differentiated massification was thus the possible resolution to the contradiction between equity and development.

The NCHE accepted the massification argument; however, in a rereading of the 1996 report[8], it is clear that it could have done a much better job of explaining and promoting it. If truth be told, the Commission itself was not that clear about how it should be done and what the implications could be. And, of course, with the strong presence of the vice-chancellors of the historically black universities, differentiation was a taboo topic.

Massification was rejected by both the Ministries of Finance and Education. It is disappointing that neither the Hegelian liberals nor the Marxist revolutionaries could grasp the dialectic. The 1997 White Paper instead proposed planned growth – a decision which had serious, unanticipated consequences. The first consequence was that the higher education system in South Africa remained elite. Overall gross participation increased from around 14% in 1996 to only 19% in 2011.[12-15] While this figure puts South Africa third in sub-Saharan Africa (behind Mauritius and Botswana), only South Africa and India are under 20% amongst the BRICS countries.[16] Countries in the World Economic Forum innovation (knowledge) economies are now almost all at over 60% post-secondary participation rates and many, like South Korea, are at over 80%.[16]

The one consequence of a low overall participation rate is that, even if the proportion of overall enrolments grows, the participation rate does not necessarily increase significantly. Figure 1 shows that for black Africans, head count enrolments increased from 53% in 1996 to 69% in 2011, while the participation rate only increased from 9% to 16%.[12-15] In contrast, for whites, the enrolment percentage declined from 34% to 19% but the participation rate only dropped from 57% to 56%.[12-15]

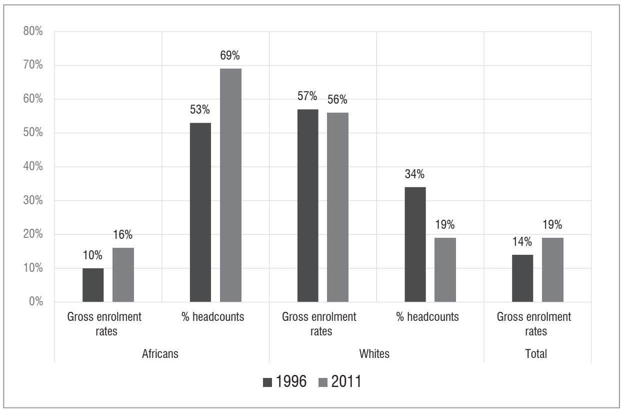


Figure 1: Gross enrolment rates in higher education for black African and white students, in 1996 and 2011.

When looking at participation rates, it is important to take changes in population growth into consideration. The white population in the 20- to 24-year age cohort declined from 349 102 in 1996 to 316 262 in 2011[12-15] – a 10% decline (Table 1). In contrast, the black African population in this age cohort increased by 912 444[12-15] or 29% (Table 1). To increase the participation rate of black Africans to be at the same level as that of the white population (56%) in 2011[12-15], an additional 1.63 million black African students would have needed to be enrolled in 2011. This means that the system would have needed 2.8 times its current capacity. With the current size of the South African higher education system, even if *all* the students were black, their participation rate would be only 23%!

Table 1: Gross enrolment rates in higher education by race, in 1996 and 2011

	1996			2011		
	Enrolments	**20–24 year olds**	**Gross enrolment rate**	**Enrolments**	**20–24 year olds**	**Gross enrolment rate**
African	308 104	3 153 083	10%	646 829	4 065 527	16%
White	198 904	349 102	57%	177 365	316 262	56%
Indian	37 118	103 123	36%	59 312	110 667	54%
Coloured	32 742	344 373	10%	54 698	404 336	14%
Total	**576 868**	**3 982 353**	**14%**	938 204	4 896 792	19%

The definition of the gross enrolment rate as determined by UNESCO is:

$$\frac{\textit{Total number of enrolments in higher education}}{\textit{Population size in the 20–24-year-old age cohort}} \textit{ x 100\%}$$

The real indicator of equality is participation rate and not percentage of enrolments. As a result, having a significant improvement in percentage of enrolment does not reflect a major improvement in equality. The damning Council on Higher Education report shows that more than half of all first-year entrants never graduate, which means that for greater equality it is not only the actual numbers who enrol, but also the success rate that needs to be considered.[17]

The Govinder et al.[1] formula uses a version of participation rate. However, in my view, it is used incorrectly as it is only applied at an institutional level. An improvement in participation rate (equality) is both a system and an institutional issue, and will not be corrected by identifying a few individual, institutional scapegoats.

The most disastrous unintended consequence of planned growth was revealed in a 2009 CHET[18] report which showed that, in 2007, there were 2.7 million young people in the 18–24-year-old cohort who were not in employment, education or training. By 2011, this figure had grown to around 3 million or about 40% of the cohort, and there are more than three times more young people not in employment, education or training than the 950 000 students in the public and private universities.[12-15]

The leader article in the *South African Journal of Science* May/June 2013 issue,[19] addressing the problem of Generation Jobless, concludes:

> *Here, then, is the timeline: In 2009, CHET reported that 2.7 million young people between the ages of 18 and 24 were NEETs. The immensity of the problem was covered in the local and international press (including the New York Times). In 2011, the number of NEETs had grown to 3.2 million, by which time work was in progress on what was hoped would turn out to be a relevant Green Paper. Now, 4 years after the problem was identified and made public, nothing practical has been done by the Department of Higher Education and Training to implement current solutions. The numbers of NEETs continue to grow and there is nothing available to address the present problem. The solution proposed for the future will take, at best, many more months to finalise and a good number of years, and large sums of state funds, to implement. So many years wasted; so many opportunities wasted. Time for the Ministry to focus more earnestly on the well-being of young people and the economy.*

The above quotation demonstrates fairly dramatically that the more serious problem is systemic rather than individual institutional change. Furthermore, the Department of Education – despite all its rhetorical or ‘symbolic’ policy[20] – has yet to implement a policy plan to incentivise or sanction the enrolment of black students in South African universities. In overall figures, the rather remarkable increase in the enrolments of black African students was achieved through individual institutional strategies, aided by the first recommendation of the NCHE to establish a national student financial aid scheme, along with the substantial expansion of the scheme by the Department of Higher Education and Training and the inclusion of further education and training colleges in the funding scheme.

In a significant departure from previous Department of Education polices, the National Planning Commission background paper[21] and the subsequent National Development Plan 2030[22] came out categorically in favour of South Africa joining the knowledge economy through massification and differentiation. It proposed a dramatic increase in post-secondary school enrolments, mainly in the further education and training college sectors. The National Development Plan[22] envisages a 30% participation rate for universities by 2030, with enrolments at around 1.62 million by that time. It recommends a participation rate of 25% in further education and training colleges, which would accommodate about 1.25 million enrolments compared to the current 300 000.

The task is thus to build a new post-secondary differentiated higher education system with built-in quality checks. This system should include a mix of research-led universities, universities that are mainly undergraduate teaching institutions, a further education and training college sector that is mainly post-matric and vocationally orientated, and a private sector that is market driven.

It is the development of a differentiated and massified post-secondary system that will dramatically expand participation for the majority and provide skills[23] for an economy that needs increasingly larger numbers of people with post-matric education. The unintended consequence of the Equity Index of the Transformation Oversight Committee[24] could be an over-focus on equity for a privileged elite at precisely the moment that the central challenge for higher education is to support development, with increased equity, as outlined in the new vision of the National Development Plan.

Acknowledgements

Dr Charles Sheppard of Nelson Mandela Metropolitan University compiled the statistics used in this review.

References

1. Govinder KS, Zondo NP, Makgoba MW. A new look at demographic transformation for universities in South Africa. S Afr J Sci. 2013;109(11/12), Art. #2013-0163, 11 pages. http://dx.doi.org/10.1590/sajs.2013/20130163
2. Cloete N, Pillay P, Badat S, Moja T. National policy and regional response in South African higher education. Cape Town: David Philip; 2004.
3. Centre for Higher Education Transformation. 2013 South African public higher education key statistics. Cape Town: CHET; 2012.

4. Cloete N, Mouton J, editors. The doctorate in South Africa. CHET. Forthcoming 2014.

5. African Doctoral Academy [homepage on the Internet]. No date [cited 2013 Dec 07]. Available from: http://sun025.sun.ac.za/portal/page/portal/Arts/ADA

6. Castells M, Himanen P, editors. Reconceptualising development in the global information age. Oxford University Press. Forthcoming 2014.

7. Wickham S. Report on the CHEC/PGWC Joint Regional Seminar on student performance [document on the Internet]. c2009 [cited 2013 Dec 07]. Cape Town: School of Public Health, University of the Western Cape. Available from: http://www.chec.ac.za/reports/Reports%20PDFs/10%20Report%20on%20Student%20Perfomance%20seminar%202009.pdf

8. National Commission on Higher Education. *An overview of a new policy framework for higher education transformation.* Report of the National Commission on Higher Education to the Minister of Education [document on the Internet]. c1996 [cited 2013 Dec 07]. Available from: http://www.polity.org.za/html/govdocs/policy/educ.html

9. Nowak M, Ricci LA. Post-apartheid South Africa: The first ten years. Washington DC: International Monetary Fund; 2005.

10. Scott P. The meanings of massification. Buckingham: Oxford University Press; 1995.

11. Fisher W. The metamorphosis of higher education in the UK – is there an identity crisis? Hertfordshire: CELT, University of Hertfordshire; 2006. Available from: www.lancaster.ac.uk/fss/events/hecu3/documents/william_fisher.doc

12. Department of Education. South African post-secondary information (SAPSE) database. Pretoria: Department of Education; 1996.

13. Department of Higher Education and Training. Higher Education Management Information System (HEMIS). Pretoria: Department of Higher Education and Training; 2012.

14. Statistics South Africa. Census 1996. Pretoria: StatsSA; 1998.

15. Statistics South Africa. Mid-year population estimates for 2001 to 2010. Pretoria: StatsSA; 2011.

16. World Economic Forum. Human capital report 2013: Insight report [document on the Internet]. c2013 [cited 2013 Dec 07]. Available from: http://reports.weforum.org/human-capital-index-2013/#=

17. Damning CHE report into university performance. Mail & Guardian. 2013 Aug 20. Available from: http://mg.co.za/article/2013-08-20-damning-che-report-into-university-performance

18. Cloete N, editor. Responding to the educational needs of post-school youth (synthesis of final report). Cape Town: CHET/FETI; 2009.

19. Butler-Adam J. Generation J. S Afr J Sci. 2013;109(5/6), Art. #a0021, 1 page. http://dx.doi.org/10.1590/sajs.2013/a0021

20. Jansen JD. Political symbolism as policy craft: Explaining non-reform in South African education after apartheid. J Educ Policy. 2002;17(2):199–215. http://www.tandf.co.uk/journals/titles/02680939.asp

21. National Planning Commission. National Development Plan: Vision for 2030 [document on the Internet]. c2011 [cited 2013 Dec 07]. Available from: http://www.npconline.co.za/medialib/downloads/home/NPC%20National%20Development%20Plan%20Vision%202030%20-lo-res.pdf

22. National Planning Commission. National Development Plan 2030. Our future - Make it work [document on the Internet]. c2012 [cited 2013 Dec 07]. Available from: http://www.npconline.co.za/MediaLib/Downloads/Downloads/NDP%202030%20-%20Our%20future%20-%20make%20it%20work.pdf

23. Bhorat H, Jacobs E. An overview of the demand for skills for an inclusive growth path. Johannesburg: Development Bank of Southern Africa; 2010.

24. Nzimande BE. Announcement of the Oversight Committee on the transformation of South African universities [media statement]. 2013 Jan 24 [cited 2013 Dec 07]. Available from: http://www.dhet.gov.za/LinkClick.aspx?fileticket=xfQF1KMbS8I%3D&tabid=36

Mathematical errors, smoke and mirrors in pursuit of an illusion: Comments on Govinder et al. (2013)

AUTHOR:
Tim Dunne[1]

AFFILIATION:
[1]Department of Statistical Sciences, University of Cape Town, Cape Town, South Africa

CORRESPONDENCE TO:
Tim Dunne

EMAIL:
tim.dunne@uct.ac.za

POSTAL ADDRESS:
Department of Statistical Sciences, PD Hahn Building, Upper Campus, University of Cape Town, Rhodes Gift 7707, South Africa

KEYWORDS:
equity index; demographic divergence; measurement

HOW TO CITE:
Dunne T. Mathematical errors, smoke and mirrors in pursuit of an illusion: Comments on Govinder et al. (2013). S Afr J Sci. 2014;110(1/2), Art. #a0047, 6 pages. http://dx.doi.org/10.1590/sajs.2014/a0047

The 'Equity Index' (EI) as gratuitously labelled by Govinder and Makgoba in a recent paper[1] is not an equity index. It is actually simply a demographic divergence index (DDI), one of many possible mathematical alternatives which warrant the name DDI. The invocation of the word 'equity' in the original name is a deliberate but implicit claim of moral and ethical authority for the construct. This claim needs to be tested before the label 'equity' is admitted as a meaningful description. A thorn by any other name is not a rose, and proximity is not provenance.

The DDI is a simple case of a long-known mathematical device to attribute numerical distances between pairs of points in a multidimensional space (dimension = n). The index is not new in itself. Its mathematical structure is well known. However, its applicability to the setting described in the paper of Govinder and Makgoba[1] is both logically incorrect for the intended purpose and morally dubious. The error is compounded in a second paper by Govinder, Zondo and Makgoba[2].

This critique addresses the mathematical adequacy of the DDI for its intended purpose. At the heart of the critique is the fact that some numbers do not admit arithmetic, essentially because they are only labels (e.g. digits on a motor licence plate or in a cell number). Other numbers may admit addition and subtraction under appropriate conditions, and perhaps multiplication and division under further conditions. Applying arithmetic where it is not valid will yield meaningless numbers as outcomes.

At its heart, the argument of Govinder and Makgoba[1] invokes a single mathematical formula or structure. The gravamen of a mathematical formula is the implied source of unquestionable rational authority. Subrational application of the formula is then assumed to be objectivity, rather than error. The objectivity is inferred from the mathematical replicability of the error across all contexts. This objectivity is then applied to representations of South African universities,[2] but its wider application to other social institutions and conundrums is extravagantly but explicitly envisaged by the authors.

By a further assumption of a universal reference demographic profile, postulated as an exclusive and complete notion of equity, the mathematical structure of the DDI is invoked in the first paper to make value judgements about the states of institutions. In the second paper the extreme simplicity of single criterion decision-making is explicitly advocated, and the notion of institutional punishment for demographic divergence is sketched as means of steering social policy outcomes. The whole artifice is then predicated as a model for general application, in all nations, and described as an unprecedented first mathematical engagement with inequity.

What is not explicitly stated is the intended range of institutional types to which this conceptual device is to be applied in South Africa or elsewhere. There are hints from the authors which might suggest applicability to the staff and the beneficiaries and the services of schools, hospitals, welfare institutions, businesses and perhaps also government departments and non-profit organisations. However, the imperative of the authors, namely conformity with their sublimely narrow notion of equity, is the core rationale for the apparent innovation. Their particular urgency is exasperation with some universities with larger DDI values than their counterparts. On this basis these universities are perceived and asserted as intransigent on the issue of transformation.

The danger of erroneous thinking rooted in a putative exclusive concern for moral purpose and social accountability is that any underlying logical or mathematical errors are too easily excused by the imputation of vested interest and *mala fides* to those who contest the dubious mathematics. Contrary voices can easily be caricatured as at least impervious or at worst opposed to the claimed moral purpose. Indeed, one of the hypotheses offered by the authors is that several universities (other than their own university – the University of KwaZulu-Natal, UKZN) are currently impervious to equity objectives.

There is a need therefore to clarify upfront that there are indisputably terrible and consequential residues of the apartheid past and all its evil consequences, in every aspect of South African society. Some of this residue of persistent inequality and suffering is in part a consequence of preserved privilege, unjust advantage, obdurate structural inequalities, culpable indifference, wilful ignorance, lack of compassion, hypocrisy, greed and plain incompetence. Some suffering has more recent origins of a similar kind. Inescapably, suffering in South Africa has a racial and gendered face.

Universities cannot and should not be immune from the probing and critique that exposes the current extent and the likely progress of their own transformation within the society. Holding universities to account for their internal structures and their external impacts is both a legitimate and necessary act of citizenship. But social phenomena and processes are inherently more complex in their causal and contextual relationships than their counterparts in the natural and physical sciences, precisely because of the inherent agency of every human participant and stakeholder. We cannot afford pseudoscience posturing itself as relevance and objectivity in social science domains, by virtue of a single mathematical device and the numbers which a formula generates.

The resort to the achievement of measurement for evidence in the physical sciences has great power, but is limited in extent to the particular contexts in which measurement is possible. Nonetheless measurement is an engine of technological progress, within the simplicities and regularities that order our experience of the physical world. Measurement is a worthy pursuit and a magnificent achievement. This achievement arises from three key elements. Firstly, the definition of a replicable unit of extent of a characteristic common to many objects in every salient context must be clarified and exhibited. Secondly, a replicable mechanism has to be discovered or constructed, by which the extent of the characteristic can be compared with the chosen unit, and elicit a ratio outcome that is reliable to some explicitly chosen degree of accuracy, in specified contexts. Thirdly, the concatenation of the extent of objects should elicit ratios that are consistent with the properties of arithmetic, to the same degree of accuracy. Thus, to borrow a term favoured by Govinder et al, we may assert there is no cheap or *mahala* measurement of any characteristic, least of all from mere invoking of a formula.

In the human sciences there is no analogue of scientific measurement. There may be stochastic rather than deterministic analogues of measurement instruments, but such instruments are fiendishly difficult to develop or achieve or validate or verify.

Measurement instruments in the natural sciences have to be accurate and reliable under environmental conditions. In the human sciences the instruments have to transcend the observer, the observed, their complex interactions and the entire set of all relevant milieux. Before any quantitative approach is ventured in these humanities domains, a sound and plausible qualitative conceptual and methodological framework of understanding has to be postulated and critically examined.

In the realm of human sciences, we are not simply concerned with natural phenomena, which would be difficult enough. We have also to deal with perception, motive, choice, belief, conscience, mutuality and relations of power, agency and efficacy. Thus any proposal for a mathematical panacea in the social sciences should properly evoke deep and vigorous scepticism, and robust debate. We have an obligation to dignify postulated nonsense by rigorously exploring its implicit and explicit foundations, so as to expose its seductive weaknesses for what they are.

The mere assignation of a number by a conceptual or arithmetic device, even if such a formula is centuries old, does not of itself offer any objectivity, coherence or relevance. Further, the structure and pertinence of any rule for assigning numbers is open to scrutiny.

Especially in this matter of equity, we have to contest the hidden assumptions imposed on the method and context of enquiry, when the root sum of squared differences (RSSD) is engineered and purported as a final arbiter of the state and fate of universities. This caveat will also apply to any rankings derived from flawed numbers (with or without decimals) within any sphere of application.

When there is clarity about what the notion of number can and cannot offer in this debate, we still have to contend with contrasting appearances, compositions and outputs of institutions. There we will have to address the cry of the poor and yearnings of those who may be victims of our own ongoing privileges of every kind in all walks of life.

Mathematical considerations

In mathematics a multidimensional space of interest may often involve dimensions for comparable measurements in a single common measurement unit for each dimension, such as length, breadth and height (e.g. in metres), of points in three-dimensional space. The distance measure is in the same units (metres). There are several other natural mathematical distances between pairs of points (with coordinates all in the same units). These various distances would all be admissible as alternatives to the specific RSSD. The distance measures all have different utilities.

Extensions of mathematical distance measures are well known throughout science. These measures have origins deeply embedded in the history of science. One variety involves giving different weightings rather than equal weightings to the separate dimensions of the space. The measures are all applicable in contexts where each dimension is essentially unconstrained, so that technically infinite differences and distances may arise, but need not.

Every such mathematical measure would be usable as a plausible distance measure for any context involving units of the same kind on every dimension. However, a declared common specific measurement unit would be required on each dimension, before distance is meaningfully invoked in that unit.

The so-called EI (hereafter just DDI) offered in the paper is different from the mathematical distances, although it borrows one of the formulae. The DDI discards any dimensions of infinite extent. It discards continuous measurement and is simply a function of counts, not measurements. These properties are not necessarily faults but mathematical limitations, which render measurement impossible.

Although measurements invariably involve decimal fractions as multiples of a defined physical unit, the mere appearance of decimals in numbers does not constitute evidence that measurement has occurred. It is seductive, but misleading, to impute the authority of scientific measurements to numbers derived from pure counts, just because the counts have proportional or percentage forms which include decimal components.

The DDI involves subcounts of some finite countable number of persons, in precisely n defined categories. After defining the n categories and all the associated subcounts, all inferences are drawn upon the basis that every person within any nominated category is fully described by that category. For all intents and purposes related to the counts, the persons within a category are equivalent and mutually exchangeable. This fact is a consequence of the act of reducing the persons to objects in categories that are subject to particular counting arrangements. Any act of counting is not inherently wrong, but that very act has limiting consequences. The issues of exchangeability and equivalence of persons within a count will be discussed further later.

The DDI involves the category counts but first reduces them to proportions (summing to 1.00) or percentages (summing to 100.00) with some minor rounding of decimals. The purpose of using only unit-free proportions or percentages is to introduce a constructed comparability between category counts from several distinct sources (e.g. 23 separate universities). This construction that assumes the total sizes of the sources has no relevance for the nature of the intended comparisons.

Next, each institution is located in an n-dimensional space. The number of relevant dimensions (n) may vary, depending upon the choice by the observer about the number of categories to be used as a means of partitioning the observations. For the authors of the DDI paper, this dimension has been reserved to be $n=2$ (gender), $n=4$ (race) or $n=8$ (gender within race), by an appeal to the authority of their particular interpretation of the South African constitution. Other additional categories would be admissible, such as age, location, competences, experience and qualifications, but are deliberately excluded.

Each institution is then allocated n coordinate values that reflect its profile of category counts. The sum of the values within location coordinates must be 1.00 for proportions, or 100% for percentages, whatever the choice of n.

Thus it might be coherent, but not necessarily useful, to record pseudodistances between profiles using the underlying RSSD, as in the DDI. But no inference about the pseudodistances in any hyperspace carries through into any reduced or extended set of dimensions.

The DDI by construction seeks to operate only on a surface, called the simplex plane of non-negative numbers summing to 1.00, in a particular n-dimensional space. These spaces are nested within one another in the same way that many two-dimensional surfaces are nested within our familiar three-dimensional space. Thus, these DDI measures are not comparable across distinct values of n, but possibly only within a fixed value of n. The authors of the DDI paper have apparently acknowledged that fact, but ignored its consequences.

The geometry of these simplex hyperspaces is peculiar, or at least unfamiliar in our usual ways of thinking. Firstly, these hyperspaces of dimension n have all possible subspaces of dimension m nested within them, providing $m < n$.

Each n-dimensional hyperspace has a central point whose n coordinate values are all equal, namely $n^{-1}=1/n$. This central point has a common pseudodistance, RSSD=sqrt[(n-1)/n], from each of the extremal points in its hyperspace. For n=3, this hyperspace would have the appearance of an equilateral triangle, joining points at (1; 0; 0), (0; 1; 0), and (0; 0; 1). The central point is at pseudodistance sqrt(2/3) = 0.816497 from the extremal points.

All pairs of extremal points have a common pseudodistance RSSD=sqrt(2)=1.414214 between them. This extremal pseudodistance applies unchanged across all n-dimensional spaces. No pair of points in any n-dimensional simplex hyperspace can be further apart in RSSD than the common RSSD between all extremal points. This notion of maximal RSSD is discussed again later.

The DDI notion involves the assumption that a single specifiable point on the hyperplane has both a mathematically and a contextually significant position. It is legitimate to nominate a reference point mathematically, but the intended meaningfulness of the reference point must be argued from beyond mathematics (e.g. arguments from equity or other criteria), along with the meaningfulness of the number of dimensions. Choices of n and of reference points are contestable. In particular, national demographics may be too narrow a set of n categories to address the complexity of any issue in question.

The notion of RSSD pseudodistance is not a notion of inequity unless some reference point is hypothesised on the simplex hyperplane. That reference point is, by assumption, an ideal point that is relevant in and of itself, but also completely adequate for a purpose at hand. Hence the reference point can only be ideal in the particular n-dimensional space if no other space of smaller or larger dimension is deemed to matter at all.

This limitation implies that any use of the DDI in n dimensions necessarily discards the intrusion of any other source of information of any kind about the persons involved. The notion of equity is reduced in this context to a notion of deliberate and sustained ignorance about all other possible contributions to the choice of a reference other than those embodied in the chosen reference point. Rather than being a strength of the DDI method, as viewed by the authors' agenda, this feature constitutes a severe fragility for the DDI in all applications, including their applications.

In the applications cited by the first paper, we would have to infer that only the race and gender issues mattered as selection outcomes at each level of application, e.g. the senior administrative level at UKZN or Rhodes University. Moreover, the reference point is next subjectively defined as a fixed set of national demographic proportions. The DDI calculates a pseudodistance from that reference.

At the moment of definition of the reference, the profile (of any university) does not correspond to the ideal. Thus one is faced with a choice, either to discard the current cohort of leadership in senior positions and immediately replace them by new selections, or to permit passage over time towards the reference point, through controlled demographic selection.

The RSSD pseudodistance might be useful if our process for selecting these incoming university administrators was to randomly select such new appointments from a suitable pool. The preferred pool of the authors must be constituted precisely and only by the reference race and gender proportions, with no regard to any other characteristic that might be specific to the human resource requirements of an incumbent, in a prospective senior administrative appointee. The function of the reference point is to penalise all other considerations for appointment.

The DDI might then serve as an indicator of the randomness of the process of selection, if randomness from the desired demographic profile was the only criterion required. Any leeway to select on criteria other than demographics alone will necessarily permit, and even perhaps require, deviations from the defined target.

The same objective of randomness of selection can again be assured by the use of the DDI reference point at every level of aggregation (academic staff, technical staff, service staff, students, etc.). Use of randomness as the single selection criterion for new appointments from a pool of candidates already satisfying the reference profile will generate, over time, a series of appointments which will eventually satisfy the same intended profile of incumbents, at every level of aggregation. If one exchanges the incumbents often enough, then random selection from the reference pool will steadily approximate the chosen reference profile.

In the sense that all selections from the pool will be random, the process and its replications will be fair (free of any selection bias). The utility of these selections would still require demonstration. The question may arise as to whether or not the use of any national demographic profile can be legitimately characterised as random selection. The legitimacy of this description is motivated later.

The RSSD pseudodistance might conceivably be adopted as a confirmatory criterion of the appointments processes over the period, beginning from the first moment when randomness of selection from the idealised pool is deemed appropriate. UKZN would conceivably congratulate itself on this assured journey to achievement of a reference point by the innovative device of iteratively ignorant blind choice from the entire population.

The utility of the reference point is moot for another reason. Unless the idealised point is rendered mathematically tractable, by rounding conventions, all configurations in all positions will be short of the reference (they will be at some pseudodistance, even in the putatively salient UKZN environment). The authors adopt a notion of tolerance to address this issue.

It is quite another matter whether such a pseudodistance from randomness is ever meaningful on instantaneous states (e.g. current occupants of the positions) rather than only on the process changes (e.g. new appointments) at each specific level. The authors have noted this limitation.

What the provision of a formula hides is the misconception that counts can be handled mathematically as if they are interchangeable with measures. The fact that we may count people does not make them equivalent and exchangeable. A principal, a registrar and a dean will count as three people in leadership positions, but we do not believe we can switch them arbitrarily, not even at UKZN. A person is not a unit of measurement. On the other hand, the metre in terms of which we measure height is equivalent to the metre by which we measure length. The fact that proportions and percentages can be written to some degree of accuracy as decimal numbers does not make either the proportions or percentages measurements.

If one wishes to ascertain how much the actual count profile of changes at staff selection differs from a desired set of random probabilities, then a formal randomisation test can be invoked. An approximate but correct method for checking compliance with the idealised profile is a chi-square goodness of fit test. This test is available in first-year texts and is easily calculated using software such as Microsoft Excel.

As in all statistical analysis and evidence collection, the use of any formula, such as chi-square, may elicit a signal from data. The signal indicates that at least one of the underlying assumptions we have made does not fit with the message from the data. The subject domain expert then has to take a view on whether or not the discernible signal constitutes evidence of some consequential violations of assumptions, possibly followed by decisions and actions. No statistic can replace the role of the thinking scientist in either the natural or human sciences.

The paper of Govinder and Makgoba[1] in the *South African Journal of Science* is remarkable. It will in time become a frequently cited paper. The citations will not be to celebrate its elegance, simplicity or profundity – it has none of these characteristics to warrant citation. Instead it will gradually become cited for its errors and less scholarly characteristics. One such infelicity is its implicit argument for randomness as the principal criterion to distinguish one candidate from another, as the long-term strategy of a university to reach and maintain an ill-conceived idealised profile.

The issues of equity and redress are too important to be trivialised by allowing ourselves to be intimidated by the sequestered word 'equity' and the torrid outcomes of mathematical orchestration.

Confusion thrice confounded

In a follow-up paper, Govinder et al.[2] claim to extend the original DDI apparatus into 'an important policy tool in steering the system towards a notion of transformation that connects, rather than disconnects, equity, development and differentiation'. They further aver 'The index may also become a useful universal measurement of equity in higher education (and other) systems globally.'

In support of this set of claims, they report 10 sets of applications of the formula to 23 universities in South Africa. These 10 sets cover seven employment categories, enrolment, graduation and a new indicator – their equity-weighted research output. These 10 sets of 23 indices give rise to 10 rankings in their analysis.

They proceed to consider arithmetic on some of these indices using subtraction and ratios (reported as percentages). They claim to explore relationships between equity and quality by the device of partitioned scatter plots involving DDI values and publication counts.

What appears to be unstated is that the entire set of 10 analyses reported are based upon four race categories alone, although there is bracketed comment: (ignoring gender imbalances). The consequence of this offhand remark is that the entire analysis appears to take $n=4$ rather than the claimed constitutional imperatives of race and gender, with $n = 8$. All the analyses appear to be implicitly referenced to Table 2 in the paper.

The effect of choosing as small an n as 4 is likely to be a very much exaggerated range of plausible DDI values than might be the case for $n=8$ or larger. Given the deep concern about equity that presumably motivates the paper, the analyses with gender included may well have been conducted, but are not reported. The applications for $n=8$ may give rise to an artefact: reduced DDIs in regions close to reference points.

Clarity on this matter of the reportage was sought by a request for the data and spreadsheet calculations on which the reported analyses were based. The advice received from the authors was to consult the sources specified in the references for the necessary data. Further, through the editor, a clarification was received: 'As far as our personal spreadsheets are concerned, we do not believe that it is appropriate to release them as this was obtained as a result of considerable work on our part.'

Such a position in a matter as consequential as this debate has severe ethical and scientific weaknesses. It is also open to several unfortunate interpretations. The reader is denied the opportunity to assess the data and the claimed calculations. Such an attitude contrasts with values of openness and transparency, and of the replicability of allegedly scientific methods and processes.

The StatsSA source of data reported in the references gives only the aggregate percentages for race ($n=4$), as determined by the 2011 census. The census outcomes have been announced and publicly contested. The 10% sample from the census has yet to be released, despite the controversy about the post-enumeration survey allegedly being resolved by a disciplinary process that has not yet been heard. Nonetheless, we may currently regard one part of Table 2 (labelled 'Overall') as being sufficiently coherent with official figures for an exploration of the DDI.

The authors have indicated that foreign visitors and permanent residents in South Africa apparently constituted 0.5% of the census population and they and their constitutional rights are ignored in the analyses. This approach might be constitutionally awkward, but unwelcome foreigners can be mathematically eliminated by a minor upward correction of the population percentages.

However, Table 2 and subsequent discussions introduce further errors. Briefly, these errors involve the maximal RSSD, problems with acceptable RSSD levels, and misunderstanding of the notion of quintiles.

For $n=4$, and its associated overall population percentages embodied in an ideal, a maximally contrasting profile for an institution would arise from an only-Indian composition, and yield a DDI value of 126.5 with minor rounding approximation permitted, using four percentages summing to 100. An only-foreigner institution ($n=5$) has a corresponding approximate value of 127.8, using five percentages summing to 100. As previously noted, the maximum RSSD between any two extremal (single population group) institutions is 141.4 for any value of n. This maximal value of around 141.4 for extremal RSSDs contrasts with the repeated error in all four columns of Table 2 which report impossible maximal values of RSSD for the South African data.

The erroneous maxima are next partitioned into intervals of common width, equal to one-fifth of the reported maximal values. The one-fifth segments are further erroneously labelled as quintiles.

The fifths of an interval do not correspond to quintiles of a distribution except in one circumstance (uniform density of RSSD values over the entire correct permissible range). That necessary circumstance cannot possibly apply under the conditions of percentages summing to 100 as required here. This emperor has no clothes.

The language of the paper describes a tolerance of 5% of each target value. If we presume this tolerance, we still have to take into account that the permitted variations have to balance each other out. Thus 5% of the target for each of the three smallest of four racial categories in use, will maximally combine to 5% of their 19.3% total, about 1%. This maximal combined tolerance then also applies as the maximal tolerance for the complementary single largest racial category.

After adjusting for the eliminated foreigners, this calculation will permit a deviation of, at most, about 1% from the 79.2% recorded alongside the category Black African. The subsequent RSSD value is 1.20%. This number is very different from the reported value 5.3%. An Excel spreadsheet is available for the curious.

Other interpretations of the wording used for tolerances were explored. None of these gave rise to the tolerance quaintly labelled 'quintile zero' in Table 2, noted as 5.3%.

Despite all these difficulties, the paper goes on to claim the utility of being able to report both the pseudoquintile, and even changes in pseudoquintiles, as evidence of achievement and progress.

A further source of mathematical astonishment is the use of subtraction in Table 3. This operation generates the new and allegedly profound efficiency DDI by subtraction of graduation DDI from enrolment DDI across 23 institutions. The hidden assumption is that the RSSD functions behave additively or linearly for any fixed n. This assumption is false. For example, two persons both 3 units distant from their destination may be anywhere between 0 and 6 units distant from one another.

The same false assertion of additivity is again applied in the construction of Figure 2, in which the various DDIs for overall staff and the seven staff component categories are aggregated by concatenation, against a vertical axis for cumulative DDI values. Indeed, a new mathematical faux pas: summing of both the whole and the sum of some of its parts.

The RSSD is not a quantity of measurement in terms of a reliable unit of any kind. It is not even a count. The RSSD values cannot therefore be claimed to admit a valid arithmetic of addition or subtraction. They also do not admit ratio comparisons within or across institutions.

It is correct to treat RSSD as an ordinal feature, and hence we can admit rankings as offered extensively in the paper. We would be able to infer that on some rankings one university has a higher RSSD than another, but we would have difficulty in explaining what such difference in ranking meant per se for any decision-making. Further criteria from beyond mathematics would have to be argued and debated, and their fitness for purpose examined.

There is a lurking hint that RSSD values should be tracked over time, and that universities should be able to exhibit trajectories towards lower values. Again, we can make such comparative judgements over time within single institutions, on the basis of ordinality, but the judgements do not have the power to inform decision-making, except as self fulfilments.

Some final paragraphs of Govinder et al. impute intransigence in the higher education sector on the grounds that after some 23 years since the visible fall of apartheid, the universities have not yet reached adequate national profiles for these authors. A litany of allegations is neatly composed: passive resistance, denial of failure, abuse of autonomy, abhorrence of accountability, failure of government to steer or monitor, the state cowed by the privileged and impervious to the voice of the disadvantaged, conservatism.

All these allegations are worthy of debate, but it is a form of intellectual bullying to hide behind a mathematical formula as the justification for unspecified 'extraordinary measures'. The intended punitive actions assume that all playing fields prior to the imposition of the reference profile are level, and that the location of the problem of inequity lies singly and only in the universities themselves.

The imperative to adopt national reference profiles does not ameliorate in any way the profile of school-leavers apparently eligible for university entrance and technically capable of graduation. The rationale for urgent measures purports that the obstacles to better profiles are solely the fault of universities, and that no other constraints or preconditions or simultaneous imperatives apply.

The RSSD does not address the notions of real distance from home to institution, of term-time accommodation, of local travel costs and constraints, of access to books and technology, of adequate preparation, of emotional support, of scholarships, or of differential living costs across rural and urban settings.

In respect of prospective employees, the RSSD does not take into account the composition of pools of available candidates; the effects of competing positions in commerce, business and industry; or varying forms of family responsibility and cultural preferences of the candidates themselves.

These issues too are worthy of debate. It is not possible by mere *fiat* for universities to set aright the suffering of this society, by admission, graduation, research and employment profiles that match a national reference profile. Indeed, it has not been possible for democratic government in South Africa to achieve corresponding laudable goals in housing, education, nutrition, health, transport and employment in 20 years. It is legitimate to argue that some of the outcomes reflect difficult initial conditions rather than dereliction, fault or animosity within universities.

All societal change is contextual and inherently unpredictable. What is necessary is debate about mechanisms that work and the necessary conditions for their success. In such debates we may hold all role players mutually accountable for processes that eliminate or moderate suffering and injustice. It is a dubious principle to rule out regional objectives on the grounds that they reflect imbalances and injustices of the past. Contextualisation is not ipso facto a reneging on justice.

If we may not contextualise and if only DDI conformity matters, we can only comply by ensuring no criteria other than the national demographics alone, intrude into our decisions. There is only one way of verifying that conformity, by being able to demonstrate that only random selection from the national profile (and nothing else) is exercised at the level of every decision-making concerning individuals at universities. We require demonstrable random selection from the reference race and gender groups for admission, selection, passing, graduation, employment and promotion.

This argument is not a trite parody of the arguments of the DDI authors; it is unfortunately the essence of their position. It is also the basis upon which they diagnose culpable indifference, or worse, at the universities.

The issues of equity, development and democracy need robust engagement. They require open minds and open hearts. The DDI should be left in the Euclidean cupboard. There are too many flaws to warrant prolonged discussion. Let us rather debate the injustices and the needs authentically, and clarify the nature of processes and resourcing which will have some chance of offering a better future for all.

The great flaw in the DDI as a stand-alone methodology is that it permits only a partial view of outcomes of complex processes. The method focuses upon one set of outputs – demographics – but ignores all inputs and all process characteristics that precede and lead to those limited outcomes. Such an approach cannot claim equity as a hallmark of its achievement. Yet the approach of these authors also predicates a whole white box of cause-and-effect relations dominated entirely by the leadership of institutions, as if no other role players exercised either effect or judgement.

Dancing with other divergence demons

In the latter segment of the second paper, the authors seek to expose recourse to quality (and extent) of scholarly output as an apparent disguise for intransigence, often invoked in their view by several target universities. The methods of the paper seek to correct for advantageous effects arising through retaining privileged DDI profiles, within various aggregate and per capita indices of research output. Partitioned scatter plots contrast the locations of the universities.

Scatter plots and their partitions may be meaningful as depictors of relationships between characteristics but only to the extent that the underlying coordinate systems are meaningful. Even then the plots have an inherent limitation. When we reduce, say 23 universities, to only the two characteristics in use within the scatter plot, all emerging graphical insights are filtered through the AOTBE (all other things being equal) lens.

In science, especially in human sciences, we have to take into account the distortions of this lens. We seldom mean that all other factors have been eliminated or effectively controlled by suitable balancing for equivalences. We usually mean that all other factors are ignored because the task of observing them and taking them into account is too difficult or too costly in time or money, or perhaps impossible.

Our lens and inferences must in almost every case be modified from AOTBE to AAOTBEU (almost all other things being equally unknown). This term describes a qualitatively different set of conditions, and alerts us to the practice that distinguishes scholarship from slippery reasoning and sleight of hand. That practice is to declare explicitly any ignorance or unknowability or limitations.

In the scatter plots (Figures 3 and 4) of the second paper, no such caveat is offered. All 23 points for 2011 data are plotted in each case. The choices to partition the scatter plots are admissible, but both the relevance and the adequacy of two sets of four groupings are open to challenges based upon other information.

We note that the weighted research output is a counting device. This count aggregates all papers published and all degrees awarded. The count does so in a manner that notes the existence but does not distinguish between any levels of quality of the publications and theses. All these elements are regarded as interchangeable in their weight classes.

The weighted count numbers are a bureaucrat's attempt to quantify scholarship, and remain subjects of debate, even as they are also sources of funding. These numbers appear in column 2 of Table 6.

Again a false assumption of admissible arithmetic is imposed on these numbers. The DDI overall numbers of column 2 in Table 5 are divided into the bureaucrat counts of Table 6. The results enter Table 6 at column 8, labelled 'equity-weighted research output'.

The inadmissibility of these various arithmetics is papered over implicitly by the loose use of the word index. What valid and honest scholarship requires is a contestation around the observable phenomena, not mathematical smokescreens.

Several grave dangers of the DDI as methodology have now been made further apparent in a press release from UKZN.[3] Reported recommendations, apparently accepted and approved already by a Ministerial Transformation Oversight Committee, chaired by one of the authors, are drawn from the second paper. These elements include 'realistic targets for high-level knowledge production linked to equity', in respect of which Table 6 column 8 of the second paper conveniently asserts UKZN in the first rank.

The press report also makes various claims for time periods to attainment of demographic profiles by named institutions. Their source is allegedly a seminal study report published in the *South African Journal of Science*, for which neither the first nor the second paper provides any formal evidence. Thus we note a new confusion has been introduced into public life.

This confusion is an assertion that rigorous estimation of the passage of time from some current profiles to the attainment of DDI = 0 can be offered. The estimates of the periods specified range from 40 years for academic staff to 43 years for overall staff of the institutions generally. For particular institutions, the estimated periods include 261 years and 382 years for Stellenbosch University and the University of Cape Town staff numbers, respectively. No estimated standard errors accompany these estimates – an interesting omission.

In the latent scatter plots for 23 institutions over time, there will be fewer points than 23 in earlier scatter plots. There will be 23 distinct scatter plots of two (perhaps more) time points each. Private correspondence indicates there are precisely two time points, but ongoing data collection is expected to produce DDI values for more retrospective time points. Thus we currently have 23 time series analyses, one for each institution, based on exactly two observed values and one observed difference over time!

Whether the seminal study or the author of the unexpurgated press release is responsible for the time series analysis is as yet unclear. But a ministerial committee apparently believes in the AOTBE approach applied to two consecutive data points. They buy into inferences of periods spanning between at least 40 and at most 382 years before the required demographic profiles are reached, and without indications of imprecision. This type of pervasive foresight can only be matched by the prognostications of astrology, but unfortunately not by the application of scholarly methods.

The danger is that such perverse conclusions will determine policy, predicated on an assumption that scholarship has driven these inferences.

A further recommendation apparently specifies '20% of each institution's block grant must be reprioritised to address equity transformation [because] there is no cheap or *mahala* [free] transformation'.

Conclusions

The various DDI manifestations thus far offered in pursuit of an illusion speciously labelled as equity should be rejected outright as invalid and misleading in name and content and implied authority. The DDI may be more fully debated. However no DDI will yield measurement in a scientific sense. Thus, for any specified set of counts (students, staff, etc.) the choice of DDI applied may be used as an ordinal variable, and can support rankings only. DDI values cannot support arithmetic, either within or across indices.

The contrasts between notions of divergence and notions of equity need to be clarified. The debate about equity, including its meaning and attainment, has to embrace the reality of suffering and injustice in South Africa. This debate may include the universities, but the other institutions also warrant attention, preferably of a rational rather than pejorative kind. The universities have a dual part in this debate, as objects of enquiry and voices of observers.

Many processes may be required to eliminate injustice and promote more rapid access to better life circumstances. Elimination of injustice cannot be adjudicated by evidence only from a mere calculation. Both the legitimacy and role of any arithmetic have to be firmly clarified. Otherwise the invocation of one or more indices becomes a vehicle of bureaucratic self-gratification, rather than a series of ordinal indicators, each indicative of only one possible objective at a time.

This position does not exonerate universities from accountability. It affirms a collective obligation of an examination of conscience in robust debate. However it also claims that true transformation is a matter of the heart and an issue of complexity, which warrants authentic scholarship rather than fumbling mathematical conjuring.

References

1. Govinder KS, Makgoba MW. An Equity Index for South Africa. S Afr J Sci. 2013;109(5/6), Art. #a0020, 2 pages. http://dx.doi.org/10.1590/sajs.2013/a0020

2. Govinder KS, Zondo NP, Makgoba MW. A new look at demographic transformation for universities in South Africa. S Afr J Sci. 2013;109(11/12), Art. #2013-0163, 11 pages. http://dx.doi.org/10.1590/sajs.2013/20130163

3. Seminal study devises Equity Index to measure the pace of transformation in South African universities [UKZN press release]. 2013 Oct 24 [cited 2013 Dec 10]. Available from: http://www.ukzn.ac.za/news/2013/10/24/seminal-study-devises-equity-index-to-measure-the-pace-of-transformation-in-south-african-universities

Flaws in the approach and application of the Equity Index: Comments on Govinder et al. (2013)

AUTHORS:
Tom A. Moultrie[1]
Rob E. Dorrington[1]

AFFILIATION:
[1]Centre for Actuarial Research (CARe), University of Cape Town, Cape Town, South Africa

CORRESPONDENCE TO:
Tom Moultrie

EMAIL:
tom.moultrie@uct.ac.za

POSTAL ADDRESS:
Centre for Actuarial Research,
University of Cape Town, Private
Bag X3, Rondebosch 7701,
South Africa

KEYWORDS:
transformation; demographics; equity profile; Euclidean distance formula; higher education; South Africa

HOW TO CITE:
Moultrie TA, Dorrington RE. Flaws in the approach and application of the Equity Index: Comments on Govinder et al. (2013). S Afr J Sci. 2014;110(1/2), Art. #a0049, 5 pages. http://dx.doi.org/10.1590/sajs.2014/a0049

Transformation of South African society post-apartheid, which includes higher education institutions (HEIs), is a national imperative, but it is a complex process which cannot, and possibly should not, be encapsulated in a single index, particularly one as poorly conceived, applied and interpreted as that presented by Govinder, Zondo and Makgoba[1] (and Govinder and Makgoba[2] before that). While these publications and some extreme interpretations of the results by the authors at various fora may have the benefit of stimulating debate, there is a very real risk that such reductionist monitoring could very well undermine, rather than encourage, the process of transformation.

Although the concept of an 'Equity Index' (EI) was initially presented in Scientific Correspondence in this journal by Govinder and Mokgoba[2], the idea was presented and expanded upon in more detail in the more recently published paper by Govinder et al.[1], and it is on this paper that most of our comments focus. The expanded paper applies the index to determine the extent of the deviation from a benchmark demographic profile of staff and students at 23 HEIs in South Africa in 2011. Apart from ranking the institutions on various components of staff and students, they also compute what they refer to as an 'equity-weighted research index' which purportedly adjusts research output for equity. In terms of instructional/research professional (IRP, i.e. 'academic') staff, which was the major focus of their paper, they found that although no institution had a satisfactory EI, by splitting the ranges of the EIs and of output per capita of the HEIs into two halves, only two institutions – the University of Fort Hare (UFH) and the University of KwaZulu-Natal (UKZN) – were located in the best quadrant (that represents the lowest EI and highest output per capita). In terms of student profiles, they argue that, apart from no university having a satisfactory EI, all but three universities had higher (worse) graduation EIs than enrolment EIs, singling out five as showing a 'dramatic worsening'. From these results, the authors draw a number of often extreme and ill-reasoned conclusions.

Unfortunately, apart from the questionable mission of the paper, the authors make a number of algebraic, computational and conceptual errors in the implementation and interpretation of their index, which all but negate most of the arguments for and from the index. In fact it is quite surprising that a paper with so many flaws was deemed good enough to be published.

Algebraic and statistical errors

The EI proposed by Govinder et al. is based on a simple measure of distance, the root sum of squared differences (RSSD), which can be expressed more meaningfully than they do as follows:

$$EI = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} (p_{ij} - p_{ij}^{b})^2}$$

where p_{ij} represents the percentage of the institution's population in population group i = {African; Coloured; Indian; White} and sex j = {male; female}, and p_{ij}^{b} represents the corresponding percentages in the benchmark population.

A detailed critique of the inappropriateness of using percentages or proportions as a measure of distance in an RSSD index, including the problem that it treats departures from the benchmark percentage in each category as being equivalent, and other criticisms have been offered by Dunne[3]. We do not intend to cover these criticisms. Instead, we confine ourselves to still further problems with the EI as implemented by the authors.

The first problem with the calculation of the EI is that, instead of the formulation above, the authors have calculated the EI as

$$EI = \sqrt{\sum_{i=1}^{4} (p_{i,\cdot}^{r} - p_{i,\cdot}^{b})^2 + \sum_{j=1}^{2} (p_{\cdot,j}^{r} - p_{\cdot,j}^{b})^2}$$

Despite the authors' assertion that either method of calculating the index 'is relevant'[1], their approach double counts and is not mathematically correct. Correcting for this error has the effect of reducing all the EIs by approximately 30%. If the primary concern is simply the ranking of the universities (as opposed to drawing any other conclusions) the correction does not change those rankings, but it does impact, inter alia, the maximum determined values of the EI; the relationship between research output per capita and EI; and the demarcation of the quadrants, and hence the conclusions drawn from partitioning the plot area into quadrants.

To compound their computational error, the authors excluded from the benchmark what they assumed were 'foreigners' (an issue dealt with in greater detail below) without re-estimating the percentages so that the percentages in the benchmark do not sum to 100%. This approach makes it impossible for the EI to be zero.

More substantive issues arise when the authors apply the derived EIs and attempt to draw inferences and implications therefrom. For example, they treat the EIs as additive (they are not) and thus produce what they term an 'equity efficiency index' to measure change by differencing the indices of enrolled students and graduates; or, even more absurdly, they sum the EIs for the various categories of employees (presented in Figure 2 of their paper) to compare universities. The absurdity is compounded by the small numbers of staff in several categories – a problem of which they are aware, but appear to ignore in this instance. For example, they calculate an EI for the 'crafts and trades' group across the 23 HEIs, despite there being fewer than 20 such staff at 13 of the 23 institutions. The resulting proportional distributions by population group and sex are not robust and cannot convey any useful insights.

However, the most worrying misuse of the EI by the authors is their construction of what they call their 'equity-weighted research output' (i.e. total research output divided by the staff (or IRP staff) EI). Apart from the fact that they do not standardise for the size of the institution (using total research output rather than output per capita in the numerator) in their measure, the metric is difficult to interpret, and certainly does not 'weight' for EI. (In passing, it is worth questioning whether it is coincidental that using the equity-weighted research output for IRP staff ranks UKZN as the top university when controlling for size drops it down to eighth.) The difficulty can best be understood by considering what will happen to the equity-weighted research output of an institution that produces some (even the most modest amount) research as the EI approaches the ideal (i.e. zero) – it will tend to infinity. This is clearly neither desirable nor sensible.

Measuring transformation: Who is eligible?

If the proposed EI is to be used to track transformation, the data used in its construction should be presented according to the requirements of South African labour legislation. The *Employment Equity Act of 1998* (amended by regulations promulgated in May 2006[4]) adopts a very strict definition of who counts as a member of the 'designated group': non-South Africans who might otherwise be regarded as 'designated' are expressly excluded, as are any foreigners who became naturalised South African citizens after 1994.

The Act's definition of the designated group thus goes far beyond a simple test of citizenship, and the vast majority of foreign-born South African citizens who obtained South African citizenship after 1994 are excluded from the designated group. The proportional distributions used in the calculation of the EI should be determined by the ten categories identified in the Act, but neither the 2011 census data nor the Higher Education Management Information System (HEMIS) database on staff employed at South African HEIs permit the appropriate analysis of the data in this form. It is therefore not possible for the EI to be calculated from the data available in a manner consistent with South African legislation. (The HEMIS data are available through the Higher Education Data Analyser. Using the data available we are able to reproduce very closely the results produced for 2011 by Govinder et al.[1] Somewhat curiously, we cannot reproduce at all closely the results presented for 2007 in the authors' presentation to Parliament's Higher Education Committee.[5])

Restricting the calculation of the EI to only South African citizens (possible with both the 2011 census and HEMIS data) would go some way to aligning the EI more closely with the requirements of South African labour legislation, although those who took citizenship after 1994 would be erroneously included as a result. However, there are two further problems with doing this. Firstly, other comparisons based on the EI, particularly the analysis of research output per capita, would be invalidated as this research output would then also have to be restricted to include only that produced by South Africans. Secondly, excluding foreign nationals from the EI entirely would mean that HEIs could increasingly employ staff from this group (so they constitute an increasing proportion of all staff) and still improve their equity profile by ensuring that the (diminishing proportion) remaining are increasingly drawn from the designated groups. This too would not encourage meaningful transformation of staff bodies.

Problems with the data on higher educational institutions

In addition to the algebraic and statistical errors outlined above, there are three further problems inherent to the data on staff at South African HEIs as presented in the HEMIS database.

Firstly, it is not clear how Govinder et al. dealt with staff for whom population group was unclassified. For example, approximately 12% of IRP staff at UKZN did not have their population group recorded in 2011. Likewise, in the data for the University of Cape Town (UCT) for 2007, more than a quarter of IRP staff were 'unclassified'. It would appear that those staff were either ignored (as was done nationally) or the percentages in the classified groups increased to sum to 100%, neither of which are satisfactory.

Secondly, because the HEMIS data rely on information submitted by institutions, the data are heavily affected by the classifications used by those institutions, and by institutional policies relating to outsourcing of certain activities. As two examples, 17% of all staff at HEIs in South Africa in 2011 classified as 'Executive/Administrative or Management Professionals' would appear to have been employed at a single institution – the University of Pretoria – while nearly a third (31%) of all 'Technical' staff employed by universities in 2011 were at UKZN. In a similar vein, UCT has one of the smallest complements of service staff (140 reported in 2011), no doubt the result of that institution outsourcing most of this work. The utility of including clearly problematic categories such as those described above in any comparative measure of transformation is questionable.

Problems with the benchmark(s) used

In addition to the problem that the 2011 census data do not allow for data to be extracted that conforms to the requirements of South African labour legislation, there are further conceptual issues with the benchmarks proposed.

Instructional/research professional staff

For their benchmark for instructional/research professional (academic) staff, Govinder et al. propose using the distribution by population group and sex of the population as a whole aged 24–65 from the 2011 census. There are several important aspects to these data which would appear not to have been considered by the authors.

Firstly, they have confused the census questions relating to citizenship and population group, implying that the 0.6% of people in the census who are coded as population group 'Other' are not South African citizens. This assumption is incorrect. Nearly half of all people aged 25–64 coded as 'Other' in terms of population group are South African citizens, while over a million non-citizens aged 25–64 classified themselves in terms of population group.[6] (Statistics South Africa's online database does not permit the extraction of citizenship status of those aged 24–65 (as used in the original paper), only that in conventional 5-year age groups. The effect of the slight difference in classification of ages is likely to be trivial.) The census data suggest that around 4.6% of the population aged 25–64 are foreign nationals, with a further 2.3% not being classifiable in terms of population group and/or citizenship. The resulting distributions underlying the benchmark proportions are rather different from those presented in Govinder et al.'s paper (Table 1).

It must be re-emphasised that the figure of 4.6% for foreign nationals would be an underestimate of the proportion deemed 'not designated' in terms of the *Employment Equity Act*, as most of the foreign-born who obtained South African citizenship after 1994 should be included in the 'Foreign' category for employment equity purposes, as described earlier. It is also important to note that 2.3% of the census population aged 25–64 could not even be accurately classified in terms of citizenship and population group, let alone in terms of the finer gradations required by the legislation.

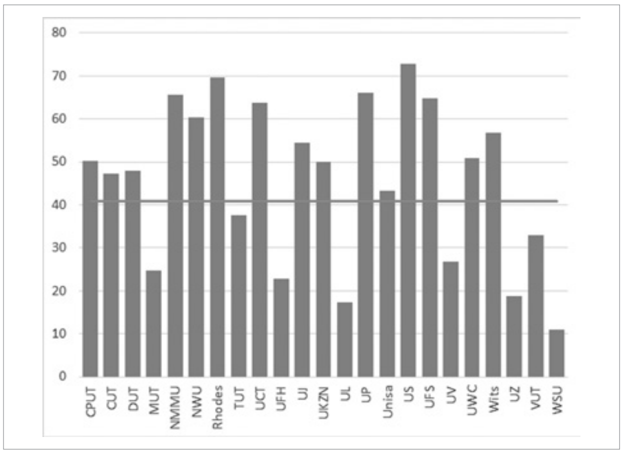
Table 1: Benchmark percentages by population group based on a citizenship test, and those used by Govinder et al.

	African	Coloured	Indian	White	Foreign	Unclassifiable
Percentage based on citizenship (25–64 years)	70.6	9.3	2.8	10.4	4.6	2.3
Govinder et al. (24–65 years)	76.1	9.4	3.0	10.9	0.6	

While nationally the proportion of foreigners may be unremarkable, many HEIs have a significant component of ‘International’ staff. At UCT, for example, in each of 2010, 2011 and 2012, a quarter of academic staff are classified as ‘International’ in terms of labour legislation.[7] This observation brings into sharp focus the tension caused by the proposed EI between institutions hiring the most suitable candidate for a post, globally, with a view to offering the best possible training to students and best advancing the development of South Africa and the continent, and a much narrower, parochial, mindset that the EI would seek to engender, because any institution with a large complement of ‘International’ staff will, by definition, perform poorly based on the EI. A naïve focus on the EI would serve to punish institutions that seek to recruit the globally best staff.

Secondly, the authors assert that the chosen benchmark is further justified because ‘universities are responsible for ensuring that this age group is suitably qualified’[1]. This assertion ignores the long-term consequences of apartheid education: matric exemption rates (which determine the enrolment profile of first-year students –‘passing’ matric does not automatically confer university exemption, a point neglected by the authors) differ markedly by population group. Even if the ongoing crisis in primary and secondary education (as evidenced by the recently released results of the Annual National Assessment[8]) was resolved immediately, there would be a lag of nearly two decades before those entering school in 2014 had completed a postgraduate education.

Thirdly, although a benchmark based on the proportion of the population by sex and population group in possession of a higher degree might seem a sensible alternative, it is not really. Those with higher degrees in South Africa are overwhelmingly White (and male), so benchmarking against this population would have the perverse effect of penalising institutions that have fewer White staff and are closer to the national demographic profile. Nonetheless, the EI of this alternative against the national benchmark offers an approximate indication of the extent of the structural/systemic component of the institutional EIs. The EI for the population with higher degrees relative to the benchmark population aged 24–65 is 40.9, about half the range of the correctly calculated EIs of universities. As can be seen from Figure 1, 15 of the 23 HEIs have an EI greater than this value, suggesting that – even relative to that alternative benchmark – they are relatively ‘untransformed’. It is important to note, however, that Figure 1 also suggests that a large part of the ‘transformation problem’ described by Govinder et al.’s analysis results from structural rather than institutional causes, and that the structural nature of who currently has higher degrees in South Africa contributes more to the overall index than institutional specificities do.



Abbreviations of universities' names as in Govinder et al.[1]

Note: Values of the Equity Index (EI) have been corrected for algebraic errors in the originally proposed EI.

The horizontal line shows the EI of the population aged 24–65 with masters or doctoral degrees relative to the benchmark (EI=40.9).

Figure 1: Equity Indices for 23 higher education institutions corrected for errors in the initial formulation, and national Equity Index based on qualifications.

Comparing the distribution of IRP staff by level of qualifications at the eight institutions with EIs that fall below the line, combined, against that of IRP staff at the eight institutions with the highest EIs, it is evident that the staff at the ‘more transformed’ institutions are much less qualified than those at the ‘less transformed’ institutions (Table 2).

This finding suggests that universities could achieve greater levels of ‘transformation’ simply by employing less-qualified staff – an approach unlikely to be in the long-term developmental interests of the country as a whole, or of the higher education sector specifically.

Table 2: Percentage of instructional/research professional (IRP) staff with master’s and doctoral degrees at the eight ‘most transformed’ and eight ‘least transformed’ higher education institutions in 2011

	Percentage of IRP staff with		
	Master’s degree	**Doctoral degree**	**Master’s or doctoral degree**
Eight ‘most transformed’	39.1	20.0	59.1
Eight ‘least transformed’	31.9	51.1	83.0

Student profiles and administrative staff

The authors argue that the national rather than provincial demographic profile of people aged 17–40 who have a matric should be used for the calculation of EIs for the student body at HEIs. The HEMIS data on full-time equivalent (FTE) student enrolments at HEIs in 2011 suggest otherwise. Universities draw their students primarily from the region in which they are located.

For example, while around 3.5% of the national population aged 17–40 with a matric in 2011 were Indian (Table 2 from Govinder et al.), Indians comprised about 5.6% of the national FTE student enrolment in 2011, and, at UKZN, Indians in 2011 accounted for no less than 29.6% of the FTE student population. In no possible way is this more ‘a microcosm of the nation rather than the region’ as Govinder et al.[1] claim, mentioning UKZN specifically. Likewise, according to their reasoning, Coloured students are hugely over-represented at all institutions in the greater Cape Town area and at Nelson Mandela Metropolitan University. The controversy engendered by former Director-General of Labour Jimmy Manyi’s comments in 2010[9] about the ‘over-concentration’ of Coloureds in the Western Cape springs to mind.

A similar observation applies to the argument that national rather than provincial benchmarks should be used in the calculation of the Equity Index for administrative staff: HEIs are most likely to draw such staff from their surrounding communities. The recent decision of the Cape Town Labour Court[10] in the case brought by certain prison warders against the Department of Correctional Services determined that regional, rather than national, demographic profiles must be taken into account in setting employment targets.

Research output, quality and transformation

The authors understand that there could be tension between output/productivity and transformation of staff. They seek to investigate this tension by considering the relationship between research output and the EI. However, apart from the problems with their efforts, mentioned above, the authors assume all research output units are equivalent. This is evidently false. Thus, considering accredited research outputs, the assumption is made that a qualifying, but very brief, article in, for example, the *South African Journal of Science* is the equal of a 20-page article in *Nature.* Also implicit in their analysis is the assumption that all postgraduate degrees are of equal quality regardless of the awarding institution (never mind what proportions find employment or proceed to higher degrees, or whether standards are being maintained when an institution triples or quadruples output over the span of 4 or 5 years without significant changes to staff, etc.). Clearly neither is true, nor can the metric used offer any indication of the potential contribution to the national project. Thus one notes that the only university to appear in the 'best' quadrant (lower than mid-range (correctly calculated) EI and higher than mid-range output per capita), fails to make any impression on international systems of ranking of HEIs.

The assessment of performance allowing for any impact of moving towards national racial profiles as envisaged by the EI requires classifying output by whatever equity categories are used. However, unless one also controls for rank or age, this approach would serve only to bolster the argument that transformation takes place at the expense of output, as output is a function (to a large extent) of seniority.

It is also worth noting that Govinder et al.[1] identify a cluster of 8–11 institutions with the lowest EIs which they claim 'adds no value to national development', and are characterised by 'equity transformation without quality'. Their statements are problematic for a number of reasons. Firstly, these are apartheid legacy institutions, which have not 'transformed' but have rather historically always employed a high proportion of Black South Africans and have never produced large numbers of graduates with higher degrees or much research. Secondly, universities of technology have a markedly different role to play relative to comprehensive and traditional universities, and should not really be compared with these HEIs. Nonetheless, if the institutions that are 'most transformed' are indeed those that add least to national development, then it is not clear at all that a lower EI is something to which to aspire.

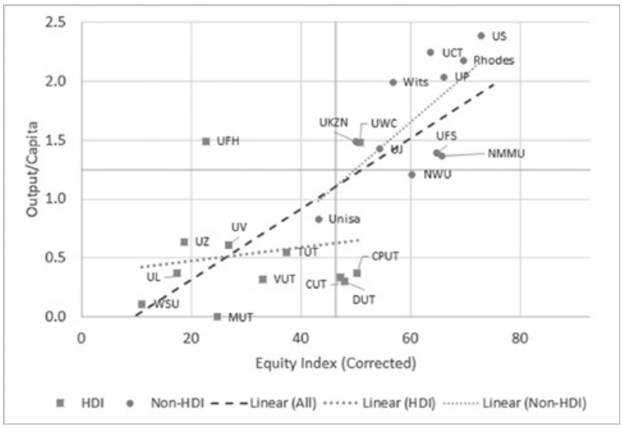
Misuse and misinterpretation of the Equity Index

The authors claim that 'there is no direct linear correlation between EI ranking and research productivity'. Inspection of Figure 2, which plots the output per capita of each university against the correctly calculated EI, shows that this claim is patently false. There is a clear linear trend, not only considering all universities, but, perhaps more pertinently, after excluding the group of '8–11 universities' that the authors argue 'adds no value to national development'.

A second feature which can be noted from Figure 2 is that if we split the plot area into four quadrants using the EI and output per capita of all institutions combined to determine the quadrants rather than the more arbitrary approach used by Govinder et al.[1], only the University of Fort Hare lies in the most desirable quadrant (bottom half EI, top half output per capita). In addition, the five institutions (which coincidentally include UKZN) identified by the authors as having 'the greatest potential for exhibiting good equity and high productivity' because they were found to be clustered around the centre of the plot, no longer cluster around the centre of the plot.

Turning to their application of the EI metrics to the student population, the authors argue, on the basis that the 'enrolment EI' exceeds the 'graduation EI' for all but three universities, that there is 'a definite equity profile of students dropping out of universities nationally'. However, while selective dropping out may well be occurring, one cannot be sure that this is the explanation for results presented in the paper. Because the comparison is not by year of enrolment (i.e. considering the progress of a cohort of students), this conclusion could be entirely erroneous. In particular, a university which has improved the EI of its student enrolment over recent years would find that its enrolment EI exceeded its graduation EI, irrespective of any selective dropping out.

An application of the EI not covered in the paper but presented elsewhere by the authors and given significant media coverage,[11] is the estimation of the time each institution is expected to take to transform, by calculating the EI at two time points (in their case 4 years apart) and extrapolating the trend linearly to determine when the EI reaches zero. Apart from the inappropriateness of extrapolating, in some cases several hundred years into the future, from only two, potentially arbitrarily chosen, points only four years apart, the authors seem completely oblivious of the fact that employees have rights, academics have tenure, and staff turnover and growth in staff at the most productive universities are low, so to a large extent the demographic profile of the staff is confined to the replacement of older retirees by younger new recruits.



HDI, historically disadvantaged institution; abbreviations of universities' names as in Govinder et al.[1]

Figure 2: Output per capita versus the corrected Equity Index by institution.

Conclusions

The need for transformation of higher education in South Africa is undisputed. However, a single summary index of the kind proposed by Govinder et al. may obfuscate more than it illuminates. Transformation of the demographic profile of HEIs is but one aspect of a much bigger transformation project. That project should encompass the national development priorities of the country, an understanding of the role and function of HEIs in that developmental agenda, and the role that South African universities, being generally the best in sub-Saharan Africa, can play in facilitating and driving development across the continent. By focusing on winners and losers (via the ranking system) it is probable that efforts will be made by HEIs to 'game' the index, at the expense of the other broader aspects of transformation. Rather, the focus of any metric of this sort should be on the progress of each institution within its circumstances, rather than on how it compares with other institutions.

We reiterate our concern that the EI as proposed, and its extensions to incorporate research output, offers little meaningful in terms of engaging with the question of quality of higher education offered in South Africa. In addition, the proposed EI is conceptually and algebraically flawed, the benchmarks chosen are problematic and the data from which the EI is calculated do not allow it to be presented in a form consistent with South African labour legislation. Finally, there is the possibility that the EI, by benchmarking and through the suggestion that transformation will only be complete when the EI is close to zero, is inconsistent with the recent judgement from the Supreme Court of Appeal which again confirmed that employment quotas contravene the *Employment Equity Act.*[12]

While a metric which tracks progress in transformation of staff and students may be of some use in monitoring progress (as opposed to producing league tables comparing institutions), much more work is required to develop a meaningful index. Not only will such an index have to account for more dimensions than the index proposed by Govinder et al., but productivity/output and the quality of that output would need to be taken into account, because, when all is said and done, sight of the primary objective of HEIs, namely, the production of suitably trained/qualified graduates and research to meet the needs of the country, should never be lost. Any effort to measure progress in transforming the institution should never undermine this primary objective. The development of such an index takes on even greater urgency in the light of the perceived utility of the completely inadequate current attempt to measure transformation at HEIs in South Africa.

Acknowledgements

While we have benefitted much from the input of colleagues at the University of Cape Town, the views presented here are solely those of the authors and do not necessarily reflect those of the University.

References

1. Govinder KS, Zondo NP, Makgoba MW. A new look at demographic transformation for universities in South Africa. S Afr J Sci. 2013;109(11/12), Art. #2013-0163, 11 pages. http://dx.doi.org/10.1590/sajs.2013/20130163

2. Govinder KS, Makgoba MW. An Equity Index for South Africa. S Afr J Sci. 2013;109(5/6), Art. #a0020, 2 pages. http://dx.doi.org/10.1590/sajs.2013/a0020

3. Dunne T. Mathematical errors, smoke and mirrors in pursuit of an illusion: Comments on Govinder et al. (2013). S Afr J Sci. 2014;110(1/2), Art. #a0047, 6 pages. http://dx.doi.org/10.1590/sajs.2014/a0047

4. Employment Equity Act of 1988 (Act 55 of 1998). Amendments to the Employment Equity Regulations. Government Gazette. 2006;28858. Pretoria: Government Printer; 2006.

5. Parliamentary Monitoring Group. Equity Index in South African universities: Briefing by Deputy Minister of Higher Education and Training and the Transformation Oversight Committee [homepage on the Internet]. c2013 [updated 2013 Oct 23; cited 2013 Dec 13]. Available from: http://www.pmg.org.za/report/20131023-equity-index-in-south-african-universities-briefing-deputy-minister-higher-education-and-training-and

6. Statistics South Africa. Census 2011 [database on the Internet]. c2011 [cited 2013 Dec 13]. Available from: http://interactive.statssa.gov.za/superweb/login.do

7. University of Cape Town. UCT Teaching and Learning Report. Cape Town: University of Cape Town; 2013.

8. Department of Basic Education. Report on the Annual National Assessment of 2013 [document on the Internet]. c2013 [cited 2013 Dec 13]. Available from: http://www.education.gov.za/LinkClick.aspx?fileticket=Aiw7HW8ccic%3D&tabid=36

9. Manyi: ‘Over-supply’ of Coloureds in Western Cape. Mail and Guardian [online]. 2011 Feb 24 [cited 2013 Dec 13]. Available from: http://mg.co.za/article/2011-02-24-coloureds-overconcentrated-in-wcape-says-manyi

10. Solidarity and Others v. Department of Correctional Services and Others (C 368/2012, C968/2012) (2013) ZALCCT 38 (2013 Oct 18) [case on the Internet]. [cited 2013 Dec 13]. Available from: http://www.saflii.org/za/cases/ZALCCT/2013/38.html

11. Masondo S. 382 years for SA’s top 5 research universities to transform. City Press [online]. 2013 Oct 23 [cited 2013 Nov 28]. Available from: http://www.citypress.co.za/news/382-years-sas-top-5-research-universities-transform/

12. Solidarity obo Barnard v. SAPS (165/2013) (2013) ZASCA 177 (2013 Nov 28) [case on the Internet]. [cited 2013 Dec 13]. Available from: http://www.justice.gov.za/sca/judgments/sca_2013/sca2013-177.pdf

Responsible conduct of research: Global trends, local opportunities

AUTHORS:
Theresa M. Rossouw[1,2]
Christa van Zyl[3]
Anne Pope[4]

AFFILIATIONS:
[1]Department of Family Medicine, Faculty of Health Sciences, University of Pretoria, Pretoria, South Africa

[2]Department of Immunology, Faculty of Health Sciences, University of Pretoria, Pretoria, South Africa

[3]Research Coordination, Ethics and Integrity, Human Sciences Research Council, Pretoria, South Africa

[4]Private Law Department, Law Faculty, University of Cape Town, Cape Town, South Africa

CORRESPONDENCE TO:
Theresa Rossouw

EMAIL:
theresa.rossouw@up.ac.za

POSTAL ADDRESS:
Department of Family Medicine, University of Pretoria, PO Box 667, Pretoria 0001, South Africa

DATES:
Received: 11 Apr. 2013
Revised: 04 Sep. 2013
Accepted: 10 Sep. 2013

KEYWORDS:
responsible; conduct; research; integrity; misconduct

HOW TO CITE:
Rossouw TM, Van Zyl C, Pope A. Responsible conduct of research: Global trends, local opportunities. S Afr J Sci. 2014;110(1/2), Art. #2013-0103, 6 pages. http://dx.doi.org/10.1590/sajs.2014/20130103

Instances of research misconduct reported in the lay and scientific literature as well as international efforts to encourage research integrity and the responsible conduct of research are currently receiving considerable attention. In South Africa, however, the topic has not featured prominently in public debate and clear evidence of a national, coordinated effort to address the problem of research misconduct seems to be lacking. Given increasing globalisation of research efforts, the need exists to promote standardised approaches to interpretation and implementation of the principles and values that underlie responsible conduct of research as well as to create guidelines and structures to promote integrity in research in the country. We explore the notions of research misconduct and research integrity, focusing on initiatives that promote responsible conduct of research, and propose a framework for the South African context.

Introduction

The topic of responsible conduct of research is currently receiving considerable attention. A reason for this attention is the increasing concern about revelations of fraud and other inappropriate behaviour in the research context.[1-3] Such revelations often surface in the lay media as high-profile cases and attract negative publicity that not only highlights the harm caused by the individual perpetrator, but also casts doubt on the integrity of the institution or scientific discipline within which the research was conducted. Another reason for the attention is the globalisation of the scientific enterprise and the burgeoning opportunities to work in interdisciplinary, transdisciplinary and applied research environments. In this global context, the need to foster a shared understanding of principles and values that form the foundation of research integrity, and to implement standardised approaches to designing, planning, conducting and administering research, ought to be clear.

The notion of 'responsible conduct of research' is distinguishable from both 'research integrity' and 'research ethics'. 'Research ethics' usually includes the processes in terms of which the proposed research study is scrutinised to assess compliance with the desired values and principles that are part of ethical research. 'Research integrity', on the other hand, has a broader meaning and may be understood to also incorporate implementation of the research processes and the conduct of the researchers. 'Responsible conduct of research' is an umbrella term that includes notions like authorship, plagiarism, research misconduct, whistle-blowing, research ethics guidelines, codes of conduct, conflict of interest, research ethics and other training. The distinction drawn between 'research integrity' and 'responsible conduct of research' is increasingly fading in practice, as is evident in the ensuing discussion in which we employ these terms interchangeably.

All of these concepts, when properly articulated and explained in solid policy and procedure documentation, serve to support and facilitate research conduct so that risks of harm are minimised. Various national and international bodies are currently developing or updating guidelines and policies to promote the responsible conduct of research (National Institutes of Health 2012, InterAcademy Council 2012, European Science Foundation 2011, Council of Canadian Academies 2010, Australian Government 2007, Organization for Economic Cooperation and Development 2007, to name but a few). In 2007, 2010 and 2013, consecutive World Conferences on Research Integrity brought together key role players including researchers, research managers, funders and journal editors in a global effort to foster responsible research. The 2010 Singapore Statement on Research Integrity,[4] published at the conclusion of the Second World Conference, emphasises the principles and professional responsibilities regarded as essential for integrity in research. While research institutions are held responsible for creating a climate conducive to desirable behaviour, the focus of the Singapore Statement is largely on the researcher: appropriate attitudes and behaviours expected of researchers as professionals are spelled out.[4]

The South African research community has by no means been free from instances of research misconduct.[5,6] Even if such cases have been dealt with quite decisively by institutions that employ the alleged perpetrators, the misconduct is not, as a rule, made known publicly, or subjected to external scrutiny or censure. There is no oversight body or association of interested entities that has taken on the role of drafting guidelines on responsible conduct of research and no entity mandated to deal directly with allegations or cases of misconduct beyond the institutional level. The National Health Research Ethics Council (NHREC) has an oversight role in the more specialised area of health research ethics, and may deal with complaints or appeals in this field. Even this role is not well known in disciplines beyond health research. Some individual academic and research institutions have endorsed the Singapore Statement, but others are still debating the efficacy of adopting or adapting the Statement and very limited efforts have been made to raise awareness on a national level.[7]

Whereas national and international standards for research ethics have arguably reached levels of maturity in terms of guidelines and application, we point to a need for similar work in related areas that aims to promote the responsible and accountable conduct of research. We explore the notions of research misconduct and research integrity and their importance within the context of responsible conduct of research. We describe the perceived international prevalence of and examine some African researchers' views about research misconduct and why it may occur and also briefly explore initiatives for promoting responsible conduct of research in other countries. Finally, we propose a framework for responsible research conduct for the South African context.

Background

Responsible conduct of research, or research integrity, is the cornerstone of excellent research; it also is a prerequisite for a flourishing academic research environment. Various definitions of research integrity exist, including 'the coherent and consistent application of values and principles essential to encouraging and achieving excellence in the search for, and dissemination of, knowledge'[8]. The Singapore Statement acknowledges different cultural and national standards for scientific research, but maintains that certain principles and professional responsibilities are fundamental to the integrity of research, whatever the context. In other words, the researcher has a personal and professional responsibility to behave ethically and responsibly and to conduct research with integrity. The Singapore Statement articulates four basic principles: *honesty* in all aspects of research; *accountability* in the conduct of research; *professional courtesy and fairness* in working with others; and *good stewardship of research* on behalf of others.[4] The principles of honesty and trust are emphasised as the golden threads found throughout the scientific enterprise: society should be able to trust the integrity, accuracy and honesty of scientific results, while researchers should be able to trust the meticulous and honest data capturing, analysis and reporting of results by colleagues.[9] At an institutional level, integrity is a 'commitment to creating an environment that promotes responsible conduct by embracing standards of excellence, trustworthiness, and lawfulness'[8].

Conceptually, research integrity requires adherence to *ethical principles* and *values* deemed essential for responsible research conduct, as well as adherence to *professional standards* set down by oversight bodies such as governmental entities, funding agencies, professional associations and employers. The ideal of research integrity is attained when individual researchers adopt the principles and practices of their profession as a personal credo, rather than merely accept them as impositions.[10]

Similarly, no uniform definition exists for research misconduct, i.e. behaviour that deviates from the accepted standards of research conduct. Early definitions were broad, such as 'non-adherence to rules, regulations, guidelines, and commonly accepted professional codes or norms'[11] or 'fabrication, falsification, plagiarism and other serious deviations from accepted practice'[12]. 'Other serious deviations' were taken to refer to diverse acts such as intentional protocol violations, dropping outliers from a data set or falsification of a biosketch or résumé. Concerns about the vagueness of 'other serious deviations' have led to recent definitions that restrict research misconduct to fabrication, falsification and plagiarism, which are regarded as the key concepts, akin to scientific fraud. For instance, the Office of Science and Technology Policy in the USA defines research misconduct as 'fabrication, falsification or plagiarism in proposing, performing, or reviewing research, or in reporting research results'[13].

In our view, such a narrow definition is unsatisfactory as other questionable behaviours could also bring the research profession into disrepute. These behaviours can be grouped together under the term 'questionable research practices' and defined as 'actions that violate traditional values of the research enterprise and that may be detrimental to the research process [but do not] directly damage the integrity of the research process'[14]. We concur with the proposal to separate research practices into three categories: deliberate misconduct – including fabrication, falsification and plagiarism (FFP); questionable research practices (QRP); and responsible conduct of research (RCR).[15] Here, RCR, or research integrity (RI), represents the expected or ideal standard, FFP denotes very serious transgressions and QRP falls somewhere in between.

Exploring the status quo of research misconduct

Globally, a variety of studies and analyses has attempted to describe and explain the perceived prevalence of, and the causes and costs of research misconduct. Possible preventative measures and remedies are also proposed within these studies. Of particular interest are initiatives that aim to promote RCR in an integrated and coordinated manner, and to establish clear procedures to manage allegations of non-adherence to expected standards. South Africa does not have a national coordinating entity that serves to promote RCR in the country.

A review of existing and emerging initiatives elsewhere was undertaken in order to propose a possible framework for the promotion of RCR in South Africa. The available literature, including policy documents and online resource materials, was surveyed and analysed against the backdrop of recent policy developments in South Africa. Our own experience in supervising research and developing support structures to address or promote RCR added to the contextualisation of the analysis.

Prevalence and impact of scientific misconduct

The analysis revealed increased interest in the field of research misconduct, yet little agreement about its prevalence. It was further noted that most studies and surveys have methodological concerns and other limitations. Until recently, it was a fairly widely held belief in the research community that FFP occurs rarely – mostly estimated at less than 1% – and that QRP, even though more prevalent, was usually not considered serious enough to warrant official action. Indeed, confirmed research fraud cases in the USA are variably reported as between 1 in 10 000 to 1 in 100 000 scientists, depending on the methodology used for calculation.[15,16] Furthermore, most of the profession held the view that scientific research could regulate itself through mechanisms such as peer review.[15] However, more recent data reveal that such views might have been overly optimistic.

A study published in 2008 estimated the rate of serious research misconduct or FFP at 3/100 researchers per year.[17] A recent meta-analysis showed that an average of 1.97% of scientists admitted to have fabricated, falsified or modified data at least once and a further 33.7% admitted to other QRP. When asked about the behaviour of colleagues, 14.12% suspected their colleagues of serious research misconduct and 72% thought colleagues were guilty of QRP.[18] Although much higher than the previously reported prevalence of <1%, these data may still be an underestimation of the actual prevalence of research misconduct. Indeed, a recent study in Nigeria – a first for the African continent – reported that 68.9% of 133 researchers participating in the study admitted to at least one of eight listed forms of scientific misconduct, 42.2% admitted to FFP and the most common QRP (affecting 36.4% of the sample) was reported as disagreement about authorship matters.[19]

Retractions from *PubMed* because of scientific misconduct indicate that up to 0.2% of published papers contain some form of fraud.[20] A comprehensive study by Grieneisen and Zhang[21] confirms that, since 2001, retractions have increased dramatically across a range of disciplines, with close to 20% explicitly citing scientific misconduct as the reason for retraction. Even though this figure represents a very small number of publications (<1%), the implications are serious. In particular, suspicion about the trustworthiness of published scientific data is given room to grow. Similarly, the implications also point to the potential risk posed by a new generation of technologically advanced researchers entering the scene. Consequently, the opportunities for research misconduct might increase exponentially, necessitating a more informed and vigilant approach on the part of editors, researchers and research institutions.

Policymakers, researchers and clinicians are also negatively affected by research misconduct, especially fabrication and falsification, as is evidenced by the recent case published in the *British Medical Journal*.[3] In this case, a researcher, considered by his peers to be a world leader, was found to have fabricated a very large amount of data and even whole studies. His work had been highly regarded and strongly influenced the formulation of policies and clinical practice. Both policies and practices are now in doubt because reanalysis of the meta-analyses after exclusion of his data radically changed the findings. Consequently, policies influenced by the fabricated work are no longer valid and clinical procedures based on his work also have to be revisited.

Possible causes of research misconduct

While one can only speculate as to reasons for the difference between the US and Nigerian data, it might be explained partly by lack of awareness, institutional support and regulatory oversight in the developing world. In addition, different cultures, generations and disciplines in research may interpret aspects of research misconduct in different ways and accordingly report differently on the prevalence thereof.[22] Having said that, however, it should be conceded that not much is known about the drivers and facilitators of research misconduct in any context.

In the Nigerian study, 73% of researchers sampled cited the need for publications as a push factor,[19] a view shared by research coordinators in the USA.[23] The worldwide drive to improve university rankings, resulting in part in the pressure to publish, is regarded as one cause of an increased incidence of plagiarism and QRP, such as so-called 'salami-slicing'.[24,25] More than 50% of the Nigerian researchers sampled thought that competition for external funding, the need for peer recognition, as well as insufficient explicit censure of misconduct strongly influence the prevalence of questionable behaviour.[19] This finding is important because it implies that the research community colludes in the facilitation of wrongdoing by remaining silent instead of speaking out. Not only does this finding highlight the need for a solid ethical foundation for researchers, but it also underscores the imperative of creating national standards for promoting RI and enabling institutional oversight.

In addition, pressure to gain tenure, unclear definitions of misconduct, financial conflicts of interest as well as the level of involvement of the principal investigator in the enrolment of human participants were considered by the majority of the Nigerian sample to have 'some influence' on questionable behaviour patterns. The level of interest shown by the principal investigator in study enrolments and outcomes, the number of open or current research studies for which the principal investigator is responsible, and the belief that the risk of harm for participants is low in a study also contribute to lax attitudes towards questionable behaviour patterns.[19]

It is vital that further context-specific research into possible facilitators of research misconduct be conducted in order to inform future training and guide policy frameworks.

Initiatives to promote responsible conduct of research

Recently, national and international bodies have developed policy documents that focus on RCR as an integrated construct that includes all phases and aspects of the research endeavour. In these documents, specific responsibilities are allocated to individual researchers, to research institutions and to oversight bodies. Examples of such initiatives can be found across the globe. Approaches to oversight range from countries with national guidelines and government-sponsored offices for statutory reporting and oversight, to national or regional interest groups promoting coordination between institutions, to countries where no national guidelines or oversight bodies exist.

An example of a country with a national guideline is Australia. The Australian Code for the Responsible Conduct of Research was developed through a collaborative effort of the National Health and Medical Research Council, the Australian Research Council and Universities Australia. The purpose of the Australian Code is to guide institutions and researchers in responsible research practices. The Australian Code includes specific guidance on how to manage breaches of the Code and allegations of research misconduct, maintain research data and material, publish and disseminate research findings, attribute authorship, conduct effective peer review, and manage conflicts of interest.[26] Another example of a country with a well-developed national guideline is Canada, where the Tri-Council policy statement fulfils a similar function to the Australian Code.[27]

More recently, triggered by the globalisation of research, international declarations, such as the Singapore Statement on Research Integrity[4] and the 2012 policy work of the global network of science academies and the IAP[22], have been formulated. These documents reflect consensus amongst the individual researchers present, e.g. participants at the Second World Conference on Research Integrity contributed to the creation and adoption of the Singapore Statement. The relatively short and succinct Statement calls for the globalisation of a basic understanding of RCR and is the 'first international effort to encourage the development of unified policies, guidelines and codes of conduct'[4]. The aim of the Statement is to initiate a global discussion on RI and the development of a set of international norms and standards that can serve as the basis for national and regional ethics guidelines, while acknowledging and accommodating national differences. The Statement can be a valuable tool to countries such as South Africa, for which no national guidelines or oversight bodies exist.

The South African context

Despite instances of local research fraud receiving international attention, South Africa has not introduced a system of formal scrutiny or censure. A National Health Research Ethics Council (NHREC) was established in terms of the *National Health Act 61/2003*.[28] Its role is to provide guidance for researchers in health research, principally regarding research ethics and not RI as such. For instance, the NHREC is responsible for ensuring that up-to-date research ethics guidelines are available and accessible, and for assisting with complaints, queries and capacity building for research ethics committees. The NHREC has a Complaints and Advisory Disciplinary Committee, mandated to 'refer to the relevant statutory health professional council matters involving the violation or potential violation of an ethical or professional rule by a health care provider' and to 'institute such disciplinary action as may be prescribed against any person found to be in violation of any norms and standards, or guidelines, set for the conducting of research' [28,29]. However, this committee's jurisdiction is limited to protection of human participants and animals used in research, hence cases involving allegations such as plagiarism or data fabrication fall outside its mandate.

In the absence of national guidelines or codes of conduct specific to RCR, much reliance is placed on professional codes of conduct guiding the behaviour of individual researchers, or on institutional guidelines to promote the responsible and ethical conduct of research. The NHREC also favours the approach that institutions should first attempt to resolve matters internally before referring them to the NHREC. This view is consistent with the administrative law rule that one exhausts domestic remedies before seeking outside adjudication. This approach supports the autonomy of institutions and their power to govern themselves, which is statutorily mandated. On the other hand, this approach may inadvertently disadvantage institutions that lack the know-how, resources and infrastructural support to deal with allegations and cases of misconduct. It is further possible that at individual institutions there might be reluctance to engage effectively with researchers, especially high-profile researchers, alleged to have committed research misconduct, because of an inherent conflict of interest.[30] Such attitudes could lead to examples of unequal treatment of similar incidents between and even within institutions. A centralised office might be better placed to ensure a fair and equitable approach by establishing national guidelines on the appropriate procedures to be followed in cases of alleged and confirmed misconduct. Even though it is preferred that institutions deal with such matters internally, a centralised office could also function as an adjudicatory or referring body for instances when individual institutions are unable to resolve cases, procedures are disputed or conflicts of interest exist.

However, the NHREC is mandated to deal only with matters concerning health research. In increasingly complex research arenas, it seems desirable to cross institutional and disciplinary boundaries by promoting harmonised approaches to interpretation and implementation of principles and values underlying RCR. A problem for South Africa is that the various documents outlining science policy for the country (e.g. Department of Arts, Culture, Science and Technology White Paper 1996, National Research and Development Strategy 2002, Ten-Year Plan for Innovation 2008, Ministerial Review Report 2012) are silent on key aspects of research ethics and integrity, and miss the opportunity to create guidelines and structures to promote the quality and responsible conduct of research in the country. The 2012 Ministerial Review Report suggested that a national oversight body for science policy should be established. Such a body could establish guidelines for a national system or structure(s) to promote RCR, and to respond to possible research misconduct. This body could provide the needed transdisciplinary leadership for guiding the prevention, investigation and correction of research misconduct, while ensuring that research institutions remain primarily responsible for management of individual cases.

Whether a centralised approach is desirable has not been debated in public fora in South Africa. Informally, it seems that opinions are divided: some think that a central bureaucracy is desirable given the attraction of consistency, standardisation and institutional support; others favour the autonomy of research institutions in dealing with allegations of misconduct expeditiously, discreetly and effectively. Bureaucracies are not easily able to act in this way. Furthermore, a centralised approach has considerable specific human and financial resource requirements and is likely to not be the most time- and cost-effective approach in a country with limited resources.

On an individual level, South African researchers have contributed to the international debate. Representatives from the National Research Foundation (NRF) and the Academy of Science of South Africa attended the 2010 World Conference on Research Integrity and the NRF subsequently promoted broader awareness of the Singapore Statement on Research Integrity.[7] It seems that some South African institutions of higher learning and other research entities are adopting this document as a reference point, which might signal the dawn of a shared understanding of RI in the country.

Proposed framework for South Africa

A culture of research excellence and RI requires contributions from actors and interest groups that operate at different levels and in a range of roles and responsibilities. Initiatives to promote RCR should therefore be multipronged and aimed at multiple and diverse role players. Using the categories included in the 2012 InterAcademy Council policy document[22] as a point of departure, we propose the inclusion of the following actors in the South African framework:

- *Individual researchers*: Researchers are the foundation of sound scientific practice and need to be imbued with strongly developed moral and ethical reasoning skills. They are required to embrace, uphold and promote professional standards of research excellence in their own research work and when reviewing the work of others. In addition, the notion of RCR requires them to actively promote appropriate conduct in their teaching and mentoring relationships.

  The South African research community would benefit from improved communication systems designed to provide guidance and support in order to promote RCR. In particular, researchers should reflect on their own knowledge and understanding of RCR and contribute to discussions, at least at their own institutions. In this way, the valuable intellectual capital of our researchers may be shared and invested in sustaining the research enterprise in South Africa.

- *Research and academic institutions*: Within institutional settings, an environment conducive to RCR may be fostered by having appropriate and clear policies and procedures. These policies and procedures should stipulate how cases of alleged or suspected misconduct should be managed. This recommendation does not call for more regulation or for an infringement of institutional autonomy or academic freedom. Rather, we recommend frank and rigorous discussions, and careful and thoughtful reflections, so that where systems may be weak, support can be given and best practices may be shared.

- *Editorial boards and publishers of scientific journals*: Researchers who submit research outputs for publication in scholarly journals should provide proof of ethics approval of the research undertaken, as well as a clear indication of authorship allocations. This requirement is already specified by many journals. Furthermore, plagiarism-checking software has improved the ability to trace incidents of plagiarism and should be used routinely. Recently, after the case of research misconduct reported in the *British Medical Journal*, some commentators are also now calling for all authors to verify that they have seen the original data.[3]

- *National and international professional organisations:* In the past decade, these organisations have done much to give prominence to responsible research conduct and the promotion thereof. Examples include recent reports of interacademy councils in Canada (2010)[8], Europe (2010)[31] and internationally (2012)[22]. World conferences dedicated to RI are supported by funding bodies, science academies and professional associations and can act as a platform to raise awareness and share resources. South African institutions should be encouraged to actively participate in and contribute to these international debates.

- *National governments and government departments:* Government is in a position to provide support for, or actively promote, the RCR. Although South Africa has national guidelines and oversight in relation to ethics in health-related research (i.e. *The National Health Act*[28] and the South African Good Clinical Practice Guidelines[32]), it lacks similar guidelines and oversight to promote and provide guidance in relation to other aspects of RCR. Government departments can play a role by making funding available to support the development and sharing of guidelines that can become a national repository of best practice and resource for independent guidance when required. In particular, we recommend that the Singapore Statement be incorporated into the national research ethics collection of guidelines, either by reference or by a statement endorsed by various research bodies.

- *Funding bodies:* The exact role of funders in promoting RI is controversial. While there are justified concerns about the perils of funders setting the research agenda and issues around effective and appropriate oversight, funders may play a role by insisting, as a prerequisite to funding eligibility, that minimum standards for policies and procedures to promote RCR are in place. Funding bodies may also introduce additional mechanisms aimed at promoting responsible research practices, such as independent post-submission reviews of reports which may include standard checks for possible examples of plagiarism or standard contractual requirements dealing with authorship agreements, or open access to data or research findings. Furthermore, funding bodies are in a position to support research about aspects of RI, as well as the development of training materials and programmes. It should be noted, however, that funding bodies do not have direct jurisdiction over the quality and integrity of research not funded by them, hence a case being made for generic norms and standards to be adopted at national or international level.

  The NRF could potentially play such an enabling role in South Africa. The NRF has already indicated that it will be investigating ways of implementing the Singapore Statement, for instance, by making the implementation of its principles a prerequisite for grant agreements with grantholders. The NRF has further committed itself to 'translate the statement into South Africa's eleven official languages and to disseminate it widely amongst research institutions and government departments'[7].

- *Research networks and ethics training programmes*: Research networks and professional associations provide platforms for the exchange of information and shared learning. There are well-established South African examples of relevant professional and training networks, such as the South African Research Ethics Training Initiative (SARETI), the Southern African Research and Innovation Management Association (SARIMA) and Advancing Research Ethics Training in Southern Africa (ARESA) which replaces the International Research Ethics Network for Southern Africa (IRENSA).

  These networks currently focus specifically on research ethics, as do most of the online training programmes that have been developed by international institutions and consortiums, such as Training and Resources in Research Ethics Evaluation (TRREE) and the African Malaria Network Trust (AMANET). Some of these programmes deal with RCR, but the focus is mostly tangential. If this aspect could be further developed and strengthened, these programmes could become a valuable resource for all researchers, and institutions could consider incorporating them into continued professional development curricula.

Conclusion

Revelations of research misconduct are embarrassing for journals and academic institutions, deleterious to the research enterprise in general and disastrous for the researcher who is found to have cheated. The advent of easy access to information through electronic media, coupled with a new generation of technologically advanced researchers faced with ever increasing pressure to publish and secure funding, mean that opportunities for research misconduct might increase in future. In addition, the globalisation of the scientific enterprise and increasing opportunities for interdisciplinary and international research, highlight the need for a shared understanding of the principles and values that inform the notion of RI.

South Africa, despite being home to some high-profile instances of research misconduct, has not yet escalated the problem of research misconduct to the level of public debate and has not yet embarked on a national and unified effort to put systems in place for its regulation. Although individual researchers and research institutions have been party to the international debate and have endorsed the Singapore Statement, many institutions lack the resources and infrastructure to follow suit. We therefore support the development of coordinated approaches to designing, planning, conducting and administering research, in conjunction with appropriate and clear policies and procedures stipulating how cases of alleged or suspected misconduct should be managed. We do not call for more regulation or for an infringement of institutional autonomy or academic freedom, but rather for the establishment of policies and guidelines that should form part of a national repository of best practice and resource for independent guidance when required.

We propose that multiple role players be involved in establishing a South African framework for RCR, including research and academic institutions, editorial boards, professional organisations, government departments, funding bodies and research networks. Such a framework could then be maintained and implemented by a centralised body, such as an Office of Scientific Research Integrity. The vision of an honest and trustworthy research enterprise can, however, only be realised when individual researchers have been imbued with the strongly developed moral character needed to embrace, uphold and promote professional standards of research excellence.

Authors' contributions

T.R. conceived the idea and wrote the first draft. C.v.Z. and A.P. gave significant intellectual input; C.v.Z. specifically contributed significantly to the section on international initiatives and the proposed framework for South Africa. C.v.Z. and A.P. assisted in revising and refining the draft into its current format. All three authors read and approved the final version.

References

1. Marc Hauser engaged in research misconduct. Harvard Magazine [serial on the Internet]. 2012 Sep 05 [cited 2013 Mar 05]. Available from: http://harvardmagazine.com/2012/09/hauser-research-misconduct-reported.
2. Pelley S. Deception at Duke: Fraud in cancer care? [homepage on the Internet] c2012 [cited 2013 Mar 05]. Available from: http://www.cbsnews.com/8301-18560_162-57376073/deception-at-duke-fraud-in-cancer-care/
3. Wise J. Research misconduct. Boldt: The great pretender. BMJ. 2013;346:f1738.
4. 2nd World Conference on Research Integrity. Singapore statement on research integrity [homepage on the Internet]. c2010 [cited 2012 Oct 12]. Available from: www.singaporestatement.org.
5. Gottlieb S. Breast cancer researcher accused of serious scientific misconduct. BMJ. 2000;320(7232):398.
6. Sidley P. Another AIDS "cure" scandal hits South Africa. BMJ. 1998;316(7146):1696. http://dx.doi.org/10.1136/bmj.316.7146.1696f
7. National Research Foundation Emerging Researchers Network. Singapore Statement on research integrity receives thumbs up [homepage on the Internet]. c2011 [cited 2013 March 21]. Available from: http://ern.nrf.ac.za/control/ViewFeatureArticle?contentId=11188&featureContentId=11188&articleType=mainFeature.
8. Council of Canadian Academies. The expert panel on research integrity. Honesty, accountability and trust: Fostering research integrity in Canada. Ottawa: Council of Canadian Academies; 2010.
9. National Academy of Sciences, National Academy of Engineering, Institute of Medicine, Committee on Science, Engineering, and Public Policy. On being a scientist: Responsible conduct in research. 2nd ed. Washington DC: National Academy Press; 1995.
10. Korenman SG. Teaching the responsible conduct of research in humans [homepage on the Internet]. c2006 [cited 2012 Oct 10]. Available from: www.ori.hhs.gov/education/products/ucla/chapter1/page02.htm.
11. Broome ME, Pryor E, Habermann B, Pulley L, Kincaid H. The scientific misconduct questionnaire–revised (SMQ-R): Validation and psychometric testing. Account Res. 2005;12(4):263–280. http://dx.doi.org/10.1080/08989620500440253
12. Buzzelli DE. The definition of misconduct in science: A view from NSF. Science. 1993;259(5095):584–585. http://dx.doi.org/10.1126/science.8430300
13. United States Office of Science and Technology Policy. Federal Policy on Research Misconduct. Federal Register. 2000;65(235):76260–76264.
14. National Academy of Science, Committee on Science Engineering and Public Policy, Panel on Scientific Responsibility and the Conduct of Research. Responsible science: Ensuring the integrity of the research process. Washington DC: National Academy Press; 1992.
15. Steneck NH. Fostering integrity in research: Definitions, current knowledge, and future directions. Science and Engineering Ethics. 2006;12:53–74.
16. Marshall E. Scientific misconduct – how prevalent is fraud? That's a million-dollar question. Science. 2000;290:1662–1663. http://dx.doi.org/10.1126/science.290.5497.1662
17. Titus SL, Wells JA, Rhoades LJ. Repairing research integrity. Nature. 2008;453(7198):980–982. http://dx.doi.org/10.1038/453980a
18. Fanelli D. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. PLoS One. 2009;4(5):e5738. http://dx.doi.org/10.1371/journal.pone.0005738
19. Okonta P, Rossouw T. Prevalence of scientific misconduct among a group of researchers in Nigeria. Dev World Bioeth. 2013;13(3):149–157. http://dx.doi.org/10.1111/j.1471-8847.2012.00339.x
20. Claxton LD. Scientific authorship. Part 1. A window into scientific fraud? Mutat Res. 2005;589(1):17–30. http://dx.doi.org/10.1016/j.mrrev.2004.07.003
21. Grieneisen ML, Zhang M. A comprehensive survey of retracted articles from the scholarly literature. PLoS ONE. 2012;7(10):e44118. http://dx.doi.org/10.1371/journal.pone.0044118
22. InterAcademy Council / IAP. Responsible conduct in the global research enterprise: A policy report. Amsterdam: Bejo Druk & Print, Alkmaar; 2012. Available from: http://www.interacademies.net/File.aspx?id=19789.
23. Fletcher SW, Fletcher RH. Publish wisely or perish: Quality rather than quantity in medical writing. Ann Acad Med Singapore. 1994;23:799–800.
24. Jefferson T. Redundant publication in biomedical sciences: Scientific misconduct or necessity? Sci Eng Ethics. 1998;4(2):135–140. http://dx.doi.org/10.1007/s11948-998-0043-9
25. Roig M. Avoiding plagiarism, self-plagiarism, and other questionable writing practices: A guide to ethical writing [document on the Internet]. No date [cited 2013 Mar 12]. Available from: http://ori.dhhs.gov/education/products/plagiarism/plagiarism.pdf.
26. Australian Government. Australian code for the responsible conduct of research [document on the Internet]. c2007 [cited 2012 Oct 15]. Available from: http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/r39.pdf.

27. Canadian Institutes of Health Research, Natural Sciences and Engineering Research, Council of Canada, and Social Sciences and Humanities Research Council of Canada. Tri-Council policy statement: Ethical conduct for research involving humans [document on the Internet]. c2010 [cited 2013 Mar 22]. Available from: http://www.ethics.gc.ca/pdf/eng/tcps2/TCPS_2_FINAL_Web.pdf.

28. Republic of South Africa. National Health Act No. 61 of 2003. Government Gazette. 2004;469(26595).

29. National Health Research Ethics Council. Guidelines for the management of complaints. Complaints and Advisory Disciplinary Committee [document on the Internet]. c2012 [cited 2013 Aug 26]. Available from http://www.nhrec.org.za/wpcontent/uploads/2013/guideline_mngcomplnt.pdf.

30. Ana J, Koehlmoos T, Smith R, Yan LL. Research misconduct in low- and middle-income countries. PLoS Med. 2013;10(3):e1001315. http://dx.doi.org/10.1371/journal.pmed.1001315

31. European Science Foundation. Fostering research integrity in Europe: A report by the ESF Member Organisation Forum on research integrity [homepage on the Internet]. c2010 [cited 2013 Mar 05]. Available from http://www.esf.org/activities/mo-fora/publications.html.

32. Department of Health. Guidelines for good practice in the conduct of clinical trials with human participants in South Africa. Pretoria: Department of Health; 2006.

Inflammation and cancer: The role of the human neutrophil

AUTHORS:
Ronald Anderson[1]
Gregory R. Tintinger[2]
Charles Feldman[3]

AFFILIATIONS:
[1]Medical Research Council Unit for Inflammation and Immunity, Department of Immunology, Faculty of Health Sciences, University of Pretoria, Pretoria, South Africa

[2]Department of Internal Medicine, Steve Biko Academic Hospital, Faculty of Health Sciences, University of Pretoria, Pretoria, South Africa

[3]Division of Pulmonology, Department of Internal Medicine, Charlotte Maxeke Johannesburg Academic Hospital, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

CORRESPONDENCE TO:
Ronald Anderson

EMAIL:
ronald.anderson@up.ac.za

POSTAL ADDRESS:
Department of Immunology,
University of Pretoria, PO
Box 2034, Pretoria 0001,
South Africa

DATES:
Received: 02 July 2013
Revised: 12 Sep. 2013
Accepted: 12 Sep. 2013

KEYWORDS:
angiogenesis; chemokines; cytokines; hydrogen peroxide; proteinases; redox signalling; reactive oxygen species

HOW TO CITE:
Anderson R, Tintinger GR, Feldman C. Inflammation and cancer: The role of the human neutrophil. S Afr J Sci. 2014;110(1/2), Art. #2013-0207, 6 pages. http://dx.doi.org/10.1590/sajs.2014/20130207

Chronic inflammation of both infective and non-infective origin has been implicated in the aetiology of approximately 30% of all human epithelial malignancies. The primary carcinogens are reactive oxygen species (ROS) derived from activated, infiltrating cells of the innate immune system, especially neutrophils, which inflict oxidative damage on the DNA of bystander epithelial cells. The consequence is gene modifications which initiate cellular transformation. The process of tumourigenesis is exacerbated by the sustained generation of pro-proliferative ROS, as well as by the release of neutrophil-derived cytokines and proteases, all of which contribute to tumour promotion and progression. It is now well recognised that, in addition to inflammation causing cancer, many cancers per se induce an inflammatory response, with a high magnitude of neutrophil influx being indicative of a poor prognosis. In this setting, CXC chemokines produced by tumours not only promote neutrophil influx and hyperreactivity, but also cause autocrine activation of the proliferation of the chemokine-producing tumour cells. These various mechanisms of inflammation-mediated tumourigenesis are the primary focus of this review, together with a consideration of neutrophil-targeted anti-inflammatory strategies with potential as adjunctive cancer therapy.

Introduction

The link between cancer and inflammation was described 150 years ago by the distinguished German pathologist Rudolph Virchow.[1,2] Based on epidemiological studies, it is now recognised that up to 30% of all cancers may have underlying inflammation-associated aetiology, triggered by chronic infection or other types of non-infective unresolved inflammation.[3-6] In this setting, reactive oxygen and nitrogen species (ROS/RNS) released by activated phagocytes at sites of inflammation inflict oxidative damage on neighbouring bystander cells, especially epithelial cells, initiating tumourigenesis.[7,8] Subsequent promotion/progression and metastasis involve not only ongoing oxidative stress, but also the release of pro-proliferative and pro-angiogenic/pro-metastatic, phagocyte-derived cytokines/chemokines and proteases.[1-17]

Although inflammation is a major and primary cause of cancer, many cancers per se also activate an inflammatory response, resulting in infiltration by various types of myeloid cells of the innate immune system. These cells include neutrophils, monocytes/macrophages, dendritic cells, and so-called myeloid-derived suppressor cells of both granulocytic and monocytic origin.[18-23] Although this tumour-associated inflammatory response is potentially protective, at least in the case of neutrophils, monocytes/macrophages and dendritic cells,[19,20] it may also contribute to tumour progression and metastasis through the mechanisms alluded to above.[9]

In this review, we focus on the pro-tumourigenic potential of the neutrophil, which, amongst other types of immune and inflammatory cells, is abundant in tumours, appearing to be an important, independent predictor of poor outcome in many,[3,21] but not all, types[24] of malignancy. The major themes addressed here are the roles of neutrophil-derived/-associated ROS, chemokines/cytokines, proteinases and adhesion molecules in tumour initiation, promotion, progression and metastasis, as well as the potential role of anti-inflammatory strategies in cancer prevention and therapy.

Neutrophils and carcinogenesis

As described by Weitzman and Gordon in their seminal review[7] and, more recently, by Knaapen et al.[10], the propensity for cancers to develop at sites of inflammation is well recognised, the association being supported by compelling epidemiological and experimental evidence. Examples of inflammation-associated cancers, primarily epithelial, of both infective and non-infective origin are shown in Tables 1 and 2, respectively. In the setting of inflammation-associated cancer, phagocyte-derived ROS – produced and released extracellularly by infiltrating neutrophils – have been identified as the primary offenders.[7,10] These indiscriminate, toxic agents are potent carcinogens, posing the potential hazard of oxidative damage to the DNA of bystander, host structural cells at sites of inflammation and resulting in the gene modifications which precede cellular transformation.

Convincing evidence demonstrating the carcinogenic potential of ROS was derived from experiments in which eukaryotic structural cells and lymphocytes were exposed to activated neutrophils, to cell-free enzymatic ROS-generating systems, or to the relatively stable, cell-permeant ROS hydrogen peroxide (H_2O_2) in vitro, which resulted in severe oxidative stress and damage to the genetic material of these cells.[7,10] In all of these systems, direct oxidative damage to DNA appears to involve intracellular conversion of H_2O_2 to a highly potent and reactive ROS – hydroxyl radical – probably by Fenton-type mechanisms involving electron donation by heavy metals. The types of ROS-mediated damage include: (1) gross chromosomal damage (sister chromatid exchanges), (2) single- and double-DNA strand breaks and (3) oxidative damage to the bases in DNA.[7,10] In the case of the latter, the signature of oxidative damage is conversion of guanosine to 8-hydroxydeoxyguanosine, although the other DNA bases are also vulnerable to oxidative damage.[7] RNS produced predominantly by macrophages result in the formation of reactive aldehydes and malondialdehydes which also induce point mutations.[8]

In addition to the direct, DNA-damaging activities of phagocyte-derived ROS, these oxidants also inhibit the activities of several DNA repair enzymes, thereby exacerbating oxidative damage to genetic material.[10] In this context, it is noteworthy that hypochlorous acid generated via the H_2O_2-dependent oxidation of chloride ions by myeloperoxidase (MPO), the neutrophil/monocyte primary granule enzyme, has been reported to interfere with the base excision repair enzyme poly (ADP-ribose) polymerase.[25] Other DNA repair enzymes which are susceptible to oxidative inactivation include the glycolase Ogg1 and topoisomerase II, which are inactivated by nitric oxide and H_2O_2, respectively, compromising repair of 8-hydroxydeoxyguanosine moieties and strand scission/ligation.[10,26,27]

Table 1: Examples of inflammation-related malignancies of chronic infective origin

Type of malignancy	Associated infective agent
Squamous cell carcinoma of the bone, sinuses and skin	Chronic osteomyelitis most commonly caused by *Staphylococcus aureus*
Urinary bladder cancer	*Schistosoma haematobium*
Ovarian cancer	Pelvic inflammatory disease most commonly caused by *Chlamydia trachomatis* and *Neisseria gonorrhoeae*
Gastric cancer	Gastritis caused by *Helicobacter pylori*
MALT lymphoma	*Helicobacter pylori*
Lung carcinomas	Chronic and recurrent pulmonary infection as a result of various bacterial pathogens
Testicular cancer	Orchitis caused by mumps virus
Hepatocellular carcinoma	Hepatitis viruses B and C
Cervical cancer	Human papilloma virus
Kaposi's sarcoma	Human herpes virus type 8

Sources[1–9]

Table 2: Examples of inflammation-related malignancies of chronic non-infective origin

Type of malignancy	Associated condition
Colon carcinomas	Inflammatory bowel disease (Crohn's disease, colitis)
Urinary bladder cancer	Long-term indwelling catheters, stones
Gall bladder cancer	Chronic cholecystitis, cholelithiasis
Oesophageal squamous cell carcinoma and adenocarcinoma	Chronic exposure to chemical irritants and acid reflux oesophagitis, respectively
Lung carcinomas	Cigarette smoking, pulmonary fibrosis, sarcoidosis
Mesothelioma	Asbestos inhalation
Head and neck cancer	Cigarette smoking
Skin cancer (basal cell/squamous cell carcinoma, melanoma)	Exposure to sunlight

Sources[1–9]

Clearly, ROS-mediated direct damage to DNA, together with oxidative dysfunction of DNA repair enzymes, predisposes to gene modifications, which, particularly in the case of mutations to tumour suppressor and promoter genes, may lead to cellular transformation. However, these mechanisms are not the only ones by which phagocyte-derived ROS contribute to carcinogenesis. Other mechanisms include: (1) oxidative conversion of pre-carcinogenic chemicals/xenobiotics to complete carcinogens[10,28], (2) redox activation of intracellular signalling mechanisms, which not only promote aberrant cellular proliferation, but also intensify inflammation-related oxidative stress[29-34] and (3) oxidative suppression of the proliferative activity of anti-tumour T lymphocytes[35,36]. In addition to these mechanisms, several neutrophil-derived chemokines/cytokines, proteinases and adhesion molecules also contribute to tumourigenesis via their pro-proliferative, pro-angiogenic and pro-metastatic activities.

Neutrophil-mediated oxidative activation of pre-carcinogens

Neutrophil-derived ROS, specifically those generated by the MPO/H_2O_2/halide system, have been implicated in the transformation to carcinogens of chemical pollutants generated by industrial, motor vehicle and household combustive processes, as well as those present in cigarette smoke.[10] Examples of the former include aromatic and heterocyclic amines, especially polycyclic aromatic hydrocarbons, while benzo(a)pyrene in cigarette smoke undergoes oxidative conversion to BPDE (bay-region diol expoxides), which is mutagenic via formation of covalent adducts with guanine.[10] In addition, MPO-derived ROS have been reported to convert the anti-cancer drug etoposide to its potentially mutagenic phenoxy radical, which may explain the increased frequency of secondary myeloid leukaemia in cancer patients treated with this agent.[28]

Redox activation of cellular proliferation

Unlike other ROS (such as superoxide anion, singlet molecular oxygen, hydroxyl radical and hypochlorous acid), H_2O_2 – because of its relative stability, cell permeability and ability to target proteins – can function efficiently as an intracellular signalling molecule.[29-31] Indeed, it is well established that H_2O_2 can modulate cellular differentiation, proliferation, survival and synthesis of inflammatory mediators via the oxidative modification of key cysteine residues in various enzymes, including phosphatases and kinases, especially mitogen-activated protein kinases (MAPKs), as well as transcription factors.[29-32] Under controlled conditions, H_2O_2 generated intracellularly in various types of cells by the ubiquitous, stringently regulated Nox (NADPH oxidase) family of enzymes, contributes positively to the maintenance of cellular homeostasis.[29-32] However, when structural cells, especially epithelial cells, are subjected to intense oxidative stress, whether directly as a consequence of protracted activation of Nox enzymes or indirectly because of influx of extracellular H_2O_2 as a result of proximity to activated phagocytes, or both, then cell proliferation as a consequence of dysregulated intracellular signalling may ensue. Although disputed by those who believe that over-exposure to H_2O_2 is more likely to drive the cells into apoptosis,[29-32] this scenario is countered to some extent by the following lines of evidence from experimental sources: (1) exposure of a Barrett's oesophagus adenocarcinoma cell line to low concentrations of H_2O_2 resulted in cell proliferation which was associated with sequential activation of extracellular regulated kinase 2, MAPK, the transcription factor, nuclear factor kappa B (NFκB) and Nox 5-S[33]; and (2) exposure of human oral cancer cells to the H_2O_2-producing microorganism *Enterococcus faecalis* resulted in catalase-inhibitable activation of the epidermal growth factor receptor and cell proliferation, underscoring the association between infection with this bacterial pathogen and oral carcinogenesis.[34]

ROS-mediated inactivation of tumour-targeted T cells

Although H_2O_2 at low concentrations can trigger the proliferation of epithelial cells via intracellular, redox signalling mechanisms, at higher concentrations this ROS can also promote the oxidative inactivation of the protective activities of T lymphocytes.[35,36] In the setting of murine models of experimental tumourigenesis, infiltrating phagocytes, most notably a subset of activated neutrophils, have been found to inhibit the protective responses of tumour-targeted T-lymphocytes. The mechanism of immunosuppression involves intimate cell-cell contact mediated by the neutrophil β2-integrin Mac-1, exposing the T cells to high concentrations of neutrophil-derived H_2O_2.[35,36]

Neutrophil-derived cytokines in tumourigenesis

Although originally believed to have a very short lifespan and an extremely limited biosynthetic capacity, the survival time of neutrophils in the circulation of healthy humans has recently been reported to be 5.4 days.[37] Following extravasation to sites of infection, tissue injury or cancer, this time may be considerably longer because of exposure to anti-apoptotic cytokines, especially granulocyte colony-stimulating factor (G-CSF) and granulocyte/macrophage colony-stimulating factor (GM-CSF).[38] Extended survival of neutrophils is associated with acquisition of the capacity, albeit limited, to synthesise cytokines or chemokines,[14,39] some of which are already stored in pre-synthesised, rapidly mobilisable form in cytoplasmic secondary and tertiary granules.[40,41] These include: (1) the pro-angiogenic growth factor vascular endothelial growth factor (VEGF), (2) the chemokine interleukin (IL)-8 and the cytokines IL-1β, IL-6 and TNF, all of which interact to promote neutrophil extravasation, accumulation and activation and (3) IL-12 which links innate and adaptive immunity, promoting cell-mediated immune responses involving T helper 1 lymphocytes.[14,39-41] With respect to tumour promotion/progression, the most prominent of these are VEGF, which promotes tumour neovascularisation, and IL-8, which not only sustains neutrophil influx and activation, but also promotes tumour cell proliferation by the autocrine and paracrine mechanisms described under 'Chemokines and tumourigenesis'.

Although the evidence is somewhat less compelling than that for VEGF and IL-8, several other neutrophil-derived cytokines have been implicated in tumour promotion/progression and angiogenesis. Because these have been extensively reviewed recently,[14] they are considered only briefly here.

APRIL (also known as 'a proliferation-inducing ligand') and BAFF (B cell activation factor, BLyS), both of which belong to the TNF ligand family, interact with several receptors, especially BMCA (B cell maturation antigen), TAC1 and BAFF receptor, inducing B cell proliferation and survival. Both of these cytokines are produced by tumour-infiltrating neutrophils, and have been implicated in tumour promotion in malignancies such as diffuse large B cell lymphoma and multiple myeloma.[14]

Oncostatin M (OSM) and hepatocyte growth factor (HGF) are cytokines which are both stored and synthesised by tumour-infiltrating neutrophils. OSM appears to mediate tumour progression via induction of detachment of tumour cells and activation of synthesis of pro-angiogenic VEGF and fibroblast growth factor by endothelial cells, while HGF induces an invasive phenotype.[14]

These various cytokines and their reported roles in tumourigenesis are summarised in Table 3, with the exception of IL-8 which is discussed in detail below.

Table 3: Neutrophil-derived cytokines implicated in tumourigenesis

Cytokine	Pro-tumourigenic action
*Vascular endothelial growth factor (VEGF)	Tumour neovascularisation; pro-metastatic
*APRIL ('a proliferation-inducing ligand')	Tumour promotion; implicated in the aetiologies of diffuse large B cell lymphoma and multiple myeloma
*B cell activation factor (BAFF)	Tumour promotion; also implicated in the aetiology of B cell malignancies
*Oncostatin M (OSM)	Tumour progression
*Hepatocyte growth factor (HGF)	Tumour progression

**Recently reviewed by Tecchio et al.[14]*

Chemokines and tumourigenesis

Chemokines are a group of low molecular weight chemotactic cytokines which promote the receptor-mediated migration of cells of the innate and adaptive immune systems. In the case of neutrophils, these cells are attracted to sites of tissue injury or infection by members of the sub-family of CXC/ELR-motif-positive chemokines (CXC denotes the presence of an intervening amino acid, X, between the first two conserved cysteine residues, while the ELR motif is a glu-leu-arg sequence preceding the first conserved cysteine residue). The various members of this chemokine sub-family are shown in Table 4, with the potent neutrophil chemoattractant IL-8 (CXCL8) predominating.

Table 4: Examples of neutrophil-targeted CXC/ELR$^+$ chemokines

Systematic name	Alternative name
CXCL1	Growth-related oncogene (GRO)-α
CXCL2	Growth-related oncogene (GRO)-β
CXCL3	Growth-related oncogene (GRO)-γ
CXCL5	Epithelial neutrophil-activating peptide-78 (ENA-78)
CXCL6	Granulocyte chemotactic protein-2 (GCP-2)
CXCL7	Neutrophil-activating peptide-2 (NAP-2)
CXCL8	Interleukin-8 (IL-8)

Notwithstanding production by cells of the innate and adaptive immune systems, CXC/ELR$^+$ chemokines, especially IL-8, are also produced by various types of structural cells, including epithelial and endothelial cells, fibroblasts and smooth muscle cells. The major counter-receptor for these CXC/ELR$^+$ chemokines is CXCR2 (IL-8 also interacts with CXCR1), which is expressed not only on neutrophils and mast cells, but also on epithelial and endothelial cells.[12]

Importantly, and aside from their primary role in neutrophil mobilisation, CXC/ELR$^+$ chemokines and CXCR2 are also expressed by a diverse range of human cancers, including cancers of the breast, bladder, cervix, colon, liver, lymphatics, oesophagus, ovary, prostate and skin.[12,13,42] In this setting, these chemokines drive tumour expansion via both autocrine and paracrine pro-proliferative interactions with CXCR2-expressing tumour cells.[12,13,42,43] In the case of oesophageal squamous epithelial cells, and probably other tumour cell types, CXCR2-mediated proliferation results from activation of the transcription factor early growth response-1.[11] In addition, tumour neovascularisation is mediated via the pro-angiogenic activities of these chemokines, especially IL-8,[12] while the chronic influx of inflammatory cells exacerbates ROS-mediated oxidative damage to DNA and immunosuppression.

Neutrophil-derived proteases in tumour angiogenesis and metastasis

Neutrophil-derived proteases – specifically elastase and matrix metalloproteinase-9 (MMP-9) stored in primary and secondary/tertiary granules, respectively – have also been implicated in inflammation-associated tumour neovascularisation and invasion. Elastase has been reported to degrade the intercellular adhesion molecule cadherin,[44] while MMP-9 is a potent inducer of angiogenesis and tumour metastasis.[16,17]

Neutrophil adhesion molecules in tumour metastasis

Notwithstanding the expression of counter-receptors for endothelial adhesion molecules by some types of tumours,[2] pro-adhesive interactions between circulating neutrophils, albeit in an animal model of experimental liver metastasis, have also been reported to mediate delivery of tumour cells to distant sites.[45] In this setting, neutrophil/tumour cell adhesion is mediated via interactions of the β2-integrin Mac-1 on neutrophils, with its counter-receptor, intercellular adhesion molecule-1 (ICAM-1), on tumour cells.[45]

The aforementioned mechanisms of neutrophil/inflammation-mediated tumourigenesis are summarised in Figures 1 and 2.

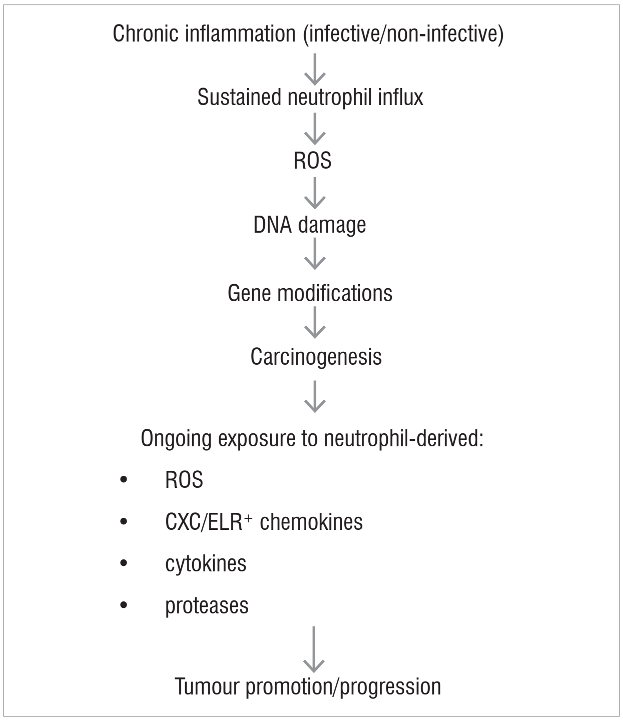

Chronic inflammation (infective/non-infective)
Sustained neutrophil influx
ROS
DNA damage
Gene modifications
Carcinogenesis
Ongoing exposure to neutrophil-derived:
• ROS
• CXC/ELR+ chemokines
• cytokines
• proteases
Tumour promotion/progression

Figure 1: Proposed mechanism by which chronic inflammation leads to oxidative damage to the DNA of bystander tissue cells as a result of the sustained release of reactive oxygen species (ROS) from infiltrating neutrophils. Tumour initiation is followed by promotion and progression as a result of ongoing exposure to neutrophil-derived pro-proliferative and pro-angiogenic mediators.

Inflammation-targeted chemotherapy and immunotherapy in cancer

The chemopreventive potential of aspirin in particular, and possibly other non-steroidal anti-inflammatory drugs (NSAIDs), in reducing the incidence of colorectal cancer, and possibly other cancers, such as those of the liver, lung, oesophagus and stomach, is well recognised.[2,19,46-48] Although the underlying mechanism is presumed to be anti-inflammatory in origin, other mechanisms, such as attenuation of prostaglandin E2-mediated inhibition of tumour-targeted T lymphocytes, have also been proposed.[2] In the case of therapy, the potential of NSAIDs as adjuncts to conventional anti-cancer therapies remains largely unknown, a possible exception being the use of aspirin in the treatment of colorectal cancer associated with *PIK3CA* gene mutations.[49,50]

Other potential pharmacological strategies include the use of inhibitors of MMP-9, although these have proved disappointing in phase II/III clinical trials in various types of malignancy,[51] and, perhaps the most promising strategy albeit unproven in the clinical setting, the use of pharmacological antagonists of CXCR2[43] and possibly dual antagonists of CXCR1/CXCR2. In addition to these, other categories of pharmacological agent which target the pro-inflammatory activities of neutrophils include 14/15-membered macrolide antibiotics and inhibitors of type 4 phosphodiesterase (PDE), the predominant PDE in human neutrophils. Unlike corticosteroids, which have limited efficacy in controlling neutrophilic inflammation, macrolides and PDE4 inhibitors possess a range of neutrophil-targeted anti-inflammatory activities which have recently been described in detail elsewhere.[52] Although untested with respect to their adjunctive potential as anti-inflammatory agents in cancer chemotherapy, it is noteworthy that novel macrolides and PDE4 inhibitors are currently under investigation for their direct anti-tumour activities.[53,54]

Monoclonal antibody-based anti-inflammatory therapies include those which target VEGF in various types of metastatic cancer, a strategy which has enjoyed variable success.[55-57] Although monoclonal antibodies which target neutrophil-mobilising cytokines such as TNF, IL-8 and, more recently, IL-17A[58,59] have been proposed as adjunctive anti-inflammatory strategies in the therapy of cancer, inhibitors of CXCR2 appear to be a superior option.

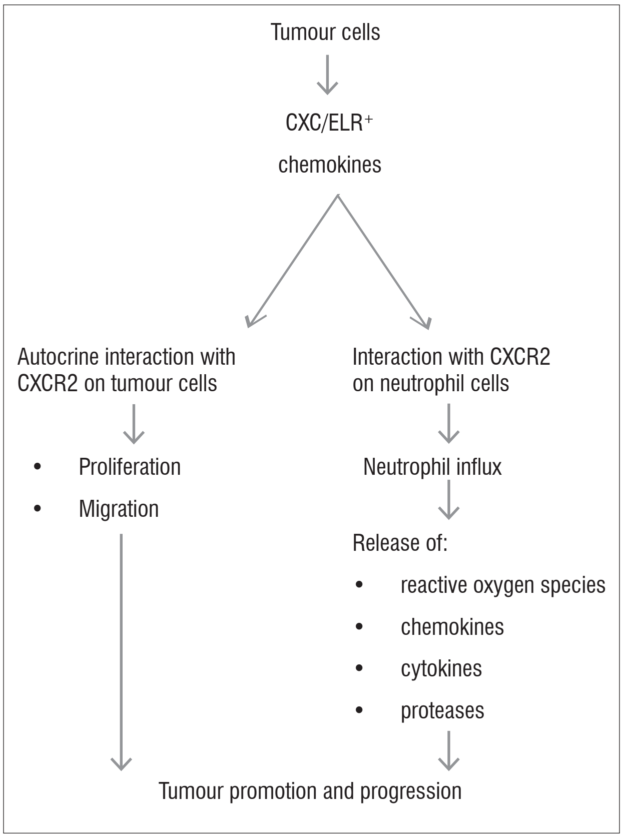


Figure 2: Proposed mechanism by which tumour-derived CXC/ELR$^+$ chemokines exacerbate promotion and progression via the autocrine induction of proliferation and metastasis, as well as by recruitment of pro-tumourigenic neutrophils.

Conclusions

Although they are key players in innate host defence, human neutrophils are also inadvertent participants in the aetiology of inflammation-related cancers via the release of carcinogenic ROS and other mediators which contribute to tumour promotion and progression. Other types of cancer, which are not inflammatory in origin, also utilise inflammatory mechanisms to enhance their proliferative and invasive potential. The most significant of these mechanisms is the production of CXC/ELR$^+$ chemokines. These chemokines not only recruit pro-tumourigenic neutrophils, but are also pro-proliferative and pro-metastatic via their autocrine interactions with CXCR2 expressed on tumour cells. These important insights into inflammation-associated mechanisms of tumourigenesis have enabled identification of potential anti-inflammatory adjunctive strategies to complement conventional anti-cancer therapies. However, given the range of neutrophil- and tumour-derived inflammatory mediators which contribute to tumourigenesis, selective targeting of a single mediator is unlikely to be successful. Although unproven in the clinical setting, selective antagonists of CXCR2, which target both neutrophils and tumour cells, represent a possible exception, as do NSAIDs, particularly aspirin. Preventive strategies include routine intake of low-dose NSAIDs, immunisation against cancer-causing viral pathogens, early aggressive antimicrobial chemotherapy to eradicate chronic inflammation caused by microbial pathogens, and avoidance of pro-inflammatory aspects of lifestyle such as cigarette smoking and excessive exposure to ultraviolet radiation.

Authors' contributions

All authors were involved in the overall conception, planning, design and writing of the manuscript.

References

1. Balkwill F, Mantovani A. Inflammation and cancer: Back to Virchow. Lancet.2001;357(9255):539–545. http://dx.doi.org/10.1016/S0140-6736(00)04046-0

2. Coussens LM, Werb Z. Inflammation and cancer. Nature. 2002;420(6917):860–867. http://dx.doi.org/10.1038/nature01322

3. Amulic B, Cazalet C, Hayes GL, Metzier KD, Zychlinsky A. Neutrophil function: From mechanisms to disease. Annu Rev Immunol. 2012;30:459–489. http://dx.doi.org/10.1146/annurev-immunol-020711-074942

4. Chaturvedi AK, Gaydos CA, Agreda P, Holden JP, Chatterjee N, Goedert JJ, et al. Chlamydia pneumoniae infection and risk for lung cancer. Cancer Epidemiol Biomarkers Prev. 2010;19(6):1498–1505. http://dx.doi.org/10.1158/1055-9965.EPI-09-1261

5. Shebl FM, Engels EA, Goedert JJ, Chaturvedi AK. Pulmonary infections and risk of lung cancer among persons with AIDS. J Acquir Immune Defic Syndr. 2010;55(3):375–379. http://dx.doi.org/10.1097/QAI.0b013e3181eef4f7

6. Shiels MS, Albanes D, Virtamo J, Engels EA. Increased risk of lung cancer in men with tuberculosis in the alpha-tocopherol, beta-carotene cancer prevention study. Cancer Epidemiol Biomarkers Prev. 2011;20(4):672–678. http://dx.doi.org/10.1158/1055-9965.EPI-10-1166

7. Weitzman SA, Gordon LI. Inflammation and cancer: Role of phagocyte-generated oxidants in carcinogenesis. Blood. 1990;76(4):655–663.

8. Perwez Hussain S, Harris CC. Inflammation and cancer: An ancient link with novel potentials. Int J Cancer. 2007;121(11):2373–2380. http://dx.doi.org/10.1002/ijc.23173

9. Rakoff-Nahoum S. Why cancer and inflammation. Yale J Biol Med. 2006;79(3–4):123–130.

10. Knaapen AM, Güngör N, Schins RPF, Borm PJA, Van Schooten FJ. Neutrophils and respiratory tract DNA damage and mutagenesis: A review. Mutagenesis. 2006;21(4):225–236. http://dx.doi.org/10.1093/mutage/gel032

11. Wang B, Khachigian LM, Esau L, Birrer MJ, Zhao X, Parker MI, et al. A key role for early growth response-1 and nuclear factor-κB in mediating and maintaining GRO/CXCR2 proliferative signaling in esophageal cancer. Mol Cancer Res. 2009;7(5):755–764. http://dx.doi.org/10.1158/1541-7786.MCR-08-0472

12. Verbeke H, Struyf S, Laureys G, Van Damme J. The expression and role of CXC chemokines in colorectal cancer. Cytokine Growth Factor Rev. 2011;22(5–6):345–358. http://dx.doi.org/10.1016/j.cytogfr.2011.09.002

13. Verbeke H, Karel G, Van Damme J, Struyf S. The role of CXC chemokines in the transition of chronic inflammation to esophageal and gastric cancer. Biochim Biophys Acta. 2012;1825(1):117–129.

14. Tecchio C, Scapini P, Pizzolo G, Cassatella MA. On the cytokines produced by human neutrophils in tumors. Sem Cancer Biol. 2013;23(3):159–170. http://dx.doi.org/10.1016/j.semcancer.2013.02.004

15. Shimizu M, Tanaka T, Moriwaki H. Obesity and hepatocellular carcinoma: Targeting obesity-related inflammation for chemoprevention of liver carcinogenesis. Semin Immunopathol. 2013;35(2):191–202. http://dx.doi.org/10.1007/s00281-012-0336-6

16. Ardi VC, Kupriyanova TA, Deryugina EI, Quigley JP. Human neutrophils uniquely release TIMP-free MMP-9 to provide a potent catalytic stimulator of angiogenesis. Proc Natl Acad Sci USA. 2007;104(51):20262–20267. http://dx.doi.org/10.1073/pnas.0706438104

17. Bekes EM, Schweighofer B, Kupriyanova TA, Zajac E, Ardi VC, Quigley JP, et al. Tumor-recruited neutrophils and neutrophil TIMP-free MMP-9 regulate coordinately the levels of tumor angiogenesis and efficiency of malignant cell intravasation. Am J Pathol. 2011;179(3):1455–1470. http://dx.doi.org/10.1016/j.ajpath.2011.05.031

18. Dumitru CA, Lang S, Brandau S. Modulation of neutrophil granulocytes in the tumor microenvironment: Mechanisms and consequences for tumor progression. Sem Cancer Biol. 2013;23(3):141–148. http://dx.doi.org/10.1016/j.semcancer.2013.02.005

19. Gregory AD, Houghton AM. Tumor-associated neutrophils: New targets for cancer therapy. Cancer Res. 2011;71(7):2411–2416. http://dx.doi.org/10.1158/0008-5472.CAN-10-2583

20. Brandau S, Dumitru CA, Lang S. Protumor and antitumor functions of neutrophil granulocytes. Semin Immunopathol. 2013;35(2):163–176. http://dx.doi.org/10.1007/s00281-012-0344-6

21. Brandau S. The dichotomy of neutrophil granulocytes in cancer. Sem Cancer Biol. 2013;23(3):139–140. http://dx.doi.org/10.1016/j.semcancer.2013.02.008

22. Toh B, Abastado J-P. Myeloid cells: Prime drivers of tumor progression. OncoImmunol. 2012;1(8):1360–1367. http://dx.doi.org/10.4161/onci.22196

23. Evans A, Costello E. The role of inflammatory cells in fostering pancreatic cancer cell growth and invasion. Frontiers Physiol. 2012;3(270):1–7.

24. Carus A, Ladekari M, Hager H, Pilegaard H, Nielsen PS, Donskov F. Tumor-associated neutrophils and macrophages in non-small cell lung cancer: No immediate impact on patient outcome. Lung Cancer. 2013;81(1):130–137. http://dx.doi.org/10.1016/j.lungcan.2013.03.003

25. Van Rensburg CE, Van Staden AM, Anderson R. Inactivation of poly (ADP-ribose) polymerase by hypochlorous acid. Free Radic Biol Med. 1991;11(3):285–291. http://dx.doi.org/10.1016/0891-5849(91)90125-M

26. Jaiswal M, La Russo NF, Nishioka N, Nakabeppu Y, Gores GJ. Human Ogg1, a protein involved in the repair of 8-oxoguanine, is inhibited by nitric oxide. Cancer Res. 2001;61(17):6388–6393.

27. Cai YJ, Lu JJ, Zhu H, Xie H, Huang M, Lin LP, et al. Salvicine triggers DNA double-strand breaks and apoptosis by GSH-depletion-driven H_2O_2 generation and topoisomerase II inhibition. Free Radic Biol Med. 2008;45(5):627–635. http://dx.doi.org/10.1016/j.freeradbiomed.2008.05.017

28. Vlasova II, Feng W-H, Goff JP, Giorgianni A, Do D, Gollin SM, et al. Myeloperoxidase-dependent oxidation of etoposide in human myeloid progenitor CD34$^+$ cells. Mol Pharmacol. 2011;79(3):479–487. http://dx.doi.org/10.1124/mol.110.068718

29. Forman HJ, Maiorino M, Ursini F. Signaling functions of reactive oxygen species. Biochemistry. 2010;49(5):835–842. http://dx.doi.org/10.1021/bi9020378

30. Gough DR, Cotter TG. Hydrogen peroxide: A Jekyll and Hyde signalling molecule. Cell Death Dis. 2011;2:e213. http://dx.doi.org/10.1038/cddis.2011.96

31. Veal E, Day A. Hydrogen peroxide as a signaling molecule. Antioxid Redox Signal. 2011;15(1):147–151. http://dx.doi.org/10.1089/ars.2011.3968

32. Runchel C, Matsuzawa A, Ichijo H. Mitogen-activated protein kinases in mammalian oxidative stress responses. Antioxid Redox Signal. 2011;15(1):205–218. http://dx.doi.org/10.1089/ars.2010.3733

33. Zhou X, Li D, Resnick MB, Behar J, Wands J, Cao W. Signaling in H_2O_2–induced increase in cell proliferation in Barrett's esophageal adenocarcinoma cells. J Pharmacol Exp Ther. 2011;339(1):218–227. http://dx.doi.org/10.1124/jpet.111.182352

34. Boonanantanasarn K, Gill AL, Yap Y, Jayaprakash V, Sullivan MA, Gill SR. *Enterococcus faecalis* enhances cell proliferation through hydrogen peroxide mediated epidermal growth factor receptor activation. Infect Immun. 2012;80(10):3545–3558. http://dx.doi.org/10.1128/IAI.00479-12

35. Kusmartsev S, Nefedova Y, Yoder D, Gabrilovich DI. Antigen-specific inhibition of CD8$^+$ T cell response by immature myeloid cells in cancer is mediated by reactive oxygen species. J Immunol. 2004;172(2):989–999.

36. Pillay J, Kamp VM, Van Hoffen E, Visser T, Tak T, Lammers JW, et al. A subset of neutrophils in human systemic inflammation inhibits T cell responses through Mac-1. J Clin Invest. 2012;122(1):327–336. http://dx.doi.org/10.1172/JCI57990

37. Pillay J, Den Braber I, Vrisekoop N, Kwast LM, De Boer RJ, Borghans JA, et al. In vivo labeling with ^2H_2O reveals a human neutrophil lifespan of 5.4 days. Blood. 2010;116(4):625–627. http://dx.doi.org/10.1182/blood-2010-01-259028

38. Milot E, Filep JG. Regulation of neutrophil survival/apoptosis by Mcl-1. Scientific World J. 2011;11:1948–1962. http://dx.doi.org/10.1100/2011/131539

39. Cassatella MA. Neutrophil-derived proteins: Selling cytokines by the pound. Adv Immunol. 1999;73:369–509. http://dx.doi.org/10.1016/S0065-2776(08)60791-9

40. Gaudry M, Brégerie O, Andrieu V, Benna JE, Pocidalo M-A, Hakim J. Intracellular pool of vascular endothelial growth factor in human neutrophils. Blood. 1997;90(10):4153–4161.

41. Lacy P, Stow JL. Cytokine release from innate immune cells: Association with diverse membrane trafficking pathways. Blood. 2011;118(1):9–18. http://dx.doi.org/10.1182/blood-2010-08-265892

42. Lazennec G, Richmond A. Chemokines and chemokine receptors: New insights into cancer-related inflammation. Trends Mol Med. 2010;16(3):133–144. http://dx.doi.org/10.1016/j.molmed.2010.01.003

43. Jamieson T, Clarke M, Steele CW, Samuel MS, Neumann J, Jung A, et al. Inhibition of CXCR2 profoundly suppresses inflammation-driven and spontaneous tumorigenesis. J Clin Invest. 2012;122(9):3127–3144. http://dx.doi.org/10.1172/JCI61067

44. Grosse-Steffen T, Giese T, Giese N, Longerich T, Schirmacher P, Hänsch GM, et al. Epithelial-to-mesenchymal transition in pancreatic tumor cell lines: The role of neutrophils and neutrophil-derived elastase. Clin Dev Immunol. 2012;2012:720768. http://dx.doi.org/10.1155/2012/720768

45. Spicer JD, McDonald B, Cools-Lartigue JJ, Chow SC, Giannias B, Kubes P, et al. Neutrophils promote liver metastasis via Mac-1-mediated interactions with circulating tumor cells. Cancer Res. 2012;72(16):3919–3927. http://dx.doi.org/10.1158/0008-5472.CAN-11-2393

46. Vandramini-Costa DB, Carvalho JE. Molecular link mechanisms between inflammation and cancer. Curr Pharm Des. 2012;18(26):3831–3852. http://dx.doi.org/10.2174/138161212802083707

47. Levy IG, Pim CP. An aspirin a day: The allure (and distraction) of chemoprevention. J Natl Cancer Inst. 2012;104(23):1782–1784. http://dx.doi.org/10.1093/jnci/djs462

48. Sahasrabuddhe VV, Gunja MZ, Graubard BI, Trabert B, Schwartz LM, Park Y, et al. Non-steroidal anti-inflammatory drug use, chronic liver disease, and hepatocellular carcinoma. J Natl Cancer Inst. 2012;104(23):1808–1814. http://dx.doi.org/10.1093/jnci/djs452

49. Liao X, Lochhead P, Nishihara R, Morikawa T, Kuchiba A, Yamauchi M, et al. Aspirin use, tumor PIK3CA mutation, and colorectal cancer survival. N Engl J Med. 2012;367(17):1596–1606. http://dx.doi.org/10.1056/NEJMoa1207756

50. Pasche B. Aspirin – from prevention to targeted therapy. N Engl J Med. 2012;367(17):1650–1651. http://dx.doi.org/10.1056/NEJMe1210322

51. Zuker S, Cao J. Selective matrix metalloproteinase (MMP) inhibitors in cancer therapy: Ready for prime time? Cancer Biol Ther. 2009;8(24):2371–2373. http://dx.doi.org/10.4161/cbt.8.24.10353

52. Tintinger GR, Anderson R, Feldman C. Pharmacological approaches to regulate neutrophil activity. Semin Immunopathol. 2013;35(4):395–409. http://dx.doi.org/10.1007/s00281-013-0366-8

53. Napolitano JG, Daranas AH, Norte M, Fernández JJ. Marine macrolides, a promising source of antitumor compounds. Anti-Cancer Agents Med Chem. 2009;9(2):122–137. http://dx.doi.org/10.2174/187152009787313800

54. Sengupta R, Sun R, Warrington NM, Rubin JB. Treating brain tumors with PDE4 inhibitors. Trends Pharmacol Sci. 2011;32(6):337–344. http://dx.doi.org/10.1016/j.tips.2011.02.015

55. Wagner AD, Thomssen C, Haerting J, Unverzagt S. Vascular-endothelial-growth-factor (VEGF) targeting therapies for endocrine refractory or resistant metastatic breast cancer. Cochrane Database Syst Rev. 2012;7:CD008941.

56. Gianni L, Romieu GH, Lichinitser M, Serrano SV, Mansutti M, Pivot X, et al. AVEREL: A randomized phase III trial evaluating bevacizumab in combination with docetaxel and trastuzumab as first-line therapy for HER2-positive locally recurrent/metastatic breast cancer. J Clin Oncol. 2013;31(14):1719–1725. http://dx.doi.org/10.1200/JCO.2012.44.7912

57. Makhoul I, Klimberg VS, Korourian S, Henry-Tillman RS, Siegel ER, Westbrook KC, et al. Combined neoadjuvant chemotherapy with bevacizumab improves pathologic complete response in patients with hormone receptor negative operable or locally advanced breast cancer. Am J Clin Oncol. In press 2013. http://dx.doi.org/10.1097/COC.0b013e31828940c3

58. Charles KA, Kulbe H, Soper R, Escorcio-Correia M, Lawrence T, Schultheis A, et al. The tumor-promoting actions of TNF-alpha involve TNFR1 and IL-17 in ovarian cancer in mice and humans. J Clin Invest. 2009;119(10):3011–3023. http://dx.doi.org/10.1172/JCI39065

59. Oshiro K, Kohama H, Umemura M, Uyttenhove C, Inagaki-Ohara K, Arakawa T, et al. Interleukin-17A is involved in enhancement of tumor progression in murine intestine. Immunobiology. 2012;217(1):54–60. http://dx.doi.org/10.1016/j.imbio.2011.08.002

A centrifugal microfluidic platform for point-of-care diagnostic applications

AUTHORS:
Suzanne Hugo[1]
Kevin Land[1]
Marc Madou[2]
Horacio Kido[2]

AFFILIATIONS:
[1]Materials Science and Manufacturing, Council for Scientific and Industrial Research, Pretoria, South Africa
[2]Department of Mechanical and Aerospace Engineering, University of California, Irvine, CA 92697, USA

CORRESPONDENCE TO:
Suzanne Hugo

EMAIL:
shugo@csir.co.za

POSTAL ADDRESS:
Materials Science and Manufacturing, Council for Scientific and Industrial Research, PO Box 395, Pretoria 0001, South Africa

DATES:
Received: 02 Apr. 2013
Revised: 08 Aug. 2013
Accepted: 02 Sep. 2013

KEYWORDS:
microfluidics; centrifugal microfluidics; lab-on-a-disc; point-of-care; diagnostics

HOW TO CITE:
Hugo S, Land K, Madou M, Kido H. A centrifugal microfluidic platform for point-of-care diagnostic applications. S Afr J Sci. 2014;110(1/2), Art. #2013-0091, 7 pages. http://dx.doi.org/10.1590/sajs.2014/20130091

Microfluidic systems enable precise control over tiny volumes of fluid in a compact and low-cost form, thus providing the ideal platform on which to develop point-of-care diagnostic solutions. Centrifugal microfluidic systems, also referred to as lab-on-a-disc or lab-on-a-CD systems, provide a particularly attractive solution for the implementation of microfluidic point-of-care diagnostic solutions as a result of their simple and compact instrumentation, as well as their functional diversity. Here we detail the implementation of a centrifugal microfluidic platform – the first of its kind in South Africa – as a foundation for the development of point-of-care diagnostic applications for which both the need and impact is great. The centrifugal microfluidic platform consists of three main components: a microfluidic disc device similar in size and shape to a CD, a system for controlling fluid flow on the device, and a system for recording the results obtained. These components have been successfully implemented and tested. Preliminary test results show that microfluidic functions such as pumping and valving of fluids can be successfully achieved, as well as the generation of monodisperse microfluidic droplets, providing a complete centrifugal microfluidic platform and the building blocks on which to develop a variety of applications, including point-of-care diagnostics. The lab-on-a-disc platform has the potential to provide new diagnostic solutions at the point-of-need in health- and industry-related areas. This paves the way for providing resource limited areas with services such as improved, decentralised health-care access or water-quality monitoring, and reduced diagnosis times at a low cost.

Introduction

The technology of microfluidics entails the precise and automated control of very small volumes of fluids, usually on a nanolitre scale. A number of comprehensive reviews detail the advances that have been made in microfluidic technologies over the last 30 years.[1,2] Microfluidic systems are often referred to as lab-on-a-chip systems or micro-Total-Analysis-Systems (microTAS), and are well suited to the development of point-of-care diagnostics[3-5] as these systems utilise a small sample to provide a compact and low-cost solution.

Centrifugal microfluidic systems, also referred to as lab-on-a-disc or lab-on-a-CD solutions, provide a particularly attractive solution for the implementation of microfluidic point-of-care diagnostic systems, specifically for biomedical applications.[6] Centrifugal microfluidic technology makes use of a disc, similar in size and shape to a CD or DVD, to house microfluidic channels and features. A motor is used to rotate the microfluidic disc, transporting fluid radially outwards through the microfluidic device, and manipulating fluid by means of various microfluidic functions and features on the disc. Functions such as valving, mixing, pumping and separation of fluids can be readily achieved in centrifugal microfluidic systems by exploiting the forces responsible for fluidic control. Fluidic control in lab-on-a-disc microfluidics depends on centrifugal forces, Coriolis forces and capillary action.

Centrifugal force is the driving force for fluid transport in lab-on-a-disc systems and causes fluid to flow radially outward from the centre of the disc to the outer circumference. The centrifugal force can be mathematically described by

$$f = -\rho\omega \times (\omega \times r) \qquad \text{Equation 1}$$

where ρ is the mass density of the fluid, ω is the rotational frequency and r is the radial position along the disc. The centrifugal force is particularly useful for facilitating fluid flow through a system, as well as for sedimentation or separation of a sample (centrifugation) and for compression of fluid.

The Coriolis force is the force moving perpendicular to the centrifugal force and is described by

$$f = -2\rho\omega \times v \qquad \text{Equation 2}$$

where ρ is the mass density of the fluid, ω is the rotational frequency and v is the vector of flow velocity. The Coriolis force can be used for mixing of samples in a lab-on-a-disc microfluidic device, as well as for flow switching, or moving of a sample into a specific channel path (e.g. moving a sample into the left channel where a fork occurs in the main channel).

Capillary action is often used to balance the centrifugal force of a lab-on-a-disc system in order to create valves. Once the burst frequency – the rotational speed at which the centrifugal and capillary forces balance each other – is reached and the centrifugal force exceeds the capillary force, the valve will burst, releasing the fluid. The capillary force is described by

$$f = 2\sigma/r \cos \theta \qquad \text{Equation 3}$$

where σ is the surface tension, Θ is the contact angle and *r* is the radius of the capillary. Capillary action is useful for valving of fluids as well as for volume metering, which is important in many biological assays. The reader is referred to a number of excellent reviews[6-8] for more detailed explanations of the theory surrounding these forces and the application thereof to microfluidic functions.

Centrifugal microfluidic systems are well suited to integrated point-of-care diagnostic systems – and have a number of advantages over existing microfluidic and other point-of-care diagnostic methods.[7,9] The lab-on-a-disc platform eliminates the need for active elements such as pumps, actuators and active valves. These components present a complex and costly challenge in many microfluidic systems.[7,9] In centrifugal microfluidic systems, pumps, valves and other fluidic functions are achieved primarily using centrifugal forces, with only a small motor required to power the system. A high degree of parallelisation is also offered by centrifugal microfluidics, as numerous devices can be implemented on one disc as a result of radial symmetry. Examples of centrifugal microfluidic applications for biomedical diagnostics have been presented and include blood plasma separation[10] and a variety of biological assay implementations.[11-13]

The simple, low-cost and multiplex nature of the lab-on-a-disc platform is further strengthened by the low-cost and rapid fabrication techniques that can be utilised to realise the disc devices. Simple layered designs manufactured from plastics and adhesives can be used to fabricate microfluidic discs quickly and effectively. Centrifugal microfluidic systems enable a variety of components from sample preparation through to detection to be implemented efficiently into an integrated microfluidic solution for point-of-care diagnostic applications.[14]

In addition to the low-cost implementation of the lab-on-a-disc platform, centrifugal microfluidics have the added benefit of an accelerated route to market, as they can be viewed as microfluidic applications that are compatible with various existing and commercially available technologies.[15] Existing equipment such as CD players, DVD drives and laboratory centrifuges can be used to drive the microfluidic discs and analyse the results, eliminating the need for extensive development on the reader/actuator component of the point-of-care device. The compatibility of lab-on-a-disc devices with commercially available readers is of particular benefit for developing countries, as this compatibility enables a readily accessible solution where it is needed most.

The work presented here details the implementation of a centrifugal microfluidic platform – the first of its kind to be established in South Africa – with the aim of enabling the development of point-of-care diagnostic applications. It is believed that this platform will have the potential to assist in addressing health-related issues in developing countries.

Initial applications of the centrifugal microfluidic platform are designed to showcase the platform and test the various components which are required for this implementation. This includes sample preparation steps for performing blood analyses, as well as investigation of diverse microfluidic applications such as droplet generation.

The final goal of this platform development is to enable research and development activities towards the realisation of microfluidic-based point-of-care diagnostic solutions for South Africa.

Centrifugal microfluidic platform

The lab-on-a-disc platform consists of three main components: a microfluidic disc device, a system for controlling fluid flow on the device and a system to record the results obtained. These components have been successfully implemented into an integrated system including programmable spin cycles and both macro imaging and microscopy. The integrated components provide a complete centrifugal microfluidic platform on which to develop new and novel applications in fields such as point-of-care health diagnostics, environmental diagnostics and chemical and biological production.

Microfluidic disc design, manufacture and assembly

Centrifugal microfluidic disc devices can be designed using a computer aided design (CAD) program such as Solidworks or DesignCAD and manufactured in-house. The microfluidic discs were made from polycarbonate sheeting (Naxel, supplied by Maizey Plastics, Pretoria, South Africa) and pressure-sensitive adhesive (Flexmount DFM 200 clear V-95 150 Poly H-9 V-95 400 H-9, Flexcon, supplied by Synchron, Johannesburg, South Africa), assembled in layers. The various features of the microfluidic disc, including channels and chambers, were machined using different materials and methods. The polycarbonate layers were machined using a milling machine (ProtoMat S63, LPKF Laser and Electronics, supplied by Cadshop (Pty) Ltd, Johannesburg, South Africa), while the pressure-sensitive adhesive layers were cut out using a vinyl cutter plotter (Roland GX-41 CAMM 24” x 12”, supplied by Telpro Management, Johannesburg, South Africa). Individual pieces were then assembled and pressed together using a cold roll laminator (ML25, Drytac, supplied by MyBinding, Hillsboro, OR, USA) to produce the finished microfluidic disc device.

Figure 1 shows the microfluidic disc manufacture process and the relevant equipment and materials required.

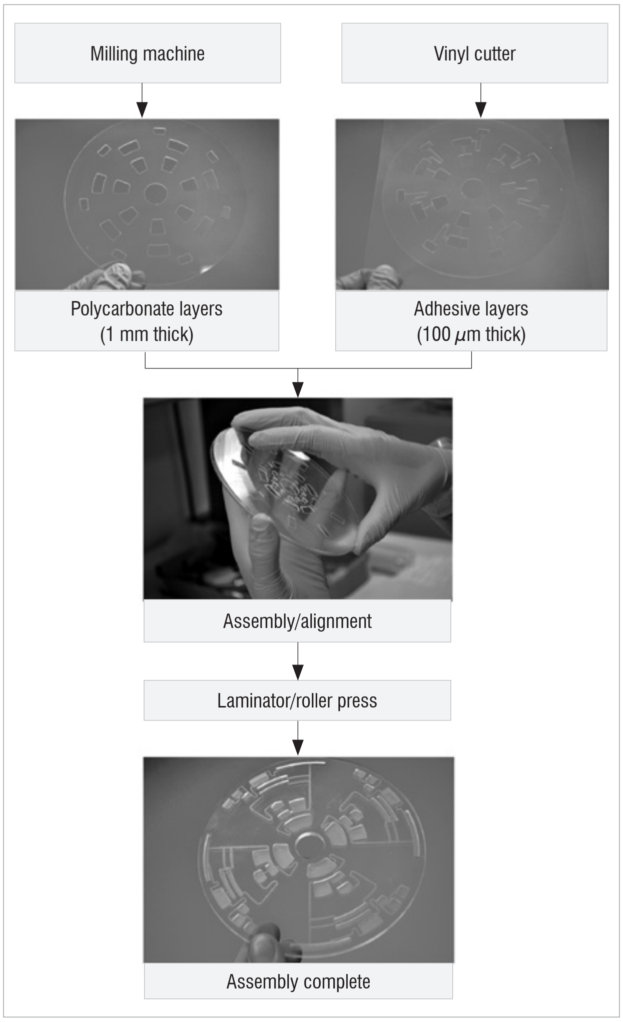


Figure 1: Illustration of microfluidic disc manufacture and assembly process.

Fluid control and analysis of disc

After assembly of the device, the disc was tested using a system that consists of a motor to rotate the disc, as well as an image capturing unit that allows for a picture of an area of interest to be captured for each revolution of the disc. Different rotational speeds and timing cycles were used to implement various fluidic functions (including valving, mixing, sedimentation, separation and compression) by exploiting centrifugal forces.

Figure 2 shows the disc testing set-up that was assembled to enable fluid control on the microfluidic disc and imaging of the device as it rotates to enable results of the fluidic functions on the disc to be recorded. A schematic representation of the test set-up is given in Figure 2a, with a photograph of the laboratory set-up given in Figure 2b.

A motor and controller (Smartmotor SM3450D, supplied by Animatics, Memmingen, Germany) were used to control the rotation of the microfluidic disc. An imaging set-up, consisting of an optical sensor (D10DPFP, Banner, supplied by RET Automation Controls (Pty) Ltd, Johannesburg, South Africa), fibre optic cable (0.5 mm Fiber Plastic, Banner, supplied by RET Automation Controls), a CMOS camera (DFK 22BUC03, supplied by The Imaging Source, Bremen, Germany) and lens (C1614-M(KP), The Imaging Source), as well as a strobe light (DT-311A Stroboscope, Shimpo, supplied by Resonance Instruments (Pty) Ltd, Johannesburg, South Africa), was constructed. The optical sensor and fibre optic cable served as a trigger to the camera and the strobe light to allow for a clear still image to be captured each time the disc completed a revolution. A small piece of reflective tape was attached to the microfluidic disc to be tested to allow the transmitted light from the optical sensor to be reflected into the receiver of the optical sensor, in turn triggering the camera to capture an image, and triggering the strobe light to illuminate the region of interest on the microfluidic disc, ensuring that a clear still image was captured.

The user controls the rotation of the microfluidic disc or spin cycle via a user interface on a PC. The user can program the speed, acceleration, deceleration and timing cycles of the disc to automate fluidic functions on the microfluidic disc.

Platform and scale-up costs

The lab-on-a-disc platform allows for rapid, low-cost prototyping and testing of microfluidic disc devices, as the equipment and materials required are low cost and/or available in-house. Equipment such as the milling machine and vinyl cutter are common items in many mechanical laboratories. Excluding the equipment which was already available in-house, the costs to produce a complete centrifugal microfluidic system amounted to R25 000. The cost of materials for the disc devices amounted to R500/m^2 and R10 per prototype disc device.

A comparison of system integration criteria for various microfluidic technologies[16] shows that centrifugal microfluidic systems rank highly as viable, low-cost solutions for integrated lab-on-a-disc systems.[16] Sin et al.'s figure 9 illustrates that centrifugal microfluidic systems are second only to paper or capillary-based microfluidic systems in terms of cost, and perform well in other important areas such as throughput and diversity, which would be favourable for point-of-care devices.

Although the lab-on-a-disc system is in the early stages of development, scale-up of the system is an ongoing consideration. Low-cost fabrication of plastic disc devices in high volumes is achievable through readily available manufacturing technologies, while the device for controlling the disc and performing tests is envisioned as a compact system housing a simple motor and measurement mechanism. Scale-up will continue to be considered and developed based on the desired end application of the system.

To ensure the successful development of the lab-on-a-disc system into a viable medical diagnostic product, medical device regulatory requirements will be an important consideration. Role players in the regulatory environment are currently being engaged to determine the requirements for the South African market.

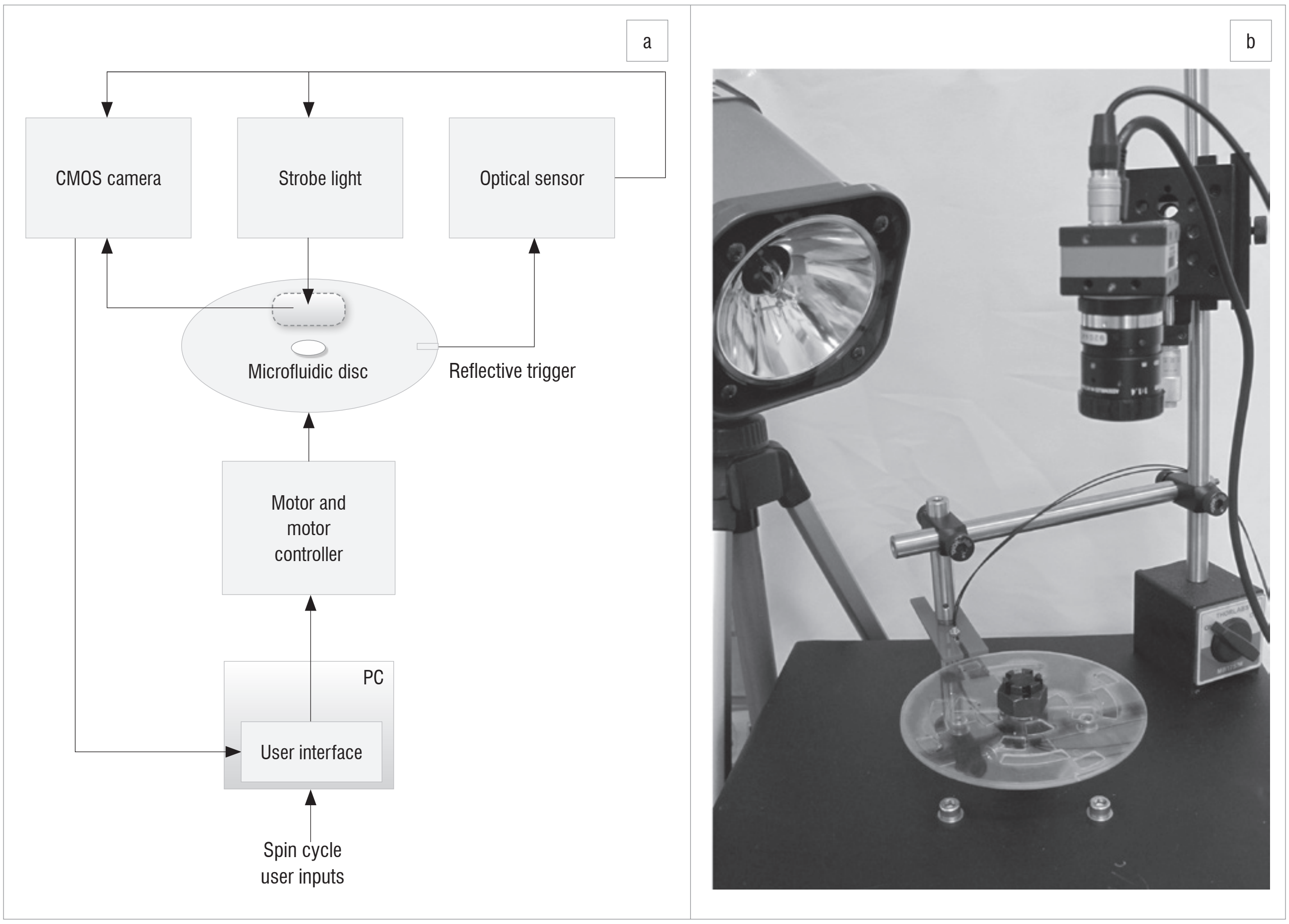


Figure 2: (a) Schematic of the components required for fluidic control and imaging of the disc device and (b) the integrated testing system set-up.

Results

Initial applications of the complete centrifugal microfluidic platform were implemented to illustrate the process carried out from design to analysis of a lab-on-a-disc system. The first example demonstrates basic fluidic functions on the disc such as introduction, valving and combining of fluids, and illustrates potential diagnostic applications for manipulation of biological samples such as blood. The second example demonstrates microfluidic droplet generation using the centrifugal microfluidic platform.

Basic fluidic functions

To demonstrate basic fluidic functions, a simple microfluidic disc design was formulated to allow for a sample and a sample reagent to be introduced separately, added together at different times, and combined, with an overflow chamber for excess solution. For the purposes of illustration, a yeast solution was used to simulate blood, while the reagent was a staining solution commonly used to stain blood cells for visualisation and performing manual blood cell counting. The use of a yeast solution as a proxy also allowed the sedimentation or separation of particles in fluids to be illustrated by the centrifugal microfluidic system.

Figure 3 shows the microfluidic features of the disc design used to achieve the desired fluidic functions. Four identical microfluidic systems were designed and manufactured on one disc. The microfluidic channels are 1 mm wide and 100 μm deep, while the chambers have a depth of 1.2 mm and vary in width and length. The vent holes have a diameter of 1 mm.

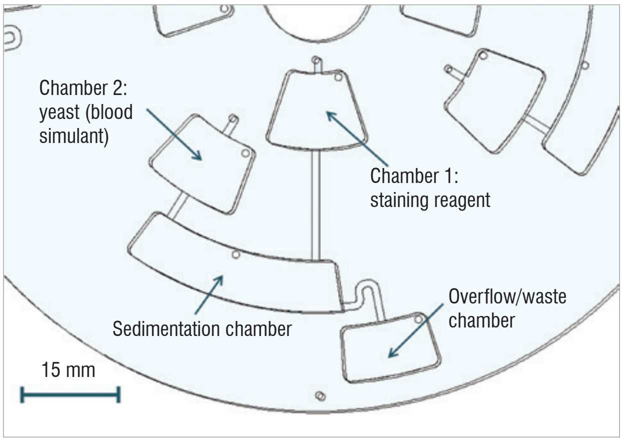

Chamber 2: yeast (blood simulant)
Chamber 1: staining reagent
Sedimentation chamber
Overflow/waste chamber
15 mm

Figure 3: Microfluidic disc design to illustrate the introduction, combination and sedimentation of samples and reagents, with applications for blood testing.

Figure 4 shows the complete composition of the microfluidic disc manufactured from three layers (first, third and fifth layers) of 1-mm thick polycarbonate and two layers (second and fourth layers) of 100-μm thick pressure-sensitive adhesive. Sample inlet holes and air vent holes are situated on the top polycarbonate layer, the chambers are situated in the middle polycarbonate layer and both the pressure-sensitive adhesive layers and the channels are situated in the bottom pressure-sensitive adhesive layer.

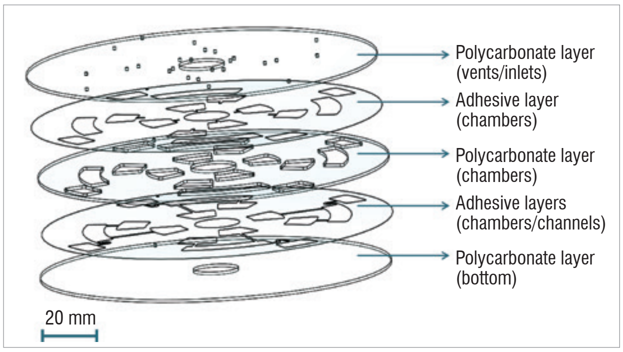
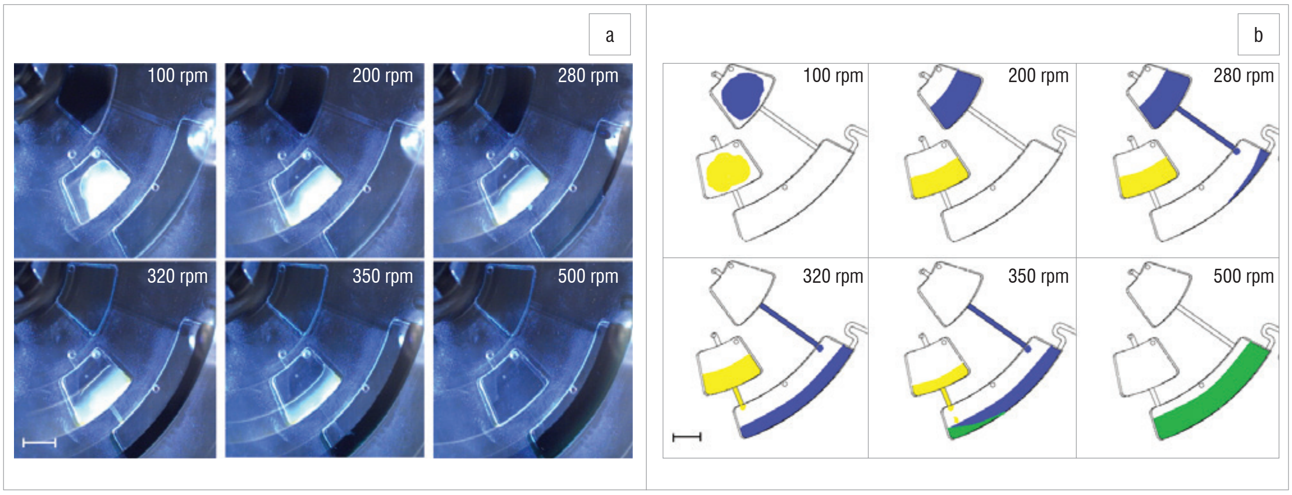

Polycarbonate layer (vents/inlets)
Adhesive layer (chambers)
Polycarbonate layer (chambers)
Adhesive layers (chambers/channels)
Polycarbonate layer (bottom)
20 mm

Figure 4: Illustration of the multilayered centrifugal microfluidic disc device.

The blood simulant solution was made from 10 mg dry baker's yeast in 100 mL deionised water to yield a similar concentration of cells to that of white blood cells found in a human blood sample. The staining reagent was a 2% acetic acid solution with 1 mg crystal violet in 100 mL deionised water – a standard white blood cell reagent commonly used to lyse red blood cells and stain the nuclei of white blood cells for manual white blood cell counting.

Approximately 70 μL of both the staining solution and the yeast solution were pipetted into chambers 1 and 2, respectively, via the inlet holes on top of the chamber openings (Figure 3). The microfluidic disc was then placed on the motor spindle of the centrifugal microfluidic platform set-up for testing of the fluid functions.

The motor was controlled through the SmartMotor Interface software issued with the motor hardware. The motor was set to operate at a constant velocity to enable continuous rotation of the disc on the motor spindle. For each change in the speed of the rotating disc, an acceleration of 350 rpm^2 was used. Figure 5a shows a sequence of images of the microfluidic disc rotating at various speeds to achieve different fluidic functions. Figure 5b shows the corresponding sketches of each of the images in Figure 5a to illustrate the fluidic operations for each step. In Figure 5a, the yeast solution is visible as a bright white colour, while the staining reagent is a deep blue/purple.

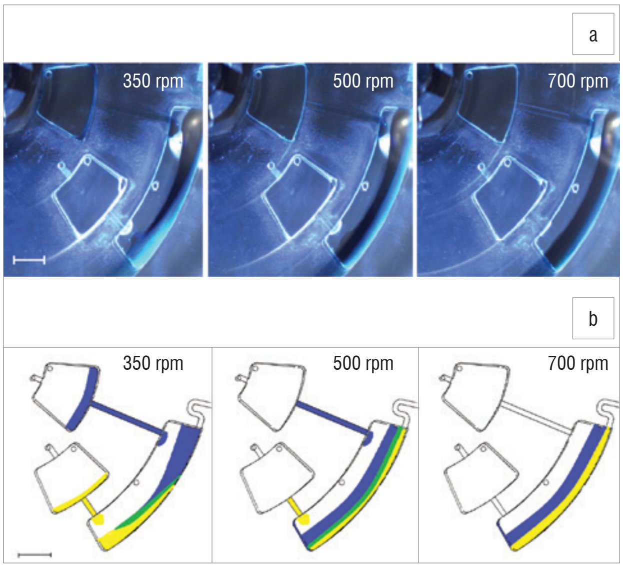


Scale bar = 3 mm

Figure 5: The microfluidic disc at various spin speeds to illustrate timed valving and combining of fluids: (a) images of the disc device captured using the experimental set-up and (b) corresponding sketches to illustrate the fluid interactions for each of the images in (a).

The motor was initially set to rotate at a speed of 100 rpm. Referring to Figure 5, it can be seen that, at this speed, no fluid movement occurs and both the yeast solution and the staining reagent stay in the inlet chambers into which they were introduced. At 200 rpm, the fluid in both the inlet chambers starts to compress and is pushed to the bottom of the chambers. At a slight increase in rotational speed up to 280 rpm, the staining solution from chamber 1 is released via a channel into the sedimentation chamber. The fluid is released as a result of the centrifugal force exceeding the capillary force – commonly referred to as the burst frequency. Increasing the speed further to 320 rpm causes the yeast solution from chamber 2 to prime the connecting channel to the sedimentation chamber. At a slightly higher speed of 350 rpm, the yeast solution from chamber 2 is released fully into the sedimentation chamber, combining with the staining reagent. At 500 rpm, the inlet chambers have been completely emptied and the fluid is combined in the sedimentation chamber.

Figure 6 illustrates the sedimentation of fluids in the microfluidic disc, again by making use of the yeast solution as it contains cells or particulate matter. Fluids were introduced into the same disc design in the same manner as in Figure 5. In this example, the yeast solution used was a higher concentration (approximately 10 g dry baker's yeast in 100 mL deionised water) for ease of visualisation of the sedimentation process. This concentration is also similar to the concentration of both red and white blood cells found in a sample of human blood. The staining reagent used was again a 2% acetic acid solution with 1 mg crystal violet in 100 mL deionised water.



Scale bar = 3 mm

Figure 6: The microfluidic disc at various spin speeds to illustrate sedimentation of fluids: (a) images of the disc device captured using the experimental set-up and (b) corresponding sketches to illustrate the fluid interactions for each of the images in (a).

A sequence of images from the rotating disc device are shown in Figure 6a, with corresponding sketches of the fluidic operations for each of these images illustrated in Figure 6b. At 350 rpm, both the yeast solution and the staining reagent are in the process of being released into the sedimentation chamber, as in Figure 5. However, Figure 6 clearly illustrates, as a result of the higher concentration of yeast, how the fluids combine in the sedimentation chamber. Although the yeast solution is released after the staining reagent, the yeast solution starts to move to the bottom of the sedimentation chamber as a result of the centrifugal forces. At an increased speed of 500 rpm, sedimentation of the yeast solution from the staining reagent is clearly visible, and at 700 rpm the inlet chambers have been completely emptied into the sedimentation chamber and compressed sedimentation of the yeast solution is visible. Again, the acceleration used for the adjustment of each rotational speed was 350 rpm².

Microfluidic droplet generation

After successfully demonstrating the implementation of basic fluidic functions from design through to analysis on the centrifugal microfluidic platform, microfluidic droplet generation using the centrifugal microfluidic platform was also investigated. Droplet generation in microfluidic devices is an extensive area of research with a vast number of applications, and has been summarised in a number of useful reviews.[17,18]

A large poly(methyl methacrylate) (PMMA) disc was designed and manufactured to house existing droplet generation devices (Figure 7). The droplet generation devices, which produce monodisperse droplets, are currently being used for the production of self-immobilised enzymes, which would find application in chemical, food, textile and other industries.

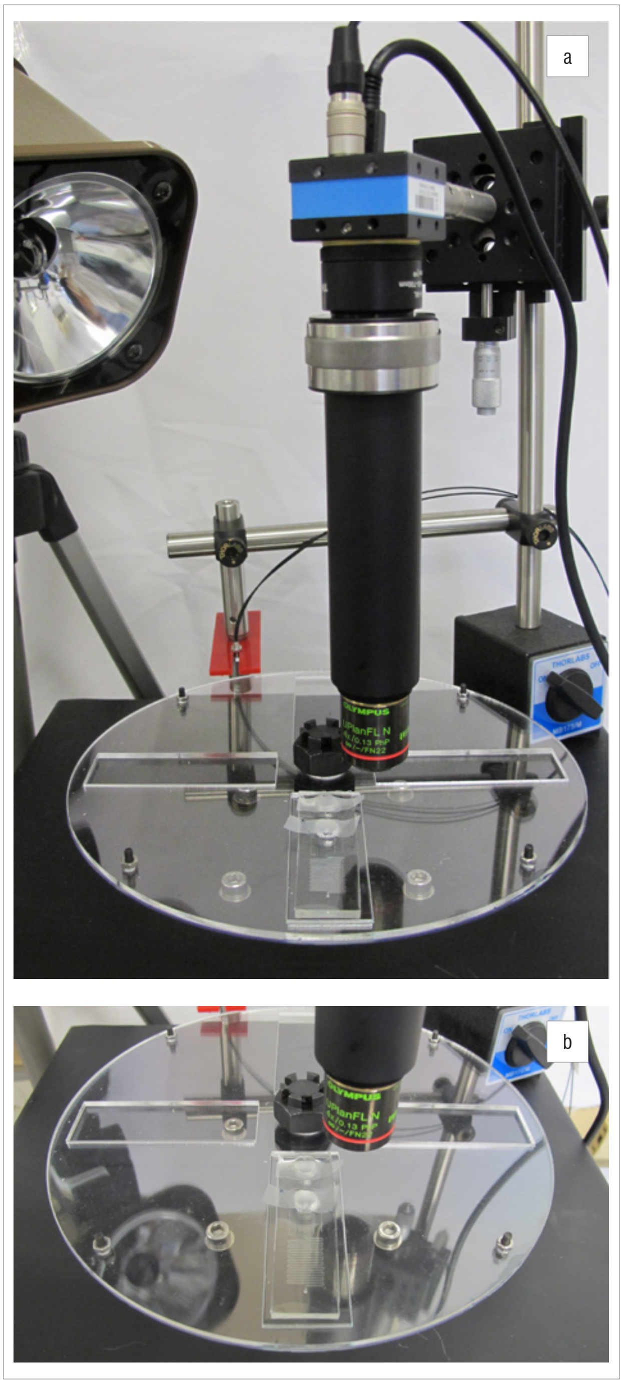


Figure 7: (a) Microscope set-up for capturing images of the droplet generation experiments, and (b) a close-up of the disc used to house the polydimethylsiloxane droplet-generation devices.

The existing droplet generation devices are made out of polydimethylsiloxane (PDMS) using soft lithography techniques to manufacture micro-channel features. The PDMS layer that houses the micro-channels is bonded to a glass slide to create a complete microfluidic device for testing. Typically these devices are tested using syringe pumps to introduce fluid to the devices. Desired flow rates can be programmed into the syringe pumps. For testing the PDMS droplet generation devices using the centrifugal microfluidic platform, the microfluidic devices were manufactured with relatively large reservoirs (8-mm diameters), allowing for a larger volume of fluid to be stored on the microfluidic disc and used during a droplet generation experiment.

A microscope set-up was implemented using various attachments connected to the CMOS camera (DFK 22BUC03, The Imaging Source) of the centrifugal microfluidic platform. The microscope set-up consisted of (in the order in which they were connected to the CMOS camera): a SM1 to C mount adaptor (SM1A9, supplied byThorlabs, Newton, NJ, USA), a tube lens (AC254-030-A-ML, Thorlabs), two lens tubes (SM1ZM, SM1L40, Thorlabs), an RMS adaptor (SM1A3, Thorlabs), and a microscope objective (UPLFLN 4X, Olympus, supplied by Wirsam Scientific and Precision Equipment (Pty) Ltd, Johannesburg, South Africa). This set-up enabled images of the droplet generation on the rotating microfluidic disc to be captured (Figure 7). The large PMMA disc allowed for PDMS devices to be mounted on the centrifugal microfluidic platform. The reservoirs on the PDMS devices were filled with mineral oil as the continuous phase and blue dye in deionised water as the droplet phase.

Figure 8 shows the design of the droplet generation device manufactured from PDMS. The channel widths of the PDMS device are 100 μm at the inlets and 400 μm at the outlet. The depth of the device is 70 μm.

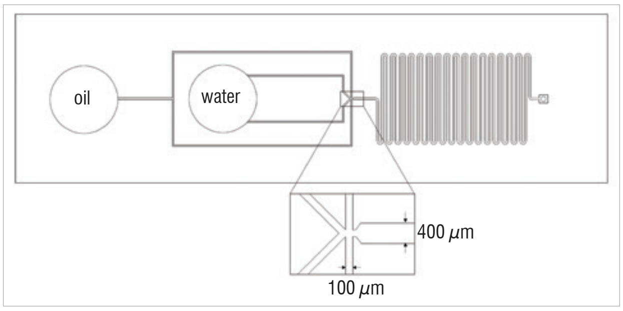
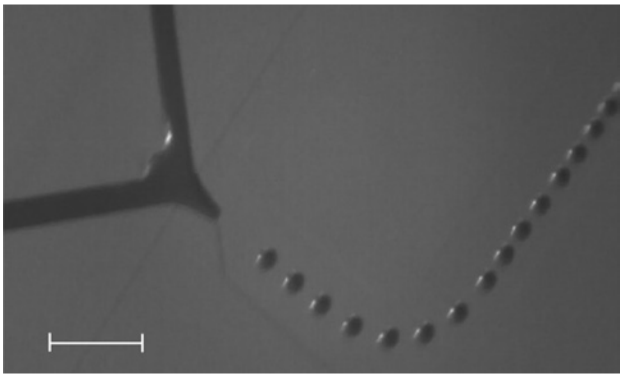


Figure 8: Design of the droplet-generation device used to generate droplets on the centrifugal microfluidic platform.

The PDMS devices were mounted to the PMMA disc with the reservoirs filled with mineral oil (with surfactant 3% by weight of Span 80) and deionised water with blue dye and observed at varying rotational speeds. At approximately 550 rpm, monodisperse water droplets in an oil phase were produced with high stability (Figure 9).



Scale bar = 400 μm

Figure 9: Water-in-oil droplet generation demonstrated on the centrifugal microfluidic platform, with the disc rotating at a speed of 550 rpm.

Discussion

The centrifugal microfluidic platform was successfully assembled. The design, manufacture and assembly processes were then successfully implemented and tested. The microfluidic disc control and analysis set-up was also successfully established, with hardware and software interfaces designed and implemented. A complete design-to-analysis example was developed, which illustrated the success of the integration of the various components of the centrifugal microfluidic platform. The ability of the centrifugal microfluidic platform to implement diverse microfluidic functions was illustrated by generating monodisperse water droplets in oil.

The results of the microfluidic disc example illustrate microfluidic functions as would be required for diagnostic applications, with particular relevance to blood tests. The microfluidic disc example illustrates that a biological sample can be added to an inlet chamber, with an appropriate sample preparation reagent – such as a lysing and/or staining reagent – contained in a separate chamber on the disc. The sample and reagent can then be added together in a controlled manner and contained for a required period of time.

Sedimentation of particles in fluids can also readily be achieved using the centrifugal microfluidic platform and could be useful in various diagnostic applications where cells need to be separated out of a sample. Sedimentation using the centrifugal microfluidic platform could be of use in blood tests in which plasma and blood cells are required to be separated, for example for the packed cell volume or haematocrit tests which form part of a full blood count, as well as for various other assays which make use of plasma as a sample.

The results of the droplet generation experiments illustrate that monodisperse droplets can be generated on the centrifugal microfluidic platform with high stability. This example also illustrates the ease with which existing PDMS microfluidic devices with fine microfluidic features can be integrated with the centrifugal microfluidic platform. A low-cost and simple microscope system was established for the centrifugal microfluidic platform, creating a basis on which to test and observe a variety of microfluidic devices at a high level of detail.

Microfluidic functions can be implemented on the centrifugal microfluidic platform with relative ease. In addition, the microfluidic disc manufacture process is simple, rapid and low cost, making it an ideal disposable component for point-of-care applications as well as allowing for rapid development of devices as a result of efficient prototyping. In addition, the radial symmetry of the microfluidic discs lends itself to multiplexed applications, where an array of tests can be carried out simultaneously on one disc. Similarly, a number of identical tests for different samples can be carried out on the same disc at the same time, increasing the throughput for the desired diagnostic application.

Fluid actuation of the lab-on-a-disc system is also simple and robust, using only a motor rotating at various speeds to achieve a vast array of functionality. The centrifugal microfluidic platform thus also has the potential to be developed into a compact, robust and simple system, ideally suited to point-of-care applications.

Conclusion

The lab-on-a-disc centrifugal microfluidic platform is a first of its kind in South Africa and has the potential to provide new diagnostic solutions at the point-of-need in health and industry-related areas. This potential can pave the way for providing resource-limited areas with services such as improved, decentralised health-care access or water-quality monitoring, and reduced diagnosis times at a low cost.

Acknowledgements

This work was made possible by the BioMEMS group at the University of California, Irvine (UCI) in the USA, who shared their expertise in the field of centrifugal microfluidics. The Council for Scientific and Industrial Research provided funding and support for this research.

Authors' contributions

S.H. was responsible for setting up the platform in South Africa and for conducting experiments and recording results. S.H. designed and manufactured the devices and platform interfacing hardware. K.L. provided guidance and conceptual contributions towards the platform design and implementation. H.K. and M.M. provided expertise and training on the platform fundamentals based on their facilities and research in the USA. S.H. and K.L. wrote the manuscript.

References

1. Haeberle S, Zengerle R. Microfluidic platforms for lab-on-a-chip applications. Lab Chip. 2007;7:1094–1110. http://dx.doi.org/10.1039/b706364b
2. Mark D, Haeberle S, Roth G, Von Stetten F, Zengerle R. Microfluidic lab-on-a-chip platforms: Requirements, characteristics and applications. Lab Chip. 2010;39:1153–1182.
3. Chin C, Linder V, Sia, S. Lab-on-a-chip devices for global health: Past studies and future opportunities. Lab Chip. 2007; 7:41–57. http://dx.doi.org/10.1039/b611455e
4. Lee W, Kim Y-G, Chung B, Demirci U, Khademhosseini A. Nano/Microfluidics for diagnosis of infectious dieseases in developing countries. Adv Drug Deliver Rev. 2010;62:449–457. http://dx.doi.org/10.1016/j.addr.2009.11.016
5. Yager P, Edwards T, Fu E, Helton K, Nelson K, Milton R, et al. Microfluidic diagnostic technologies for global public health. Nature. 2006;442:412–418. http://dx.doi.org/10.1038/nature05064
6. Gorkin R, Park J, Siegrist J, Amasia M, Lee B, Park J-M, et al. Centrifugal microfluidics for biomedical applications. Lab Chip. 2010;10:1758–1773. http://dx.doi.org/10.1039/b924109d
7. Madou M, Zoval J, Jia G, Kido H, Kim J, Kim N. Lab on a CD. Annu Rev Biomed Eng. 2006;8:601–628. http://dx.doi.org/10.1146/annurev.bioeng.8.061505.095758
8. Ducree J, Haeberle S, Lutz S, Pausch S, Von Stetten F, Zengerle R. The centrifugal microfluidic Bio-Disk platform. J Micromech Microeng. 2007;17:S103–S115. http://dx.doi.org/10.1088/0960-1317/17/7/S07
9. Sin M, Gao J, Liao J, Wong P. System integration – A major step toward lab on a chip. J Biol Eng. 2011;5(6):1–21.
10. Amasia M, Madou M. Large-volume centrifugal microfluidic device for blood plasma separation. Bioanalysis. 2010;2(10):1701–1710. http://dx.doi.org/10.4155/bio.10.140
11. Chen H, Li X, Wang L, Li P. A rotating microfluidic array chip for staining assays. Talanta. 2010;81:1203–1208. http://dx.doi.org/10.1016/j.talanta.2010.02.011
12. Duffy D, Gillis H, Lin J, Sheppard N Jr, Kellog J. Microfabricated centrifugal microfluidic systems: Characterization and multiple enzymatic assays. Anal Chem. 1999;71:4669–4678. http://dx.doi.org/10.1021/ac990682c
13. Noroozi Z, Kido H, Peytavi R, Nakajima-Sasaki R, Jasinskas A, Micic M, et al. A multiplexed immunoassay system based upon reciprocating centrifugal microfluidics. Rev Sci Instrum. 2011;82:064303. http://dx.doi.org/10.1063/1.3597578
14. Siegrist J, Peytavi R, Bergeron M, Madou M. Microfluidics for IVD analysis: Triumphs and hurdles of centrifugal platforms. Part 3: Challenges and solutions. IVD Technology DX Directions. 2010;Spring:22–26.
15. Mark D, Von Stetten F, Zengerle R. Microfluidic apps for off-the-shelf instruments. Lab Chip. 2012;12:2464–2468. http://dx.doi.org/10.1039/c2lc00043a
16. Sin MLY, Gao J, Liao JC, Wong P. System Integration – A major step toward lab on a chip. J Biol Eng. 2011;5:6. http://dx.doi.org/10.1186/1754-1611-5-6
17. Teh S-Y, Lin R, Hung L-H, Lee A. Droplet microfluidics. Lab Chip. 2008;8(2):198–220.
18. Huebner A, Sharma S, Srisa-Art M, Hollfelder F, Edel J, DeMello A. Microdroplets: A sea of applications? Lab Chip. 2008;8:1244–1254.

Shedding light on spectrophotometry: The SpecUP educational spectrophotometer

AUTHORS:
Patricia B.C. Forbes[1]
Johan A. Nöthling[1]

AFFILIATION:
[1]Department of Chemistry, University of Pretoria, Pretoria, South Africa

CORRESPONDENCE TO:
Patricia Forbes

EMAIL:
patricia.forbes@up.ac.za

POSTAL ADDRESS:
Department of Chemistry,
University of Pretoria, Private
Bag X20, Hatfield 0028,
South Africa

DATES:
Received: 05 Apr. 2013
Revised: 03 Sep. 2013
Accepted: 30 Sep. 2013

KEYWORDS:
spectrophotometry; educational; spectrophotometer; photonics; SpecUP

HOW TO CITE:
Forbes PBC, Nöthling JA. Shedding light on spectrophotometry: The SpecUP educational spectrophotometer. S Afr J Sci. 2014;110(1/2), Art. #2013-0096, 5 pages. http://dx.doi.org/10.1590/sajs.2014/20130096

Students often regard laboratory instruments as 'black boxes' which generate results, without understanding their principles of operation. This lack of understanding is a concern because the correct interpretation of analytical results and the limitations thereof is invariably based on an understanding of the mechanism of measurement. Moreover, a number of tertiary institutions in Africa have very limited resources and access to laboratory equipment, including that related to the field of photonics, which prevents students from acquiring hands-on practical experience. We address both of these challenges by describing how students can assemble a novel, low-cost spectrophotometer, called the SpecUP, which can then be used in a range of experiments. The same kind of information can be generated as that obtained with costly commercial spectrophotometers (albeit of a lower quality). With the SpecUP, students also have the opportunity to vary instrumental parameters and to observe the effects these changes have on their experimental results, allowing for enquiry-based learning of spectroscopic principles. The results obtained for some chemistry-related spectrophotometric experiments are described for each of the two operational modes of the SpecUP, although the instrument can be applied in fields ranging from physics to biochemistry.

Introduction

Spectroscopy forms an integral part of many undergraduate courses, in fields ranging from chemistry to physics to biochemistry. Spectrophotometers are therefore widely used in the practical portion of these courses. Problems in this regard arise with an increase in student numbers at tertiary institutions, especially if this increase is combined with a limited budget for capital equipment, which leads to students typically having to work in large groups during practical sessions or, at worst, to students merely observing a demonstrated experiment. Technology has progressed such that the variation of instrumental components within modern equipment has become largely electronically controlled, which has prevented students from obtaining hands-on knowledge relating to the operating principles of modern equipment.

Therefore, we designed and developed a spectrophotometer which would afford students the opportunity to discover and understand the concepts of spectrophotometry. The most important requirements for the final design included: that the instrument could be provided to universities as a kit containing components to be assembled by the students; that the settings of the components of the instrument could be manually varied by the student; that the instrument be sufficiently rugged to withstand repeated use; that the instrument generate experimental data of sufficiently good quality to be used in spectroscopy-related applied experiments; that building the electronic component of the instrument should not be the focus of the experiment; and, finally, that the cost of the instrument be significantly lower than that of commercially available instruments.

The concept of a 'build your own' spectrophotometer was based on a publication by Tavener and Thomas-Oates[1], which involved the use of a light emitting diode (LED) with a prism in front of a slit as a light source, and a light-dependent resistor (LDR) as a detector. The construction and accompanying electronics of their design are simple enough to allow for cost-effective assembly of the instrument, which made it a suitable starting point for the development of our design. However, their spectrophotometer does not allow for the movement of individual components, therefore we decided to modify the design to meet this requirement as well as to remove the focus from the construction of the electronic component of the instrument. Other spectrophotometer designs reported in the literature were far more complicated, especially with regard to the electronics, and were therefore not considered further.[2]

A first prototype was made in order to assess whether useful experimental results could be obtained from a spectrophotometer of very simple and cheap design, and to establish the components required, the spacing thereof and the size of the resulting instrument. The prototype was constructed largely out of black cardboard with plastic components (such as the lens, sample cell and slider mechanisms) (Figure 1). A LDR was placed behind the cuvette as a detector and the LED light source was built into an electric circuit based on that of Tavener and Thomas-Oates[1], as shown in Figure 2. The circuit contained a plug for the LEDs in order to allow for interchange of different coloured components. Encouragingly, useable results for a number of experiments were obtained with this model, including the generation of absorption spectra. The configuration of a more robust final prototype was then decided on, which led to the construction of the novel SpecUP.

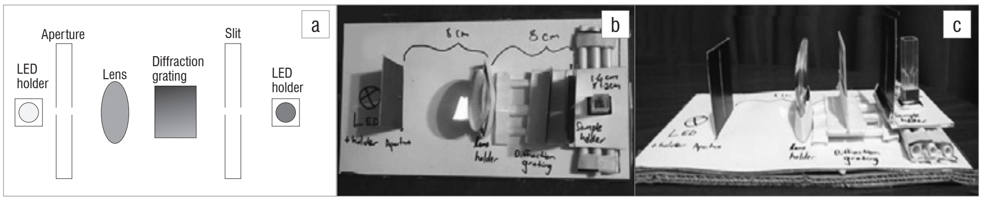


Figure 1: The first spectrophotometer prototype: (a) schematic layout of the optical plate and (b) an overhead view and (c) side view of the optical set-up.

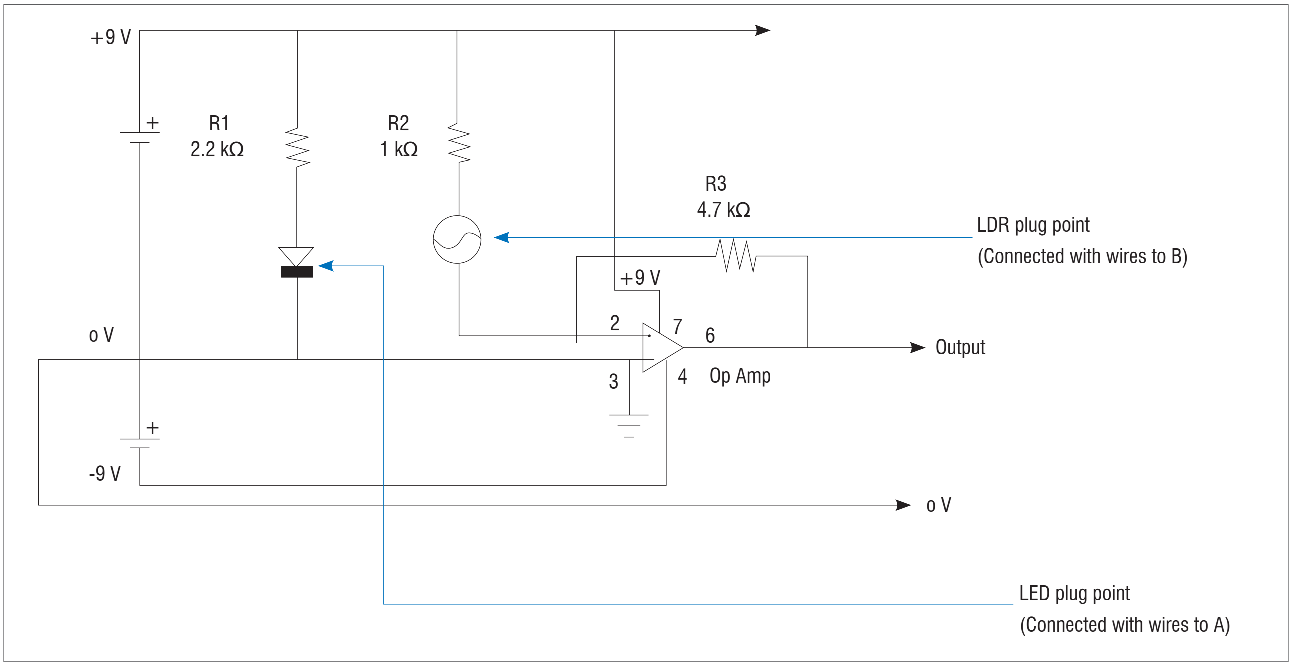


Figure 2: The electronic circuit diagram for the spectrophotometer, used for both the first prototype and the SpecUP (based on Tavener and Thomas-Oates[1]).

Design of the SpecUP

The SpecUP consists of either a white or coloured LED as the light source, a lens to focus the light, a diffraction grating with a slit which serves as a monochromator when a coloured LED is not used, and a LDR as detector (Figures 3 and 4). These components are mounted on a retractable bar consisting of three parts that can move independently: the LED on the first part, the lens and grating on the second, and the slit, cuvette holder and LDR on the third. The printed electronic circuit (Figure 2) equipped with 9 V batteries is used to supply the LED with power and to convert the resistance of the LDR to a voltage in the 1 V range. To measure the voltage, a low-cost commercial multimeter is used.

The SpecUP must be kept in the dark during operation because stray light can significantly influence the results. We therefore designed a tent-like cover of thick black cloth to fit over the SpecUP. The cover is light-tight and easily transportable and has a clear plastic window which allows for the viewing of experiments and can be covered with a flap (Figure 5).

The components of the SpecUP can be packed into a box of approximately 500 mm × 200 mm × 200 mm. Students could be responsible for assembling the SpecUP as part of their experimental report.

The estimated cost of the components is less than R500, which is significantly lower than the cost of commercial instruments, which are in the order of R30 000. The cost of many of the individual components of the SpecUP would in fact be reduced if purchased in bulk to prepare more spectrophotometer kits.

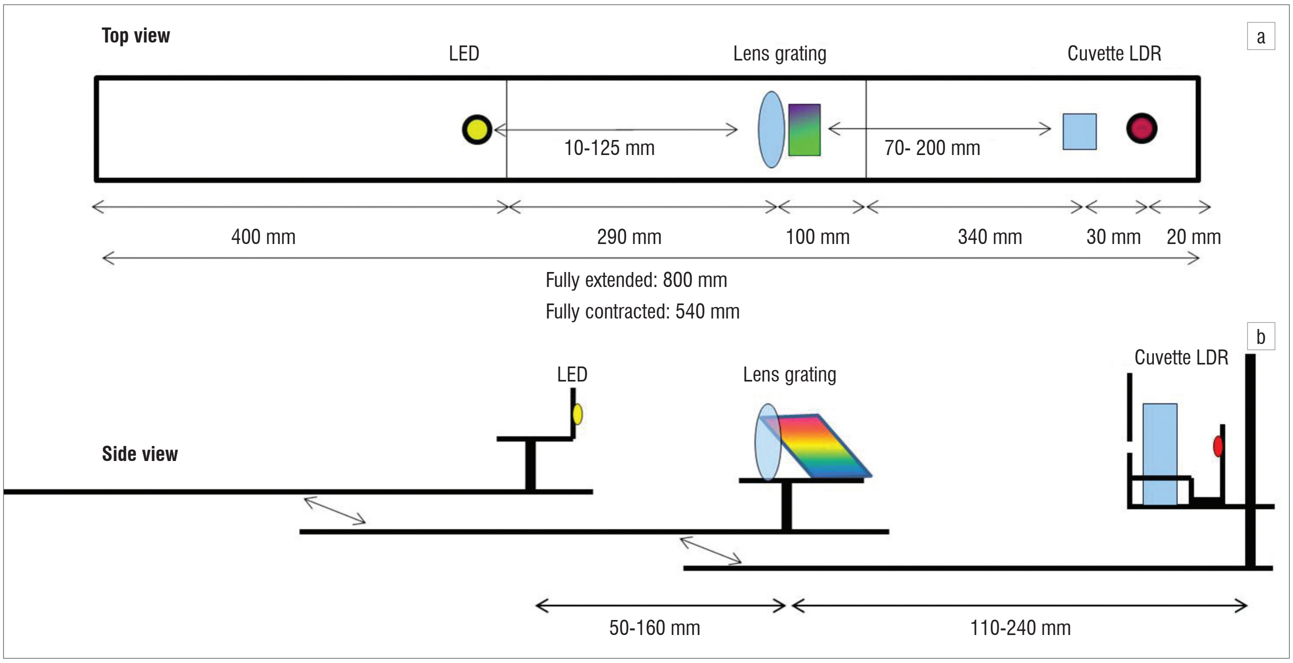


Figure 3: Schematic diagram of the (a) overhead view and (b) side view of the SpecUP.

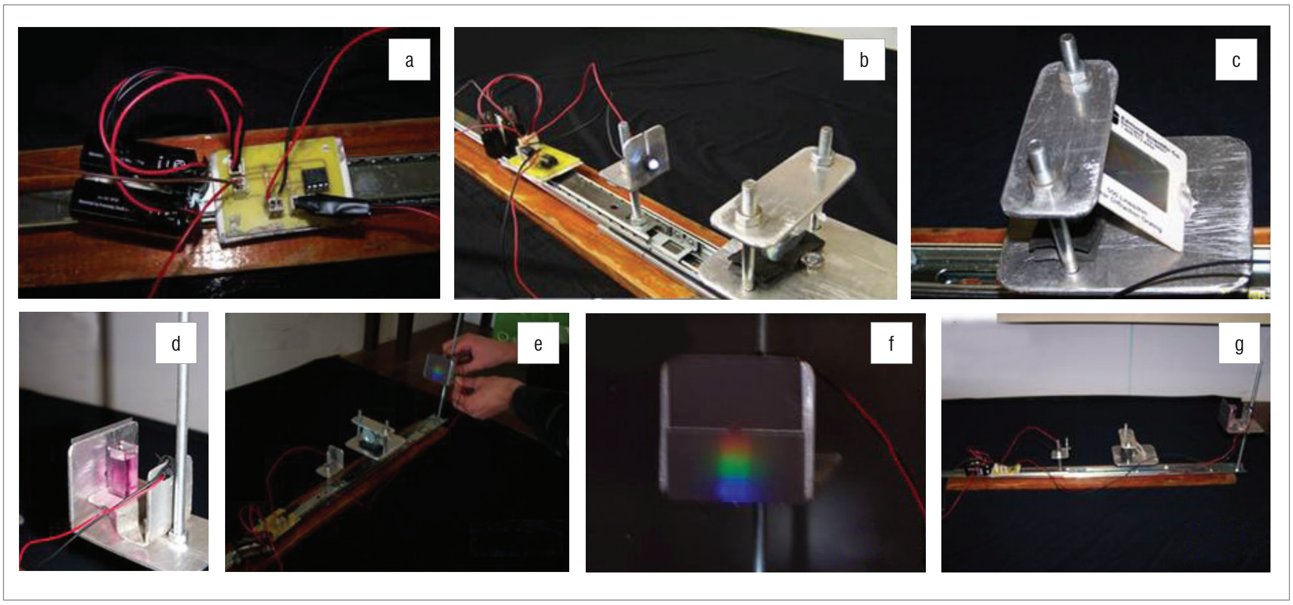


Figure 4: Components of the SpecUP spectrophotometer: (a) printed circuit board with batteries, (b) LED light source, (c) lens and diffraction grating, (d) slit and sample holder with sample, (e) adjustment of the slit position and (f) light separation at the slit. (g) Side view of the SpecUP.

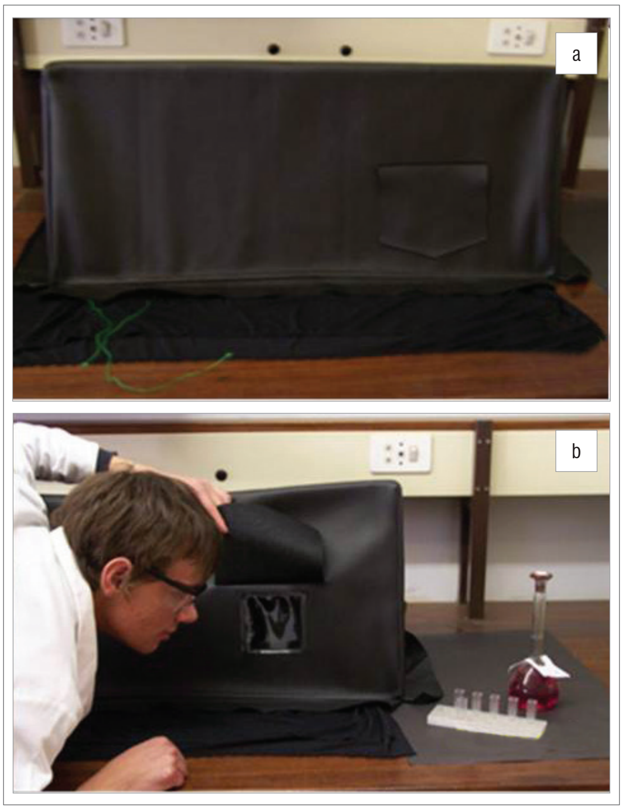


Figure 5: Housing of the spectrophotometer: (a) a black tent-like cover maintains the SpecUP in darkness during operation and (b) a clear plastic window enables the spectrophotometer components to be viewed.

Operation of the SpecUP

The SpecUP can be operated in two modes, which differ in the light source employed and the consequent requirement of incorporating a diffraction grating or not. Both modes are briefly described here and examples of experimental results achieved in each mode are provided.

Mode 1

In the first mode, no diffraction grating is needed. The colour of the sample or standard solution is judged by eye and a LED of complementary colour is used. The sample holder is lowered (or the long threaded rod is replaced with a shorter one) until the slit is directly in the path of the LED beam. The two parts of the retractable bar are shifted relative to each other until the light is focused at the slit. The cuvette is rinsed thoroughly and is then filled with distilled water (or another solvent if appropriate) and is placed in the cuvette holder. The SpecUP cover is placed carefully over the SpecUP and the bottom flaps are folded flat to prevent light ingress. The voltage ($V_{solvent}$) is measured, followed by the voltages obtained when the cuvette is sequentially filled with a range of suitable standard solutions of the coloured compound of interest in known concentrations. The voltage of a sample of the same compound of an unknown concentration is then recorded. In each case, correction is made for V_{dark}, which is the voltage reading obtained when the LED is off. The results obtained from the standard solutions are used to construct a calibration curve, and the concentration of the unknown sample is determined from the equation of the regression line for the linear region of the curve obtained for more dilute solutions.

An example of a simple experiment based on using the SpecUP in Mode 1 is the determination of the linear range of absorbance versus concentration for aqueous solutions of food colourants, which may serve as a practical introduction to spectroscopy and to the Beer–Lambert Law. A LED of a complementary colour to the food colourant is used (e.g. a red LED is used for green food colourant solutions). Voltage readings are used to calculate the transmittance of the solutions by Equation 1 (where P_{sample} and $P_{solvent}$ are the power of the light after it has passed through the sample and solvent, respectively); transmittance is then converted to absorbance using Equation 2.

$$T = \frac{P\ sample}{P\ solvent} = \frac{V\ sample - V\ dark}{V\ solvent - V\ dark} \qquad \text{Equation 1}$$

$$A = -logT \qquad \text{Equation 2}$$

Examples of the absorbance values obtained for sequentially diluted solutions are shown in Figure 6. It is evident that excellent linearity was achieved at lower colourant concentrations (between 0 and 7 drops of colourant per 100 mL water).

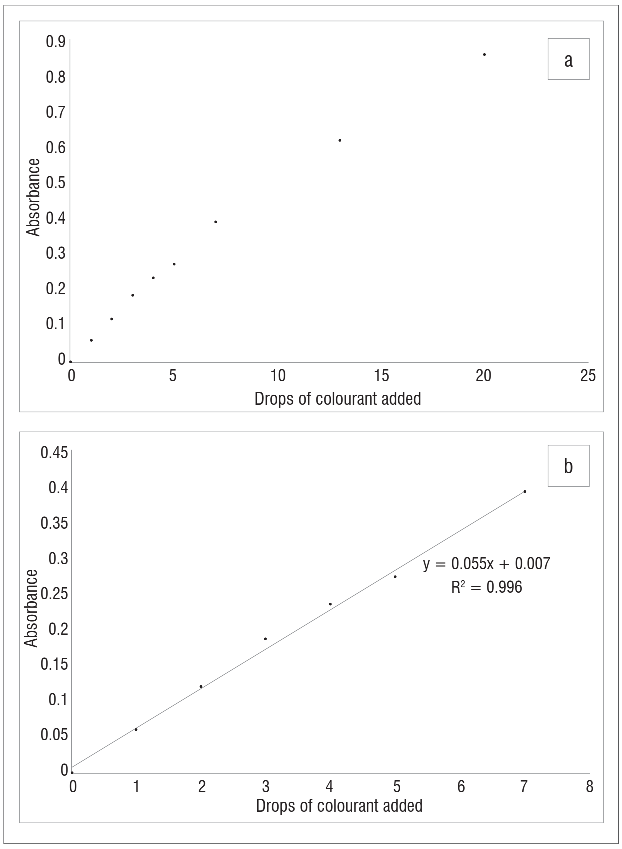


Figure 6: Absorbance values obtained for aqueous solutions of green food colourant using a red LED and the SpecUP in Mode 1: (a) linear and non-linear regions and (b) linear response obtained for lower colourant concentrations.

The SpecUP operated in Mode 1 with a blue LED was also successfully employed in the determination of the reaction kinetics of the iodine clock reaction (not reported on here).

Mode 2

In the second mode, the diffraction grating is used together with a white LED. The height of the sample holder is changed until the first light on either one of the edges of the diffraction spectrum barely shines through the slit (refer to Figure 4f). $V_{solvent}$ followed by V_{sample} are recorded and are corrected for V_{dark}. The nut on top of the sample holder (visible in Figure 4d) is loosened and the wingnut underneath the sample holder is turned through 360° so that a particular colour of light shines through the slit. The voltage of both the solvent and sample is measured for the new slit position, in order to determine the absorbance. This process is repeated until light on the far side of the spectrum barely shines through the slit. In order to calibrate the slit, the number of wingnut turns to reach various colours is noted and a calibration curve of wavelength (being the mean wavelength of the particular colour) versus number of turns is constructed (Figure 6). This calibration curve is then used to convert the number of wingnut turns (the independent variable) to wavelength. It is acknowledged that there is not a linear relationship between these variables; however, this simple, rough approximation provides useable results over the narrow range of angles employed, as is evident from Figure 7.

The SpecUP was operated in Mode 2 in order to obtain the absorbance spectrum of a solution of potassium permanganate. The absorbance of the solution for each wingnut position was obtained in the same manner as that described for Mode 1, but a white LED was used in conjunction with the diffraction grating. The spectrum of absorbance versus wavelength (Figure 8) was plotted after the number of wingnut turns had been converted to wavelength by means of a calibration graph such as that shown in Figure 7. The experiment was repeated three times in order to assess the repeatability of the method and its precision. From Figure 8 it is clear that the results achieved with the SpecUP were comparable to that obtained by a commercial instrument, although some of the fine spectral structure was lacking. The maximum absorbance wavelength can be determined from the SpecUP experiment, and used to determine the concentration of potassium permanganate in solutions of unknown concentrations by comparison of their absorbance values with those of standard solutions. The repeatability was also very good despite the rudimentary wingnut approximation employed.

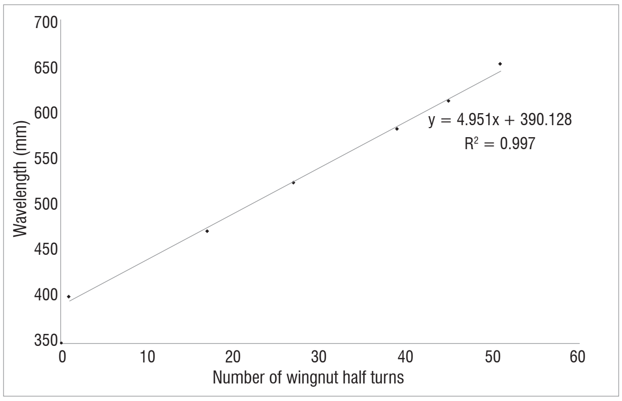


Figure 7: Calibration curve for wavelength as a function of the number of half turns of the cuvette sample holder wingnut obtained by using the SpecUP in Mode 2.

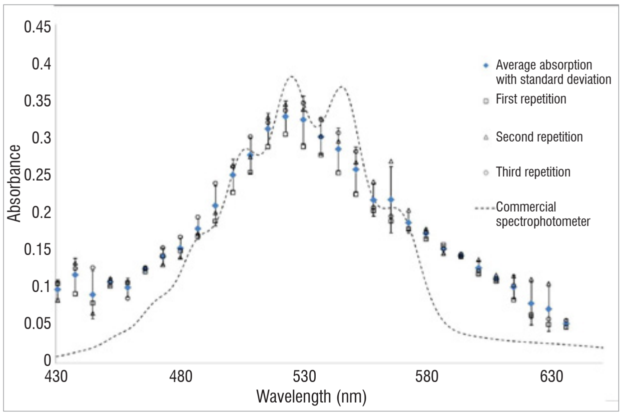


Figure 8: Comparison among the absorbance spectra of an aqueous potassium permanganate solution obtained using the SpecUP in Mode 2 (three replicates) and a commercial spectrophotometer, showing the spectral detail and repeatability obtained.

Conclusion

The SpecUP is not intended to be an accurate analytical instrument but an educational spectrophotometer. However, the sensitivity of the SpecUP is good enough to determine concentrations to within a few per cent. An absorbance spectrum of a potassium permanganate solution could be obtained which resembled that acquired with a commercial instrument; however, the finer spectral details could not be resolved. Modifications could be made to the SpecUP to enhance the experimental data which it can produce, such as improving the quality of the optical components employed.

A limitation of the SpecUP is that the path length of incident light through the sample varies with changes in the cuvette height. As the height of the sample holder is increased, the path length of the light also increases, which, according to the Beer–Lambert Law, increases the absorbance. Students should be encouraged to discover this effect and could suggest means of compensating for it as part of their experiments.

It is recommended that experimental methods include sections relating to the effect of changing experimental parameters (such as lens and sample positions) on the experimental results obtained, in order to fully utilise this important functionality of the SpecUP. The SpecUP enables students to probe topics such as accuracy, precision, resolution, sensitivity and selectivity. These variables, as well as possible sources of experimental error and the limitations of each experiment, should be addressed by students in their laboratory reports.

The illustrative experiments which were included do not require the use of expensive or toxic chemicals. Additional experiments which could be performed using the SpecUP include the determination of metal ions in water based on the formation of coloured metal complexes,[3] the investigation of reaction rates (such as the iodine clock reaction) and the determination of the end points of colourimetric titrations.[4]

The SpecUP spectrophotometer which has been developed is novel, simple and cheap to construct. In addition, the configuration allows for the movement of many of the components, which enhances the opportunity for students to follow the enquiry-based learning principle in discovering the effects of changing parameters on their experimental results. The open design of the instrument and the tent-like cover with a transparent window, allow students to get a real hands-on experience of spectrophotometry and thereby to better understand the principles involved.

The cost of the SpecUP is approximately R500, whilst commercial spectrophotometers cost ∼R30 000. It is therefore believed that by utilising the SpecUP in undergraduate laboratories, universities could afford to have more instruments than if commercial spectrophotometers were employed. Students would thus have much more of a hands-on practical experience, which would aid their understanding of spectrophotometry and hopefully encourage them to engage in photonics-related research in the future.

Acknowledgements

We thank the CSIR National Laser Centre and the African Laser Centre for financial support; Dr Paul Motalane, Prof. Andrew Forbes and Thomas du Plooy for their guidance; Prof. Thomas-Oates for her interest and comments on the project; and Leon Engelbrecht, Nico van Vuuren, David Masemula and Mrs and Prof. Nöthling for assisting with the manufacture of components. Phakama Botha is acknowledged for her contributions in the early stages of the project.

Authors' contributions

J.A.N. undertook this project as a third-year undergraduate student and was involved in all phases from the development of the design of the SpecUP to its manufacture and operation (devising and performing suitable application experiments). P.B.C.F. came up with the idea of the SpecUP, obtained funding for the project and supervised the project.

References

1. Tavener SJ, Thomas-Oates JE. Build your own spectrophotometer. Educ Chem. 2007;4:151–154.
2. Yeh T-S, Tseng S-S. A low cost LED based spectrometer. J Chin Chem Soc. 2006;53:1067–1072.
3. Hauser PC, Rupasinghe TWT. Simultaneous determination of metal ion concentrations in binary mixtures with a multi-LED photometer. Fresenius J Anal Chem. 1997;357:1056–1060. http://dx.doi.org/10.1007/s002160050304
4. Nazarenko AY. Optical sensors for manual and automatic titration in undergraduate laboratory. Spectrosc Lett. 2010;43:555–560. http://dx.doi.org/10.1080/00387010.2010.510705

AUTHOR:
Naven Chetty[1]

AFFILIATION:
[1]School of Chemistry and Physics, University of KwaZulu-Natal, Pietermaritzburg, South Africa

CORRESPONDENCE TO:
Naven Chetty

EMAIL:
chettyn3@ukzn.ac.za

POSTAL ADDRESS:
School of Chemistry and Physics, University of KwaZulu-Natal, PO Box X01, Scottsville 3209, South Africa

DATES:
Received: 06 Dec. 2012
Revised: 22 Mar. 2013
Accepted: 09 Oct. 2013

KEYWORDS:
problem-based learning; access; extended curriculum; language; study techniques

HOW TO CITE:
Chetty N. The first-year augmented programme in Physics: A trend towards improved student performance. S Afr J Sci. 2014;110(1/2), Art. #2012-0096, 9 pages. http://dx.doi.org/10.1590/sajs.2014/20120096

The first-year augmented programme in Physics: A trend towards improved student performance

Amidst a critical national shortage of qualified Black graduates in the pure and applied sciences, the University of KwaZulu-Natal has responded to a call from government for redress by launching the BSc4 Augmented Physics programme. In this paper, the methods employed to foster learning and to encourage student success in the Mechanics module of the Augmented Physics programme are described and discussed. The use of problem-based learning and a holistic learning policy that focuses on the emotional, physical and knowledge development of the student seems to have yielded higher throughput in the first semester of an undergraduate programme in Physics. Furthermore, the results point to an increase in the conceptual understanding of the student with respect to Mechanics. When appraising this success, the results of the 2007–2009 cohorts, with and without teaching interventions in place, were analysed. These initial analyses pave the way for a course designed to benefit the student and improve throughput. These methods are not unique to Physics and can be adapted for any module in any country.

Introduction

Worldwide, students have difficulty with the language of physics, be it subject-specific terminology or the use of everyday language in a physics context.[1] Thus, even definitions may give students trouble.[2] This difficulty is compounded when learning in a second or third language,[3-7] which is the case in South Africa where only 12% of students applying for tertiary education are mother-tongue speakers of English.[8]

Further difficulties in South Africa can be attributed to a dysfunctional education system as a result of the apartheid regime's under-development of Black human potential.[7] This system gives rise to the problem of talented students being unable to study further in the sciences because of inadequate schooling.[7] Furthermore, as a critical national priority, South African universities have been urged to alleviate the problem of 'scarce skills' by increasing the number of Black graduates in the natural and applied sciences.[8,9] Universities in South Africa have attempted to redress the inadequate number of Black graduates through a variety of programmes offering alternative access and support.[10-13]

Hutchings and Garraway[13] discuss in great detail the current extended curriculum position in South Africa and highlight the need for such measures in light of the current educational schooling pitfalls. In particular, they discuss the various approaches employed at other South African higher education institutions to implement extended curricula as part of the government's National Development Plan. The University of KwaZulu-Natal's (UKZN's) augmented programme is unique in this realm of extended curricula and thus its description will assist greatly in formulating plans by other institutions for implementation of extended curricula. In particular, at UKZN, the Centre for Science Access (CSA) seeks to address the needs of students from disadvantaged schools who do not meet the normal requirements for entry to the Faculty of Science and Agriculture. In the CSA, students register for a 4-year BSc in either a foundation or an augmented programme. In the BSc-4 (Foundation) programme students engage in learning that is modelled on the Science Foundation Programme which has been described by Kioko[12], Grayson[14-15] and Barnsley[16]. The Physics programme for the BSc-4 (Augmented) has not been described previously in the literature, but will be described here later.

Level 1 Physics at tertiary level is highly dependent on problem solving.[17,18] Problem solving is often a stumbling block for many students in Physics as they perceive it to be difficult.[19,20] Research also shows there is often little or no change in conceptual understanding before and after formal instruction and that students are unable to apply the concepts that they have studied to the task of solving quantitative problems.[20]

Here, teaching interventions instituted for the Physics module in the BSc-4 (Augmented) programme are described and their effectiveness in addressing students' ability to answer typical Mechanics questions, as given in first-year textbooks, is appraised.[21-22] An attempt is also made to investigate the problem-solving ability of three cohorts of students. The results, although preliminary, will be of interest to researchers in the field of extended curricula.

Context of the study

Educational context

The BSc-4 (Augmented) degree is for students from disadvantaged schools who are interested in science degrees but whose matric results are slightly below college entry requirements, although they have a full matriculation exemption or National Senior Certificate (NSC) qualification. The entry requirements for both the augmented and the mainstream programmes are listed in Table 1. Students in the Augmented Physics programme are admitted into first-year BSc but initially take fewer courses with extra tutorials and practicals, and courses in Scientific Communication and Life Skills. The first year of the degree is therefore spread over a maximum of 2 years during which students can also take some second-year modules. Thereafter, students carry the normal load for their degrees. Students thus take 4 years to complete a 3-year BSc degree, doing so more slowly, but being more assured of success.[14-16] The sudden withdrawal of support after the first year of study is of concern and its impact needs to be investigated further as it may have direct consequences on throughput and pass rates in subsequent years. One needs to be cognisant of this possibility when deciding on the exact structure of the extended curriculum that is implemented. At UKZN, various support structures have been set in place for all students in second year and above and it has been deemed appropriate to not offer a specific support structure for the students in the augmented programme.[23]

The BSc-4 (Augmented) degree is based on an integrated programme for students' first year of study.[23] In this programme all students register for two compulsory modules to support their studies and two optional modules determined by their choice of degree majors. One compulsory module is Scientific Communication, in which students are coached to improve their English language skills in the context of reading and writing on science topics. They learn the appropriate language, report structures and basic comprehension skills to help them in their chosen degree. Students must pass this module in order to be awarded their degree. This module is, however, the basis of much controversy. Such discrete communication courses have been shown to have very little impact on improving language in a specific discipline such as physics[13,24] and indeed this was the case in this Scientific Communication module. This course served to provide the students with generic communication skills that had little impact on their physics communication skills. With this in mind the Augmented Physics module implemented a mechanism of redress which included physics communication and, in particular, communication with respect to understanding, unpacking and answering problems in the problem-based learning (PBL) mode of teaching and learning.

The second compulsory module is Life Skills. This module has an attendance but no assessment requirement. It addresses issues including HIV/AIDS, food security, note taking skills, career prospects and basic study techniques through workshops and discussions conducted by qualified psychologists. These activities are designed to help students adjust to the social and academic requirements of university.[10-12,14-16,24,25] Furthermore, all students in the CSA are encouraged to meet campus psychologists for further counselling on personal, academic, career or social issues. This component has been shown to be most effective in improving the confidence of the students as well as in helping them overcome the adversities of their past without feeling victimised or alienated in their new environment.[25]

The students enrolled in the Augmented Physics module have twice the amount of contact time in Physics (but not double the workload) as those registered for mainstream Physics. Although they attend twice as many Physics lectures, students in Augmented Physics still have a lower workload than students in the mainstream programme as they only take two modules as opposed to the mainstream students who register for four modules. Augmented Physics students attend normal classes for the calculus-based Physics module (four lectures, one tutorial and a 3-h practical per week) alongside their mainstream counterparts. For these classes, they are expected to do various assignments such as answering tutorial questions. Being previously disadvantaged, students in the augmented programme have had very little exposure to laboratory work and thus they are required to attend an additional 3-h laboratory session a few days before their mainstream practical. These additional sessions differ from the mainstream ones – during these sessions, students are introduced to the apparatus, experimental techniques and analytical skills that are required for the practical later in the week.

To further support their studies in the mainstream, students in the augmented programme attend five other contact periods which the Augmented Physics lecturer uses for explaining or extending the content of the mainstream lectures, rather than simply re-lecturing the same content. However, students tend to struggle with the pace of the mainstream lectures, so when required the Augmented Physics lecturer re-lectures the mainstream content. Over and above the assignments for mainstream classes, the lecturer in Augmented Physics assigns further tutorial questions for students to prepare for the classes. The workload in Augmented Physics is initially less intensive than the mainstream programme to enable the students to adapt to the demands of tertiary physics education, but is incrementally increased until it is on a par with that in mainstream Physics.

Student context

This research was conducted among three student cohorts between 2007 and 2009 who had different secondary school backgrounds. Table 2 shows the composition of the three cohorts of students, according to the examination they wrote in their final school year.

The 2007 cohort of students followed a content-based traditional school curriculum. The curriculum statement for this system outlined clear learning objectives from which teachers should work.[26] The final examination included questions across a range of levels on Bloom's taxonomy.[26,27] In this way, the exam encouraged teachers to focus on teaching some problem-solving skills to the learners. The learners had no continuous assessment tasks and their final marks were wholly obtained from the final exam.

By contrast, students matriculating in 2008 followed the new NSC curriculum which is purported to be outcomes-based education (OBE). These learners were the guinea pigs in that they had, along with their teachers, constantly pioneered a new system over 12 years of schooling. The curriculum statement for this certificate leaves quite large areas open for teachers to interpret. Furthermore it appears to cover even more content than the curriculum in previous years,[28,29] thus creating a situation in which teachers might need to rush quickly through material in order to complete the syllabus superficially. This situation can result in learners simply plugging numbers into an equation. Consequently, there may be little time to teach learners to solve problems that require

Table 1: Comparison of entry requirements for BSc and BSc-4 (Augmented) degrees at the University of KwaZulu-Natal

	Senior Certificate (–2007)		**National Senior Certificate (2008–)**	
	Normal entry to BSc in Faculty of Science and Agriculture	**Entry to BSc-4 (Augmented)**	**Normal entry to BSc in Faculty of Science and Agriculture**	**Entry to BSc-4 (Augmented)**
Points based on overall achievement	32	28	30	22
Mathematics	HG D or SG A	HG E or SG B	Level 4	Level 3
At least one of: Physical Science(s), Biology, Biological Sciences, Agricultural Science	HG E or SG B	HG E or SG B	Level 4	Level 3
English	HG E or SG B	HG E or SG B	Level 4	Level 4
Life Orientation	Not required	Not required	Level 4	Level 4

deeper interrogation and integration of the subject. Integration deals with the extent to which teachers use examples, data and information from a variety of disciplines and cultures to illustrate the key concepts, principles, generalisations and theories in their subject area or discipline.[30] The final result for these learners includes a component of continuous assessment. The majority of the 2008 cohort came through the new curriculum, although the cohort also included a few students who matriculated in 2007 but only enrolled for BSc-4 in 2008. All of the students in the 2009 cohort matriculated through the new NSC.

Researcher context

The author taught the Augmented Physics module for the three years under study. He had 8 years experience tutoring physics at high school and tertiary level and completed his PhD during this period. The author taught the module for the first time in 2007. In a mixed modal research study (based on both qualitative and quantitative data) such as this one, the observer is part of the research process.[31-35] Therefore, despite attempting to be a disinterested observer, the author's particular perspective has undoubtedly framed this study.[32-35] In this paper, the qualitative research in which the author/researcher has been directly involved, is presented. The aim of the study was to gather data in the form of the qualitative assessments and to supplement those research findings with quantitative data in the form of rich content-based descriptions of people, events and situations by using different, especially non-structural, techniques to discover the stakeholders' views, to analyse the gathered data and, finally, to interpret the findings in the form of a concept- or contextually dependent grounding.[32-36]

Table 2: Composition of each cohort of students in the study

	2007	2008	2009
'Old' matric	25	4	0
National Senior Certificate	–	15	26
Total	25	19	26

Theoretical framework

The effectiveness of a particular educational approach (in this case the Augmented Physics programme) and its teaching effectiveness (as demonstrated by student knowledge of the subject matter and evidenced by performance in the assessments) need to be determined. The aim of this study was to determine the knowledge gained by students in terms of factual knowledge, conceptual understanding and functional proficiency in physics and, in particular, mechanics.[32-35]

The mixed-modal approach was used as it starkly provides the answers to the question of effective teaching by allowing quantitative analysis of the results obtained by the students to particular questions designed to test various levels of the revised Bloom's taxonomy.[27,36] It also relates their performance to follow-up questions and discussions (qualitative analysis). Although very useful, it must be noted that a shortcoming of this method is that it is subject to the specific teaching measures implemented.[32-35] A further justification of a quantitative approach arose from the large number of students enrolling in mainstream Physics which made conventional assessment tasks difficult to implement and tedious to mark. These numbers prompted a departmental decision to move away from asking a number of conceptual understanding questions (so called 'long questions' with proofs and derivations) in tests, to more problem-based/quantitative assessments such as multiple-choice questions (MCQs) or short numerical response questions.[32-35,37] Given the debate surrounding the use of MCQs,[38-41] the results of this research may be debated. However, the questions were designed to meet the specific aim of the researcher, which was to determine the cognitive ability of the students to answer questions at various levels of Bloom's taxonomy[27] and the revised Bloom's taxonomy[36] and the results should be interpreted in this light.

Bloom contends there are three types of educational activities[27]:

- Cognitive: mental skills (*knowledge*)
- Affective: growth in feelings or emotional areas (*attitude*)
- Psychomotor: manual or physical skills (*skills*)

Anderson and Krathwohl[36] provide a revision of this taxonomy. Changes in terminology between the two versions are the greatest differences. Bloom's six major categories were changed from noun to verb forms. The lowest level of the original – knowledge – was renamed to 'remembering'. The levels of comprehension and synthesis became 'understanding' and 'creating', respectively. In this study, the focus was on Bloom's original cognitive category and the tests were developed as part of our quantitative approach. The first few questions on the test (approximately 20%) appealed to the first level of Bloom's cognitive level, that of knowledge[27] or remembering[36]. If proficient at this level, the students should be able to provide answers to questions that simply test basic definitions and recall.[36,42]

The next set of questions on the test (approximately 40%) related to the second level of Bloom's taxonomy, that of comprehension[27] or understanding[36]. At this level, a competent student would understand the meaning, translation, interpolation and interpretation of instructions and problems.[36,37,42,43]

The final set of questions (approximately 40%) appealed to Bloom's third level – application[27] or creating[36]. This level requires students to use a concept in a new situation or, unprompted, to use an abstraction. The three highest levels of analysis, synthesis and evaluation were not considered in this study although the linking of the revised Bloom's taxonomy[36] with PBL aims to achieve this as students progress through the module[44]. Yadav[45] alludes to this form of instruction as the transference of higher-order thinking skills. Yadav refers to this skill as the ability of the student to collect, analyse and evaluate information to draw conclusions or make inferences[44,45] and his approach[45] forms the basis of the principle used in this paper. In preparing for the questions set using the revised Bloom cognitive levels, the method proposed by Serrat[46] was used. In this method, PBL is used together with Bloom's taxonomy (or the revised taxonomy) to facilitate higher-order thinking skills. Serrat's[46] approach makes provision for problem-solving techniques for use in PBL. Four of these techniques are focused on here:

1. *Affinity diagrams* encourage students (either in a group or as individuals) to organise ideas into a common theme.
2. *Brainstorming* encourages students (either in a group or as individuals) to generate a large number of ideas to solve a problem or to find ways of solving the problem.
3. *Flowcharts* are used to help students identify the aspects they do not understand with respect to content before attempting to teach problem-solving skills.
4. The '*five why's technique*' encourages students to ask at least five questions when solving a problem, which helps them to think 'out of the box'.

Fogler and LeBlanc[47] were the first to develop the concept of linking Bloom's taxonomy to PBL and as such to a scientific field (engineering in this case). From this impetus, this study was modelled.

Method

In interrogating the successful outcome of any of the teaching interventions employed in the Augmented Physics programme, it is only possible to consider examination pass rates. However, pass rates may give only a superficial evaluation and may not show any of the finer aspects relating to students' improved performance or the challenges faced by a lecturer in facilitating this change. Consequently, analyses of student performance on tests related to their problem solving ability are also included.[32,34,35]

At the start of the first semester in each year under consideration, the same pre-test was administered to each cohort in the first lecture of the Augmented Physics class, in which students worked alone and without reference material. As a formative assessment instrument, the purpose of the pre-test was to gauge the students' understanding of physics at school-leaving level, and consequently their ability to solve problems. Students were informed beforehand that their performance in the test would not affect their class mark in order to reduce test anxiety among the students which may have adversely affected their performance. Furthermore, time pressure was alleviated by allowing students up to 1 h to answer the 30-mark test which, according to common practice in the Physics Department, would normally have taken 30 min. The lecturer then collected the scripts, analysed the responses and checked the marking and mark allocation before handing the scripts back to the students. A discussion session with the students then occurred, either on a one-to-one basis or in a classroom group, to obtain data for the qualitative aspect of the research.

The test consisted of three sections with each section evaluating student performance on the first three levels of cognitive understanding according to the revised Bloom's taxonomy as described previously.[36] The first level of knowledge simply required that a student recall basic definitions. The second section, at the second level of understanding, entailed interpreting a straightforward 'story' to extract information concerning a sequence of actions. The final section was at the third level of creating[36] and required students to use a concept in a new situation or to use an abstraction unprompted, that is the student had to apply what was learnt in the classroom into novel situations. Examples of all these questions are given in Table 3.[42]

Table 3: Sample test questions for the first three levels of Bloom's cognitive levels

Bloom's cognitive level of knowledge State how a vector quantity differs from a scalar and give an example of a vector quantity. (2)
Bloom's cognitive level of comprehension Sinethemba walks 60 m north before discovering that she is lost. She then walks 20 m south to try and retrace her steps with no success. Sinethemba then walks 10 m west and 40 m east before going 35 m north and reaching her destination. Determine the distance covered by Sinethemba and her resultant displacement. (3)
Bloom's cognitive level of application Sipho is travelling at a constant velocity of 20 m/s and passes an intersection where his friend Dumisani is parked. At the instant Sipho passes the intersection, Dumisani starts up in pursuit of Sipho. If Dumisani accelerates at a constant 5 m/s, determine the distance he covers before catching up with Sipho. (5)

The number within parentheses indicates the maximum mark awarded for a question.

Results and discussion of interventions

The results given in the tables and figures that follow indicate the percentage pass rate for a particular level of the revised Bloom's taxonomy. A pass is deemed to be 50%.

Table 4 illustrates the results of the pre-test. It is interesting to note that in the mixed 2008 cohort three out of the four 'old' matric students achieved an overall pass in the pre-test. The results represented in Table 4, although poor, were not unexpected. It had been predicted by various educational sources that the new NSC would not adequately prepare students for the rigors of tertiary study, where emphasis was placed on the higher levels of Bloom's taxonomy.[48-50] Jansen[51-53] was critical of OBE even before its launch and his prediction was predicated by the decision of the education ministry to scrap OBE.[54] However, a single pre-test could not be used to conclusively judge this system because the test may have been influenced by anxiety on the part of the student, despite the precautions taken to avoid such anxiety. The students were then given advance notification that they would be writing a test a week after the pre-test. The students were also explicitly informed of the material that the test would cover. No interventions were made in the week between the tests and the lecturer conducted traditional lectures with no advanced support.

Table 4: Pass rates for the pre-test for each cohort

Bloom's level	2007 (25 students)	2008 (19 students)	2009 (26 students)
Knowledge	92%	84%	77%
Comprehension	42%	37%	38%
Application	28%	11%	8%
Overall pass rate for tests	44%	37%	31%

The result of the first formative test in the Augmented Physics module, which again was not used to contribute to the students' year mark, is shown in Table 5. This test was very similar in nature to the pre-test and tested similar concepts. The results closely resembled those of the pre-test. This time all four of the 'old' matrics in the 2008 cohort passed overall. Table 5 further shows that a vast majority of the students in all three years answered questions in the knowledge section correctly. This result is not surprising as the schooling system prepares students to memorise material so that their ability to simply recall definitions is well developed. Noticeable, however, is the appreciable drop in the percentage of students who responded correctly to the question over the years. This finding may be a consequence of the inordinate amount of group work associated with OBE in which learners are not required to take ownership of their work and studying. Instead, learners are given more opportunity to simply rely on the efforts of the more active learners in a group which can curb their own learning substantially.[48-50,52-54] It must also be noted that students were not marked on the correct use of language and grammar in this section. The researcher inferred what was conveyed by the students even though the written expressions may not have made conventional linguistic sense.

Table 5: Pass rates for formative Test One for each cohort

Bloom's level	2007 (25 students)	2008 (19 students)	2009 (26 students)
Knowledge	96%	79%	77%
Comprehension	44%	37%	38%
Application	32%	11%	8%
Overall pass rate for tests	48%	37%	31%

The results for the understanding section were in stark contrast to the researcher's perception that the questions were easy, and proved to be an immense challenge to the vast majority of students across all three years. A qualitative investigation was conducted by the researcher which involved interviewing all the students who had failed to obtain 50% or more in this section to identify the reasons for their poor performance. In all three years the students battled to interpret the questions because of the perceived complexity of the language used. This perception resulted in the students being unable to separate relevant from irrelevant information provided in the question; this finding tallies strongly with reports in the field[55-57] which highlight that student responses and incorrect answers are often linked to language difficulties as opposed to subject or content difficulties. It was clear from this part of the study that the language problem facing the students would need to be addressed in conjunction with the physics interventions and that the discrete language module failed.[58,59]

The results in Table 5 also show that many of the students were unable to solve questions in section three, the creation level. The qualitative investigation again revealed that this inability may have been attributed to the complexity of the language – by their own admission, the students were unable to dissect a question into smaller, more manageable parts. In all three years the students also admitted to attempting to solve the problem without first determining the nature of the question posed. They could, by and large, readily identify that they needed to use the kinematical equations; however, they could not see the link between two scenarios, such as the distance covered and the time of the motion.

The results of this formative test then set in motion a series of interventions for the Augmented Physics module. The first two interventions sought to address informally the language problem encountered by the student. As time was limited, a more formal approach could not be taken to address the complexities surrounding language and its impact on formal learning as it was clearly apparent that the scientific communication module was not adequately preparing the students for the language of physics. The first intervention was the introduction of dictionaries into lectures. The lecturer would often take dictionaries to lectures and students were encouraged to look up the meaning of words if they did not understand them. They were also asked to start their own physics dictionary by writing down the meanings of difficult words they encountered in the module. Across the three years this intervention varied slightly, but the main thrust remained consistent. In all three years almost 95% of the students had cellular phones with WAP capabilities, and so they were introduced to the mobile Internet, more especially to online dictionaries and e-learning sites, as a way of introducing them to technological learning, an ambit of the PBL method.[60-64]The introduction of the technological ambit of the PBL approach helps to determine what factors contribute to integration or non-integration of those constructs into the curriculum.

The second measure was the implementation of an English-only policy for lectures. Students were not allowed to communicate in their home languages during any lectures.[59,64-70] Even if the communication was not physics related the students were still required to communicate in English. This policy was initially quite difficult to implement because the students reacted quite negatively to this measure. They believed that this policy belittled their language and discriminated against their cultural beliefs. The students were then counselled on the reason behind this policy and it gradually became more acceptable. The lecturer observed on many occasions that if a student did try to stray from this policy, the other students in the class or group would simply respond to his or her question in English.[61,71] The language problem is often compounded by the fact that the educators at the students' former schools (usually semi-rural and rural schools) would often 'code-switch' – that is, switch between English and the vernacular language. Often the educator would start the lesson in English and then revert to the vernacular when a concept with a relatively high degree of difficulty was encountered. Students were also given an in-depth explanation into the usage of words such as 'describe', 'explain', 'calculate', 'determine' and 'solve'. These words had to be recorded in their dictionaries and they were then randomly asked to explain the meaning of these words to the lecturer or their colleagues during lectures.

The final measure, and probably the most influential, was the PBL approach. PBL is a student-centred instructional strategy in which students collaboratively solve problems and reflect on their experiences. Characteristics of PBL[72-74] are that:

- Learning is driven by challenging, open-ended problems
- Students work in small collaborative groups
- Lecturers take on the role as 'facilitators' of learning.

Advocates of PBL claim it can be used to enhance content knowledge and foster the development of communication, problem solving and self-directed learning skill.[47] The PBL approach should encourage students to be responsible for their own learning and understanding. In other words, PBL should discourage students from memorising content or copying solutions to assigned problems because they should see that tutorial questions were assigned to develop conceptual understanding and critical or lateral thinking rather than simply to supply a final answer. Consequently, ownership of the academic experience lies with the students as they have to demonstrate understanding of the content through PBL. PBL often forces students to study their materials (notes, textbooks, readings) from lectures intensively and to make their own notes to help with solving the problems. By splitting the class into smaller groups and assigning the groups problems to attempt using the PBL approach, students are forced to learn the associated theory before attempting the problem. Ownership of the academic experience further encourages them to think laterally and critically – a consequence of the reflection part of the PBL approach. Thus PBL is an attempt to help students to think 'out of the box'[44-47,75] while OBE is an approach to education in which decisions about the curriculum are driven by the exit learning outcomes that the students should display at the end of the course.[76]

The students were split into small groups of two or three individuals. Problems were assigned to these groups and they were required to solve them without consulting the other groups. Each group was allocated different problems with the same level of difficulty. To solve the problem, students had to define a structured, step-by-step approach to the solution. In some common traditional instruction scenarios, students are handed a cookbook list of steps to solve a given problem (e.g. Part A: solve for velocity; Part B: solve for Δx).[44-47,74,75] In such situations, students may struggle with performing some of the steps or with the implicit meaning of these steps or, worse yet, may fail to recognise how each of these steps leads to a global solution to the problem. However, in a PBL approach, students must generate their own step-by-step method to solve each problem. Thus, although difficulties will arise in carrying out a given step, no confusion exists as to the sequence or meaning of each step required. Furthermore, as students work to solve the problem, various solution paths emerge among groups. In this way, students begin to view problem solving as a creative process that can take many forms within a given set of constraints.[60] To paraphrase the late Nobel laureate Richard Feynman: Good scientists always know at least three ways to solve the same problem. Contrary to traditional problem-solving activities in which one preferred solution is usually presented, many solutions are possible to any given problem. PBL activities enable students to appreciate that problem solving is not a uniform one-size-fits-all kind of activity and that many solution paths are possible.[60]

The assigned questions often required the same conceptual understanding but of different scenarios. During these sessions students were also taught to link subjects such as Physics and Mathematics, which they often place into isolated 'mental boxes'[5,32,35,72] without accessing the concepts taught in one module to relate to the other. This method also facilitated active learning[5,32,35,72] in the sense that the students had to discover and work with content that they determined to be necessary to solve the problem.[43,74] It must be noted that the full PBL approach was not implemented because of the cognitive demand that it would have placed on the students. Fade effect[73] scaffolding was implemented in this module of PBL to reduce the cognitive load of the students. In fade effect PBL, guidance is provided by the lecturer in the early stages and later, as the students gain expertise and become more confident, this guidance is gradually reduced.[77] The students are first introduced to simple problems and are then gradually given more complex problems in which elements are added to model real-life problems or situations.[77,78] This process helps students to slowly transit from studying examples to solving problems.[79] Another modification to the PBL approach was the combining of mini-lectures on some days with in-class, small group work on problem sets later in the week.

After they had finished their problem work the students assessed themselves and each other in order to develop skills in self-assessment and the constructive assessment of peers. Self-assessment is a skill essential to effective independent learning. The objectives of the PBL approach is to produce students who will[43,44,46,77,78:]

- Engage the problems they face in their life and career with initiative and enthusiasm
- Problem solve effectively using an integrated, flexible and usable knowledge base
- Employ effective self-directed learning skills to continue learning as a lifetime habit

- Continuously monitor and assess the adequacy of their knowledge, problem solving and self-directed learning skills
- Collaborate effectively as a member of a group.

The PBL method of transference of learning using the techniques described above ensured that mutual learning took place and that the knowledge was active.[35,37,44,46] Constant evaluation took place to effectively determine the level of transference. Evaluation in the Augmented Physics module included all mainstream evaluations such as tests, assignments and practical reports, as well as specific tasks for Augmented Physics. These assessments took the form of unseen quizzes, seen quizzes, mini tests and formal tests. All assessments used the first three cognitive levels of Bloom's taxonomy as previously discussed.[27,36]

The results of the practical sessions, quizzes and mini-tests that occurred in the month leading up to the first formal assessment are not considered here as they had a very narrow focus in terms of curriculum content in order to verify the transfer of initial learnings.[20,26,35,37,42] These frequent assessments also ensured that the students attempted the prescribed tutorials and homework and kept up to date with the material being taught. The first formal assessment test was structured similarly to the pre-test and formative Test One to enable direct comparisons.

Figure 1 shows the results of these tests. They show that the student performance in the Augmented Physics test was similar to that of the formative assessment before any of the interventions were implemented. The performance of the 2008 and 2009 cohorts of Augmented Physics students in the mainstream test was, however, much worse than their collective performance in either the augmented formal Test One or the first formative test while the results for the performance of the 2007 cohort in the mainstream Test One seems consistent with their performance in the augmented assessments. A qualitative interrogation of the student performance in these years revealed that in 2008 the lecturer in the mainstream introduced a new section a few days before the scheduled test was to be written with the indication to students that it would not be tested. However, this material was included within the application component of the mainstream formal test. Students were quick to associate this with their poor performance. However, considering that this question accounted for only 5 out of a possible 40 marks (12.5%) and their equally poor performance in the comprehension portion of the test, this may well have not been the reason, although it does warrant further investigation and possible intervention.

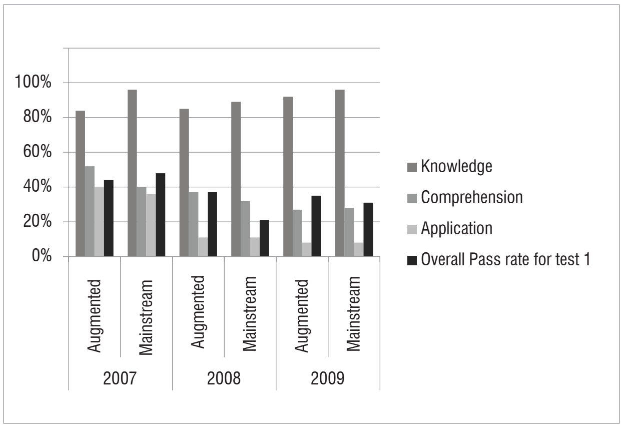


Figure 1: Formal Test One results.

The 2009 cohort investigation revealed that, for various reasons beyond the control of the students, most notably industrial action, the students had to write two mainstream tests on the same day. A Mathematics Level 1 test was written in the morning and the Physics test was written in the afternoon. Students in all three years cited a lack of time allocated to answer the questions as another major reason for their poor performance. They were not used to such stringent time demands as they often received some extra time during the augmented assessments. Considering the poor results in all three years, the researcher implemented a further two interventions and continued re-enforcing the ones introduced previously. The first new intervention was that all the students were interviewed and monitored to determine if any external factors, such as financial problems, poor living conditions at home, HIV/AIDS, alcoholism, drug addiction or teenage pregnancy, were contributing to their poor academic performance. If any such problems were identified, the student was referred to the appropriate centre for further assistance. Coupled with this intervention came the development of self-awareness in the student.[65] Part of this development occurred during the Life Skills component of the programme and was continued by the lecturer with constant pep talks, the screening of motivational video clips, and the invitation of people with similar backgrounds who were academic successes to give motivational talks.[65] Workshops on proper study techniques, both general and specific to the Physics module, were also held during the augmented practical sessions. During these workshops the students were taught how to draw up study timetables and how to plan for success. Proper time management during tests or exams was also discussed. To this end the 'mark a minute' philosophy was explained and implemented for all augmented assessments to keep in line with the mainstream. Students were at first alarmed by this idea but gradually came to accept it as something they could not change but would have to adopt. Students often asked for permission to use their cellular phones to monitor the time while attempting tutorial questions during the designated tutorial times, as some did not have watches. To further reinforce this notion all assigned questions were accompanied by a maximum mark and a clock was provided for tutorials or tests at which they could not use their cellular phones.

The Augmented Physics test was always scheduled a week before the mainstream test. It was designed to be formative and to help students evaluate their understanding of the material taught. The tests were pitched either at the same level or were harder than the mainstream tests. These tests were marked and returned to the students a few days before their mainstream tests and were discussed in detail along with the provision of model solutions. In general, the augmented test aimed to cover at least 85% of the content in the mainstream test. The Augmented Physics lecturer did not have access to the mainstream test beforehand, so as to not favour the Augmented Physics students. The students could use the test to establish areas of learning in which they had difficulty. They could then consult with the lecturer or student assistants for help in these and other areas before the mainstream test. It was hoped that these tests would provide critical feedback to the researcher regarding the success or failures of the implemented interventions.

The results of the final formal assessment for the augmented and mainstream module are shown in Table 6. These tests were heavily focused on the comprehension and application levels of Bloom's taxonomy, with the knowledge section almost entirely covered in the latter two levels. Thus the results were not broken down into three levels as before. A clear improvement was noted in all three cohorts. The glaring disparity in pass levels among cohorts from formal Test One was not evident in this comparison. In the 2008 and 2009 cohorts, the Augmented Physics students performed better in the mainstream test than they did in their augmented module assessments. In all cases their performance in the mainstream was well beyond that of the first formal assessment.

As a final means of comparison and to appraise critically the value of the interventions outlined above, student performance in the final exam is provided in Table 7. The students wrote a single theory exam and the Augmented Physics lecturer set approximately 25% of the exam in all three years under study. The mainstream marks provided in Table 7 are the results for the students registered in the mainstream programme and are provided for comparison of the Augmented Physics students' performance against mainstream norms. The mark breakdown showing the performance of the students at the various percentiles is given in Figure 2. It is evident from these results that the Augmented Physics students in the 2007 and 2008 cohorts seemed to perform on par with their mainstream counterparts. The 2009 cohort performed relatively poorly in comparison with the mainstream students. Figure 2 also clearly shows that not only did the Augmented Physics students pass, they passed well. Some of the students were able to achieve first-class passes (marks higher than 75%) and others obtained second-class passes (60–74%). Students who obtained a mark of between 40% and 49% were given the chance to write a second exam, for which they were required to pay. In all three years, no such student passed the second exam even though its level of difficulty was equivalent to that of the first exam, both in terms of content and time allocation.

Table 6: Pass rates for formal assessment Test Two for each cohort

	2007 (25 students)		2008 (19 students)		2009 (26 students)	
	Augmented	Mainstream	Augmented	Mainstream	Augmented	Mainstream
Overall pass rate	72%	76%	53%	78%	62%	65%

Table 7: Pass rates for the final exam for each cohort

	2007 (25 students)		2008 (19 students)		2009 (26 students)	
	Augmented	Mainstream	Augmented	Mainstream	Augmented	Mainstream
Overall pass rate	75%	64%	67%	83%	42%	73%

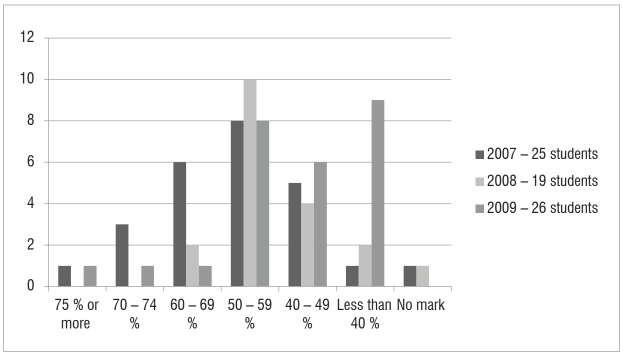


Figure 2: Final exam mark distribution.

Conclusion

The improvement (as indicated by Figure 3 in which an upward trend is clearly evidenced) in the test results for all the cohorts and the subsequent exam performance can be linked in part to the interventions set in place within the augmented module. The initial disparity between the results of the 'old' matric and the NSC cohorts, as evidenced by the pre-test and formative Test One results, is not so markedly evident in the subsequent formal Test Two results. Further, the exam performance of the 2007 and 2008 cohorts is in line with their mainstream counterparts. Taken at first impression, the exam results of the 2009 cohort are poorer than those of the mainstream students, which leaves one to conclude that the interventions may not have been enough to assist these students.

Further investigation into the performance of these students revealed that they had done relatively well in all sections taught by the researcher, achieving a pass rate of 63%. However, they performed poorly in sections taught by a replacement Augmented Physics lecturer when the researcher was away for a period during the latter half of the first semester in 2009. The incumbent had been trained by the researcher to implement all of the interventions in the style and method of instruction suitable for the Augmented Physics students; however, this training may not have translated into the classroom, which may have influenced the final exam pass rate. The cohort may have also experienced 'transitional hiccups' associated with having a new lecturer very close to their final exam. However, their performance in the tests may be used to appraise the effectiveness of the interventions. The improved performance of all the cohorts seems to indicate the effectiveness of the interventions in assisting the students to overcome their initial circumstances and lack of preparedness to perform as expected at tertiary level. We see that the interventions, such as the use of a dictionary (determined by interrogating the pass rates of the tests with qualitative analysis such as interviews with the students), English language policy and PBL, are effective tools to use during lectures, tutorials and practical sessions and greatly influence the problem-solving ability of the students, as evidenced by their higher results in the subsequent tests at the higher cognitive levels of Bloom's taxonomy.

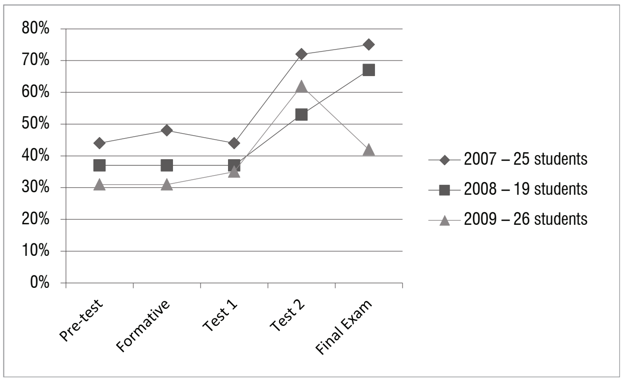


Figure 3: Semester progress in terms of pass rates for the Augmented Physics module.

A difference in the pass rates between the NSC and 'old' matric cohorts, although evident at the outset of the study, has not been clearly demonstrated by this work as this study is too narrow to judge whether the schooling system had any impact on the problem-solving ability of the students. A further study is necessary to make this distinction. These results do, however, show that, regardless of the initial school shortcomings faced by the students, a comprehensive programme such as the Augmented Physics one at UKZN does enable some students to achieve success and to move beyond their previous disadvantages.

The data suggest that in order to perform successfully in their academic endeavours, students need to be provided with more than just academic mentoring. To fully develop their cognitive skills it is imperative that the academic support provided be coupled with basic skills such as time management, language interpretation and life skills. Although not covered here, issues such as food insecurity, domestic violence, lack of accommodation and no funding are also factors that influence academic performance. However, much of these are catered to by the Life Skills module and thus no direct interventions are implemented in the Augmented Physics module itself.

The sound model of the CSA's Augmented programme contributes to improve the skills development of students within the realm of the natural and applied sciences. Clearly, the augmented programme and its approach to educational outcomes may be the answer to addressing the critical shortage of qualified Black scientists in South Africa, as outlined by government. The methods employed as described in this paper are not unique to physics or even to the natural sciences, and may be tailored to assist students to foster learning and encourage ownership of their learning.

This preliminary work will, in the near future, be extended to determine the effectiveness of the initial interventions for the students as they move into higher levels of study. Further work will also analyse possible models for interactive teaching that address the language barrier faced by many non-English mother-tongue students.

Author's note: In 2011, the Physics honours class comprised six students, four of whom were part of the 2007 cohort in the Augmented Physics module. All these students successfully passed their honours degrees and are now reading for MSc degrees at various institutions in South Africa. The four NATED (old) matric students from the 2008 cohort finished their degrees in the minimum specified time of 4 years in 2011 and two of these students have subsequently registered for honours degrees in Physics at UKZN while the other two have decided to pursue honours degrees in other fields.

Acknowledgement

I thank Mrs Sheelagh Halstead for the many hours of helpful discussions and guidance in preparing this manuscript.

References

1. Johnstone AH. Chemical education, facts, findings and consequences (Nyholm Lecture). Chem Soc Rev. 1980;9(3):365–380. http://dx.doi.org/10.1039/cs9800900365
2. Galili I, Lehavi Y. Definitions of physical concepts: A study of physics teachers' knowledge and views. Int J Sci Educ. 2006;28(5):521–541. http://dx.doi.org/10.1080/09500690500338847
3. Rollnick M, Rutherford M. The use of mother tongue and English in the learning and expression of science concepts: A classroom based study. Int J Sci Educ. 1996;18(1):91–104. http://dx.doi.org/10.1080/0950069960180108
4. Inglis M, Grayson DJ. An approach to the development of communication skills for science students: Some ideas from the Science Foundation Programme. In: Sharwood DW, editor. Proceedings of the 7th Conference of the South African Association for Academic Development; 1992 Dec 3–5; Port Elizabeth, South Africa. p. 192–200
5. Logan PF. Language and physics. Physics Educ. 1981;6:74–77. http://dx.doi.org/10.1088/0031-9120/16/2/303
6. Logan PF. Physics in paradise. Phys Teach. 1976;14:81–85. http://dx.doi.org/10.1119/1.2339315
7. Mji A, Makgato M. Factors associated with high school learners' poor performance: A spotlight on mathematics and physical science. S Afr J Educ. 2006;26(2):253–266.
8. Uys M, Van der Walt J, Van den Berg R, Botha S. S Afr J Educ. 2007;27(1):69–82.
9. University of KwaZulu-Natal (UKZN). Mission and vision statement of the University of KwaZulu-Natal as well as information for prospective students. Durban: UKZN; 2013. Available from: http://www.ukzn.ac.za
10. Rollnick M. Identifying potential for equitable access to tertiary level science: Digging for gold. Dordrecht: Springer; 2010. http://dx.doi.org/10.1007/978-90-481-3224-9
11. University of KwaZulu-Natal (UKZN). Constitution of the Centre for Science Access (CSA), University of KwaZulu-Natal. Durban: UKZN; 2013. Available from: http://csa.ukzn.ac.za
12. Kioko J. Foundation provision in South African higher education: A social justice perspective. In: Hutchings C, Garraway J, editors. Proceedings of the Rhodes Seminar; 2009; Grahamstown, South Africa.. Grahamstown Rhodes University Press; 2010. p. 40–49.
13. Hutchings C, Garraway J. Beyond the university gates: Provision of extended curriculum programmes in South Africa. Grahamstown Rhodes University Press; 2010.
14. Grayson DJ. A holistic approach to preparing disadvantaged students to succeed in tertiary science studies. Part I: Design of the Science Foundation Programme. Int J Sci Educ. 1996;18(8):993–1013. http://dx.doi.org/10.1080/0950069970190108
15. Grayson DJ. A holistic approach to preparing disadvantaged students to succeed in tertiary science studies. Part II: Outcomes of the Science Foundation Programme. Int J Sci Educ.1997;19(1):107–123. http://dx.doi.org/10.1080/0950069970190108
16. Barnsley S. Thoughts on the psychological processes underlying difficulties commonly experienced by African science students at university. In: Sharwood DW, editor. Proceedings of the 7th Conference of the South African Association for Academic Development; 1992 Dec 3–5; Port Elizabeth, South Africa.
17. Leigh G. Developing multi-representational problem solving skills in large, mixed-ability physics classes [unpublished thesis]. Cape Town: UCT; 2004.
18. Leigh G, Buffler A. Benchmarking the numeracy and expectations of first year physics technikon students as part of a new research-based teaching intervention. In: Buffler A, Laugksch RC, editors. 12th annual meeting of the Southern African Association for Research in Mathematics, Science and Technology Education; 2004 Jan 13–17; Durban, South Africa. Durban: Taylor and Francis; 2004. p. 528–535.
19. Soong Y. Handheld educational applications: A review of the research. In: Ryu H, Parsons D, editors. Innovative mobile learning: Techniques and technologies information. London: Hershey Science Reference (an imprint of IGI Global); 2009. p. 302–323.
20. Walsh L, Howard R, Bowe B. An investigation of introductory Physics students' approaches to problem solving. Level 3. 2007;5:1.
21. Young HD, Freedman RA. University physics. 12th ed. New Jersey: Addison-Wesley; 2011.
22. Halliday D, Resnick R, Walker J. Fundamentals of physics. 9th ed. New Jersey: Wiley; 2011.
23. University of KwaZulu-Natal (UKZN). BSc-4 orientation handbook. Durban: UKZN; 2013. Available from: http://csa.ukzn.ac.za
24. Salomon G, Perkins DN. Rocky roads to transfer: Rethinking mechanisms of a neglected phenomenon. Educ Psyc. 2007;24(2):113–142. http://dx.doi.org/10.1207/s15326985ep2402_1
25. Bassok M, Holyoak KJ. Pragmatic knowledge and conceptual structure: Determinants of transfer between quantitative domains. In: Detterman DK, Sternberg RJ, editors. Transfer on trial: Intelligence, cognition and construction. New York: Ablex; 1993. p. 68–98.
26. National Curriculum Statement: Physical sciences. Pretoria: Department of Education; 2003.
27. Bloom BS. Taxonomy of educational objectives, handbook I: The cognitive domain. New York: David McKay; 1956.
28. Grussendorf SJ. Umalusi maintaining standards report. Pretoria: Umalusi; 2009.
29. Edwards N. An analysis of the alignment of the Grade 12 Physical Sciences examination and the core curriculum in South Africa. S Afr J Educ. 2010;30(4):571–590.
30. Davison DM, Miller KW, Metheny DL. What does integration of science and mathematics really mean? School Science and Mathematics. 1995;95(5):226–230. http://dx.doi.org/10.1111/j.1949-8594.1995.tb15771.x
31. Lincoln YS, Guba EG. Naturalistic inquiry. Newbury Park: Sage; 1985.
32. Hillel J. Physics education research – a comprehensive study [unpublished thesis]. Toronto: University of Toronto; 2005.
33. Maxwell JA. Qualitative research design: An interactive approach. Thousand Oaks: Sage; 2005.
34. Redish EF. A theoretical framework for physics education research: Modelling student thinking. In: Redish EF, Vicentini M, editors. Proceedings of the Enrico Fermi Summer School in Physics; 2004 Jan; Varenna, Italy. p. 1–63.
35. White R, Gunstone R, Elterman E, Macdonald I, McKittric B, Mills D. et al. Students' perceptions of teaching and learning in first-year university physics. Research in Science Education. 1995;25(4):465. http://dx.doi.org/10.1007/BF02357388
36. Anderson LW, Krathwohl DR. A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives: Complete edition. New York: Longman; 2001.
37. Dave RH. Psychomotor levels. In: Armstrong RJ, editor. Developing and writing behavioural objectives. Tucson, Arizona: Educational Innovators Press; 1975.
38. Considine J, Botti M, Thomas S. Design, format, validity and reliability of multiple choice questions for use in nursing research and education. J Royal College. 2005;12(1):19–24
39. Higgins E, Tatham L. Are your students cheating or guessing on tests? Consider implementing alternate multiple-test formats such as DOMC and NRET. Learning and Teaching in Action. 2003;2(1):1–12.

40. Sanderson PJ. Multiple-choice questions: A linguistic investigation of difficulty for first-language and second-language students [PhD thesis]. Pretoria: Unisa; 2010.

41. Bradbury J, Miller R. A failure by any other name: The phenomenon of underpreparedness. S Afr J Sci. 2011;107(3–4), Art. #294, 8 pages. http://dx.doi.org/10.4102/sajs.v107i3/4.294

42. Harrow A, A taxonomy of psychomotor domain: A guide for developing behavioral objectives. New York: David McKay; 1972.

43. Hmelo-Silver CE, Duncan RG, Chinn CA. Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). Educ Psyc. 2007;42(2):99. http://dx.doi.org/10.1080/00461520701263368

44. Narayanan S, Munirathnam AM. Application of Bloom's taxonomy of educational objectives as a problem solving tool in the teaching-learning process in an 'electrical engineering technology' course. Linguistics, Culture and Education. 2012;1(1):117–140.

45. Yadav A. Problem-based learning: Influence on students' learning in an electrical engineering course. J Eng Educ. 2011;100(2):253–280. http://dx.doi.org/10.1002/j.2168-9830.2011.tb00013.x

46. Serrat O. Learning in teams – Facilitator's guide. Asian Development Bank Report. 2009;1:81–83.

47. Fogler S, LeBlanc SE. Strategies for creative problem solving. New Jersey: Prentice-Hall; 1995.

48. Mokhaba MB. Outcomes based education in SA since 1994 [Phd thesis]. Pretoria: University of Pretoria; 2004.

49. Malan SPT. The 'new paradigm' of outcomes-based education in perspective. Tydskrif vir Gesinsekologie en Verbruikerswetenskappe. 2000;28:22–28.

50. Spady WG. Outcomes based education: Critical issues and answers. Arlington: American Association of School Administrators; 1994.

51. Jansen JD. Understanding social transition through the lens of curriculum policy: Namibia/South Africa. J Curriculum Stud. 1995;27:245–261. http://dx.doi.org/10.1080/0305764980280305

52. Jansen JD. Essential alterations? A critical analysis of the states syllabus revision process. Perspect Educ 1997;17(2):1–11.

53. Jansen JD. Curriculum reform in South Africa: A critical analysis of outcomes-based education. Cambridge J. Educ. 1998;28(3):321–331. http://dx.doi.org/10.1080/0305764980280305

54. Mahlangu D. Outcomes-based education to be scrapped. TimesLIVE online edition. 2010 July 04.

55. Clerk DPP, Rutherford M. Language as a confounding variable in the diagnosis of misconceptions. Int J Sci Educ. 1998;22(7):707–717.

56. Clerk DPP. Language as a confounding variable [MSc thesis]. Johannesburg: Wits University; 1998.

57. Bharuthram S, M^{c}enna S. Students' navigation of the uncharted territories of academic writing. Afr Educ Rev. 2012;9(3):581–594. http://dx.doi.org/10.1080/18146627.2012.742651

58. Bharuthram S. Developing reading strategies in higher education through the use of integrated reading/writing activities: A study at a university of technology in South Africa [PhD thesis]. Durban: University of KwaZulu-Natal; 2007.

59. Bharuthram S. Making a case for the teaching of reading across the curriculum in higher education. S Afr J Educ. 2012;32(2):205–214.

60. Lasry N. Problem-based learning for college physics [homepage on the Internet]. c2013 [cited 2013 Mar 22]. Available from: http://rea.ccdmd.qc.ca/en/pbl/resultat.asp?action=aboutPBL&endroitRetour=9&he=600

61. Levinson SC, Wilkins DP. Grammars of space: Explorations in cognitive diversity. New York: Cambridge University Press; 2006. http://dx.doi.org/10.1017/CBO9780511486753

62. Boroditsky L. Does language shape thought? English and Mandarin speakers' conceptions of time. Cognitive Psychol. 2001;43(1):1–22. http://dx.doi.org/10.1006/cogp.2001.0748

63. Tversky B, Kugelmass B, Winter A. Crosscultural and developmental trends in graphic productions. Cognitive Psychol. 1991;23:515–557. http://dx.doi.org/10.1016/0010-0285(91)90005-9

64. So H-J, Kim B. Learning about problem-based learning: Student teachers integrating technology, pedagogy and content knowledge. Australas J Educ Technol. 2009;25(1):101–116.

65. Inglis M. Proceedings of the First Annual Meeting of the South African Association for Research in Mathematics and Science Education; 1993; Grahamstown, South Africa.

66. Ferrer V. The mother tongue in the classroom: Cross-linguistic comparisons, noticing and explicit knowledge. Teaching English Worldwide. 2002;10:1–7.

67. Khati AR. When and why of mother tongue use in English classrooms. NELTA. 2011;16(12):42–51.

68. Chetty N. Student responses to being taught physics in isiZulu. S Afr J Sci. 2013;109(9/10), Art. #2012-0016, 6 pages. http://dx.doi.org/10.1590/sajs.2013/20120016.

69. Ngidi SA. The attitudes of learners, educators and parents towards English as a language of teaching and learning [MSc thesis]. KwaDlangezwa: University of Zululand; 2007.

70. Dempster E, Zuma S. Reasoning used by isiZulu-speaking children when answering science questions in English. J Educ. 2010;50:35–58.

71. Casasanto D, Boroditsky L, Phillips W, Greene J, Goswami S, Bocanegra-Thiel I. How deep are the effects of language on thought? In: Forbus K, Gentner D, Regier T, editors. Proceedings of the 26th Annual Conference of the Cognitive Science Society; 2004 Aug 5–7; Chicago, USA. Hillsdale, NJ: Lawrence Erlbaum and Associates; 2004. p. 575–580.

72. Armstrong E. A hybrid model of problem-based learning. In: Boud D, Feletti G, editors. The challenge of problem-based learning. London: Kogan; 1991. p. 137.

73. Atkinson RK, Renkl A, Merrill MM. Transitioning from studying examples to solving problems: Effects of self-explanation prompts and fading worked-out steps. J Educ Psyc. 2003;95:774–783. http://dx.doi.org/10.1037/0022-0663.95.4.774

74. Hurren H, Klegeris A. Impact of the problem based learning in a large classroom setting: Student perception and problem-solving skills. Adv Physiol Educ. 2011;35:408–415. http://dx.doi.org/10.1152/advan.00046.2011

75. Felder R. How about a quick one. Chem Eng Educ. 1992;26(1):18–19.

76. Davis MH. Outcome-based education. J Vet Med Educ. 2003;30(3):227–232. http://dx.doi.org/10.3138/jvme.30.3.258

77. Merrill MD. A pebble-in-the-pond model for instructional design. Performance Improvement. 2002;41(7):39–44. http://dx.doi.org/10.1002/pfi.4140410709

78. Sweller J. Instructional implications of David C. Geary's evolutionary educational psychology. Educ Psychol. 2008;43:214–216.

79. Sweller J. Cognitive load during problem solving: Effects on learning. Cognitive Sci. 1988;12(2):257–285. http://dx.doi.org/10.1080/00461520802392208

The impact of accent identification errors on speech recognition of South African English

AUTHORS:
Herman Kamper[1]
Thomas R. Niesler[1]

AFFILIATIONS:
[1]Department of Electrical & Electronic Engineering, Stellenbosch University, Stellenbosch, South Africa

CORRESPONDENCE TO:
Thomas Niesler

EMAIL:
trn@sun.ac.za

POSTAL ADDRESS:
Department of Electrical
& Electronic Engineering,
Stellenbosch University, Private
Bag X1, Matieland 7602,
South Africa

DATES:
Received: 26 Oct. 2012
Revised: 06 Sep. 2013
Accepted: 09 Oct. 2013

KEYWORDS:
parallel recognition; acoustic modelling; human language technology

HOW TO CITE:
Kamper H, Niesler TR. The impact of accent identification errors on speech recognition of South African English. S Afr J Sci. 2014;110(1/2), Art. #2012-0049, 6 pages. http://dx.doi.org/10.1590/sajs.2014/20120049

For successful deployment, a South African English speech recognition system must be capable of processing the prevalent accents in this variety of English. Previous work dealing with the different accents of South African English has considered the case in which the accent of the input speech is known. Here we focus on the practical scenario in which the accent of the input speech is unknown and accent identification must occur at recognition time. By means of a set of contrastive experiments, we determine the effect which errors in the identification of the accent have on speech recognition performance. We focus on the specific configuration in which a set of accent-specific speech recognisers operate in parallel, thereby delivering both a recognition hypothesis as well as an identified accent in a single step. We find that, despite their considerable number, the accent identification errors do not lead to degraded speech recognition performance. We conclude that, for our South African English data, there is no benefit of including a more complex explicit accent identification component in the overall speech recognition system.

Introduction

Although English is used throughout South Africa, it is spoken as a first language by just 9.6% of the population.[1] Accented English is therefore highly prevalent and speech recognition systems must be robust to these different accents before successful speech-based services can become accessible to the wider population.

One solution to accent-robust automatic speech recognition is to develop acoustic models that are accent independent. For accent-independent models, no distinction is made between different accents, and training data are pooled. A different approach is to develop several separate acoustic model sets that are each designed to deliver optimal performance for a particular accent and then to combine these within a single system. In this latter case, accent identification (AID) must occur in order for the correct set of acoustic models to be chosen at recognition time. However, state-of-the-art AID is complex and adds a significant additional hurdle to the development of a speech recognition system.[2] In this study we investigated the impact that such AID errors have on the accuracy of speech recognition for the five acknowledged accents of South African English (SAE).To do this, we compared three established acoustic modelling approaches by means of a set of contrastive speech recognition experiments in which the accent of the SAE input speech is assumed to be unknown. As a baseline, we also considered the performance that was achieved when the input accent is known. Our results provide some insight into the importance of AID accuracy within the context of an SAE speech recognition system.

Accents of English in South Africa

Five major accents of SAE are identified in the literature: Afrikaans English, Black South African English, Cape Flats English, White South African English and Indian South African English.[3] The term 'South African English' is used to refer collectively to all the accents of English spoken in the country.

English itself was originally brought to South Africa by British occupying forces at the end of the 18th century. The accent known as White South African English (EE) refers to the first-language English spoken by White South Africans who are chiefly of British descent. This accent is used by approximately 3.1% of the population.[1] Afrikaans English (AE) refers to the accent used by White South African second-language English speakers of Afrikaans descent. Afrikaans has its origins in 17th century Dutch, which was brought to South Africa by settlers from the Netherlands. Although the Afrikaans vocabulary has a predominantly Dutch origin, it was influenced by Malay, Portuguese and the Bantu and Khoisan languages. White Afrikaans speakers comprise approximately 5.3% of the South African population.[1] Cape Flats English (CE) has its roots in the 19th century working-class residential areas in inner-city Cape Town where residents from many different ethnic affiliations, religions and languages came into regular contact with one another. The accent spread as residents from these mixed neighbourhoods moved or were forced to move to the Cape Flats in the 1960s and 1970s.[4] Today CE is most closely associated with the 'Coloured' community of South Africa, which comprises approximately 9.1% of the total population.[1] The connection between EE, AE and CE, all three of which have been strongly influenced by Afrikaans, has been emphasised in the literature.[3] Black South African English (BE) refers to the variety of English spoken by Black South Africans whose first language is an indigenous African language. BE speakers are overwhelmingly first-language speakers of one of the nine official indigenous African languages of South Africa and comprise approximately 74.9% of the population.[1] Indian languages were brought to South Africa by labourers who were recruited from India after the abolition of slavery in European colonies in the 19th century. These Indian languages have existed in South Africa since 1860, mainly in KwaZulu-Natal. Today Indian South African English (IE) is spoken as a first language by most of the Indian South African population. Approximately 2.5% of the South African population are considered Indian or Asian and 86.1% speak English as a first language.[1]

To obtain some initial intuition regarding the relative similarity of these five accents, we determined how similar the statistical distributions describing the acoustics of corresponding sounds in each accent were to one another. We achieved this by applying the Bhattacharyya distance, which allows a measure of similarity between two probability density functions to be computed.[5] Three-state single-mixture monophone hidden Markov models (HMMs) were obtained using the acoustic data for each accent separately. For each accent pair, the Bhattacharyya distance was subsequently computed between corresponding states of the two HMMs. The average over the three resulting distances was then determined to obtain a measure of between-accent similarity for a particular monophone. Finally, the average distance between all corresponding pairs of monophones was calculated to obtain a measure of inter-accent similarity.[6] For the five accents of SAE, an approximate representation of these distances is shown in Figure 1, where particle swarm optimisation[7] was used to find an approximate projection of the distances into two dimensions. In the figure, similarity is indicated by a geometrically shorter distance between accents. From this first analysis we conclude that AE, CE and EE are quite similar, while BE and IE are more dissimilar from the other accents and from each other.

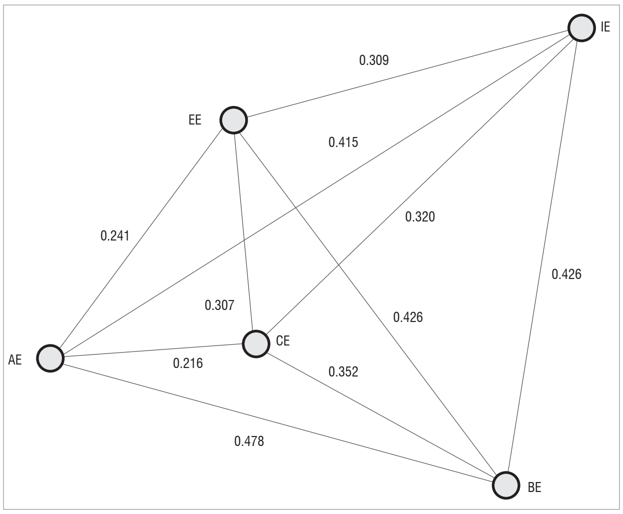


Figure 1: Graphical depiction of the average Bhattacharyya distances between the five accents of South African English: White South African English (EE), Indian South African English (IE), Black South African English (BE), Afrikaans English (AE) and Cape Flats English (CE).

Related research

Several studies have considered acoustic modelling for different accents of the same language. Approaches include the pooling of data across accents, leading to a single accent-independent acoustic model[8]; the isolation of data for each accent, leading to individual accent-specific acoustic models[9]; and adaptation techniques in which models trained on one accent are adapted using data from another[10,11]. Recently, selective data sharing across accents through the use of appropriate decision-tree state clustering algorithms has also received some attention.[6,12] These studies extend the multilingual acoustic modelling approach first proposed by Schultz and Waibel[13] to apply to multiple accents of the same language.

Most of the above studies consider the scenario in which the accent of the incoming speech is known and each utterance is presented only to the matching set of acoustic models. This approach is appropriate when the aim is to evaluate different acoustic modelling strategies without allowing performance to be influenced by the effects of accent misclassification. Because the accent is assumed to be known, we will refer to this approach as oracle AID. However, in many practical situations, the accent of the incoming speech would not be known. In such cases a single system should be able to process multiple accents.

Three approaches for the recognition of multiple accents are commonly found in the literature. One approach is to precede accent-specific speech recognition with an explicit AID step.[14] A second approach is to run a bank of accent-specific recognisers in parallel and select the output with the highest associated likelihood.[15] In this set-up, AID is performed implicitly during recognition. A third approach is to train a single accent-independent model set by pooling data across accents and thereby avoid AID altogether.[15]

These three approaches have been applied in various ways to different English accents. The recognition of non-native English from six European countries was considered by Teixeira et al.[16] They found that AID followed by recognition gave comparable performance to an oracle configuration, but both were outperformed by an accent-independent system. Chengalvarayan[15] compared the parallel and accent-independent approaches for recognition of American, Australian and British English and found that the accent-independent approach gave the best performance. Variations of the three approaches have also been considered. For example, Beattie et al.[17] proposed a parallel recognition approach in which the accent-specific recogniser is selected based on a history of scores for a specific speaker instead of only the score for the current utterance. This method proved to be superior to accent-independent modelling for three dialects of American English. We also compared the oracle, parallel and accent-independent strategies in our experiments to determine what could be learnt for the case of the accents of SAE.

Speech resources

Training and test data

Our experiments were based on the African Speech Technology (AST) databases.[18] The databases consist of telephone speech recorded over fixed and mobile telephone networks and contain a mix of read and spontaneous speech. As part of the AST project, five English accented speech databases were compiled corresponding to the five accents of SAE. These databases were transcribed both phonetically, using a common international phonetic alphabet (IPA)-based phone set consisting of 50 phones, as well as orthographically. The assignment of a speaker's English accent was guided by the speaker's first language and ethnicity.

Each of the five databases was divided into training, development and evaluation sets. As indicated in Tables 1 and 2, the training sets each contain between 5.5 h and 7 h of speech from approximately 250 speakers, while the evaluation sets contain approximately 25 min from 20 speakers for each accent. The development sets were used only for the optimisation of the recognition parameters before final testing on the evaluation data. For the development and evaluation sets, the ratio of male to female speakers is approximately equal and all sets contain utterances from both landline and mobile telephones. There is no speaker overlap between any of the sets. The average length of a test utterance is approximately 2 s.

Table 1: Training sets for each South African English accent

Accent	Number of hours of speech	Number of utterances	Number of speakers	Word tokens
AE	7.02	11 344	276	52 540
BE	5.45	7779	193	37 807
CE	6.15	10 004	231	46 185
EE	5.95	9879	245	47 279
IE	7.21	15 073	295	57 253
Total	31.78	54 078	1240	241 064

AE, Afrikaans English; BE, Black South African English; CE, Cape Flats English; EE, White South African English; IE, Indian South African English.

Table 2: Evaluation sets for each South African English accent

Accent	Speech (min)	Number of utterances	Number of speakers	Word tokens
AE	24.16	689	21	2913
BE	25.77	745	20	3100
CE	23.83	709	20	3073
EE	23.96	702	18	3059
IE	25.41	865	20	3362
Total	123.13	3710	99	15 507

AE, Afrikaans English; BE, Black South African English; CE, Cape Flats English; EE, White South African English; IE, Indian South African English.

Language models and pronunciation dictionaries

Using the SRI language modelling (SRILM) toolkit,[19] an accent-independent backoff[20] bigram language model was trained on the combined training set transcriptions of all five accents. Absolute discounting was used for the estimation of language model probabilities.[21] As part of the AST project, a separate pronunciation dictionary was obtained for each accent individually. These individual contributions were combined into a single pronunciation dictionary for the experiments presented here. These design decisions were made based on preliminary experiments which indicated that accent-independent pronunciation and language modelling outperformed the accent-specific alternatives. Language model perplexities and out-of-vocabulary rates are shown in Table 3.

Table 3: Bigram language model perplexities and out-of-vocabulary (OOV) rates measured on the evaluation sets

Accent	Bigram types	Perplexity	OOV (%)
AE	11 580	24.07	1.82
BE	9639	27.87	2.84
CE	10 641	27.45	1.40
EE	10 451	24.90	1.08
IE	11 677	25.55	1.73

AE, Afrikaans English; BE, Black South African English; CE, Cape Flats English; EE, White South African English; IE, Indian South African English.

Experimental methodology

General set-up

Speech recognition systems were developed using the HTK tools.[22] Speech audio data were parameterised as 13 Mel-frequency cepstral coefficients with their first- and second-order derivatives to obtain 39-dimensional observation vectors. Cepstral mean normalisation was applied on a per-utterance basis. The parametrised training sets were used to obtain three-state left-to-right single-mixture monophone HMMs with diagonal covariance matrices using embedded Baum-Welch re-estimation. These monophone models were then cloned and re-estimated to obtain initial cross-word triphone models which were subsequently subjected to decision-tree state clustering. Clustering was followed by five iterations of re-estimation. Finally, the number of Gaussian mixtures per state was gradually increased, with each increase being followed by five iterations of re-estimation. This yielded diagonal-covariance cross-word tied-state triphone HMMs with three states per model and eight Gaussian mixtures per state.

Acoustic modelling

When performing speech recognition of multiple accents by running separate recognisers in parallel, different acoustic modelling approaches can be followed. We considered two modelling approaches. Comparable acoustic modelling approaches have been previously considered in multilingual[13,23] as well as multi-accent[6,12,24] settings. The two approaches we used – accent-specific acoustic modelling and multi-accent acoustic modelling – are distinguished by different methods of decision-tree state clustering.

In accent-specific acoustic modelling, separate accent-specific acoustic models are trained and no sharing of data occurs between accents. Separate decision-trees are grown for each accent and the clustering process employs only questions relating to phonetic context. Chengalvarayan[15] has applied such models in a parallel configuration. In multi-accent acoustic modelling, decision-tree questions relate not only to the phonetic context, but also to the accent of the basephone. Tying across accents can thus occur when triphone states are similar, while the same triphone state from different accents can be modelled separately when there are differences. Detailed descriptions of this approach can be found in the existing literature.[6,23]

As a further benchmark we considered accent-independent acoustic modelling in which a single model set is obtained by pooling data across all accents for phones with the same IPA classification and the need for AID is side-stepped. The decision-tree clustering process employs only questions relating to phonetic context. Such pooled models are often employed for the recognition of accented speech[12,15,16] and therefore represent an important baseline.

System configuration and evaluation

Five recognisers, one tailored to each SAE accent, were configured in parallel and the output with the highest associated likelihood was selected as the final speech recognition hypothesis. This configuration is illustrated in Figure 2. The selection of the highest scoring result can be performed independently for each utterance, leading to per-utterance AID, or for each speaker, leading to per-speaker AID. The choice between these two AID schemes will depend on practical constraints and we report on the performance of both. Accent-specific and multi-accent acoustic models were employed in competing parallel systems and these were compared with accent-independent acoustic models. As a further benchmark we compared the performance of the parallel systems to that of systems in which each test utterance was presented only to the recogniser with matching accent (oracle AID). This configuration is illustrated in Figure 3. In each case the oracle configuration used the same acoustic models as the parallel configuration it was compared to. In this way the penalty as a result of AID errors occurring implicitly during parallel recognition can be analysed by direct comparison with the oracle configuration.

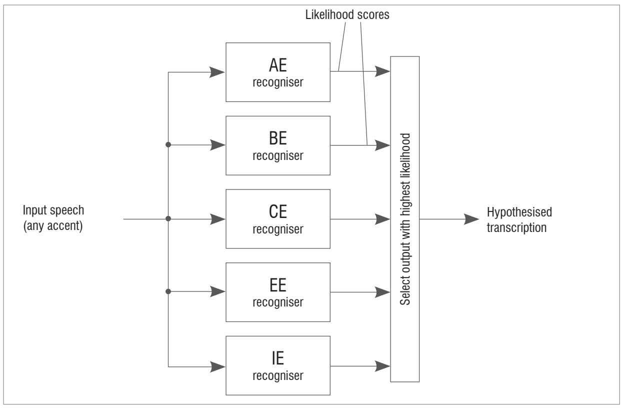

Likelihood scores
AE recogniser
BE recogniser
Input speech (any accent)
CE recogniser
EE recogniser
IE recogniser
Select output with highest likelihood
Hypothesised transcription

Figure 2: Parallel configuration in which multiple accent-specific recognisers are placed in parallel for simultaneous speech recognition in the five South African English accents: White South African English (EE), Indian South African English (IE), Black South African English (BE), Afrikaans English (AE) and Cape Flats English (CE).

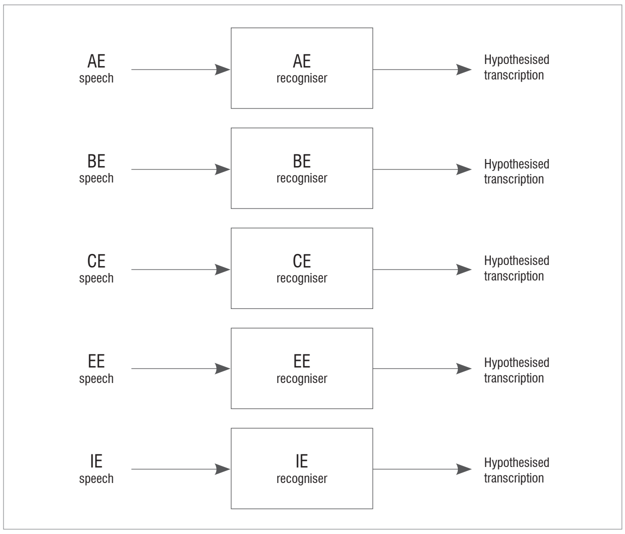

AE speech
AE recogniser
Hypothesised transcription
BE speech
BE recogniser
Hypothesised transcription
CE speech
CE recogniser
Hypothesised transcription
EE speech
EE recogniser
Hypothesised transcription
IE speech
IE recogniser
Hypothesised transcription

Figure 3: Oracle configuration in which each test utterance is presented only to the accent-specific recognition system with matching accent for speech recognition in the five South African English accents: White South African English (EE), Indian South African English (IE), Black South African English (BE), Afrikaans English (AE) and Cape Flats English (CE).

Experimental results

Table 4 shows the average word recognition and per-utterance AID accuracies measured on the evaluation sets. Implicit per-utterance AID was performed by the parallel accent-specific and multi-accent systems. Because a single recogniser was used for the accent-independent system, AID did not occur and identical results were obtained for the oracle and parallel configurations.

Table 4: Recognition performance of the oracle and parallel configurations when applying per-utterance accent identification (AID). Average word recognition accuracies (%) are given for oracle and parallel configurations and per-utterance accuracies (%) are given for AID.

Model set	Oracle	Parallel	AID
Accent-specific	81.53	81.31	67.60
Accent-independent	81.67	81.67	–
Multi-accent	82.78	82.85	65.39

The results in Table 4 indicate that the parallel configuration employing accent-specific acoustic models (with an accuracy of 81.31%) was outperformed by its corresponding oracle configuration (with an accuracy of 81.53%). In contrast, the parallel configuration employing multi-accent acoustic models (82.85%) showed a small improvement over its oracle counterpart (82.78%). This improvement was despite the fact that the accent was misclassified in 34.6% of test cases. Although the improvement in recognition performance was small and only significant at the 66% level,[25] it is noteworthy that the accent misclassifications did not lead to deteriorated accuracy. This observation indicates that some test utterances may have been better matched to the acoustic models of another accent. The results in Table 4 also confirm earlier reports in the literature of the better performance achieved by multi-accent acoustic models in relation to their accent-specific and accent-independent counterparts.[6] It is also apparent that, despite their better speech recognition performance, the multi-accent models did not lead to improved AID.

The performance of parallel configurations applying per-speaker AID is shown in Table 5. The performance of the oracle configurations and the accent-independent model set were unchanged from Table 4. A comparison between the two tables indicates that performing AID on a per-speaker basis improved speech recognition accuracy (from 81.31% to 81.69%) for the parallel systems using accent-specific acoustic models, while recognition accuracy (82.85%) was unchanged for the multi-accent systems. In both cases, AID accuracies were substantially improved to approximately 90%. Among the acoustic modelling options, the multi-accent approach continued to deliver the best performance.

Table 5: Recognition performance of the oracle and parallel configurations when applying per-speaker accent identification (AID). Average word recognition accuracies (%) are given for oracle and parallel configurations and per-utterance accuracies (%) are given for AID.

Model set	Oracle	Parallel	AID
Accent-specific	81.53	81.69	89.84
Accent-independent	81.67	81.67	–
Multi-accent	82.78	82.85	89.81

The results presented here lead us to the surprising conclusion that better AID does not necessarily lead to higher speech recognition accuracy for the accents of SAE in our databases. Despite a sizeable proportion of accent misclassifications (34.6%), the parallel per-utterance AID system employing multi-accent acoustic models showed no deterioration in accuracy compared to an oracle configuration (in which no accent misclassifications occur). When parallel recognition was performed at a per-speaker level, the proportion of AID errors reduced substantially from 34.6% to 10.2%, but the performance of the parallel multi-accent system remained unchanged.

A reason for this weak dependency of speech recognition accuracy on AID performance could be the inherent difficulty of classifying accents. Accent labels were assigned on the basis of the speaker’s first language and ethnicity, which may not be a reliable way to determine ‘true’ accent. Recent research has shown, for example, that Black speakers with an indigenous African language as a first language exhibit an increasing tendency of adopting the accent normally attributed to White South African speakers of English.[26] Furthermore, even when the accent label is uncontested, there may be considerable variation.[27,28] Hence the accent labels may in some cases not be ideal from the point of view of acoustic homogeneity.

Analysis of accent misclassifications

An accent misclassification occurs when the accent of the recogniser selected for a particular utterance during parallel recognition is different from the accent with which that utterance is labelled. We analysed these errors for the case of per-utterance AID. For each misclassified utterance, the recognition hypothesis produced by the oracle and the parallel configurations were obtained. Both these hypotheses were subsequently aligned with the reference transcription. By comparing the two alignments we determined whether the misclassification resulted in an improvement or in a deterioration in performance relative to that of the oracle configuration. The resulting effect on recognition performance of using the parallel configuration instead of the oracle configuration was also calculated.

Tables 6 and 7 present the results of this analysis for the accent-specific and multi-accent acoustic model sets, respectively. We see that, for both acoustic modelling approaches, the majority (approximately 75%) of misclassified utterances did not influence speech recognition performance. While misclassified utterances were shorter than the average evaluation set utterance (~1.7 s compared with ~2 s), those which led to improved performance were longer (~2.7 s). Misclassified utterances leading to deteriorated accuracy were of approximately average length (~2.1 s) while utterances having no effect were shorter (~1.4 s). These observations apply to both acoustic modelling approaches.

Table 6: Analysis of accent misclassifications for the parallel per-utterance accent identification system using accent-specific acoustic models

Impact of misclassification accuracy	Number of utterances	Number of tokens	Average duration (s)	Δ accuracy (%)
No effect	905	2721	1.42	0
Improvement	135	802	2.66	+1.29
Deterioration	162	773	2.22	−1.52
Total/Average†	1202	4296	1.67†	−0.23

Table 7: Analysis of accent misclassifications for the parallel per-utterance accent identification system using multi-accent acoustic models

Impact of misclassification accuracy	Number of utterances	Number of tokens	Average duration (s)	Δ accuracy (%)
No effect	1040	3213	1.46	0
Improvement	120	705	2.66	+1.19
Deterioration	124	559	2.05	−1.12
Total/Average†	1284	4477	1.63†	+0.07

Focusing first on the accent-specific systems, Table 4 indicates that the parallel configuration (81.31% accuracy) was slightly outperformed by its oracle counterpart (81.53%). Table 6 reveals that a larger number of misclassifications led to deterioration than to improvement. However, the misclassified utterances resulting in improvements were on average longer and hence the number of tokens involved in improved and deteriorated performance, respectively, was approximately equal. The effect of misclassifications leading to improvements (+1.29%) was outweighed by those leading to deterioration (−1.52%), which ultimately resulted in the 0.23% absolute drop in performance.

For the multi-accent systems, Table 4 indicates that the parallel configuration (82.85%) yielded a slight improvement over its oracle counterpart (82.78%). Table 7 indicates that, although approximately the same number of misclassifications led to deteriorated and to improved performance, the number of tokens involved in improved performance was greater. As a result, the improvement as a result of misclassifications (+1.19%) was slightly larger than the deterioration (−1.12%), leading to the small overall improvement of 0.07%.

Table 8 presents the AID confusion matrix for the parallel multi-accent system employing per-utterance AID. The table indicates that confusions are most common between AE and CE, between AE and EE, and between CE and IE. Interestingly, such closeness between the CE and IE accents has recently also been established in an independent linguistic study.[29] The diagonal of Table 8 indicates that CE, EE and AE utterances are most prone to misclassification, while IE and BE are identified correctly more often. This analysis agrees with the pattern that was depicted in Figure 1, which highlights the similarity of AE, CE and EE, and the more distinct nature of BE and IE. The AID confusion matrix for the parallel per-utterance AID system using accent-specific models indicates similar trends.

Summary and conclusions

We investigated the effect of accent misclassifications on recognition accuracy when performing parallel speech recognition of the five accents of SAE. In order to isolate the effect of AID errors, the speech recognition performance of systems employing recognisers in parallel was compared with the performance of an oracle configuration in which each test utterance is presented only to the recogniser of the matching accent. Parallel configurations were also compared with accent-independent recognition achieved by pooling training data.

Our experimental results show that a parallel configuration applying AID at a per-utterance level and employing multi-accent acoustic models, which allow selective data sharing across accents, exhibited no degradation in accuracy compared to an oracle configuration despite a considerable number of AID errors. When AID was performed at a per-speaker instead of at a per-utterance level, we found that AID accuracy improved but that the recognition accuracy remained unchanged. An analysis of accent misclassifications indicated that misclassified utterances leading to improved speech recognition accuracy were on average longer than those leading to deteriorated accuracy. However, it was found that the majority (approximately 75%) of misclassified utterances did not affect speech recognition performance.

Table 8: Confusion matrix for the parallel per-utterance accent identification system using multi-accent acoustic models. Confusions are indicated as percentages (%).

		Hypothesised accent				
		AE	BE	CE	EE	IE
Actual accent	AE	62.41	3.48	17.42	11.32	5.37
	BE	4.16	78.26	7.38	3.36	6.85
	CE	16.50	6.77	53.88	8.60	14.25
	EE	21.94	3.70	7.27	58.83	8.26
	IE	2.43	7.40	10.98	7.75	71.45

AE, Afrikaans English; BE, Black South African English; CE, Cape Flats English; EE, White South African English; IE, Indian South African English.

We conclude that accent misclassifications occurring in a parallel recognition configuration do not necessarily impair speech recognition performance and that multi-accent acoustic models are particularly effective in this regard. This conclusion is important from the perspective of practical system implementation because it suggests that there is little to be gained from the inclusion of a more elaborate AID scheme prior to speech recognition. The inclusion of such an explicit AID component would significantly increase the design and implementation complexity of the overall speech recognition system. This increased cost would be particularly keenly felt in the under-resourced South African setting, in which suitable data and associated speech resources are scarce.

Acknowledgements

This work was executed using the High Performance Computer facility at Stellenbosch University. The financial assistance of the National

Research Foundation (NRF) of South Africa is acknowledged. Opinions expressed and conclusions arrived at are those of the authors and are not necessarily to be attributed to the NRF.

Authors' contributions

Both authors designed the experiments. H.K. performed the experiments and T.R.N. was the project leader. Both authors wrote the manuscript.

References

1. Statistics South Africa. Census 2011 report 03-01-41 [document on the Internet]. c2012 [cited 2013 Sep 06]. Available from: http://www.statssa.gov.za/census2011/default.asp

2. Odyssey 2010: The Speaker and Language Recognition Workshop; 2010 June 28 – July 01; Brno, Czech Republic. Baixas, France: International Speech Communication Association; 2010.

3. Schneider EW, Burridge K, Kortmann B, Mesthrie R, Upton C, editors. A handbook of varieties of English. Berlin: Mouton de Gruyter; 2004.

4. Finn P. Cape Flats English: Phonology. In: Schneider EW, Burridge K, Kortmann B, Mesthrie R, Upton C, editors. A handbook of varieties of English. Vol.1. Berlin: Mouton de Gruyter; 2004. p. 964–984.

5. Fukunaga K. Introduction to statistical pattern recognition. 2nd ed. San Diego, CA: Academic Press; 1990.

6. Kamper H, Muamba Mukanya FJ, Niesler TR. Multi-accent acoustic modelling of South African English. Speech Commun. 2012;54(6):801–813. http://dx.doi.org/10.1016/j.specom.2012.01.008

7. Kennedy J, Eberhart R. Particle swarm optimization. In: Proceedings IEEE International Conference on Neural Networks; 1995 Nov 27 – Dec 01; Perth, Australia. IEEE; 1995. p. 1942–1948. http://dx.doi.org/10.1109/ICNN.1995.488968

8. Teixeira C, Trancoso I, Serralheiro A. Recognition of non-native accents. In: Proceedings of Eurospeech; 1997 Sep 22–25; Rhodes, Greece. Baixas, France: European Speech Communication Association; 1997. p. 2375–2378.

9. Fischer V, Gao Y, Janke E. Speaker-independent upfront dialect adaptation in a large vocabulary continuous speech recognizer. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP); 1998 Nov 30 – Dec 04; Sydney, Australia. p. 787–790.

10. Kirchhoff K, Vergyri D. Cross-dialectal data sharing for acoustic modeling in Arabic speech recognition. Speech Commun. 2005;46(1):37–51. http://dx.doi.org/10.1016/j.specom.2005.01.004

11. Despres J, Fousek P, Gauvain JL, Gay S, Josse Y, Lamel L, et al. Modeling northern and southern varieties of Dutch for STT. In: Proceedings of Interspeech; 2009 Sep 6–10; Brighton, UK. Baixas, France: International Speech Communication Association; 2009. p. 96–99.

12. Caballero M, Moreno A, Nogueiras A. Multidialectal Spanish acoustic modeling for speech recognition. Speech Commun. 2009;51:217–229. http://dx.doi.org/10.1016/j.specom.2008.08.003

13. Schultz T, Waibel A. Language-independent and language-adaptive acoustic modeling for speech recognition. Speech Commun. 2001;35:31–51. http://dx.doi.org/10.1016/S0167-6393(00)00094-7

14. Faria A. Accent classification for speech recognition. In: Renals S, Bengio S, editors. Proceedings of the Second International Workshop on Machine Learning for Multimodal Interaction (MLMI); 2005 July 11–13; Edinburgh, UK. Edinburgh: Springer; 2006. p. 285–293.

15. Chengalvarayan R. Accent-independent universal HMM-based speech recognizer for American, Australian and British English. In: Proceedings of Eurospeech; 2001 Sep 3–7; Aalborg, Denmark. Baixas, France: International Speech Communication Association; 2001. p. 2733–2736.

16. Teixeira C, Trancoso I, Serralheiro A. Accent identification. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP); 1996 Oct 3–6; Philadelphia, PA, USA. Philadelphia, PA: University of Delaware; 1996. p. 1784–1787.

17. Beattie V, Edmondson S, Miller D, Patel Y, Talvola G. An integrated multi-dialect speech recognition system with optional speaker adaptation. In: Proceedings of Eurospeech; 1995 Sep 18–21; Madrid, Spain. Baixas, France: European Speech Communication Association; 1995. p. 1123–1126.

18. Roux JC, Louw PH, Niesler TR. The African Speech Technology project: An assessment. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC); 2004 May 26–28; Lisbon, Portugal. Paris: European Language Resources Association; 2004. p. 93–96.

19. Stolcke A. SRILM – An extensible language modeling toolkit. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP); 2002 Sep 16–20; Denver, CO, USA. Denver, CO: Causal Productions; 2002. p. 901–904.

20. Katz SM. Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE T Acoust Speech. 1987;35(3):400–401. http://dx.doi.org/10.1109/TASSP.1987.1165125

21. Ney H, Essen U, Kneser R. On structuring probabilistic dependencies in stochastic language modelling. Comput Speech Lang. 1994;8(1):1–38. http://dx.doi.org/10.1006/csla.1994.1001

22. Young SJ, Evermann G, Gales MJF, Hain T, Kershaw D, Liu X, et al. The HTK book (for HTK Version 3.4). Cambridge: Cambridge University Engineering Department; 2009.

23. Niesler TR. Language-dependent state clustering for multilingual acoustic modelling. Speech Commun. 2007;49(6):453–463. http://dx.doi.org/10.1016/j.specom.2007.04.001

24. Kamper H, Niesler TR. Multi-accent speech recognition of Afrikaans, Black and White varieties of South African English. In: Proceedings of Interspeech; 2011 Aug 28–31; Florence, Italy. Baixas, France: International Speech Communication Association; 2011. p. 3189–3192.

25. Bisani M, Ney H. Bootstrap estimates for confidence intervals in ASR performance evaluation. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2004 May 17–21; Montreal, Quebec, Canada. Montreal: IEEE; 2004. p. 409–412.

26. Mesthrie R. Socio-phonetics and social change: Deracialisation of the GOOSE vowel in South African English. J Socioling. 2010;14(1):3–33. http://dx.doi.org/10.1111/j.1467-9841.2009.00433.x

27. Bekker I, Eley G. An acoustic analysis of White South African English (WSAfE) monophthongs. South Afr Linguist Appl Lang Stud. 2007;25(1):107–114. http://dx.doi.org/10.2989/16073610709486449

28. Bekker I. Fronted /s/ in general White South African English. Lang Matters. 2007;38(1):46–74. http://dx.doi.org/10.1080/10228190701640025

29. Mesthrie R. Ethnicity, substrate and place: The dynamics of Coloured and Indian English in five South African cities in relation to the variable (t). Lang Var Change. 2012;24:371–395. http://dx.doi.org/10.1017/S0954394512000178

Sunspot numbers: Implications on Eastern African rainfall

AUTHORS:
Francis Gachari[1]
David M. Mulati[1]
Joseph N. Mutuku[1]

AFFILIATION:
[1]Department of Physics, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

CORRESPONDENCE TO:
Francis Gachari

EMAIL:
regcm4@gmail.com

POSTAL ADDRESS:
Department of Physics, Jomo Kenyatta University of Agriculture and Technology, PO Box 3258, 00200 Nairobi, Kenya

DATES:
Received: 13 Mar. 2013
Revised: 28 Jun. 2013
Accepted: 12 Oct. 2013

KEYWORDS:
sunspot numbers; solar declination; lunar declination; statistical rainfall model; rainfall projection

HOW TO CITE:
Gachari F, Mulati DM, Mutuku JN. Sunspot numbers: Implications on Eastern African rainfall. S Afr J Sci. 2014;110(1/2), Art. #2013-0050, 5 pages. http://dx.doi.org/10.1590/sajs.2014/20130050

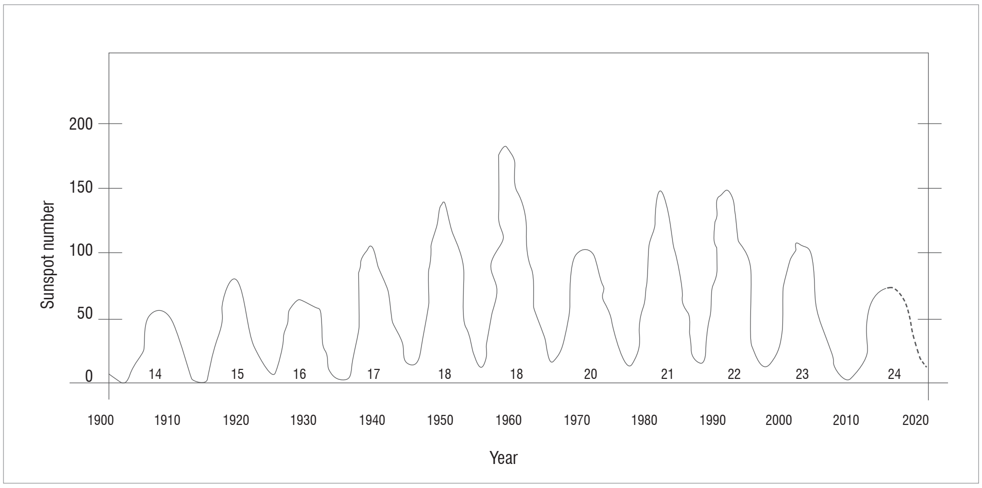
Following NASA's prediction of sunspot numbers for the current sunspot cycle, Cycle 24, we now include sunspot numbers as an explanatory variable in a statistical model. This model is based on fitting monthly rainfall values with factors and covariates obtained from solar–lunar geometry values and sunspot numbers. The model demonstrates high predictive skill in estimating monthly values by achieving a correlation coefficient of 0.9 between model estimates and the measurements. Estimates for monthly total rainfall for the period from 1901 to 2020 for Kenya indicate that the model can be used not only to estimate historical values of rainfall, but also to predict monthly total rainfall. We have found that the 11-year solar sunspot cycle has an influence on the frequency and timing of extreme hydrology events in Kenya, with these events occurring every 5±2 years after the turning points of sunspot cycles. While solar declination is the major driver of monthly variability, sunspots and the lunar declinations play a role in the annual variability and may have influenced the occurrence of the Sahelian drought of the mid-1980s that affected the Sahel region including the Greater Horn of Africa. Judging from the reflection symmetry, the trend of the current maximum and the turning point of the sunspot minimum at the end of the Modern Maximum, with a 95% level of confidence, drought conditions similar to those of the early 1920s may reoccur in the year 2020±2.

Introduction

We begin by describing briefly the three predictors we have used in this study. These are monthly values of the sunspot numbers, maximum lunar declination and solar declination.

Sunspot numbers

The number of sunspots appearing on the solar surface has been recorded each month through observations and calculation for a long time. Currently, sunspot numbers are clearly headed towards a minimum given the trends and the near symmetry of the current maximum, typically referred to as Modern Maximum, which comprises Cycles 17 to 23 (Figure 1). The current cycle, Cycle 24, will probably mark the end of the Modern Maximum, with the sun switching to a state of less strong activity. While there are three main groups of prediction methods according to Kristof[1] – precursor methods, extrapolation methods and model-based predictions – the National Aeronautics and Space Administration (NASA) and the Solar Influences Data Analysis Center (SIDC)[2] have finally used the precursor method and made their predictions for Cycle 24. The smoothed sunspot numbers and predictions can be seen in Figure 1.



Source: Solar Influences Data Analysis Center[2]

Figure 1: Smoothed sunspot numbers of Cycles 14 to 24 showing the predicted (dotted) segment.

Sunspot numbers have been associated with a change in climate, including severe climatic conditions during the Maunder Minimum – the period 1640–1705 which was characterised by a conspicuous lack of sunspots.[3] Total solar irradiance increases when the number of sunspots increases. Total solar irradiance is higher at solar maximum, even though sunspots are darker (cooler) than the average photosphere. Meehl and Arblaster[4] analysed sea surface temperatures from 1890 to 2006. They then used two computer models from the US National Center for Atmospheric Research to simulate the response of the oceans to changes in solar output. They found that as the sun's output reaches a peak, the small amount of extra sunshine over several years causes a slight increase in local atmospheric heating, especially across parts of the tropical and subtropical Pacific where sun-blocking clouds are normally scarce. The small amount of extra heat leads to more evaporation, producing extra water vapour. In turn, moisture is carried by trade winds to the normally rainy areas of the western tropical Pacific, fuelling heavier rains.

In 2008, White and Liu[5] provided evidence that the 11-year solar cycle may be the trigger for El Niño and La Niña events by using harmonic analysis on observed and model data. A model such as the one developed in this study captures interannual rainfall variability by involving sunspot numbers as predictors. The dotted line in Cycle 24 represents NASA's predicted sunspot numbers for 2012–2020.

Sunspot Cycle 24 is the last cycle of the current maximum while the dotted line shows the sunspot numbers that NASA have predicted for 2013–2020. The current prediction for Sunspot Cycle 24 gives a smoothed sunspot number maximum of about 69 in 2013. Hathaway et al.'s[6] method of predicting the behaviour of a sunspot cycle is fairly reliable once a cycle has reached about 3 years after the minimum sunspot number occurs.

Maximum lunar declination

Varying angular lunar velocity caused by the lunar node cycle is considered likely to influence natural forcing of the El Niño Southern Oscillation (ENSO) by lunar tidal forces. Cerveny and Shaffer[7] examined a possibility that lunar tidal forces act as an external forcing mechanism in regulating sea surface temperatures tied to ENSO events. They obtained a statistically significant correlation between maximum lunar declination (MLD) and both equatorial Pacific sea surface temperatures and South Pacific atmospheric pressure (the Southern Oscillation Index) for the period 1854–1999. High MLDs were associated with La Niña conditions, while low MLDs were associated with El Niño conditions. Under high MLD, circulation of the Pacific gyre is enhanced by tidal forces, inducing cold-water advection into the equatorial region that is characteristic of La Niña conditions. Under low MLD, tidal forcing is weakened, cold-water advection is reduced, and warmer sea surface conditions characteristic of El Niño prevail. Together with the solar cycle, MLDs are used to capture interannual rainfall variability.

Solar declination

Earth's axis of rotation is tilted 23.5° away from the plane perpendicular to Earth's orbit while its axis points in the same direction as Earth orbits the sun. Therefore, the solar declination angle determines the seasons, which are characterised by varying solar irradiance, varying length of day and an annual rainfall pattern. Solar declination was used to generate seasonal variation of rainfall.

Materials and methods

Data

The Kenya Meteorological Service[8] (KenMet) supplied the rainfall data (rain gauge measurements) taken at Dagoretti and Jomo Kenyatta Airport from 1959 to 2005. The Climate Research Unit of the University of East Anglia (UK) provided research data sets for the Kenyan region.[9] We extracted monthly and annual rainfall totals from 1901 to 2000. The country aggregation is based on the TYN_CY_1_1 data set. This data set is referred to here as CRU.K.

NASA provided solar and lunar declination values obtained from their *ephemeris* software.[10] The National Oceanic and Atmospheric Administration's National Geophysical Data Center provided sunspot numbers, including NASA's prediction. The international sunspot number is produced by SIDC[2] at the World Data Center for the Sunspot Index at the Royal Observatory of Belgium.

The model

Model SMS12.12 is of the form:

Response variable ~ predictor(s) Equation 1

in which the response variable is the monthly total rainfall and the three predictors are solar declination, maximum lunar declination and sunspot numbers. This model design is based on fitting a generalised linear model (GLM) of the Tweedie[11] family to Kenya's monthly total rainfall distribution values from 1951 to 1980. A fitting procedure involved obtaining beta values which satisfy the linear equation:

$$y_i = \beta_i x_i^T + e_i \qquad \text{Equation 2}$$

where y_i is the estimated monthly total rainfall values for each of the three predicting factors, x_i, and e_i is the fit error in the estimate. In this study, we fitted first-order factors and therefore $T=1$. Statistical software was used to fit a GLM and obtain the initial estimate of beta values for the fit. The fitting procedure comprised two main steps:

Step 1: Fitting a GLM

We computed an initial estimate set of beta values using the GLM of the Tweedie family. This family of exponential dispersion models is characterised by the power mean–variance relationship:

$$V(\mu) = \mu^p \qquad \text{Equation 3}$$

Thus variance V is a function of the mean (μ) and p is the power variance of the Tweedie distribution calculated by means of a routine in an R-program called *tweedie.profile*. To specify the Tweedie, the mean (μ), the dispersion parameter (ϕ) and the variance power (p) are required. Standard algorithms in R-software calculate μ and the maximum likelihood estimate is used to work out ϕ and p. A GLM fit on the rainfall distribution obtains initial estimates for a fit parameter, α, and a dispersion parameter, ϕ. At this point it is possible to use these beta values to calculate rainfall estimates for the GLM fit. However, rainfall data is correlated and therefore it is necessary to fit a generalised estimating equation to account for the correlation within the variable being fitted.

Step 2: Fitting a generalised estimating equation

To fit a generalised estimating equation, it is necessary to use a correlation matrix which best describes the manner of correlation to calculate new beta values. In this case we used estimates obtained in Step 1 for the fit parameter α, and the correlation matrix AR(1) which is defined as $\alpha^{|u-v|}$

$$R_{u,v} = \begin{cases} 1, & u = v, \\ \alpha^{|u-v|}, & otherwise \end{cases} \qquad \text{Equation 4}$$

In matrix notation this becomes

$$Ri = \begin{bmatrix} 1 & \alpha & \alpha^2 & & \alpha^{n-1} \\ \alpha & 1 & \alpha & \cdots & \alpha^{n-2} \\ \alpha^2 & \alpha & 1 & & \vdots \\ & \vdots & & \ddots & \vdots \\ \alpha^{n-1} & \cdots & & \alpha & 1 \end{bmatrix} \qquad \text{Equation 5}$$

New beta estimates are thus obtained which are then used to estimate monthly totals by use of Equation 1.

Results and discussion

Model results

Model SMS12.12 was trained on a 30-year CRU.K data set (1951–1980) and tested on two segments of data: 1901–1950 and 1981–2000. Predictors, solar declination, maximum lunar declinations and sunspot numbers used were mean values for each month. Figure 2 shows how SMS12.12 demonstrates prediction stability with time.

Methods used for avoiding artificial prediction skill included using independent training and test data sets, cross-validation and hindcasting. Forecast skill depends on the amount of lead time, the number of forecast months and the strength of the relationships between estimates and rainfall records. Each value plotted in Figure 2 represents a Pearson product-moment correlation coefficient between estimate and CRU.K value for corresponding months in the year. Correlation values remained above 0.5 throughout the 100-year period, except for 1925 (Figure 2). The probability of obtaining a correlation value above 0.5 when the model is used in estimating projected values is therefore 99 out of 100 years (0.99). An adjusted R^2-value of 0.62 was obtained between CRU.K and model estimates during the training period and reduced values of 0.52 and 0.56 were obtained from the test data sets. SMS12.12 shows stability in estimating monthly total rainfall when model monthly estimates are compared with corresponding CRU.K values. The average correlation for the period 1901–2000 is 0.8. The contributions of the individual predictors to the variability of the predictand were 59.7% for solar declination, 9.4% for sunspot numbers and 8.9% for maximum lunar declination. Thus solar declination played the dominant role in monthly rainfall variability. The remaining 22% of the variability remains unexplained.

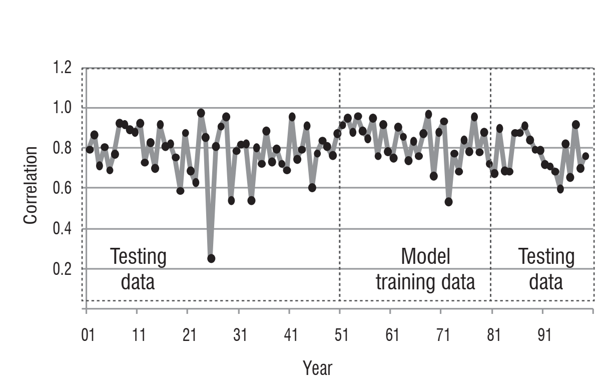


Figure 2: Correlation between SMS12.12 estimates and CRU.K monthly values for each year, showing SMS12.12 stability.

Monthly rainfall projection

SMS12.12 was then used to estimate monthly rainfall totals for the period 1901–2020. Figure 3 shows monthly estimates so obtained.

Monthly totals were then aggregated into annual values and results standardised by mean and standard deviation. The results are shown in Figure 4, in which model results have been plotted together with CRU.K values for comparison. Model SMS12.12 estimates indicate elevated monthly totals (>+1 standard deviation) in the periods 1912–1913, 1931–1932, 1951–1952, 1987–1988, 1993–1994, 1997–1998 and 2005–2006, and depressed monthly rainfall (<-1 standard deviation) in 1917, 1937–1938, 1947–1948, 1982–1985,1992–1993, 2002–2004, 2010–2011 and 2019–2020.

Model results were then validated with records from the United Nations Development Programme.[12] Below average annual total rainfall was reported to have occurred in 1928, 1933–1934, 1937, 1939, 1942–1944, 1947, 1952–1953, 1955–1957, 1975–1977, 1980–1985, 1991–1992, 1999–2000 and 2004. Other below average values were recorded by KenMet in 1965, 1973–1974, 1976 and 1992–1993. Floods recorded by KenMet occurred in 1961, 1963, 1977–1978 and 1997–1998. Projected model estimates indicate below normal rainfall in 2009–2011, 2015 and 2019–2020, with values within one standard deviation. Above normal rainfall may be expected in 2012–2014, 2016 and 2018, with values within two standard deviations. Estimates were calculated at a 95% confidence level.

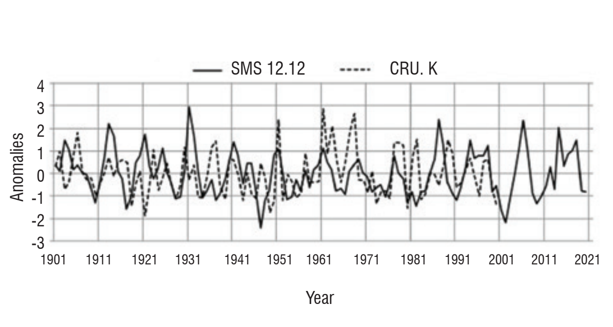


Figure 4: Projected annual total rainfall anomalies determined by model SMS12.12 for the period 1901–2020. SM12.12 estimates are plotted together with CRU.K values for comparison.

Model diagnostics

Probabilities of rainfall volumes were calculated in order to judge the accuracy of the model estimates. The results are shown in Figure 5. Estimates are comparable at all stages of model development as shown by hindcasting, training (fitting) and forecast stages as well as with the 1901–2000 climatology. Correlation values between the model and CRU.K data are shown in Table 1 for the hindcast, training and forecast stages. Model estimates are therefore reliable.

Sunspot numbers and annual rainfall

Rainfall for Kenya in the Modern Maximum indicates a peak trend that corresponds to that of sunspots, as shown in Figure 6. Figure 6 shows standardised values of annual totals of rainfall and smoothed sunspot numbers. A best fit trend line of peak annual rainfall is also shown. The trend has a peak in the 19th sunspot cycle centred around 1961. Variability in annual rainfall shows reflection symmetry in the year 1961, such that Cycles 18 and 20 are object and image, respectively, in Figure 6.

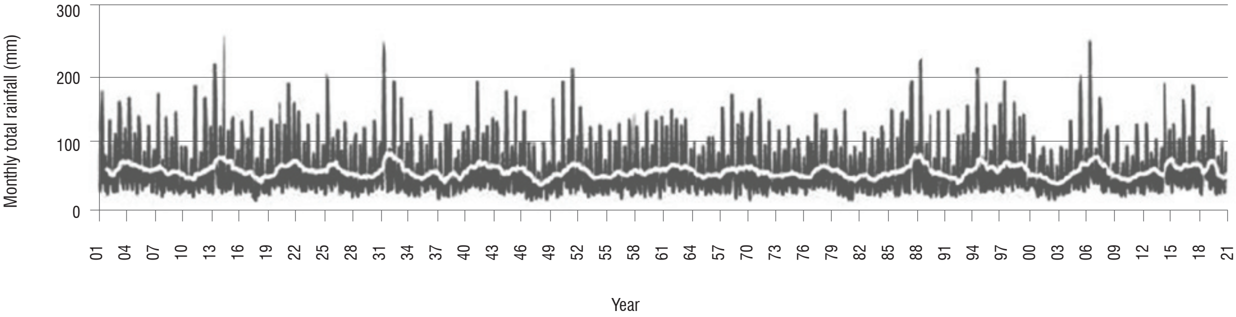


Figure 3: Projected monthly total rainfall values determined by model SMS12.12 for the period 1901–2020 with a 12-month moving average trend line.

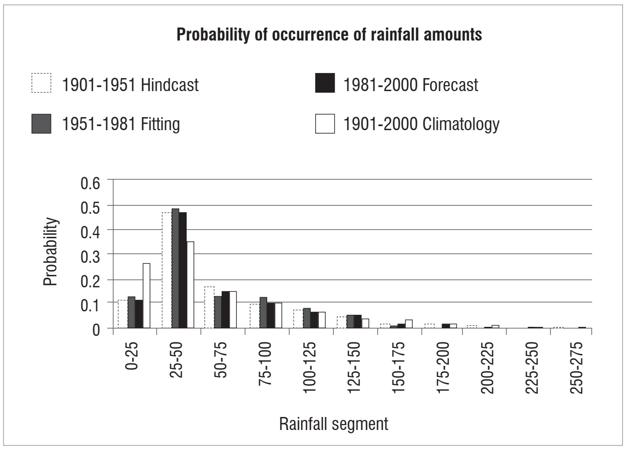


Figure 5: Probability of rainfall amounts using hindcast, SM12.12 fitting, forecast and climatology.

Table 1: Correlation coefficients between model SMS12.12 and CRU.K data for hindcast, training and forecast stages

Segment	Coefficient
1901–1951 (hindcast)	0.90
1951–1981 (training)	0.92
1981–2000 (forecast)	0.84

We refer to events occurring prior to 1961 as objects of corresponding events after 1961, which are images. Object and image pairs labelled c, d, e and f are cycle pairs: 17 and 21, 16 and 22, 15 and 23, and 14 and 24. Object and image cycle pairs have similar rainfall peak amplitudes. Reflection symmetry demands that if sunspot turning points lead rainfall events in the objects side, the reverse will happen in the image side. While the cause of the distribution symmetry is still under investigation, it is what is observed from Figure 6. At least three sunspot turning points are outstanding. The first one is the heavy rainfall of the early 1960s corresponding to the maximum in Cycle 19, the second is the drought of the mid-1970s and the minimum between Cycles 20 and 21, and the third corresponds to the great Sahelian drought after the passing of the minimum between Cycles 21 and 22. From Figure 6 one can identify a turning point for each event of severe hydrology in Kenya, suggesting that sunspot numbers had an influence on rainfall as was found by Meehl and Arblaster[4]. Now that sunspots are headed for a minimum at the end of the Modern Maximum, one may expect fewer events of high rainfall and perhaps a prolonged drought of the Sahelian type. Judging from the symmetry of the Modern Maximum, a drought of the type experienced in the early 1930s will most likely occur in 2020±2 after the passage of the current Cycle 24. This observation is also consistent with model SMS12.12 results as shown in Figure 4. Because Kenya's rainfall is influenced by the Sahel climate, it is likely that the decline in rainfall volumes may be experienced in the Eastern Africa region and perhaps the Sahel region, including the Greater Horn of Africa.

Summary and conclusions

This study was motivated by the desire to find out the physical causes of the Kenyan droughts of the early 1980s and at the turn of this century. The temporal distribution of sunspot numbers indicates that each turning point corresponds to events of severe hydrology in Kenya with a time lag of 5±2 years. Therefore, such events are predictable so long as sunspots can be predicted in advance. However, the prediction of sunspots has not been easy and the current prediction of Cycle 24 appears to be at the end of the Modern Maximum, therefore breaking the continuity. The current maximum is fairly symmetrical, increasing the confidence that sunspot activity is headed for an all time low, perhaps similar to the one at the beginning of the last century with a corresponding reduction in annual rainfall volumes. It is therefore likely that Kenya will experience reduced rainfall at the turn of Cycle 24 and around the year 2020±2.

Model estimates indicate that before 2020, above average rainfall may be expected in the period 2013–2018 and below normal rainfall in 2019–2020. No sunspot numbers are available to enable estimation beyond 2020 using the model; in addition, the behaviour of sunspots is uncertain beyond Cycle 24. However, as we head towards 2020, it is likely that the evolution of sunspots will occur in a predictable pattern so that sunspot prediction will be possible. However, this observation cannot yet be assumed for global data sets. Furthermore, we recommend that future studies be done on rainfall residuals so that the seasonality factor is eliminated and a better indication of the influence of sunspot numbers can be obtained. A comparison of the results with those obtained through statistical downscaling methods is also recommended.

Acknowledgements

We thank the following for providing data: the Kenya Meteorological Department (Nairobi, Kenya), the National Oceanic and Atmospheric Administration and the National Aeronautics and Space Administration (USA), the Climate Research Unit of the University of East Anglia (UK), and the Solar Influences Data Analysis Center of the Royal Observatory of Belgium. We also thank the following for providing software: R Foundation for Statistical Computing at the National Centre for Atmospheric Research and Unidata for the Integrated Data Viewer at the Program Center, University Corporation for Atmospheric

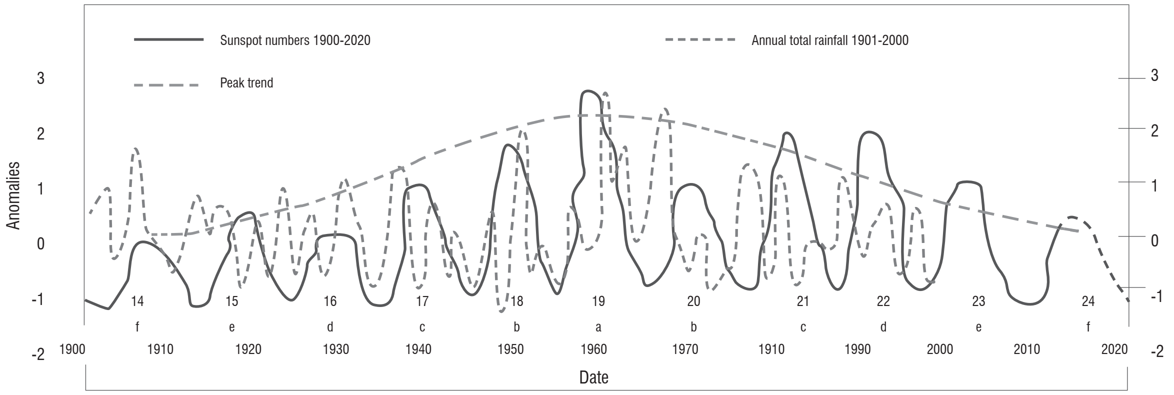


Figure 6: Smoothed sunspot numbers and total annual rainfall for the period 1901–2020 showing the peak trend.

Research (Boulder, CO, USA), ConvexDNA for the Excel Mixer (Geneva, Switzerland), and Fourmilab for HomePlanet (Switzerland).

Authors' contributions

F.G. performed the research; D.M.M. and J.N.M. supervised the work.

References

1. Kristof P. Solar cycle prediction. Liv Rev Solar Phys [serial on the Internet]. Revised 2011 Jan 05 [cited 2013 Jan 03]; 7:6. Available from: http://www.livingreviews.org/lrsp-2010-6
2. Solar Influences Data Analysis Center [homepage on the Internet]. c2013 [cited 2012 Nov 13]. Available from: http://sidc.oma.be
3. Lassen K, Friis-Christensen E. Variability of the solar cycle length during the past five centuries and the apparent association with terrestrial climate. J Atmos Terrestrial Physics. 1995;57:835. http://dx.doi.org/10.1016/0021-9169(94)00088-6
4. Meehl GA, Arblaster JM. A lagged warm event-like response to peaks in solar forcing in the Pacific region. J Climate. 2009;22:3647–3660. http://dx.doi.org/10.1175/2009JCLI2619.1
5. White WB, Liu Z. Non-linear alignment of El Niño to the 11-yr solar cycle. Geophys Res Lett. 2008;35:19607. http://dx.doi.org/10.1029/2008GL034831
6. Hathaway DH, Wilson MR, Reichmann JE. A synthesis of solar cycle prediction techniques. J Geophys Res. 1999;104:22375–22388. http://dx.doi.org/10.1029/1999JA900313
7. Cerveny RS, Shaffer JA. The moon and El Niño. Geophys Res Lett. 2001;28:25–28. http://dx.doi.org/10.1029/2000GL012117
8. Kenya Meteorological Service [homepage on the Internet]. c2005 [cited 2013 Jan 27]. Available from: http://www.meteo.go.ke/data/.
9. Mitchell TD, Hulme M, New M. Climate data for political areas. Area. 2002;34:109–112. http://dx.doi.org/10.1111/1475-4762.00062
10. Horizons. Horizons Web Interface [homepage on the Internet]. c2013 [cited 2013 Jan 20]. Available from: http://ssd.jpl.nasa.gov/horizons.cgi?s_disp=1#top.
11. Tweedie MCK. An index which distinguishes between some important exponential families. Statistics - Application and new directions. In: Ghosh JK, Roy J, editors. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference. Calcutta: Indian Statistical Institute; 1984. p. 579–604.
12. United Nations Development Programme. Kenya natural disaster profile. Enhanced security unit [homepage on the Internet]. c2004 [cited 2013 Jan 14]. Available from: http://www.gripweb.org/gripweb/sites/default/files/disaster_risk_profiles/Preliminary/

A late Holocene sea-level curve for the east coast of South Africa

AUTHORS:
Kate L. Strachan[1]
Jemma M. Finch[1]
Trevor Hill[1]
Robert L. Barnett[2]

AFFILIATIONS:
[1]Discipline of Geography, School of Agricultural, Earth & Environmental Sciences, University of KwaZulu-Natal, Pietermaritzburg, South Africa
[2]School of Geography, Plymouth University, Plymouth, United Kingdom

CORRESPONDENCE TO:
Kate Strachan

EMAIL:
kateleighstrachan@gmail.com

POSTAL ADDRESS:
Discipline of Geography, School of Agricultural, Earth & Environmental Sciences, University of KwaZulu-Natal, Private Bag X01, Scottsville 3201, South Africa

DATES:
Received: 25 Jun. 2013
Revised: 07 Oct. 2013
Accepted: 12 Oct. 2013

KEYWORDS:
foraminifera; palaeoenvironmental; intertidal; Kariega Estuary; salt marsh

HOW TO CITE:
Strachan KL, Finch JM, Hill T, Barnett RL. A late Holocene sea-level curve for the east coast of South Africa. S Afr J Sci. 2014;110(1/2), Art. #2013-0198, 9 pages. http://dx.doi.org/10.1590/sajs.2014/20130198

South Africa's extensive and topographically diverse coastline lends itself to interpreting and understanding sea-level fluctuations through a range of geomorphological and biological proxies. In this paper, we present a high-resolution record of sea-level change for the past ~1200 years derived from foraminiferal analysis of a salt-marsh peat sequence at Kariega Estuary, South Africa. A 0.94-m salt-marsh peat core was extracted using a gouge auger, and chronologically constrained using five radiocarbon age determinations by accelerator mass spectrometry, which places the record within the late Holocene period. Fossil foraminifera were analysed at a high downcore resolution, and a transfer function was applied to produce a relative sea-level reconstruction. The reconstructed sea-level curve depicts a transgression prior to 1100 cal years BP which correlates with existing palaeoenvironmental literature from southern Africa. From ~1100 to ~300 cal years BP, sea levels oscillated (~0.5-m amplitudes) but remained consistently lower than present-day mean sea level. The lowest recorded sea level of −1±0.2 m was reached between 800 and 600 cal years BP. After 300 cal years BP, relative sea level has remained relatively stable. Based on the outcomes of this research, we suggest that intertidal salt-marsh foraminifera demonstrate potential for the high-resolution reconstruction of relative sea-level change along the southern African coastline.

Introduction

Understanding past patterns of sea-level change is important on local (e.g. for coastal management and engineering), regional to national (e.g. national future sea-level predictions) and global scales (e.g. for understanding polar ice sheet history). Evidence of recent sea-level change can be derived from instrumental data such as tide gauges (for example see Douglas[1]) and satellite altimetry (for example see Nerem[2]). These records can be extended back into the Holocene by means of proxy data from archaeological sites, geomorphological features, isolation basin contacts and salt-marsh sediments.[3,4] Relative sea-level (RSL) curves have been constructed for a substantial portion of northern hemisphere coastlines[3,5-8]; however, to date, few curves have been presented for the southern hemisphere.[9,10] Recently, late Holocene RSL curves were produced for New Zealand[11] and Tasmania[12], yet no such curve exists for South Africa.

The southern African coastline has been tectonically stable throughout the late Quaternary, which means that sea-level change in this region would have been marginally affected by postglacial eustatic rise during the end of the Pleistocene and the early parts of the Holocene.[13] Holocene RSL records for the eastern and western coastlines are incomplete in extent and coarse in resolution.[13] South African sea-level research has therefore relied largely on global records as a benchmark.[14] During the last 7000 years, southern African sea levels have fluctuated by no more than ±3 m.[13,15-17] Sea-level curves based on observational data for southern Africa indicate that Holocene highstands occurred at 6000 and again at 4000 cal years BP, followed by a lowstand from 3000 to 2000 cal years BP.[13,17] The mid-Holocene highstands culminated in a sea-level maximum of approximately 3 m above mean sea level (MSL) from 7300 to 6500 cal years BP[13,18] and of 2 m above MSL at around 4000 cal years BP.[14-16] Thereafter, RSL dropped to slightly below the present level between 3500 and 2800 cal years BP.[13] Sea-level fluctuations during the late Holocene in southern Africa were relatively small (1–2 m); however, these fluctuations had a major impact on past coastal environments.[13,15-17] Evidence from the west coast suggests that there was a highstand of 0.5 m above MSL from 1500 to 1300 cal years BP, or possibly earlier (1800 cal years BP[13]), followed by a lowstand (-0.5 m above MSL) from 700 to 400 cal years BP.[17] A lowstand along the southern coast, dated to 700 cal years BP, is evident from in-situ tree stumps exposed at low tide.[19] The majority of proxy sea-level data from South Africa derive from sites on the western and southwestern coastlines (e.g. Langebaan, Knysna, Verlorenvlei and Bogenfel Pan).

High-resolution sea-level reconstructions can be achieved through the analysis of salt-marsh foraminifera, which are accurate and precise sea-level indicators as a result of their vertically zoned nature relative to tidal levels and elevation above MSL.[20,21] Salt marshes experience daily and seasonal variations in salinity, flooding frequency and suspended sediment delivery, directly linked to tidal overflow.[22] Surface elevations are predominantly controlled by tidal inundation (sediment delivery) and mean salinities (plant productivity).[23] Foraminiferal assemblages are vertically zoned along salt-marsh surface elevational gradients in relation to inundation rates and tidal elevations; thus, in temperate regions, foraminiferal assemblages are considered accurate proxies of past sea-level change.[20,24,25] By assigning indicative meanings (environmental ranges with reference to water level) to modern foraminiferal zonations, and applying these to downcore fossilised assemblages, predictions of past sea levels can be made with precisions of between ±0.05 m and ±0.2 m.[8]

In the southern African context, the application of foraminifera as biological indicators has been restricted to studies of stratigraphy[26], temperature change[27], sedimentology[28-31] and marine records[32]. The use of salt-marsh foraminifera to reconstruct relative sea-level change has been limited to a single published study at Langebaan.[33] Here we introduce an established sea-level proxy to determine proof of concept for South African sea-level research. This technique has the potential to contribute to our incomplete understanding of past sea-level change along the southern African coastline.

In this study we present a high-resolution foraminiferal analysis of a sedimentary sequence derived from the Kariega Estuary, Eastern Cape, South Africa (Figure 1). Modern foraminiferal assemblages analysed at this site exhibit clear vertical zonation across the marsh surface. Modern assemblages were used to construct a training set of foraminifera suitable for transfer function purposes which could then be applied to downcore fossilised assemblages. A late Holocene sea-level reconstruction chronologically controlled by accelerator mass spectrometry (AMS) radiocarbon dating is presented, representing the first high-resolution, continuous record of sea-level history for eastern South Africa.

Study site

The Kariega River is elongated and sinuous, with the estuary stretching approximately 18 km inland from the mouth (Figure 1). The estuary is surrounded by salt marshes, sand flats and steep slopes in the upper reaches.[34] The system is hypersaline owing to a lack of freshwater input[34]; however, scouring by tidal currents maintains a permanent connection with the sea.[35] As with many estuarine systems along the Eastern Cape shoreline, Kariega has a small tidal prism, with water levels fluctuating in response to semi-diurnal and spring/neap tidal cycles.[36] This system has a low turbidity, with little salinity and thermal stratification of the water column during any stage of the tidal cycle.[36] The system receives variable rainfall, has a small catchment area (686 km^2) and is regulated by three dams.

There are three major intertidal salt-marsh systems at Kariega: Taylors, Grants and Galpins. The intertidal creeks of Taylors and Grants marshes are relatively shallow with a narrow intertidal area, while Galpins marsh is wider and more extensive.[37,40] Several characteristics of Galpins salt marsh make it an ideal location for foraminiferal analysis: (1) the marsh is attached to a small estuarine embayment, and is thus protected from erosional dynamics associated with the main channel; (2) the intertidal zone exhibits clear vegetation zonation, dominated by *Spartina maritima* and *Sarcocornia perennis*; and (3) analysis of surface samples across the intertidal zone indicate clear vertical zonation in modern foraminiferal assemblages.

Methods

An extensive programme of coring using a 20-mm diameter gouge auger was conducted to establish the stratigraphy of the salt marsh. Core lithologies were described using notation developed by Long et al.[41], based on original classification techniques by Troels-Smith[43]. A master core was taken from the high marsh (near the level of mean high water spring tides) for palaeoenvironmental analyses. Five neighbouring, stratigraphically consistent 1-m continuous sediment cores (KAR1–5) were extracted from this point in the salt marsh (33°39’04”S; 26°39’74”E) using a 50-mm diameter gouge auger. Cores were placed into polyvinyl chloride piping, wrapped in polythene sheeting and heavy duty plastic and transported to the laboratory for cold storage and analysis. Of the five cores, one was selected for foraminiferal analysis (KAR2) and another for developing a chronology (KAR4). Remaining cores are in cold storage for potential future studies.

Stratigraphic boundaries were targeted for dating to produce a late Holocene chronology. Five samples were extracted from core KAR4 for AMS radiocarbon analysis at Beta Analytic (USA). The organic fraction

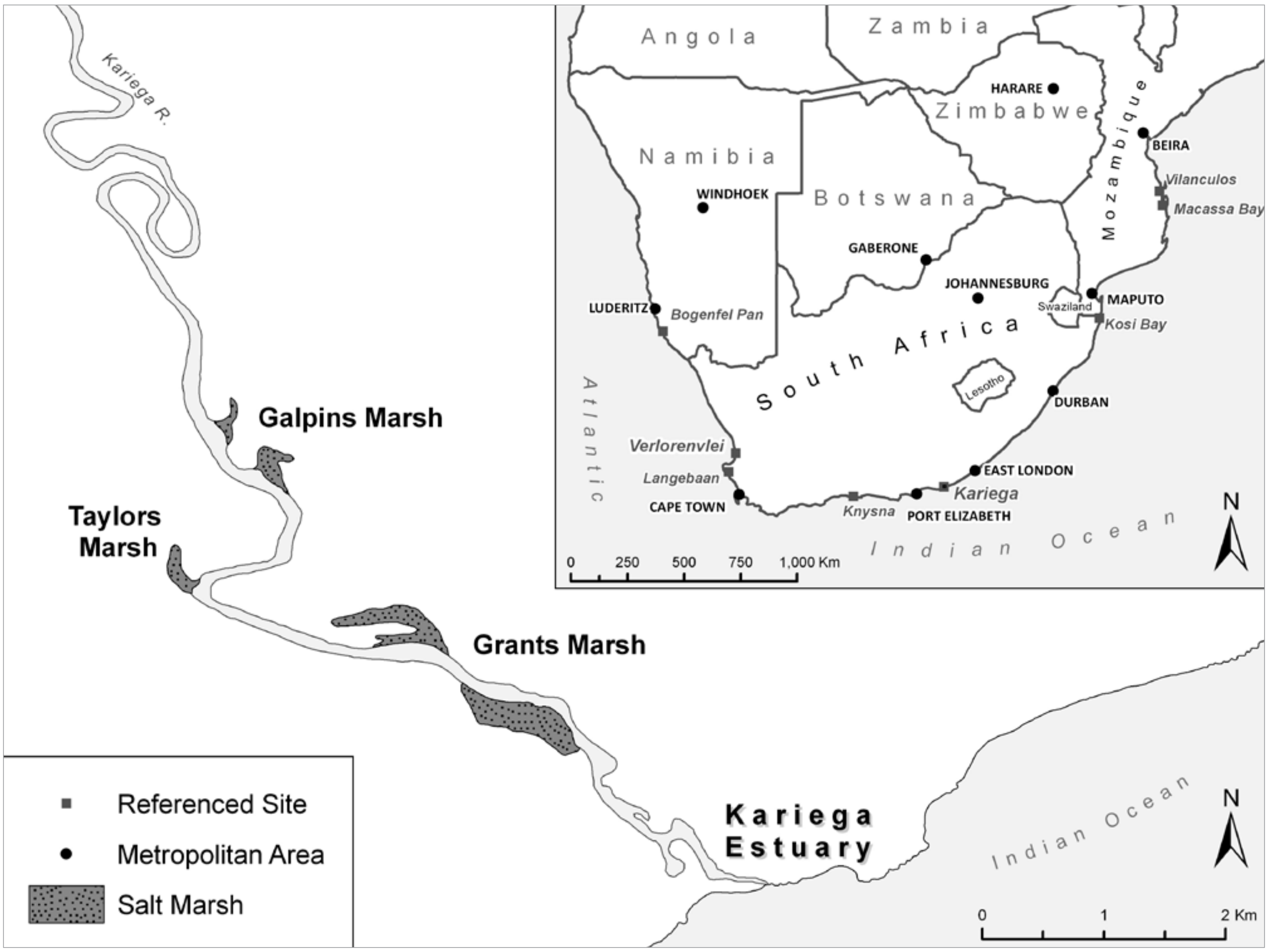


Figure 1: Location of Kariega Estuary (after Paterson and Whitfield[37]) and sites mentioned in the text: Bogenvel Pan[18]; Verlorenvlei[13]; Langebaan[17]; Knysna[19]; Kosi Bay[15]; Macassa Bay[38] and Vilanculos[39].

was dated for four bulk sediment samples and the carbonate fraction for a single shell sample at the base of the core. AMS ages for bulk sediment samples were calibrated using the calibration curve SHCal04.[43]

The marine shell-derived age was corrected using a ΔR value of 146±85.[44] This correction provided a spuriously young age, possibly as a result of shell recrystallisation. Several linear age modelling scenarios were considered, including the less parsimonious possibility that the shell age is correct and that overlying bulk sediment ages are all too old. In accordance with the principle of parsimony, the shell age was ultimately excluded and an objective Bayesian age-modelling exercise was applied to the remaining dates. An alternative explanation is that older sediments could have been washed downstream and deposited in the salt marsh. However, the consistently organic stratigraphy does not provide supporting evidence of sedimentary inwash events, as it lacks coarse sand layers. The majority of foraminiferal tests encountered were intact without etching, which does not support sediment reworking.

'Bacon' source code,[45] within the R open-source statistical environment,[46] was used to develop an age-depth model. Bacon utilises a Bayesian approach towards chronology building, which is reliant on a priori information. A mean accumulation rate of 16 cm/year was used for the model (acc.mean=16, acc.shape=2). However, as salt marsh sedimentation rates are highly variable, a low memory component was used (mem.mean=0.3, mem.strength=10). The resultant age-depth model provided age ranges at a resolution of 10-mm within 95% confidence limits. It is understood that, through using an age-depth modelling approach rather than individually dated sea-level index points, sea-level reconstruction interpretation can be compromised.[8] In this instance, however, the use of Bacon facilitates the construction of a chronologically constrained sea-level history which would not be possible without age-depth modelling.

Core KAR2 was subsampled at a 20-mm downcore resolution yielding 48 samples for foraminiferal analysis following Scott and Medioli[21] and Gehrels[47]. Samples were stored in ethanol, rinsed in distilled water and washed through 63-μm and 500-μm sieves. The larger sieve was used to remove organic matter and coarse detritus; remaining sediment on the smaller sieve was subdivided into eight aliquots using a volumetric wet splitter and retained for foraminiferal analysis. Samples were kept wet throughout the counting process to prevent aggregation of organic particles and to minimise degradation.

A Leica M205C stereomicroscope with an attached DFC295 digital camera (SMM Instruments, Durban, South Africa) was used for counting and identifying foraminiferal assemblages under 60X, 80X and 100X magnification. At least 250 individuals were counted per sample[48]; this limit was most necessary where indicator species were present in low numbers.[47] Downcore fluctuations in foraminiferal assemblages were plotted using Psimpoll Version 4.263[49]; the constrained incremental sum of squares (CONISS)[50] stratigraphically constrained ordination technique assisted in defining zonation patterns of downcore fossilised assemblages.

Detrended canonical correspondence analysis was used to determine the environmental (in this case elevation) gradient length of the modern foraminiferal data set in standard deviation (SD) units. Where gradient lengths are < 2 SD units, it is assumed species are responding linearly to the environmental variable.[51] The data set demonstrated a short gradient length of 1.4 SD units and therefore a linear-based regression model was suitable for developing a transfer function.[52] The transfer function was built using a partial least squares regression model in C_2[53] and the modern foraminifera counts of both live and dead assemblages were combined. Training sets composed of total assemblages are still occasionally used based on the assumption that the live assemblages will in time contribute to the fossil record[54]; however, it is acknowledged that there are strong arguments suggesting that the dead assemblage alone is most suitable for transfer function purposes (for example, Murray[55]). Sample-specific elevation prediction errors were calculated using bootstrapped cross validation which provided root-mean-square error of prediction values for each fossil sample. Fossil samples that were similar in composition to the training set samples were included in the reconstruction. The modern analogue technique was used to assign a minimum dissimilarity coefficient value to all fossil samples. Similarity cut-off values follow Watcham et al.[56] Samples below 0.66 m had poor fits with their closest modern analogues. These samples were manually assigned an indicative meaning associated with minimum sea-level heights only. Reconstructed elevation values for all samples were used to calculate past sea-surface elevations in relation to present MSL following Gehrels[57].

Results

Stratigraphy and chronology

The basal unit for this core (0.97–0.57 m) comprises clay, silt and sand in approximately equal quantities. At around 0.92 m, a few unidentifiable shell fragments were recovered. A clay unit containing some silt is present from 0.57 m to 0.37 m. Above this unit is a clayey, silty sand lens, which is 0.70 m thick. From 0.30 m up, the lithostratigraphy is characterised by decreasing grain size (increasing clay content) and becomes increasingly organic. Between 0.70 m and the surface, there is abundant in-situ decaying organic material. Iron oxide staining is present between 0.26 m and 0.70 m.

Accelerator mass spectrometry radiocarbon age determinations and their calibrated ages (with standard deviations) are presented in Table 1. The age-depth model created using Bacon is presented in Figure 2, and places the record within the late Holocene period. The basal date provides an age of 1331 to 1424 cal years BP at 0.895 m. The age determination at 0.185 m (970 to 1142 cal years BP) is identified by the model as a possible outlier.

Table 1: Radiocarbon dating results for the Kariega record, indicating calibrated and uncalibrated ages

Lab code	Depth (m)	Sample material	$^{13}C/^{12}C$ (‰)	^{14}C year BP	Cal years BP (2 SD)
Beta-334778	0.18−0.19	Organic sediment	−20.5	1200±30	970−1142 (92.1%)
Beta-301135	0.53−0.54	Organic sediment	−20.8	620±30	591−638 (49.7%)
Beta-334779	0.66−0.67	Organic sediment	−19.8	1460±30	1281−1363 (95%)
Beta-334780	0.89−0.90	Organic sediment	−19	1570±30	1331−1424 (69.6%)
Beta-3301136	0.92−0.93	Shell	1.5	673±30	557−654 (95%)

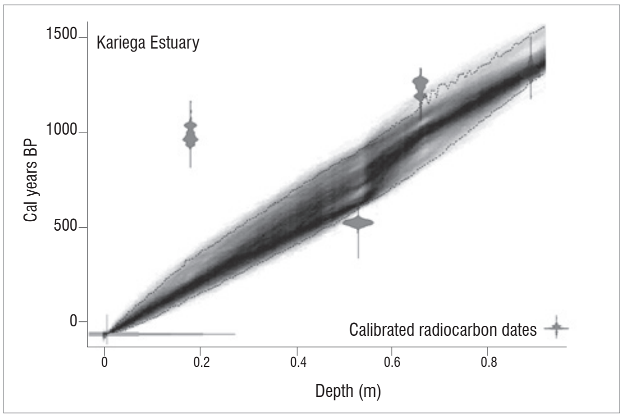


Figure 2: Age-depth model for the Kariega record based on four accelerator mass spectrometry determined ages. The surface age is assumed to represent the present day.

Foraminiferal data

Foraminifera are present and well preserved throughout the length of the core. The majority of tests identified and counted were intact, with little evidence of corrosion. A total of 16 species were identified; the main species, alongside their common environmental niches, are presented in Table 2. Low concentrations of foraminiferal tests (<100) are present between 0.38 m and 0.26 m. Average foraminiferal test concentration for the core was 213 tests.cm^3. CONISS cluster analysis identified five assemblage zones:, K-1 (0.94–0.65 m), K-2 (0.65–0.46 m), K-3 (0.46–0.39 m), K-4 (0.39–0.31 m) and K-5 (0.31–0 m). These zones differentiated major changes in foraminiferal populations through the core. Agglutinated foraminifera are most abundant in the upper reaches of the core and gradually decrease down the core, whereas the base of the core is dominated by calcareous foraminifera (Figure 3).

Table 2: Fossil assemblages identified within core KAR2

Foraminifera	Niche
Ammonia aberdoveyensis	Low marsh and tidal flats
Ammonia batavus	Low marsh and tidal flats
Fissurina sp.	Marine shelves; transported into estuarine systems
Haynesina germanica	Middle/low marsh and tidal flats
Jadammina macrescens	Brackish marshes
Miliammina fusca	High marsh and upper estuary
Quinqueloculina sp.	Estuary mouths, inner shelf and middle/low marsh
Scherochorella moniliforme	Inner shelf and estuary mouths
Spirillina vivipara	Inner shelf
Textularia earlandi	Coastal environments
Textularia gramen	Coastal environments
Trochammina inflata	High marsh and upper estuary
Trochammina ochracea	Inner shelf

Niche descriptions after Murray[58] and Horton and Edwards[59].

Zone K-1 at the base of the core is dominated by the calcareous species *Ammonia aberdoveyensis* and *Haynesina germanica,* which commonly comprise up to 90% of the assemblage. At 0.66 m there is a shift away from these species towards *Ammonia batavus* and the agglutinated salt-marsh species *Jadammina macrescens* and *Miliammina fusca* (Zone K-2). *J. macrescens* is the most abundant species in Zone K-3 which includes low numbers of *M. fusca* and *Trochammina inflata*. Zones K-4 and K-5 are dominated by *J. macrescens.* From Zone K-4 upwards there is an increasing diversity of agglutinated foraminiferal species. *T. inflata* and *Trochammina ochracea* represent up to 40% of the assemblage in Zone K-5.

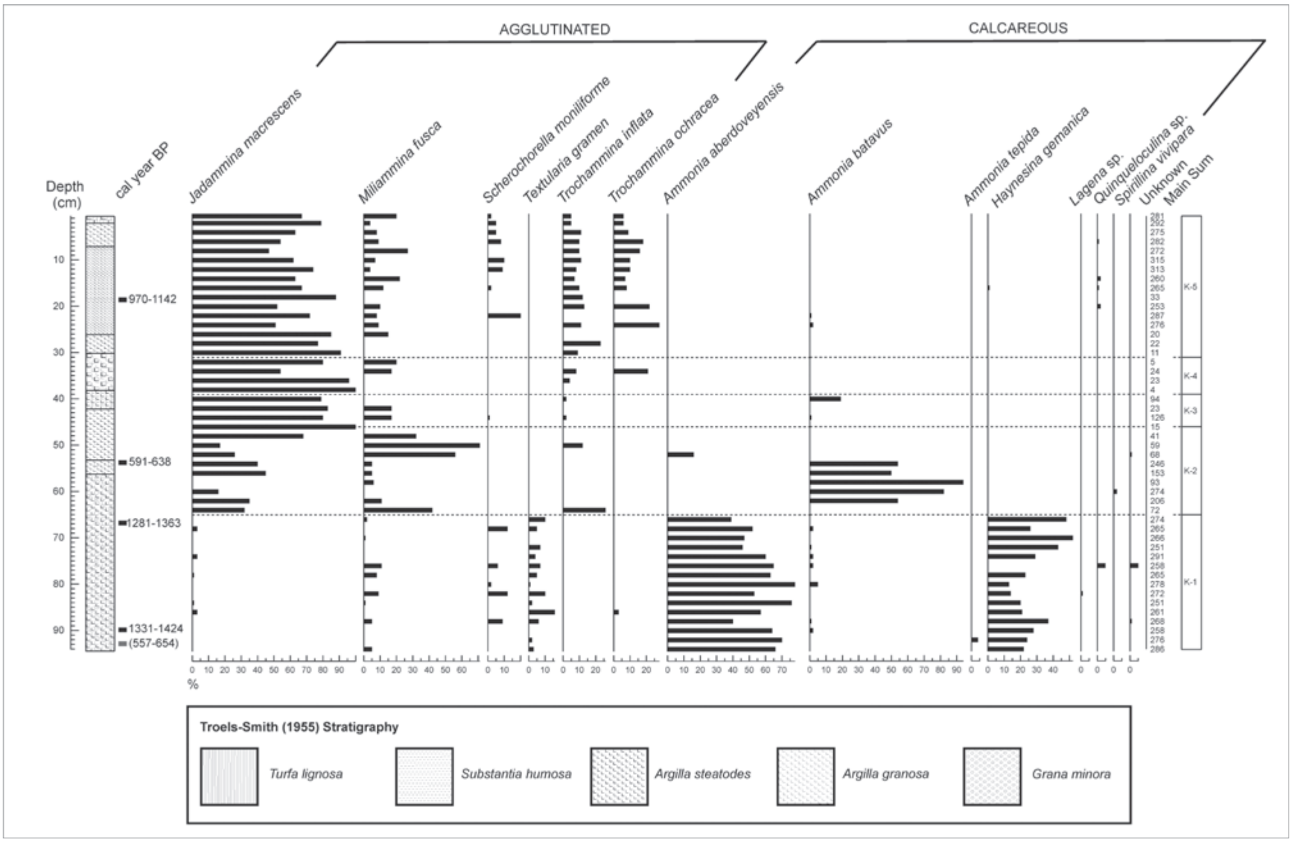


Figure 3: Foraminiferal concentrations down core KAR2. Indicated in the figure is the Troels-Smith[43] stratigraphy and the zonation derived from a constrained incremental sum of squares.

Quantitative reconstruction

Indicative-meaning based predictions of palaeomarsh-surface elevations were converted into sea-level estimates following Gehrels[57]. Samples with good or close modern analogue fits were reconstructed using the transfer function (Table 3). Those samples with poor fit represent foraminiferal assemblages which had no close modern analogues from the training set (Table 4). The presence of *H. germanica* in these samples was used to provide a sea-level estimate. This species most commonly occupies the intertidal zone[61] and therefore an indicative range from lowest astronomical tide (-1.04 m above MSL) to the lowermost sample from the training set (0.01 m above MSL) was applied to these samples. Each sample was constrained chronologically using interpolated values from the age-depth model. The reconstructed sea-level curve for Kariega Estuary (Figure 4) identifies rapid RSL fall at around 1100 cal years BP to levels down to a minimum of ~1 m below present MSL. Between 1100 and 300 cal years BP, RSL appeared to fluctuate at approximately 150-year periods, with amplitudes of around 0.5 m. After 300 cal years BP, RSL remained relatively stable.

Table 3: Performance of the partial least squares (PLS) transfer function

	Model performance			Bootstrapping		
Code	**RMSE**	***r*^{2}**	**Maximum bias**	***r*$^2_{boot}$**	**Maximum bias$_{boot}$**	**RMSEP**
PLS	0.140954	0.733101	0.200624	0.619834	0.249447	0.18225

RMSE, root-mean-square error; RMSEP, root-mean-square error of prediction

Table 4: Results from the transfer function applied to the fossil data

Depth (m)	Modern analogue fit	Modelled age (cal years BP)	Modelled age error (± year)	Indicative meaning (m above MSL)	Error (± m)	Sea level (m)
0.005	Close	-60	22.5	0.444	0.198	-0.024
0.02	Close	-19	35	0.554	0.196	-0.149
0.04	Close	29	70	0.452	0.196	-0.067
0.06	Close	83	100	0.381	0.198	-0.016
0.08	Close	121	120	0.400	0.203	-0.055
0.10	Close	171	145	0.414	0.198	-0.089
0.12	Close	209	155	0.478	0.197	-0.173
0.14	Close	245	170	0.491	0.200	-0.206
0.16	Close	295	195	0.478	0.196	-0.213
0.18	Good	338	205	0.886	0.195	-0.641
0.20	Close	373	215	0.475	0.200	-0.250
0.22	Close	404	220	0.555	0.212	-0.350
0.24	Close	436	230	0.449	0.198	-0.264
0.26	Good	470	240	0.709	0.208	-0.544
0.28	Good	499	245	0.853	0.198	-0.708
0.30	Good	532	255	0.899	0.194	-0.774
0.32	Good	564	260	0.677	0.209	-0.572
0.34	Close	603	275	0.454	0.201	-0.369
0.36	Good	629	275	0.926	0.193	-0.861
0.38	Good	664	280	0.990	0.202	-0.945
0.40	Close	696	285	0.831	0.189	-0.806
0.42	Good	732	290	0.693	0.208	-0.688
0.44	Good	766	295	0.607	0.197	-0.622
0.46	Good	807	310	0.990	0.202	-1.025
0.48	Good	834	310	0.620	0.210	-0.675
0.50	Good	865	305	0.320	0.204	-0.395
0.52	Close	889	305	0.368	0.220	-0.463
0.54	Close	923	310	0.598	0.194	-0.713
0.56	Close	956	310	0.613	0.194	-0.748
0.58	Close	998	315	0.197	0.198	-0.352
0.60	Close	1036	320	0.668	0.188	-0.843
0.62	Good	1062	315	0.522	0.196	-0.717
0.64	Close	1101	320	0.403	0.207	-0.618
0.66	Poor	1123	310	-0.515	0.525	0.28
0.68	Poor	1158	315	–	–	–
0.70	Poor	1245	355	–	–	–
0.72	Poor	1227	300	–	–	–
0.74	Poor	1262	295	–	–	–
0.76	Poor	1304	295	–	–	–
0.78	Poor	1332	285	–	–	–
0.80	Poor	1360	275	–	–	–
0.82	Poor	1404	280	–	–	–
0.84	Poor	1434	280	–	–	–
0.86	Poor	1462	265	–	–	–
0.88	Poor	1498	265	–	–	–

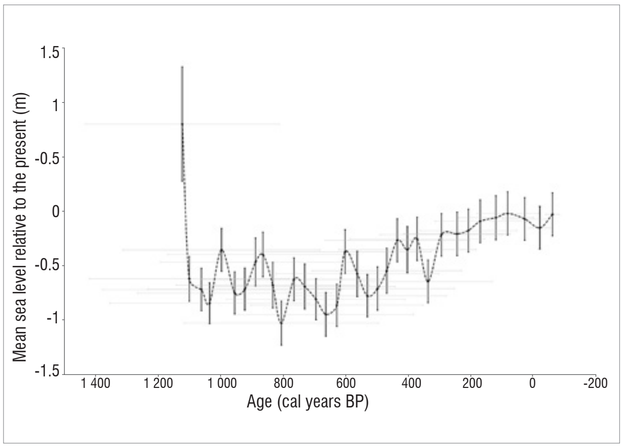


Figure 4: Reconstruction of relative sea-level change for Kariega Estuary.

Discussion

The Kariega Estuary appears to have been connected to the open ocean throughout the late Holocene, based on continuous foraminiferal presence throughout the core. The fossil microfauna suggest that prior to 1100 cal years BP, RSL was higher than it is today, in accordance with existing literature.[13,16,17] The calcareous foraminiferal assemblages from the lowest ~0.35 m of the core imply the existence of a low intertidal, or possibly shallow subtidal, environment before 1100 cal years BP. Because of the broad habitual range of calcareous foraminifera, it is difficult to determine precise indicative meanings from these assemblages.[25,48] Despite this difficulty, it is clear that the reconstruction encompasses the falling limb of rapid RSL decline out of the late Holocene highstand which is documented across South African coastlines.[17,19] Existing late Holocene sea-level data from South Africa is summarised in Table 5 and presented alongside the modelled reconstruction (Figure 5). There is strong agreement between the existing data and the RSL curve from the Kariega Estuary which provides a well vertically constrained estimate of continuous late Holocene RSL for this part of Africa.

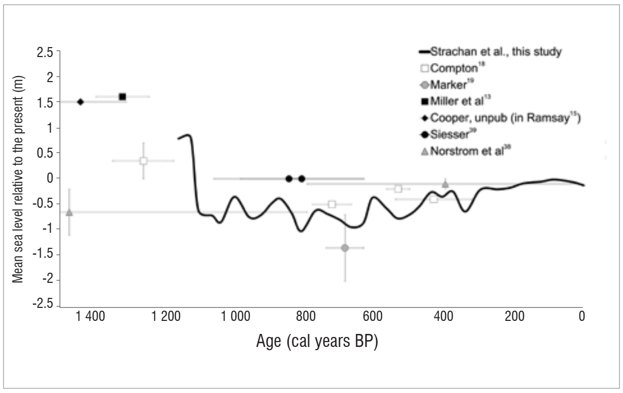


Figure 5: Late Holocene sea-level curve from Kariega Estuary compared with previously published sea-level evidence from southern Africa.

Pre-1100 cal years BP highstand

A late Holocene sea-level highstand has been reported at around 1500–1200 cal years BP[13,15,17] (Figure 5). A sea-level curve from the southwestern coast,[18] and to some extent beachrock evidence from the KwaZulu-Natal coast,[14] indicate highstands between 3000 and 1200 cal years BP. The *Zostera* facies recovered from Langebaan,[17] a pollen record from Verlorenvlei[16] and oyster shells from Langebaan[61] present further evidence that RSL was 0.5–0.7 m higher than present at around 1300 cal years BP. This highstand corresponds with the Little Climatic Optimum, a period of temperate warming which is associated with higher sea levels around this time.[62] The RSL decline out of this highstand is identified by significant changes in foraminiferal assemblages from Kariega. As RSL fell below that of modern day, the abrupt change in foraminifera populations at 0.66 m might be a result of sedimentary erosion as sea

Table 5: Summary of late Holocene sea-level indicators for southern Africa

Lab code	^{14}C year BP	Cal years BP (2 SD)	MSL relative to present (m)	Material	Locality	Reference
Pta-7589	450±70	319–537 (95)	−0.4	Bulk organic matter	Langebaan	Compton[17]
Pta-7576	560±45	497–563 (86)	−0.2	Bulk organic matter	Langebaan	Compton[17]
Pta-7579	840±45	664–774 (93)	−0.5	Bulk organic matter	Langebaan	Compton[17]
Pta-7597	1390±50	1171–1345 (94)	+0 to −0.7	Bulk organic matter	Langebaan	Compton[17]
Pta-7201	770±50	628–736 (78)	−2 to −0.7	Tree stump	Knysna	Marker[19]
Interpolated range		0–790	−0.2 to 0	Bulk organic matter	Macassa Bay	Norström et al.[38]
Interpolated range		4700–790	−1.1 to −0.2	Bulk organic matter	Macassa Bay	Norström et al.[38]
Pta-4311	1450±50	1242–1394 (92)	+1.6	Shell	Verlorenvlei	Miller et al.[13]
Unknown	910±120	631–980 (92)	0	Beachrock cement	Vilanculos	Siesser[39]
Unknown	920±140	627–1057 (92)	0	Beachrock cement	Vilanculos	Siesser[39]
Pta-4972	1610±70	1307–1569 (94)	+1.5	Coral in beach	Kosi Bay	Cooper (unpublished) in Ramsay[15]

For standardisation, ages were calibrated using the same methodology employed in this paper.

levels began to rise out of the lowstand. High precision dating would be required to determine whether a hiatus exists at this level.

1100–300 cal years BP oscillations

The Kariega record provides evidence of sea levels remaining lower (by approximately 0.5 m) than present MSL from ~1100 to ~300 cal years BP. Siesser's[39] study at Vilanculos, using beachrock cement, suggests that sea levels were equivalent to the present day from 842±215 cal years BP (Figure 5), contradicting other studies which indicate a lower RSL during this period. According to Compton[17], two lowstands of −0.5 m to −1 m took place from 700 to 400 cal years BP (Figure 5). This lowstand is evident on the south coast at 682 ± 54 cal years BP from in-situ tree stumps exposed at Knysna Estuary during low tide[19] (Figure 5). This lowstand corresponds with lower sea-surface temperatures between 500 and 400 cal years BP; therefore steric effects associated with cool water temperatures may have been a contributing factor.[17]

300 cal years BP to present

The RSL oscillations shown by this study and by Compton[17] terminate at around 300 cal years BP. Following 300 cal years BP, RSL shows a gradual rise to present MSL. Evidence from Langebaan Lagoon[17] (Figure 5) suggests that sea level has been rising to present levels since 400 cal years BP. The Kariega reconstruction shows relatively stable sea levels throughout the past 300 cal years BP. Exogenic subsidence linked with large-scale water extraction would likely result in recent (past 100 years) sea-level rise. Such evidence is absent from the Kariega reconstruction, suggesting that local subsidence has not occurred.

The Macassa Bay study indicates that lower sea levels are followed by a rise from 700 cal years BP until the present[38] (Figure 5). However, this multi-proxy data set implies that relatively low sea levels recorded at Macassa Bay were associated with a freshwater phase with little influence from marine processes.[38] A noticeable sea-level acceleration occurs in the RSL reconstruction during the middle of the 20th century following a minor RSL fall during the early 1900s. This pattern of recent sea-level change has also been documented in New Zealand and Tasmania.[12] Such similarities with other southern hemispheric RSL curves lend weight to the credibility of this study. Proposed sources for this 20th-century sea-level inflexion include northern hemispheric ice melt from Greenland.[12]

The Kariega record provides high-resolution sea-level data at a regional scale, offering insight into east coast sea-level fluctuations. The pre-1100 cal years BP highstand is supported by existing sea-level data, which provide good chronological estimation of when RSL declined out of this highstand. Observed fluctuations during the late Holocene have been small (~ 0.5 m amplitude) compared with the mid-Holocene (~2–3.5 m amplitude[15]). Reconstructed patterns of recent RSL changes are in agreement with other records from the southern hemisphere, reflecting processes contributing to sea-level change occurring on hemispheric, or greater, scales.

Few tidal gauge records for the southern hemisphere extend beyond 50 years.[22] This lack of long-term tidal gauge data hampers detailed validation of recent reconstructions. To identify and understand patterns of both recent (i.e. 19th and 20th century) and late Holocene sea-level changes, it is necessary to consider proxy data.[1,7] Further research sites should be investigated along the east coast to establish a greater modern-day analogue for South Africa and to strengthen future reconstructions. Sea-level reconstructions from neighbouring sites could present a more robust and high-resolution picture of sea-level fluctuations for the east coast of South Africa.

Conclusion

Fossil foraminiferal assemblages analysed from a ~1200-year sedimentary record from the Kariega Estuary, South Africa, identify continuous late Holocene RSL changes that show accordance with published records. The reconstruction was supported chronologically using a Bayesian age-depth model based on AMS radiocarbon age determinations. With higher resolution dating, timings of key RSL changes in South Africa could be established with previously unprecedented accuracy. This study presents the first salt-marsh foraminifera transfer function for the southern African coastline used to produce a high-resolution sea-level curve for the Kariega Estuary. A late Holocene RSL was reconstructed with mean errors of ±0.2 m. A sea-level highstand was dated to pre-1100 cal years BP. We observed fluctuations in MSL between 1100 and 300 cal years BP; yet MSL then was consistently lower than present sea levels. From 300 cal years BP, RSL increased gradually to present-day levels, and a sea-level inflexion was shown to occur at some time during the mid-20th century. The results of this study suggest that intertidal salt-marsh foraminifera demonstrate potential for high-resolution reconstruction of RSL change in South Africa.

Acknowledgements

This research was financially supported by the National Research Foundation. Angus Paterson (SAEON) facilitated access to the study site. Martin Hill and Julie Coetzee (Rhodes University) assisted in fieldwork. Jussi Baade (University of Jena) assisted with GPS data. Brice Gijsbertsen (UKZN) professionally drafted Figure 1.

Authors' contributions

All authors contributed equally to the project conceptualisation, fieldwork and write-up. K.L.S. performed the laboratory and microscopy analyses.

References

1. Douglas BC. Concerning evidence for fingerprints of glacial melting. J Coast Res. 2008;24:218–227. http://dx.doi.org/10.2112/06-0748.1

2. Nerem RS. Measuring global mean sea level variations using TOPEX/POSEIDON altimeter data. J Geophys Res. 1995;100(25):135–151.

3. Gehrels R, Horton BP, Kemp AC, Sivan D. Two millennia of sea-level data: The key to predicting change. EOS Trans Am Geophys Union. 2011;92(35):289–296. http://dx.doi.org/10.1029/2011EO350001

4. Cronin TM. Rapid sea-level rise. Quat Sci Rev. 2012;56:11–30. http://dx.doi.org/10.1016/j.quascirev.2012.08.021

5. Church JA, White NJ. Sea-level rise from the late 19th to early 21st century. Surv Geophys. 2011;32:585–602. http://dx.doi.org/10.1007/s10712-011-9119-1

6. Nicholls RJ, Marinova N, Lowe JA, Brown S, Vellinga P, De Gusmae D, et al. Sea-level rise and its possible impacts given a 'beyond 4°C world' in the twenty-first century. Philos Trans R Soc Lond A. 2011;369:161–181. http://dx.doi.org/10.1098/rsta.2010.0291

7. Siddall M, Milne GA. Understanding sea-level change is impossible without both insights from paleo studies and working across disciplines. Earth Planet Sci Lett. 2012;315–316:2–3. http://dx.doi.org/10.1016/j.epsl.2011.10.023

8. Gehrels WR, Woodworth PL. When did modern rates of sea-level rise start? Global Planet Change. 2013;100:263–277. http://dx.doi.org/10.1016/j.gloplacha.2012.10.020

9. Pirazzoli PA, Pluet J. World atlas of Holocene sea-level changes. Amsterdam: Elsevier; 1991.

10. Goodwin IAND. Did changes in Antarctic ice volume influence late Holocene sea-level lowering? Quat Sci Rev. 1998;17:319–332. http://dx.doi.org/10.1016/S0277-3791(97)00051-6

11. Gehrels WR, Hayward BW, Newnham RM, Southall KE. A 20th century acceleration of sea-level rise in New Zealand. Geophys Res Lett. 2008;35:1–5. http://dx.doi.org/10.1029/2007GL032632

12. Gehrels WR, Callard SL, Moss PT, Marshall WA, Blaauw M, Hunter L, et al. Nineteenth and twentieth century sea-level changes in Tasmania and New Zealand. Earth Planet Sci Lett. 2012;315–316:94–102. http://dx.doi.org/10.1016/j.epsl.2011.08.046

13. Miller D, Yates R, Jerardino A, Parkington J. Late Holocene coastal change in the southwestern Cape, South Africa. Quat Int. 1995;29:3–10. http://dx.doi.org/10.1016/1040-6182(95)00002-Z

14. Ramsay PJ, Cooper JAG. Late Quaternary sea-level change in South Africa. Quat Res. 2002;57:82–90. http://dx.doi.org/10.1006/qres.2001.2290

15. Ramsay P. 9000 years of sea-level change along the southern African coastline. Quat Int. 1995;31:71–75. http://dx.doi.org/10.1016/1040-6182(95)00040-P

16. Baxter AJ, Meadows ME. Evidence for Holocene sea-level change at Verlorenvlei, Western Cape, South Africa. Quat Int. 1999;56:65–79. http://dx.doi.org/10.1016/S1040-6182(98)00019-6

17. Compton JS. Holocene sea-level fluctuations inferred from the evolution of depositional environments of southern Langebaan Lagoon salt marsh, South Africa. Holocene. 2001;11:395–405. http://dx.doi.org/10.1191/095968301678302832

18. Compton JS. The mid-Holocene sea-level highstand at Bogenfels Pan on the southwest coast of Namibia. Quat Res. 2006;66:303–310. http://dx.doi.org/10.1016/j.yqres.2006.05.002

19. Marker ME. Evidence for a Holocene low sea level at Knysna. S Afr Geogr J (Special Edition). 1997;79(2):106–107. http://dx.doi.org/10.1080/03736245.1997.9713631

20. Scott DB, Medioli FS. Vertical zonations of marsh foraminifera as accurate indicators of former sea levels. Nature. 1978;272:528–531. http://dx.doi.org/10.1038/272528a0

21. Scott DB, Medioli FS. Quantitative studies of marsh foraminifera distributions in Nova Scotia: Implications for sea level studies. Cushman Foundation for Foraminiferal Research Special Publication. 1980;17:58.

22. Massey AC, Gehrels WR, Charman DJ, White SV. An intertidal foraminifera-based transfer function for reconstructing Holocene sea-level change in Southwest England. J Foramin Res. 2006;36(3):215–232. http://dx.doi.org/10.2113/gsjfr.36.3.215

23. Leorri E, Gehrels WR, Horton BP, Fatela F, Cearreta A. Distribution of foraminifera salt marshes along the Atlantic coast of SW Europe: Tools to reconstruct past sea-level variations. Quat Int. 2010;221:104–115. http://dx.doi.org/10.1016/j.quaint.2009.10.033

24. Gehrels WR. Determining relative sea-level change from salt-marsh foraminifera and plant zones on the coast of Maine, USA. J Coast Res. 1994;10(4):990–1009.

25. Woodroffe SA, Horton BP, Larcombe P. Whittaker JE. Intertidal mangrove foraminifera from the central Great Barrier Reef shelf, Australia: Implications for sea-level reconstruction. J Foramin Res. 2005;35(3):259–270. http://dx.doi.org/10.2113/35.3.259

26. McMillan IK. Foraminiferal biostratigraphy, sequence stratigraphy and interpreted chronostratigraphy of marine Quaternary sedimentation on the South African continental shelf. S Afr J Sci. 1993;89:83–89.

27. McMillan IK. Cainozoic planktonic and larger foraminifera distributions around southern Africa and their implications for past changes of oceanic water temperatures. S Afr J Sci. 1986;82:66–69.

28. Cooper JAG, McMillan IK. Foraminifera of the Umgeni Estuary, Durban and their sedimentological significance. S Afr J Geol. 1987;90(4):489–498.

29. McMillan IK. Foraminifera from the Late Pleistocene (Latest Eemian to Earliest Weichselian) Shelly Sands of Cape Town City Centre, South Africa. Ann S Afr Mus. 1990;99(5):121–186.

30. Lindsay P, Pillay S, Wright CI, Manson TR. Sedimentology and dynamics of the Mfolozi Estuary, north KwaZulu-Natal, South Africa. S Afr J Geol. 1996;99:327–336.

31. Wright CI, McMillan IK, Mason TR. Foraminifera and sedimentation patterns in St. Lucia Estuary mouth, Zululand, South Africa. S Afr J Geol. 1990;93(4):592–601.

32. Martin RA. Benthic foraminifera from the Orange-Luderitz shelf, southern continental margin. Joint Geological Survey/University of Cape Town Marine Geoscience Group Bulletin. 1981;11:75.

33. Franceschini G, McMillan IK, Compton JS. Foraminifera of Langebaan Lagoon salt marsh and their application to the interpretation of late Pleistocene depositional environments at Monwabisi, False Bay coast, South Africa. S Afr J Geol. 2005;108:285. http://dx.doi.org/10.2113/108.2.285

34. Grange N, Whitfield AK, De Villiers CJ, Allanson BR. The response of two South African east coast estuaries to altered river flow regimes. Aquat Conserv. 2000;10:155–177. http://dx.doi.org/10.1002/1099-0755(200005/06)10:3<155::AID-AQC406>3.0.CO;2-Z

35. Paterson AW, Vorwerk PD, Froneman PW, Strydom NA, Whitfields AK. Biological responses to a resumption in river flow in a freshwater deprived, permanently open southern African estuary. Water SA. 2008;34(5):597–604.

36. Taylor DI. Tidal exchange of carbon, nitrogen and phosphorus between a *Sarcocornia* salt marsh and the Kariega estuary, and the role of salt marsh brachyuran in this transfer [thesis]. Grahamstown: Rhodes University; 1987.

37. Paterson AW, Whitfield AK. Do shallow-water habitats function as refugia for juvenile fishes? Est Coast Shelf Sci. 2000;51:359–364. http://dx.doi.org/10.1006/ecss.2000.0640

38. Nörstrom E, Risberg J, Grondahl H, Holmgren K, Snowball I, Mugabe JA, et al. Coastal paleo-environment and sea-level change at Macassa Bay, south Mozambique, since c 6600 cal BP. Quat Int. 2012;260:153–163. http://dx.doi.org/10.1016/j.quaint.2011.11.032

39. Siesser WG. Relict and recent beachrock from southern Africa. Geol Soc Am Bull. 1974;85:1849–1854. http://dx.doi.org/10.1130/0016-7606(1974)85<1849:RARBFS>2.0.CO;2

40. Paterson AW, Whitfield AK. The ichthyofauna associated with an intertidal creek adjacent eelgrass beds in the Kariega Estuary, South Africa. Environ Biol Fish. 2000;58:154–156. http://dx.doi.org/10.1023/A:1007629328937

41. Long AJ, Innes JB, Shennan I, Tooley MJ. Coastal stratigraphy: A case study from Johns River, Washington, USA. In: Jones AP, Tucker ME, Hart JK, editors. The description and analysis of Quaternary stratigraphic field sections. Technical Guide No 7. London: Quaternary Research Association; 1999. p. 267–286.

42. Troels-Smith J. Characterization of unconsolidated sediments. D.G.U. IV Rekke. 1955;3(10):37–82.

43. McCormac FG, Hogg AG, Blackwell PG, Buck CE, Higham TFG, Reimer PJ. SHCal04 southern hemisphere calibration, 0–11.0 cal years BP. Radiocarbon. 2004;46:1087–1092.

44. Dewar G, Reamer PJ, Sealy J, Woodborne S. Holocene marine radiocarbon reservoir correction (ΔR) for the west coast of South Africa. Holocene. 2012;22(12):1438–1446. http://dx.doi.org/10.1177/0959683612449755

45. Blaauw M, Christen JA. Flexible paleoclimate age-depth models using an autoregressive gamma process. Bayesian Anal. 2011;6(3):457–474. http://dx.doi.org/10.1214/ba/1339616472

46. Team RDC. A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2010. Available from: http://www.R-project.org.

47. Gehrels WR. Intertidal foraminifera as palaeoenvironmental indicators. In: Haslett SK, editor. Quaternary environmental micropalaeontology. London: Oxford University Press; 2002.

48. Horton BP, Murray JW. The roles of elevation and salinity as primary controls on living foraminiferal distributions: Cowpen Marsh, Tees Estuary, UK. Mar Micropaleontol. 2007;63:169–186. http://dx.doi.org/10.1016/j.marmicro.2006.11.006

49. Bennett KD. Psimpoll: C programs for plotting pollen diagrams and analysing pollen data. Villavgen: Uppsala Universitet; 2005.

50. Grimm EC. CONISS: A FORTRAN 77 program for stratigraphically constrained cluster analysis by the method of incremental sum of squares. Comput Geosci.1987;13:13–35. http://dx.doi.org/10.1016/0098-3004(87)90022-7

51. Ter Braak CJF, Prentice IC. A theory of gradient analysis. Adv Ecol Res. 1988;18:271–317. http://dx.doi.org/10.1016/S0065-2504(08)60183-X

52. Birks HJB. Quantitative palaeoenvironmental reconstructions. In: Maddy D, Brew JS, editors. Statistical modeling of Quaternary science data. Technical Guide No 5. Cambridge: Quaternary Research Association; 1995.p. 161–236.

53. Juggins S. C^2 Version 1.3: Software for ecological and palaeoecological data analysis and visualization. Newcastle upon Tyne: Department of Geography, University of Newcastle; 2003. p. 69.

54. Booth RK, Lamentowicz M, Charman DJ. Preparation and analysis of testate amoebae in peatland palaeoenvironmental studies. Mires Peat. 2010;7:1–7.

55. Murray JW. The enigma of the continued use of total assemblages in ecological studies of benthic foraminifera. J Foramin Res. 2000;30(3):244–245. http://dx.doi.org/10.2113/0300244

56. Watcham EP, Shennan I, Barlow NLM. Scale considerations in using diatoms as indicators of sea-level change: Lessons from Alaska. J Quaternary Sci. 2013;28(2):165–179. http://dx.doi.org/10.1002/jqs.2592

57. Gehrels WR. Middle and late Holocene sea-level changes in eastern Maine reconstructed from foraminiferal salt marsh stratigraphy and AMS 14C dates on basal peat. Quat Res. 1999;52:350–359. http://dx.doi.org/10.1006/qres.1999.2076

58. Murray JW. British near shore foraminiferids: Key and notes for the identification of the species. London: Academic Press; 1979.

59. Horton BP, Edwards RJ. Quantifying Holocene sea-level change using intertidal foraminifera: Lessons from the British. Cushman Foundation for Foraminiferal Research Special Publication. 2006;40:97.

60. Murray JW. Ecology and application of benthic foraminifera. Cambridge: Cambridge University Press; 2006.

61. Flemming BW. Depositional processes in Saldanha Bay and Langebaan Lagoon. Marine Geoscience Bulletin. 1997;8:215.

62. Nunn PD. Sea-level changes over the past 1000 years in the Pacific. J Coast Res. 1998;14:23–30.

Dynamics of arable land requirements for food in South Africa: From 1961 to 2007

AUTHORS:
Yibo Luan[1]
Xuefeng Cui[1]
Marion Ferrat[1]
Reshmita Nath[2]

AFFILIATIONS:
[1]State Key Laboratory of Earth Surface Processes and Resource Ecology, College of Global Change and Earth System Science, Beijing Normal University, Beijing, China
[2]Center for Monsoon System Research, Institute of Atmospheric Physics, Chinese Academy of Science, Beijing, China

CORRESPONDENCE TO:
Xuefeng Cui

EMAIL:
xuefeng.cui@bnu.edu.cn

POSTAL ADDRESS:
No.19 Xinjiekouwai Street,
Haidian District, Beijing
100875, China

DATES:
Received: 29 Nov. 2012
Revised: 06 Jun. 2013
Accepted: 12 Oct. 2013

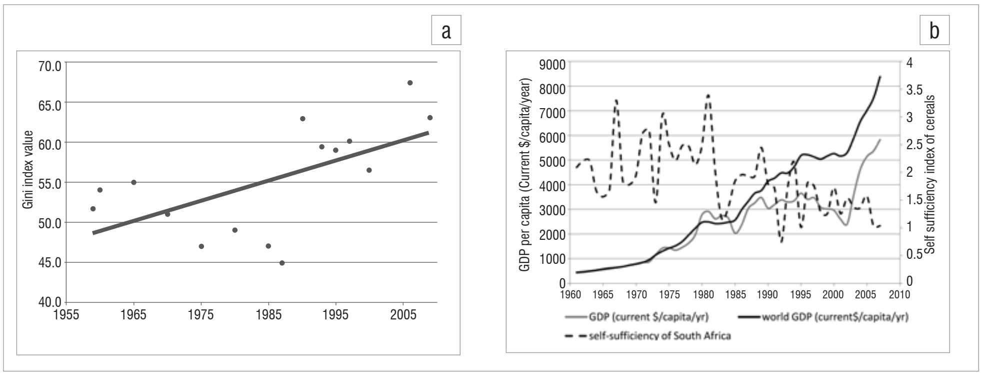
KEYWORDS:
dietary pattern; historical analysis; land requirements for food; South Africa; apartheid

HOW TO CITE:
Luan Y, Cui X, Ferrat M, Nath R. Dynamics of arable land requirements for food in South Africa: From 1961 to 2007. S Afr J Sci. 2014;110(1/2), Art. #2012-0088, 8 pages. http://dx.doi.org/10.1590/sajs.2014/20120088

Food consumption puts pressure on natural resources and arable land. In the present study, we examined the dynamics of land requirements for food in South Africa from 1961 to 2007 and investigated the relationships between dietary patterns, yield, cropping intensity, population and the area of required land using the thought experiment method. Strong population growth and the development of agricultural technology (indicated by yield) accounted for more than a 2.5-fold increase in the total land requirements for food from 1961 to 2007. Before the 1990s, the increase in crop yields enabled constant land requirements, whereas, after the 1990s, the combined effect of agricultural technology and population growth, together with a small contribution from dietary changes, led to an increase in the land requirements for food. Our findings confirm that the variation in land requirements for food is a complex, non-linear function of agricultural production techniques, population growth and dietary patterns and show that the complex relationship between dietary pattern changes, and economic development challenges future predictions of land requirements for food in South Africa.

Introduction

Despite being the wealthiest country on the African continent, as evidenced by its striking growth in gross domestic production (GDP) from 1961 to 2007 (Figure 1), the Republic of South Africa ranks as one of the most socially unbalanced developing countries, with a Gini coefficient[1] of 63.1 in 2009 (Figure 1a). Food self-sufficiency (defined as total production divided by total consumption) in South Africa decreased significantly from above 2 in 1961 to approximately 1 in 2007, indicating that the future of food security in South Africa in terms of self-sufficiency is not optimistic.[2] The evaluation of the food security situation in South Africa has been the subject of considerable attention in the scientific community, with a particular focus on exogenous factors affecting food systems. For example, studies have shown how climate change (mainly in terms of precipitation and temperature variations) affects agriculture,[3,4] particularly crop yields, which are an indicator of productivity.[5,6] Other investigations have been conducted from the perspective of government policy and society[7,8]; possible solutions to mitigate food insecurity in some regions have been suggested but few studies have focused on consumption issues.[9,10]



Note: The line indicates the trend and the dots represent the data valued by different people or groups.

Data resources: UNU-WIDER World Income Inequality Database, Version 2.0c, May 2008.

Figure 1: Corresponding information about South Africa: (a) trend of the Gini index of South Africa and (b) trends of GDP and self-sufficiency in South Africa compared with the GDP world average.

The extent of suitable land for agriculture is limited by several internal and external factors, including physical conditions and competition with other types of land use. Climate variability also plays a major role in the decrease in agriculturally suitable land area.[11] Given the restrictions on land resources, determining the amount of land required for food production is essential. Productivity is determined by techniques focusing on factors such as yield and cropping intensity. Population size and consumption per capita also indirectly influence land requirements via the total food consumption of a country. It is therefore necessary to investigate the relationship between land requirements and all of these factors to determine the area of land required for food. Apart from the effects of environmental changes and natural hazards on crop yield and land resources, it is crucial to explore the possible pressure exerted on land resources by food consumption and population growth to provide a comprehensive evaluation of future trends in food security in South Africa.

Studies on the land requirements for food (LRF) have been conducted in other regions (e.g. the Netherlands),[10,12-14] and have revealed that wherever economic growth occurs, people turn to more nutrient-rich food, which requires more arable land.[12,13] However, the development pattern of LRF varies from region to region, and its drivers play different roles. A study focused on South Africa is necessary to help understand the food security situation and frame future government policy in this nation.

We aimed to analyse LRF in South Africa by investigating the individual effects of population, technology and dietary pattern on the development of LRF. We back-calculated the arable land requirements of each food item consumed per capita, conversion factors, yield and cropping intensity. In this process, we focused only on arable land requirements to meet the physical health demand without considering other uses such as export, seed, feed and fuel use.[14] We hope that this work will inspire other studies on the projection of future agricultural land use in South Africa by analysing the historical development of LRF in South Africa.

Methods

Study area

The Republic of South Africa is located at the southern tip of Africa. It is the 25th largest country in the world by area and the 24th most populous country with a population of approximately 50 million. South Africa is ranked as an upper-middle income economy and is the largest economy in Africa. However, approximately one-quarter of the population is unemployed and lives on less than USD1.25 a day, making South Africa one of the top 10 countries in the world with income inequality.[15] The Gini index is a measure of the inequality of a distribution.[1,16] South Africa's Gini index has been over 40 (the internationally recognised warning line) throughout most of the past 50 years (Figure 1a), which indicates the instability of South Africa's economy at the household level, as confirmed by other surveys.[17,18]

Population growth has been very rapid in South Africa, with a 2.7-fold increase from 1961 to 2007 (Table 1), including the influx of approximately 5 million illegal immigrants (pooled estimated number).[19-21] The agricultural industry has contributed approximately 10% to formal employment over the past 10 years[22] and contributes approximately 2% currently to the GDP of the nation[23]. The shares of trade in the GDP have increased markedly since the opening of the economy in 1994.[24]

Only 13.5% of the land in South Africa can be used for crop production, and only 3% is considered high-potential land because of the restriction on limited natural resources.[25] Over the past 50 years, cultivated land has increased by only one-third, resulting in a reduction of half of the arable land per capita in 2007 compared with that in 1961.[26] Although productivity is improving (as demonstrated by the yield of cereals), South Africa's self-sufficiency in cereals has dropped from approximately 2.5 to below 1 (Figure 1b).

Data sources and methods

The data used in this work are from the FAOSTAT data set and include information regarding food supply, food consumption, yield, import and production from 1961 to 2007.[26] A detailed description of the data source and the calculation methodology is presented below.

Linking food consumption to LRF

Our approach was based on the methodology described by Kastner and Nonhebel [27] who calculated the historical arable land requirements for food of the Philippines from 1910 to 2003. We used food supply data as an indication of food consumption and obtained land requirements for that consumption. The calculation of LRF is separated into two parts: the vegetal part ($\mathrm{LRF}_{vegetal}$) and the animal-based part (LRF_{animal}).

Calculating the vegetal part

The vegetal LRF was computed as follows:

$$\mathrm{LRF}_{vegetal} = \sum_{i=1}^{n} \frac{consumption_i * convertionFactor_i}{yield_i * croppingIntensity} \qquad \text{Equation 1}$$

where *i* stands for a crop item, *conversionFactor$_i$* is the index for each food item converted to its primary crop equivalent, and *yield$_i$* is the crop yield for each crop. *CroppingIntensity* was obtained from the ratio of the total harvested area to the total area of arable land and permanent cropland. *Consumption$_i$* is defined as the total consumption of each food item by the entire population in 1 year.

The details of each data set and the steps of the calculation are as follows:

(1) Consumption to primary crop equivalents

Food consumption, reflecting average dietary patterns, is translated to primary crop equivalents by the respective conversion factors for each food item. The food consumption data supplied by the food balance sheets of FAOSTAT[26] 'are constructed for primary crops, livestock, and fish commodities up to the first stage of processing in the case of crops and to the second (and sometimes the third) stage of processing in the case of livestock and fish products'[28]. Because this study was limited to arable land, we calculated food consumption only for 18 categories covering 80 types of food items, excluding aquatic food items. Referring to the study conducted by Kastner and Nonhebel[27], the original 18 categories were aggregated into six broad groups: (1) cereals, (2) fruits and vegetables, (3) sugar and sugar crops, (4) vegetable oil and oil crops, (5) other vegetal food items and (6) animal products.

Table 1: Overview of South Africa

	1961	1970	1980	1990	2000	2007
Population (millions)	17.85	22.50	29.08	36.79	44.76	48.84
Population density (person/ha)	0.15	0.19	0.24	0.30	0.37	0.40
Arable land and permanent crops (million ha)	12.88	13.21	13.25	14.30	15.71	15.45
Per capita (ha/person)	0.72	0.59	0.46	0.39	0.35	0.32
Permanent meadows and pastures (million ha)	89.15	82.97	81.42	82.50	83.93	83.93
Per capita (ha/person)	4.99	3.69	2.80	2.24	1.88	1.72
Per capita GDP (current USD /person)	432.53	793.45	2775.93	3048.41	2961.26	5819.64
Share of agriculture in GDP (%)	11.54	7.16	6.20	4.63	3.27	3.37

Note: Population and land use from FAO FAOSTAT data sets; GDP from http://data.worldbank.org/data-catalog/world-development-indicators.

Generally, there are two methods used to calculate crop equivalents. One method is based on caloric equivalents and the other on the extraction rate. The conversion factors we used in this paper were based on caloric equivalents and were derived from the data supplied by Kastner and Nonhebel[27]. The value for food items used in this paper refers to supply at the household level.[28] The processing of food products creates losses and rest streams. For instance, using the extraction rate of cane sugar in South Africa (11%[29]), 9.1 tons of sugarcane are needed to obtain 1 ton of cane sugar. However, the losses and rest stream created in this process usually have other functions, such as animal feed. Thus, in calculating this figure, duplicate calculation occurs and the calculated primary crop equivalents are not reasonable. A conversion factor based on caloric equivalents was therefore used in the present study. For example, to produce 100 g of sugar containing 373 kcal, 1243.3 g of sugarcane per 100 g containing 30 kcal is needed. This method avoids duplicate calculations and excludes waste.

(2) Primary crop equivalents to total LRF and LRF per capita

The total harvested area is the sum of the harvested areas of each food item. The area that can maintain one person's average food demand was thus calculated by dividing the harvested area by the cropping intensity and population.

The following formula for calculating cropping intensity was derived from Siebert et al.[30]:

$$CI = \frac{AH}{CE} \qquad \text{Equation 2}$$

where *CI* is the cropping intensity, *AH* is the area harvested and *CE* is the extent of cropland that is left temporarily fallow.

Calculating animal-based foods

The situation is more complex in calculating livestock products. In this work, we simplify the computation of LRF_{animal} as follows:

$$\mathrm{LRF}_{animal} = \frac{Consumption_{animal\ (kcal)} * LRF_{vegetal}}{Consumption_{vegetal\ (kcal)}} * 3 \qquad \text{Equation 3}$$

where $Consumption_{animal(kcal)}$ is the total animal-based food items consumed in a year, $Consumption_{vegetal(kcal)}$ is the total vegetal food items consumed in a year, and $\mathrm{LRF}_{vegetal}$ is the total land requirement for vegetal food items as calculated above.

Based on a number of studies associated with LRF for animal products,[10,14,30,31] we assumed that one calorie of food of animal origin requires three times the amount of arable land needed for an average calorie of food of plant origin. This assumption is relatively low for beef, which represents the main production livestock system in South Africa. However, crude assumption is sufficient for our purpose, which focuses on the trend of total LRF and the individual effects of different factors rather than exact numbers. Given this aim, the method of calculating land requirements for animal products is therefore less important. Unifying the methods used to calculate the land requirements for vegetal products and animal products will be helpful in performing further analyses.

Assumptions and conditions

The methodology outlined above is suitable for the purpose of our study but requires that several assumptions be made and conditions satisfied:

1. Food consumption excludes aquatic food that is not directly related to land resources.
2. With a large proportion of pasture and meadows, the livestock industry, as the backbone of South African agriculture,[31] is not intensively managed.
3. Although the systems of crop plantation and animal husbandry are different, we unify crop products and animal products into the same land-use types.

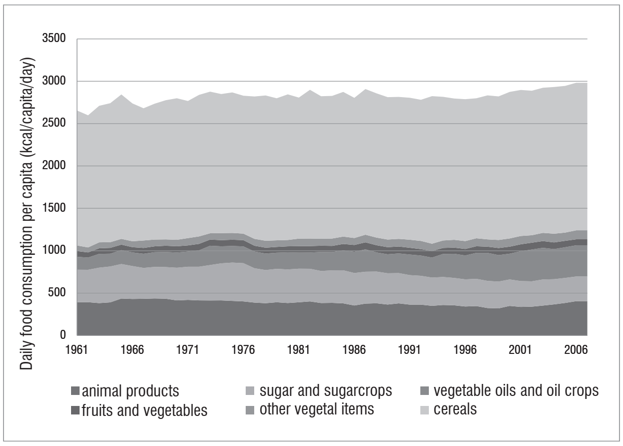
Individual effects of LRF factors

To study the individual effects of population, average dietary pattern, yield and cropping intensity on total LRF, we used the thought experiment method.[27,32,33] In this experiment, total LRF is a function of population, diet and technology. To visualise the impacts of these three drivers on LRF, we initially assumed constant population, diet and cropping intensity at the 1961 levels to qualify the impact of changes in yield only. The effect of cropping intensity was subsequently added. Then, the LRF induced by actual population development was taken into account, still assuming a constant 1961 diet. Finally, the impact of changes in the average diet was incorporated into the calculation to arrive at the total national LRF in South Africa throughout the entire period analysed. Through this method, the individual contribution of each driver can be assessed.

Results

Change in dietary pattern from 1961 to 2007

The food consumption per capita in South Africa from 1961 to 2007 for the listed categories is shown in Figure 2. Cereals, being the main dietary source, accounted for 58% or more of the total food consumption throughout the study period. Among all six categories, the most noticeable changes occurred in sugar and sugar crops and vegetable oils and oil crops. Whereas the proportion of sugar and sugar crops gradually decreased from over 14% in 1961 to below 10% in 2007, that of vegetable oils increased over the same period from 5.7% in 1961 to 12.3% in 2007. However, from 1961 to 2002, the share of animal products decreased from approximately 15% to 11.7% and then increased to 13.5% in 2007.



Source: FAO[26]

Figure 2: Changes in food consumption in South Africa from 1961 to 2007 divided into six main categories.

The total daily food consumption showed a slightly increasing trend from approximately 2670 kcal per capita per day to approximately 3000 kcal/cap per day. The increased consumption of vegetable oils and oil crops contributed most to this development with approximately 216 kcal/cap per day from this source. Cereals, as the second largest contributor with 148 kcal/cap per day, remained almost constant throughout. The only negative variation observed was the decrease in the share of sugar and sugar crops of approximately 92 kcal/cap per day over the 46-year period.

Overall, the dietary pattern in South Africa did not exhibit a significant change over the study period. The total consumption increased by only 12.3%, but that of food categories representing more nutrient-rich food items decreased. When standards of living are low, an increase in income favours the consumption of foods of animal origin and reduces the consumption of cereals and roots.[34] In terms of the absolute consumption per capita, the living standards in South Africa are not lagging far behind the world average.[26] However, the dietary pattern has not converted to the consumption of more nutrient-rich food items, as indicated by the decreasing trends observed for animal products and sugar and sugar crops.[13,35,36] We interpret the underlying driving factors to be income inequality[37] and unstable social development in South Africa (Figure 1b). The fourth migration wave, that is, the rural–urban migration (from the mid-1970s to the late 1980s) of Black populations, provided people with opportunities to live a more affluent lifestyle. The changes in population composition, as a result of more illegal immigrants and fewer White individuals, and the net emigration rate influenced dietary patterns at the country level.[24] The rapid growth in the GDP per capita and the higher levels of per capita food consumption overshadow the fact that large fractions of the population in South Africa are still living below the poverty line.[17,37] Poverty, lack of education, social instability and the long commuting distances of employed city dwellers have led many in South Africa to acquire an unhealthy dietary pattern.[38] Thus, the change in diet is the result of a synergy of changes in South Africa's low and high standards of living.[39,40] Moreover, although the dietary pattern did not change significantly over time, yield and population growth have significantly impacted the total LRF in South Africa.

Total LRF in South Africa from 1961 to 2007

The evolution of the total LRF in South Africa from 1961 to 2007 is shown in Figure 3a. The total LRF increased from 11.5 million ha to 29.2 million ha; cereals comprised the largest share, followed by animal products. Cereals, particularly maize, accounted for approximately half of the total LRF in 1961, a proportion that gradually decreased to below 40% in 2007 as a result of the increase in yields. For animal products, although the LRF increased significantly after 1990, the proportion remained almost constant over the 46-year period. However, LRF for vegetable oils and oil crops grew the most, almost 10-fold, accounting for 4.6% in 1961 and growing to 18.3% in 2007.

Another notable feature is that the change in LRF in South Africa appears to be highly unstable, with extreme peak values (Figure 3a). South Africa is unusually vulnerable to climate variability, and thus to a reduction in surface waters.[6] Maize, as the country's most important crop, accounts for 36% of South Africa's total field crop production. A serious decline in the productivity of maize and some other crops occurred because of a regional drought, resulting in a significant increase in the required land area, shown as peaks in Figure 3a. For example, as a consequence of the drought in 1992, maize yield declined from 2257.3 kg/ha to 785.3 kg/ha, causing this single crop's LRF to increase 5.4-fold (without considering changes in cropping intensity).

The changes in LRF per capita (Figure 3b) seem to follow a trend different from that of the total LRF itself. At the outset, LRF per capita first decreased from 0.65 ha/person per year in 1961 to 0.34 ha/person per year in 1989 because of increasing crop yields. The subsequent increase in LRF, which reached 0.59 ha/person per year in 2007, was the result of a reduction in cropping intensity and changes in dietary structure.

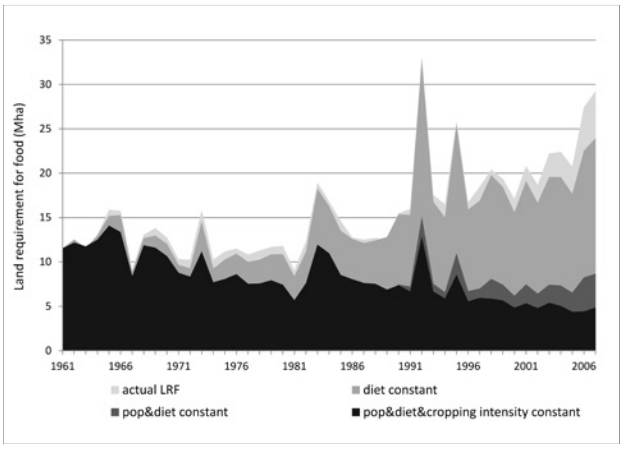
Looking at the contribution of different food items, the yield increase benefitting from the Green Revolution[41] contributed to a distinct decrease in LRF per capita for cereals from nearly 0.35 ha/person per year to approximately 0.2 ha/person per year, whereas for vegetable oils and oil crops, the LRF per capita increased from 0.03 ha/person per year to 1.11 ha/person per year over the study period – an increase which can be attributed to the combined impact of increasing consumption and increasing yield. For sugar and sugar crops, decreasing yield combined with declining cropping intensity led to an increase in land requirements.

LRF cannot reflect the actual amount of land required; nonetheless, the upward trend of LRF indicates a probable future increase in the amount of land needed to feed people, especially considering the growth in the population (Table 1).

Individual effects of population, diet and yield on LRF

The combined effects of population development, dietary change, cropping intensity and productivity on LRF were discussed above. The available data are sufficient to assess the individual effects of these drivers.

Based on the thought experiment method,[33] it is possible to quantify the effect of individual drivers on total LRF by keeping certain input factors constant (Figure 4). With solo-crop yield improvements, LRF decreased from over 11 million ha in 1961 to 6.6 million ha in 1980. However, LRF became relatively stable after 1980, mainly because of the decline in cropping intensity. The extremely high values of LRF coincided with extremely low yields, indicating the influence of the latter. Considering population growth, LRF showed a decreasing trend from 1961 to 1980 but at a much flatter rate than before. This behaviour clearly indicates that the increase induced by constant population growth was counteracted by the increase in yield. After 1980, LRF greatly increased from 10.9 million ha/year to 23.9 million ha/year, implying a greater impact of population growth over this period.



Note: The upper boundary of the black area represents the hypothetical LRF if population, cropping intensity and diet are constant at the 1961 levels, thus accounting only for changes in yield; the top of the darker grey area shows the effects of cropping intensity development and the changes in agricultural technology development with yield; the top of the lighter grey area shows the effect of population development;and the top of the lightest grey area shows the effect of dietary change.

Figure 4: Effect of changes in yield, cropping intensity, population and diet on the land requirements for food (LRF) in South Africa, 1961–2007.

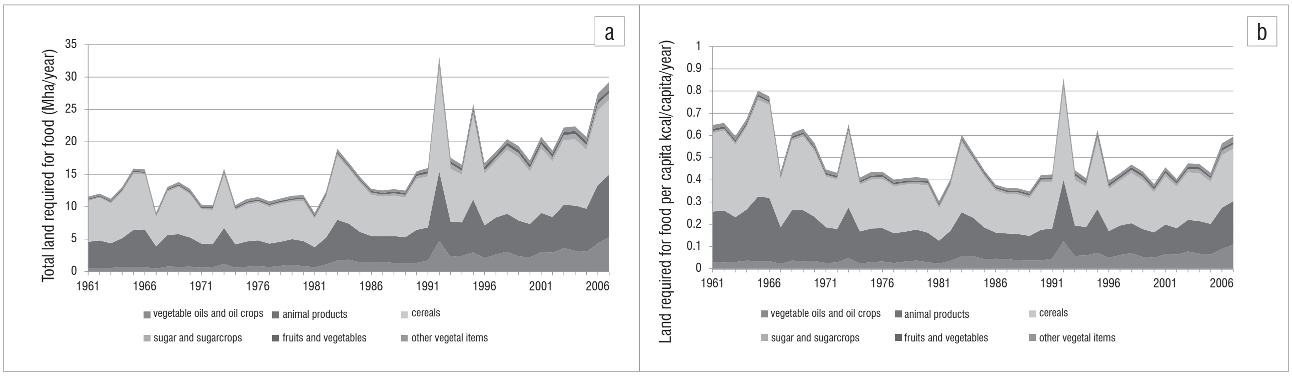


Figure 3: Land requirements for food in South Africa from 1961 to 2007: (a) total and (b) per capita.

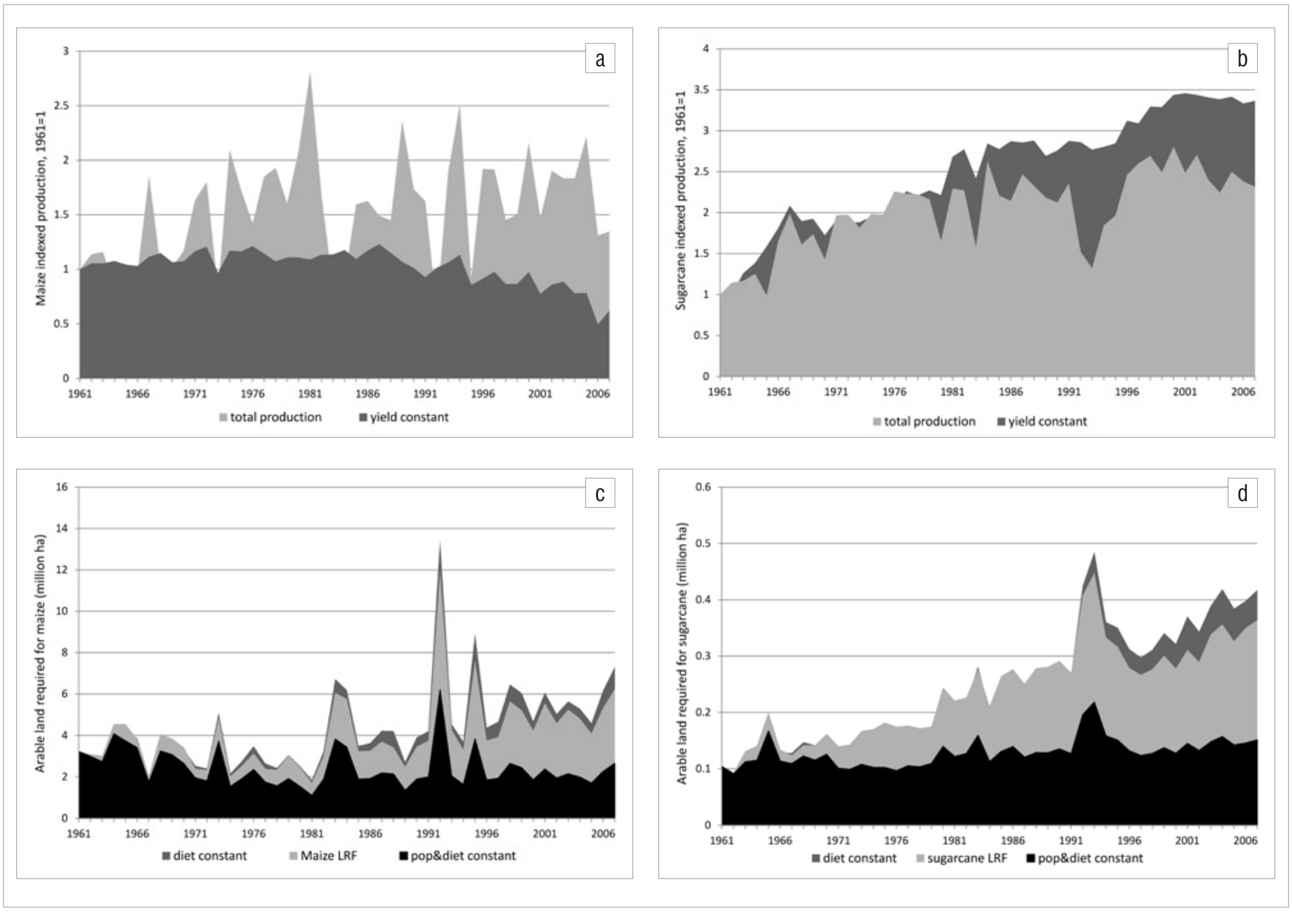
The effect of dietary pattern on LRF clearly changed after 1990. However, the underlying reason is much more complex than this trend suggests. During the apartheid era (e.g. 1961 to 1990), the populations categorised as 'non-Whites' in South Africa were severely restricted within major cities and metropolises in accordance with the 'dense settlement' patterns. Internal migration led to situations in which non-Whites and Whites worked under the same conditions, improving, to some extent, the standard of living of non-Whites; the peak of this phenomenon was reached in 1986.[42] After 1990, as discrimination was gradually abolished, a notable increase in the contribution of dietary pattern to LRF was observed.

Our thought experiment method revealed that different drivers are responsible for the calculated trajectory in LRF at different periods of time. Technology and population, being the crucial factors in the calculation of LRF, developed in two different directions. Agriculture intensification led to the decline of LRF at the outset; however, over the last three decades of the study period, the LRF stabilised. As a consequence of the demographic factor, LRF rapidly increased after 1980. Although dietary pattern seems to have played a minor role in determining the total LRF, the increasing trend over the last two decades of the study period characterises the importance of dietary pattern in future changes in LRF. The development of each driving factor is a result of complex systems that complicate further analyses. The uncertainties in technological development, the demographic transition caused by unforeseen policies and the influence of AIDS and the global economic and trade network make forecasting future LRF more difficult.

Differences between crops

The overall pattern of LRF is the sum of the LRF of various food items produced from different primary crops. However, the development of LRF for different crops varies. The available data for individual crop area, crop yield and cropping intensity allowed us to conduct a second thought experiment. In this section, we choose maize and sugarcane as case studies. Maize is the country's most important crop: it is a dietary staple, a source of livestock feed and an export crop. It accounts for 36% of the gross value of South Africa's field crops but has experienced the negative effects of climate variability. Although overall maize productivity increased, the harvested area declined over the 46-year period (from 4.1 million ha to 2.6 million ha). South Africa is the world's 10th largest producer of sugarcane, an important commercial crop. However, the yield of sugarcane continuously decreased over the 46-year study period. Assuming that production is a function of both yield and harvested area, we calculated the individual effects of these two drivers on maize and sugarcane. Based on Figure 5a and 5b, the increase in maize yield was counteracted by the decrease in maize harvest area. Changes in yield had a dramatic influence on the total production of maize, reflecting the instability of South Africa's maize production. The combination of the decrease in yield of sugarcane and the expansion of the harvested area led to relatively stable production of sugarcane, especially from 1985 onwards.

To examine the individual effects of population growth, yield and dietary pattern on the LRF for each crop, the thought experiment method was used once again.



Note: Development is in accordance with the changes in area and yield; as cropping intensity is not determined for each crop, it was not considered.

The upper boundary of the black area represents the hypothetical land requirements for food (LRF) if population and diet are constant at the 1961 levels; thus, only the changes in yields are accounted for; the top of the dark grey area denotes the effects of population development; and the top of the light grey area indicates the effect of dietary change, which is the total relative LRF.

Figure 5: Development of indexed dry matter and the land requirements of maize and sugarcane in South Africa from 1961 to 2007: (a) indexed development of dry matter maize production (1961=1); (b) indexed development of dry matter sugarcane production (1961=1); (c) relevance of changes in yield, population and diet in land requirements for maize; and (d) relevance of changes in yield, population and diet in land requirements for sugarcane.

The consumption of both maize and sugarcane decreased because of changes in dietary pattern (Figure 5c and 5d) – a finding which appears to be inconsistent with the development of total LRF for all food items. A reduction in maize consumption can be observed throughout the entire study period, whereas sugarcane consumption declined from 1990 onwards. Thus, the drivers of the changes in LRF have affected the LRF for individual crops and for the whole country in different ways.

Share of imports in total LRF

The procedure for calculating the LRF considers production from inside (produced locally) or outside (by importing) a nation's territory; thus, it is possible for us to assess the share of LRF from imports. We have already excluded production destined for export. In the calculation, net import production is also assigned to South African yield values because the underlying goal of our work is to determine how much domestic land is needed to meet domestic consumption.

The share of imports in LRF exhibited a significantly increasing trend (Figure 6). In the first three decades, imports contributed negatively to the LRF of the country, indicating that South Africa was mainly an export-based country. During this period, the combination of the development of technology and population growth led to only a slight increase in total LRF. However, from 1990 onwards, the share of imports increased from approximately 0 in 1990 to over 0.33 in 2007, indicating the nation's increasing dependence on imports. In other words, the increases in total LRF in recent decades required more land outside of South Africa's territory. Relevant studies indicate that South Africa's regional economic growth (Figure 1a) is clearly related to the shares of imports.[43] The nation's World Trade Organization membership and the opening of its economy in 1994[24] might have played a crucial role in this development.

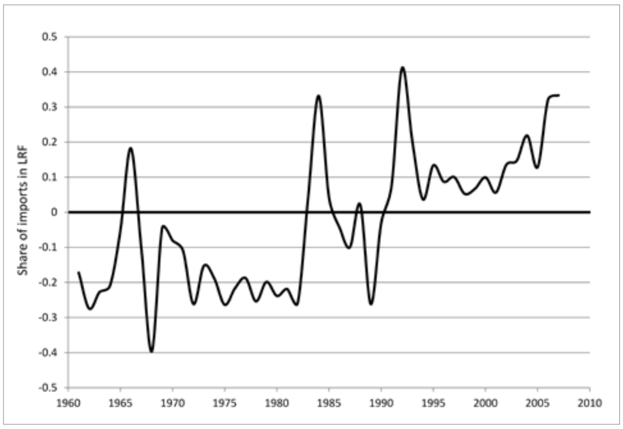


Figure 6: Share of the total national land requirements for food (LRF) met by net imports in South Africa from 1961 to 2007.

The variation in import shares is significantly associated with the fluctuation in total LRF. Imports accounted for a significantly higher proportion at peak points in some years more than in others. These peak years possibly correspond with the years of low production during droughts or social unrest, for example.[44] For instance, because of a regional drought in the early 1990s, maize production dropped to just over 3 million tons in 1992. Therefore, approximately 5 million tons of maize was imported during that year.

The trend of import share is negatively correlated with South Africa's self-sufficiency (-0.72), indicating that the nation increasingly depended on imports to meet LRF and that increases in total LRF in recent decades were mainly met by land outside South Africa's territory. Moreover, during years of drought, imports dramatically increased and accounted for a high share of total LRF. These two findings imply that South Africa has become more vulnerable to world economic and trade market fluctuations and that it has a weak climate disaster response capability. Under the scenarios of projected population growth in the future or the poor's living standards improving, the issue of South Africa's food self-sufficiency will become more serious. Based on previous research, the situation can be improved in two possible ways: importation and agricultural intensification. The former depends on internal and outer economic circumstances, and the latter would cause a series of environmental problems. Considering the negative impact of climate variability[45-47], land degradation[48-51], imbalanced domestic economic development and the wide disparity between the rich and the poor in this country, the consequences of both these solutions require further investigation.

Conclusions

In this study, we first analysed the dietary structure in South Africa from 1961 to 2007, based on FAO data, and found that the dietary pattern did not change towards more nutrient-rich food items as expected from the increase in GDP per capita. This phenomenon may be a result of the significant income gap in South Africa and the continuation of demographic transition. We then analysed the historical trend of South Africa's LRF, the contribution of different crops and the impact of individual factors on LRF. The results indicate that the total LRF in South Africa from 1961 to 2007 increased, whereas LRF per capita decreased before 1990 and increased thereafter. These trends indicate that in the near future the country may need more land to achieve self-sufficiency. The contribution of LRF differs for different crops. The arable land requirements for animal products as well as vegetable oil and oil crop consumption are the two main factors that contributed to the increase in total LRF.

Our historical assessment of LRF in South Africa reveals that LRF is driven by different factors at different times. Before 1980, increases in yield led to a slight decrease in LRF, reflecting the impacts of population growth. After 1980, the striking increase in population, combined with the decrease in cropping intensity, led to a continuing increase in LRF. In more recent decades, dietary pattern changes again led to an increase in LRF. These factors acted together to induce non-linear changes in LRF over the past 50 years. Thus, forecasting the relationship between the drivers of change and LRF will not be easy because it is not a simple linear relation. Finally, imports accounted for the increasing proportion of total LRF, especially since 1990. This result implies that the nation has increasingly depended on imports to meet LRF and that increases in total LRF in recent decades have been met mainly by land outside South Africa's territory, notably during disaster years. Such dependence makes South Africa vulnerable to the variation in global food price.

In this work, the calculated LRF is lower than the actual arable land needed in South Africa because it does not account for losses during food processing or seed need and assumes that the land will produce three calories of vegetal food to produce one calorie of animal food. Moreover, we employed a single cropping intensity for all crops. Given these limitations, caution is required in considering the calculated LRF to be reflective of the actual arable land area needed.

Acknowledgements

We are grateful for the comments of the editor and anonymous reviewers which greatly improved the quality of the manuscript. This work was supported by the National Basic Research Development Program of China (Grant Nos. 2011CB952001 and 2010CB428502), National Science Foundation of China (Grant No. 41271542), the Program for New Century Excellent Talents in University (Grant No. NCET-09-0227), the Fundamental Research Funds for the Central Universities and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

Authors' contributions

All authors contributed to the writing of the manuscript; Y.L. performed the calculations and wrote the body of the manuscript; M.F. and R.N. modified the manuscript's structure and English expression; and X.C. was responsible for the project.

References

1. Yitzhaki S. Relative deprivation and the Gini coefficient. Q J Econ. 1979;93(2):321–324. http://dx.doi.org/10.2307/1883197

2. Labadarios D, McHiza ZJR, Steyn NP, Gericke G, Maunder EMW, Davids YD, et al. Food security in South Africa: A review of national surveys. Bull World Health Organ. 2011;89(12):891–899. http://dx.doi.org/10.2471/BLT.11.089243

3. Benhin JKA. South African crop farming and climate change: An economic assessment of impacts. Global Environ Chang. 2008;18(4):666–678. http://dx.doi.org/10.1016/j.gloenvcha.2008.06.003

4. Gbetibouo GA, Hassan RM. Measuring the economic impact of climate change on major South African field crops: A Ricardian approach. Global Planet Change. 2005;47(2–4):143–152. http://dx.doi.org/10.1016/j.gloplacha.2004.10.009

5. Walker NJ, Schulze RE. Climate change impacts on agro-ecosystem sustainability across three climate regions in the maize belt of South Africa. Agr Ecosyst Environ. 2008;124(1–2):114–124. http://dx.doi.org/10.1016/j.agee.2007.09.001

6. Blignaut J, Ueckermann L, Aronson J. Agriculture production's sensitivity to changes in climate in South Africa. S Afr J Sci. 2009;105(1–2):61–68.

7. Bryan E, Deressa TT, Gbetibouo GA, Ringler C. Adaptation to climate change in Ethiopia and South Africa: Options and constraints. Environ Sci Policy. 2009;12(4):413–426. http://dx.doi.org/10.1016/j.envsci.2008.11.002

8. Baiphethi MN, Jacobs PT. The contribution of subsistence farming to food security in South Africa. Agrekon. 2009;48(4):459–482. http://dx.doi.org/10.1080/03031853.2009.9523836

9. Wirsenius S, Azar C, Berndes G. How much land is needed for global food production under scenarios of dietary changes and livestock productivity increases in 2030? Agr Syst. 2010;103(9):621–638. http://dx.doi.org/10.1016/j.agsy.2010.07.005

10. Gerbens-Leenes W, Nonhebel S. Food and land use. The influence of consumption patterns on the use of agricultural resources. Appetite. 2005;45(1):24–31.

11. Zhang X, Cai X. Climate change impacts on global agricultural land availability. Environ Res Lett. 2011;6(1), Art. 014014, 8 pages.

12. Gerbens-Leenes PW, Nonhebel S, Krol MS. Food consumption patterns and economic growth. Increasing affluence and the use of natural resources. Appetite. 2010;55(3):597–608. http://dx.doi.org/10.1016/j.appet.2005.01.011

13. Godfray HCJ, Crute IR, Haddad L, Lawrence D, Muir JF, Nisbett N, et al. The future of the global food system. Philos T Roy Soc B. 2010;365(1554):2769–2777. http://dx.doi.org/10.1098/rstb.2010.0180

14. Kastner T, Rivas MJI, Koch W, Nonhebel S. Global changes in diets and the consequences for land requirements for food. Proc Natl Acad Sci USA. 2012;109(18):6868–6872. http://dx.doi.org/10.1073/pnas.1117054109

15. Distribution of family income - Gini index [document on the Internet]. No date [updated 2008 Jan 24; cited 2012 Nov 29]. Available from: http://data.worldbank.org/indicator/SI.POV.GINI

16. Sadras V, Bongiovanni R. Use of Lorenz curves and Gini coefficients to assess yield inequality within paddocks. Field Crops Research. 2004;90(2–3):303–310. http://dx.doi.org/10.1016/j.fcr.2004.04.003

17. Rose D, Charlton KE. Prevalence of household food poverty in South Africa: Results from a large, nationally representative survey. Public Health Nutr. 2002;5(3):383–389. http://dx.doi.org/10.1079/PHN2001320

18. Higgs NT. Measuring and understanding the well-being of South Africans: Everyday quality of life in South Africa. Soc Indic Res. 2007;81(2):331–356. http://dx.doi.org/10.1007/s11205-006-9012-3

19. Posel D. Have migration patterns in post-apartheid South Africa changed? J Interdisc Econom. 2004;15(3/4):277–292.

20. Solomon H. Contemplating the impact of illegal immigration on the Republic of South Africa. J Contemp Hist. 2001;26(1):1–29.

21. Fin24. More illegals set to flood SA [homepage on the Internet]. c2006 [cited 2012 Apr 13]. Available from: http://www.fin24.com/Economy/More-illegals-set-to-flood-SA-20061123.

22. Banerjee A, Galiani S, Levinsohn J, McLaren Z, Woolard I. Why has unemployment risen in the New South Africa? Economics of Transition. 2008;16(4):715–740. http://dx.doi.org/10.1111/j.1468-0351.2008.00340.x

23. South Africa Yearbook 2012/2013. 20th ed. Pretoria: Government Communication and Information System; 2013.

24. Arora V. Economic growth in post-apartheid South Africa: A growth-accounting analysis. In: Nowak M, Ricci LA, editors. Post-apartheid South Africa: The first ten years [Internet]. Washington DC: International Monetary Fund; 2005. p. 13–22.

25. Mohamed N, Cousins B, editors. Greening land and agrarian reform: A case for sustainable agriculture. Proceedings: At the crossroads: Land and agrarian reform in South Africa into the 21st century; 1999 July 26–28; Broederstroom, South Africa. Cape Town: Institute for Poverty, Land and Agrarian Studies (PLAAS); 2000.

26. Food and Agriculture Organization. FAOSTAT statistical database [database on the Internet]. c2010 [cited 2012 Nov 29]. Available from: http://faostat.fao.org/.

27. Kastner T, Nonhebel S. Changes in land requirements for food in the Philippines: A historical analysis. Land Use Pol. 2010;27(3):853–863. http://dx.doi.org/10.1016/j.landusepol.2009.11.004

28. Food and Agriculture Organization (FAO). Food balance sheet – A handBook. Rome: FAO Electronic Publishing Policy and Support Branch; 2001. p. 99.

29. Food and Agriculture Organization (FAO). Technical conversion factors for agricultural commodities. Rome: FAO; 2003.

30. Siebert S, Portmann FT, Doll P. Global patterns of cropland use intensity. Remote Sensing. 2010;2:19.

31. Palmer T, Ainslie A. Country pasture/forage resource profiles – South Africa [homepage on the Internet]. c2006 [cited 2012 Apr 13]. Available from: http://www.fao.org/ag/AGP/AGPC/doc/Counprof/southafrica/southafrica.htm#4.

32. Ehrlich PR, Holdren JP. Impact of population growth. Science. 1971;171(3977):1212–1217. http://dx.doi.org/10.1126/science.171.3977.1212

33. Brown JR, Fehige Y. Thought experiments. Stanford Encyclopedia of Philosophy [serial on the Internet]. c1996 [updated 2011 Jul 29; cited 2012 Nov 29]. Available from: http://plato.stanford.edu/archives/fall2011/entries/thought-experiment/.

34. Grigg D. Income, industrialization and food-consumption. Tijdschr Econ Soc Ge. 1994;85(1):3–14. http://dx.doi.org/10.1111/j.1467-9663.1994.tb00669.x

35. Bruinsma J. World agriculture: Towards 2015/2030: An FAO perspective. London: Earthscan/James & James; 2003.

36. Popkin BM. The dynamics of the dietary transition in the developing world. In: Caballero BP, Popkin BM, editors. The nutrition transition: Diet and disease in the developing world. Amsterdam: Academic Press; 2002. p. 111–128. http://dx.doi.org/10.1016/B978-012153654-1/50008-8

37. Lind JT, Moene K. Miserly developments. J Dev Stud. 2011;47(9):1332–1352. http://dx.doi.org/10.1080/00220388.2010.514332

38. Bourne LT, Lambert EV, Steyn K. Where does the black population of South Africa stand on the nutrition transition? Public Health Nutr. 2002;5(1a):157–162. http://dx.doi.org/10.1079/PHN2001288

39. Temple NJ, Steyn NP. The cost of a healthy diet: A South African perspective. Nutrition. 2011;27(5):505–508. http://dx.doi.org/10.1016/j.nut.2010.09.005

40. Temple NJ, Steyn NP, Fourie J, De Villiers A. Price and availability of healthy food: A study in rural South Africa. Nutrition. 2011;27(1):55–58. http://dx.doi.org/10.1016/j.nut.2009.12.004

41. Sanchez PA, Denning GL, Nziguheba G. The African Green Revolution moves forward. Food Secur. 2009;1(1):37–44. http://dx.doi.org/10.1007/s12571-009-0011-5

42. Oosthuizen K. Demographic changes and sustainable land use in South Africa. Genus. 2000;56(3):81–107.

43. Narayan S, Narayan PK. Estimating import and export demand elasticities for Mauritius and South Africa. Aust Econ Pap. 2010;49(3):241–252. http://dx.doi.org/10.1111/j.1467-8454.2010.00399.x

44. Misselhorn AA. What drives food insecurity in southern Africa? A meta-analysis of household economy studies. Global Environ Chang. 2005;15(1):33–43. http://dx.doi.org/10.1016/j.gloenvcha.2004.11.003

45. Schlenker W, Lobell DB. Robust negative impacts of climate change on African agriculture. Environ Res Lett. 2010;5(1), Art. 014010, 8 pages.

46. Muller C, Cramer W, Hare WL, Lotze-Campen H. Climate change risks for African agriculture. Proc Natl Acad Sci USA. 2011;108(11):4313–4315. http://dx.doi.org/10.1073/pnas.1015078108

47. Lobell DB, Burke MB, Tebaldi C, Mastrandrea MD, Falcon WP, Naylor RL. Prioritizing climate change adaptation needs for food security in 2030. Science. 2008;319(5863):607–610. http://dx.doi.org/10.1126/science.1152339

48. Smith P, Gregory PJ, Van Vuuren D, Obersteiner M, Havlik P, Rounsevell M, et al. Competition for land. Philos T Roy Soc B. 2010;365(1554):2941–2957. http://dx.doi.org/10.1098/rstb.2010.0127

49. Ramankutty N, Foley JA, Norman J, McSweeney K. The global distribution of cultivable lands: Current patterns and sensitivity to possible climate change. Global Ecol Biogeogr. 2002;11(5):377–392. http://dx.doi.org/10.1046/j.1466-822x.2002.00294.x

50. Du Preez CC, Van Huyssteen CW, Mnkeni PNS. Land use and soil organic matter in South Africa 1: A review on spatial variability and the influence of rangeland stock production. S Afr J Sci. 2011;107(5–6):27–34.

51. Du Preez CC, Van Huyssteen CW, Mnkeni PNS. Land use and soil organic matter in South Africa 2: A review on the influence of arable crop production. S Afr J Sci. 2011;107(5–6):35–42.

King Penguins (Aptenodytes patagonicus) *on Trypot Beach, Marion Island. In an article on page 3, Ansorge et al. report that the Prince Edward Islands support ecosystems that are extremely sensitive to perturbations and provide an ideal natural laboratory for studying how these ecosystems respond to a changing climate (photo: Anne Treasure).*

APPLYING SCIENTIFIC THINKING IN THE SERVICE OF SOCIETY

Our vision is to be the apex organisation for science and scholarship in South Africa, internationally respected and connected, its membership simultaneously the aspiration of the country's most active scholars in all fields of scientific enquiry, and the collective resource for the professionally managed generation of evidence-based solutions to national problems.

ASSAf
ACADEMY OF SCIENCE OF SOUTH AFRICA

T +27 12 349 6600/21/22 | F +27 86 576 9514

WWW.ASSAF.ORG.ZA