SOUTH AFRICAN
Journal of Science

volume 112
number 11/12

Hassina Mouri
Department of Geology,
University of Johannesburg

Johann Mouton
Centre for Research on Science and
Technology, Stellenbosch University

Maano Ramutsindela
Department of Environmental &
Geografihical Science, University of
Cape Town

Cover caption
The seafloor around the
Prince Edward Islands in the
Southern Ocean. The image
was taken by Dr Charles
von der Meden from the
South African Environmental
Observation Network (SAEON)
using SAEON's SkiMonkey III
deep-sea camera. Dr Von
der Meden led the 2015 relief
voyage of the research vessel
SA *Agulhas II*, during which
Olsen and colleagues collected
marine sponges for analysis of
cytotoxic activity (page 111).

Research Article

Research Letter

Resolving fractured debates about fracking? The shale gas industry in South Africa

Lancashire, England:

> *Preston New Road Action Group, a local anti-fracking group, said it was 'devastated' by the decision* [to go ahead with fracking]. *'This is a sad day as it is clear to all that this* [UK] *government neither listens, nor can it be trusted, to do the right thing for local communities. It is deplorable that a* [fracking] *industry that has been rejected on every level has inflicted itself on Preston New Road,'* *said Pat Davies, the group's chair.*[1]

Eastern Cape, South Africa:

> *Chief Khomotsoana Lebenya has made a solemn vow to his more than 20 000 subjects: fracking will not happen on his land. 'Fracking will not happen here. It will not happen, I promise you. We will chase them.'*[2]

The extraction of shale gas is often said to have considerable benefits as an alternative and relatively new source of energy and as a creator of jobs, with the consequent social and economic advantages that follow. Why, then, would communities as different and geographically far apart as those in Lancashire and the Eastern Cape be so determinedly opposed to fracking? The reasons are numerous: the extraction process is water intensive in the face of water scarcity; the potentially carcinogenic chemicals used in the process may escape and contaminate ground water; air pollution is also common; heavy transport (for equipment and water) will have substantial environmental impacts; and the activity can ruin valuable or tourist-intensive landscapes. In short, the extraction of shale gas may well present significant environmental, technical, social and economic challenges – along with any benefits that it might bring.

It is for these reasons that the Department of Science and Technology (DST) approached the Academy of Science of South Africa (ASSAf) to undertake an assessment of South Africa's technical readiness to support hydraulic fracturing. ASSAf responded by using the well-tested and effective approach of establishing a consensus study in which a panel of experts undertook the various tasks needed to help to answer the core question. ASSAf launched the resulting consensus report[3] on 12 October – which has resulted in a substantial number of media reports on the panel's findings and recommendations.

Not surprisingly, the report is extraordinarily comprehensive and covers, in detail, a wide range of background material. This material includes international perspectives, factors to be considered in the case of the Karoo (not forgetting possible impacts on astronomy), a detailed analysis and the presentation of factors and the elements of readiness to be considered in the production phase, conclusions that can be drawn, and all-encompassing recommendations.

Amongst the conclusions reached by the panel, the following are perhaps amongst the most critical. Firstly, there is a need for South Africa to assess the extent of technically recoverable shale gas resources and to commit to a balanced long-term gas exploitation strategy, taking account of the security of supply, efficiency of extraction, environmental protection and effective communication to society.

Secondly, it is essential that controls be identified and implemented regarding externality costs associated with mines and abandoned mines, and that these controls be in place even before the implementation phase begins. Third on the list is the need for a rigorous environmental impact assessment of both upstream and downstream shale gas processes and determination of the most economically, socially and environmentally optimal gas source. Fourthly, water availability and use, as well as the impact of methane emissions must be assessed and monitored – another facet of environmental, social and health assessment.

The fifth conclusion drawn by the panel focuses, appropriately enough, on the potential impacts of mining on the astronomical work in progress and moving forward in Sutherland. Fracking and its supporting activities present a real risk to the scientific operations and performance of the Square Kilometre Array (SKA) and its complementary research utilities and functions. The extreme sensitivity of the SKA means that even the weakest of human-made radio signals is detectable at some level, and in some part of the radio frequency spectrum across which the SKA will operate. To minimise the potential impact of this risk, careful management and coordination with stakeholders is needed, along with the establishment of safety limits.

Sixth, comes the critical matter of the social and economic impacts of, and implications for, the mining. So far, much of the focus at a broad level has centred on the wider economic impacts and benefits to the national economy and energy balance, but completely inadequate consideration has been given to the localised effects of the impacts and consequences that will be faced and experienced in local environments. Then too, comes the seventh set of conclusions, critical if for no other reason than that they have been neglected in South Africa for over a hundred years: what happens when the mining operations come to a material or economic end? Ensuring complete maintenance throughout the operational life of wells, and after their closure, must form an essential part of any shale gas mining operation.

The report sets out three further conclusions relating to baseline studies prior to implementation; to the distribution of the gas; and to the importance of capacity and related skills development. These are clearly of equal importance.

These conclusions lay out essential steps that must be taken and actions that must be implemented if the fracking goes ahead. There is little doubt that the DST will take them seriously, as it is a department fortunate enough to have excellent leadership. It is very clear, however, that the Ministries of Mineral Resources and of Energy appear not to benefit from the same quality of leadership or commitment to good, honest practice.

Which brings us decisively back to the community activists. Should fracking proceed, to a greater or lesser degree, then the panel's conclusions (and recommendations) must be the essential, unchallenged foundations for the process. In this case, the activists still have an unquestionably critical (and possibly even more important) role to play: that of vociferously and persistently holding the state and its various arms to full account.

References

1. Vaughan A. Fracking given UK go-ahead as Lancashire council rejection overturned. The Guardian. 2016 October 06;Environment. Available from: https://www.theguardian.com/environment/2016/oct/06/uk-fracking-given-go-ahead-as-lancashire-council-rejection-is-overturned

2. Savides M. I'll make sure it doesn't happen – chief. Sunday Times. 2016 October 09;News. Available from: http://www.timeslive.co.za/sundaytimes/stnews/2016/10/09/Ill-make-sure-fracking-doesnt-happen---chief

3. Academy of Science of South Africa (ASSAf). South Africa's technical readiness to support the shale gas industry. Pretoria: ASSAf; 2016. http://dx.doi.org/10.17159/assaf.2016/0003

AUTHORS:
John Butler-Adam¹ [iD]
Nadine Wubbeling¹ [iD]

AFFILIATION:
¹Academy of Science of
South Africa, Pretoria,
South Africa

CORRESPONDENCE TO:
John Butler-Adam

EMAIL:
j.butler-adam@gmail.com

Associate Editor awarded Science-for-Society Gold Medal

Professor Brian van Wilgen received the Academy of Science of South Africa (ASSAf) Science-for-Society Gold Medal award at ASSAf's annual prestigious awards ceremony on 12 October 2016. Gold Medals are the apex awards of ASSAf and the South African science system and are awarded in recognition of outstanding achievements by individuals in scientific thinking for the benefit of society.

How to describe Brian van Wilgen? There are too many ways to be counted.

Forester; botanist; ecologist; specialist in invasive alien plants; expert in fynbos, grassland and savanna ecosystems; writer on the development of sound management strategies; professor... The list goes on.

But from the perspective of the *South African Journal of Science* (SAJS), most importantly, as a knowledgeable, incisive and constructive Associate Editor of Organismal Biology. Brian has been the Associate Editor: Organismal Biology since the inception of the role in 2008 until his resignation in October 2016.

Brian's management of the peer-review process has been efficient and effective – and, when necessary, persistent. His messages to authors, whether rejecting, conditionally accepting, or accepting a manuscript were pre-eminently supportive.

Brian studied Forestry at Stellenbosch University and graduated with a BSc degree in 1974 and an honours degree in 1977. He went on to complete a master's and then a doctoral degree in Botany at the University of Cape Town. He worked for the South African Forestry Research Institute and, when the Institute was incorporated into the Council for Scientific and Industrial Research (CSIR), he moved to the CSIR campus in Stellenbosch.

On retiring from the CSIR in 2013, he was appointed as a professor at Stellenbosch University and worked in the DST-NRF Centre of Excellence for Invasion Biology. By that time, Brian had been a member of the SAJS editorial team for 5 years and continued as an active member in 'retirement' for a further 3 years.

Brian is the author of over 250 publications, including three books. In SAJS, he has published three News and Views, nine Book Reviews, one Commentary and three Review Articles. He was also Guest Editor of the January 2004 special issue on South Africa's Working for Water Programme.

He has earned many scientific achievements and has received numerous awards. He was awarded a DSc degree by Stellenbosch University in 2008; he has also been awarded the National Science and Technology Forum award for an outstanding individual contribution to science, engineering and technology and the South African National Parks award for contributions to conservation, both in 2010.

Appropriately enough, in his generous and modest acceptance speech on being awarded the ASSAf Gold Medal, Brian confessed that, as a young boy, he knew exactly what he wanted to be: a game ranger. Fortunately, he pointed out, there were no degrees in game-ranging offered at the time, and so he studied forestry – which became the foundation for his career, and for his contributions to the SAJS.

Above all, Brian is a thoughtful, congenial and supportive colleague. The Journal is honoured to have had him as a long-standing member of its editorial team; we congratulate him on his well-deserved award of the Gold Medal – and we wish him well as he moves into the next stage of his active 'retirement'.

Brian van Wilgen receives the Science-for-Society Gold Medal from Dr Phil Mjwara, Director-General of the Department of Science and Technology, at ASSAf's annual awards ceremony on 12 October 2016.

ASSAf 1996–2016
Commemorative article

ASSAf and young scientists: Transforming the future of science in South Africa

AUTHOR:
Bernard Slippers[1]

AFFILIATION:
[1]Department of Genetics, Forestry and Agricultural Biotechnology Institute, University of Pretoria, Pretoria, South Africa

CORRESPONDENCE TO:
Bernard Slippers

EMAIL:
Bernard.Slippers@up.ac.za

Many academies globally interact with young scientists (viewed here as all those who have just completed postgraduate studies to mid-career scientists) only on an ad-hoc basis through prizes or once-off engagements. This substantial group of active knowledge producers is often not treated as a core part of the scientific community or as leaders who could shape science and society – despite the fact that they often drive new developments, and will be the leaders of the scientific community in as little as 5–10 years. It is often argued that they should be spending their time 'in the lab' and writing grants and papers, rather than busy themselves with the type of engagement and policy work that academies typically undertake. Consequently, the voices of the next generation of science leaders are often silent in policy discussions, even when the policy is about young scientists or about the future of science. It is thus not surprising that a recent special feature in *Nature* concluded: 'Academia is more difficult than ever for young scientists. That's bad for them, and bad for science.'[1]

This paradigm of interaction between academies of science and young scientists is changing in South Africa, thanks to visionary leadership in the Academy of Science of South Africa (ASSAf) over the last few years. ASSAf today is amongst the most active academies globally in their efforts to promote young scientists. There is no other body in South Africa that is currently doing more to provide platforms for their interaction and engagement, and to give a voice to this critical part of the science community.

While ASSAf has had a focus on recognising and supporting young scientists through annual Young Scientists' prizes together with The World Academy of Sciences (TWAS), the Department of Science and Technology or the African Union, and the Sydney Brenner fellowship, it has also developed a range of dedicated programmes over the last 6 years that touches a large cohort of young scientists across the country. Here I highlight some of the most significant of the activities through which ASSAf has supported young scientists.

International networks and platforms

A critical role that ASSAf has played for young scientists is 'upward connection or mentorship', by opening doors to top quality networks and organisations internationally. As an example, I first interacted with ASSAf in 2009 through a joint programme with the IAP: Global Network of Science Academies to send young scientist representatives to the annual World Economic Forum (WEF) meeting in China. As a direct result of this coming together of a global cohort of young scientists, which also included other South African scientists thanks to ASSAf, the Global Young Academy was launched in 2010. Soon after, the South African Young Academy of Science (SAYAS) was launched in 2011. In this way, ASSAf has ensured that a number of South African scientists have the opportunity to be involved at the forefront of stimulating the global Young Academy movement, which is now established in 23 countries, and with many more in development. South African young scientists continue to benefit annually from interaction with other young scientists from across the world at the WEF annual meeting in China, but also through linkage to various other global platforms, such as networks of academies of science, the Lindau Nobel Laureate meetings, international science fora, the International Network of Government Science Advisors, and many more.

ASSAf's influence extends beyond South African young scientists, through the large number of South Africans who have been involved with, and who have led the Global Young Academy, including hosting its second annual conference in South Africa in 2012. This reach is also particularly relevant in Africa where these fellows have been key partners to other Global Young Academy members in supporting the establishment and growth of other national young academies, including involvement in the Africa Science Leadership Programme, African Science and Society meetings, and continent-wide National Young Academy meetings (including one just completed in Mauritius), to name but a few.

Local networks and platforms

At a national level, ASSAf has been the key supporting organisation for the establishment and development of SAYAS. Following initial calls for support for this process, ASSAf convened a small committee of young scientists to drive the process; realising from the start that this process needed to be 'bottom up' and not 'top down'. ASSAf has continued to engage with SAYAS as an autonomous body. SAYAS operates independently from ASSAf, with a Memorandum of Understanding guiding their active and positive approach to partnership and engagement. Following the launch of SAYAS, ASSAf has also provided a physical home and administrative support for SAYAS through a shared appointment of a Liaison Officer. ASSAf also manages core funding to SAYAS from its parliamentary grant. As an active participant in management structures of SAYAS, I can attest to continuous and critical support from ASSAf without pressure or interference for the work and views of the members of SAYAS.

Such initiatives matter, not only for the individuals involved, but for young scientists in general in the country. Today SAYAS is the premier platform that provides 'a voice for young scientists' in South Africa, and that serves as a point of interaction with other young academies around the world, and with formal structures of government and science.

Young scientists are often best to serve as mentors and role models for other young scientists and school learners. By creating and supporting young scientist structures that can advance such goals, the impact of ASSAf reaches much further than the formal structures and programmes described above. For example, SAYAS is participating in projects such as '1000 Girls, 1000 Futures', science spazas that support and develop science clubs in underprivileged schools, studies on the experience of young scientists in South Africa, a PhD blog through which

students describe their journey through their PhD, and more. These projects reach cohorts of young South Africans that ASSAf would not have been able to do on its own.

An apt illustration of the unique interaction and perspectives that ASSAf and SAYAS bring are the recent statements from these organisations about the #FeesMustFall related events on campuses across the country. At the height of this crisis, with campuses being closed or engrossed in protests, there was a growing desire for these academies to give a perspective as the biggest collection of academics representing all institutions. The first statement was issued jointly from both organisations and captured a core message of concern and an offer to contribute towards solutions. While it was authoritative and strong, the statement could not address some of the more contentious issues. Soon thereafter at the General Assembly of SAYAS, the young scientists felt that there were further – and more specific – issues they wanted to address, and so issued a second statement. Together these two statements capture a richer perspective of the views of South African academia, and contribute to a more nuanced 'voice of scientists'.

Since 2010, ASSAf has hosted an Annual Young Scientist Conference, in partnership with the National Research Foundation, Department of Science and Technology and SAYAS. At these conferences, ASSAf has brought together a broad range of young scientists from across Africa to network and present their work to peers in a truly interdisciplinary setting. Importantly, these conferences have addressed a range of topical issues, from biodiversity, to the role of science in empowering women in Africa and, most recently, human rights. By engaging young scientists on these topics, ASSAf has impacted more than a 1000 future leaders in science. There is no doubt that these intersections and engagements will have deep and lasting impacts on the sensitivity and activism amongst the future leaders of the scientific community on these issues.

SAYAS and other young scientist engagements are important vehicles for ASSAf to promote diversity and transformation in the South African research community. Through this active engagement with young scientists from all institutions and groups in South Africa, and with a specific focus on underrepresented groups, ASSAf is providing the exposure and opportunity for a truly representative cohort of future leaders of science in the country to emerge.

Young scientists and ASSAf publications

ASSAf has an important national role as a custodian of key scholarly publications and in overseeing publication platforms. Herein lies opportunities to further engage meaningfully with young scientists. The *South African Journal of Science* has been an outlet for members of SAYAS from early on. Other than research articles, SAYAS members have been able to raise their voice through opinion articles on a variety of topics, from the reflections on science's role in sustainable development[2], to the research experience of young scientists in the country[3].

This is a feature that ASSAf can help enhance in future, possibly via a dedicated 'Future Voices' collection of opinion pieces on a regular basis (as *Science* does every few months through their Next Gen Voices feature). SAYAS inaugurates 10 new members every year – short opinion pieces from each of these leading young scientists on key developments in their field, especially giving a South African or African perspective, or on other topical issues (e.g. the impact of #FeesMustFall) could add rich insight into the direction of science development in the country.

Young scientists are also actively involved in another of ASSAf's publications, *Quest*. This high-quality 'popular science' magazine provides an ideal outlet for SAYAS to promote a closer interaction between science and society, which is one of its core objectives. *Quest* has covered numerous articles about the work of SAYAS members in the past 5 years. This relationship too could be explored even further through dedicated features, for example, by linking to the PhD blog run by SAYAS or by specific partnership during activities run by SAYAS.

Transforming the future

At 20, ASSAf is a young academy compared with many around the world. The next 20 years of its development will no doubt be directly influenced by the work it started amongst young scientists in the last few years. SAYAS members are already actively involved in various structures of ASSAf, providing a fresh perspective in standing committees and other structures. The pool of experienced young scientists from which to draw for these functions will continue to grow in coming years. This is an important outcome of ASSAf's efforts, especially in the light of the fact that such a resource was not available even 5 years ago.

Furthermore, the first SAYAS alumni have recently been incorporated into ASSAf (although this is not an automatic process), and we are likely to see a steady increase in such ASSAf members over the next few years. By the next 20-year mark, a substantial group within ASSAf is likely to have been members of SAYAS. These new members will bring with them experiences of different approaches from their involvement with young academy activities. Moreover, they represent a network of transformation-minded science leaders, ready to work with the ASSAf structures to engage society, our scientific community in South Africa and beyond.

Over the last few years, ASSAf has laid a foundation for the promotion, engagement and capacity development of young scientists. If these activities and platforms are supported (and expanded) into the future, they have the potential to transform the research landscape in South Africa, through a more engaged, representative and transformation-minded scientific community.

References

1. Early-career researchers need fewer burdens and more support. Nature. 2016;538:427. http://dx.doi.org/10.1038/538427a

2. Slippers B, Kassen R. Young scientists reflect on how to effect real change for Rio+40. S Afr J Sci. 2012;108(7/8), Art. #1340, 2 pages. http://dx.doi.org/10.4102/sajs.v108i7/8.1340

3. Schutte AE, Wright CY, Langdon G, Lochner C, Myers B. What is the research experience of young scientists in South Africa? S Afr J Sci. 2013;109(11/12), Art. #a0040, 2 pages. http://dx.doi.org/10.1590/sajs.2013/a0040

Sex and gender transformation in Africa

AUTHOR:
Zak Yacoob[1]

AFFILIATION:
[1]Retired Judge of the Constitutional Court, Johannesburg, South Africa

CORRESPONDENCE TO:
Zak Yacoob

EMAIL:
zak.yacoob@gmail.com

On 6 and 7 October 2016, the Academy of Science of South Africa (ASSAf) hosted the 2016 Annual Young Scientist Conference in partnership with the Department of Science and Technology; the South African Young Academy of Science; Gender in Science, Innovation, Technology and Engineering (GenderInSITE); and the South African Chapter of the Organization for Women in Science for the Developing World. The conference explored and discussed human rights in Africa in many of its dimensions to celebrate 2016 as the 'African Year of Human Rights with Particular Focus on the Rights of Women'.

A highlight of the event was a roundtable session on 'Transformation towards sex and gender equality in Africa: Where are we?' The roundtable was facilitated by Justice Richard Goldstone, retired judge of the Constitutional Court of South Africa, who chaired a panel of four people: Prof. Christof Heyns, a member of the United Nations Human Rights Committee; Ms Janet Love, National Director of the Legal Resource Centre; Prof. Barney Pityana, former Chairperson of the South African Human Rights Commission; and myself, a retired judge of the Constitutional Court.

This contribution sets out my own reflections of the roundtable and what I learnt from it. Sex and gender transformation is really about more than just the achievement of true equality between women and men and the elimination of discrimination against and the empowerment of women in society. The ultimate object is the achievement of equality of all of us as human beings regardless of our sexual orientation or gender identity. All of us should recognise as equal and respect all human beings. This is a difficult area of transformation, more especially in the light of entrenched attitudes and societal assumptions, together with our socialisation and religious beliefs or convictions. It was acknowledged, as a matter of context, that mass rape and all forms of violence against women, including sexual violence, are no longer classified merely as crimes of war or crimes against humanity; these are in essence crimes against women arising from societal discrimination, or (perhaps more accurately) discrimination by men against women.

A fundamental premise of universal human rights thinking today is that discrimination based on morally irrelevant considerations or immutable human characteristics of race, sex, religion, age, disability and the like cannot be justified and are unacceptable. This abiding and undeniable principle applies with equal logical and ethical force to discrimination on the grounds of sexual orientation and gender identity against people who are lesbian, gay, bisexual, transgender or intersex (LGBTI individuals).

Non-discrimination on these grounds lies at the heart of the worldwide human rights project, and is also central to the dignity, equality and freedom of people in Africa.

Discrimination based on sexual orientation and gender identity comes in many forms and levels of intensity. It may render people in this category subject to varied punishment in support of the prohibition – the nature of the punishment being as extreme as the death penalty which may still be imposed for same-sex relationships in a small number of African countries. Prejudice against LGBTI individuals also manifests itself when serious criminal violence deliberately perpetrated on such individuals is not properly investigated, if at all. This is usually because law enforcement authorities believe that people in this category are undeserving of and therefore not entitled to protection of their right to bodily integrity. Authorities are even more reluctant to investigate verbal abuse and infringements of dignity, which come from official or private sources, with the acquiescence of the state. Access to socio-economic rights like the right to education, housing or health care may in reality be out of reach of this category of person for the same reasons. Violence against transgender and intersex persons is underpinned by societal stigma, sex-binarism, transphobia and intersexphobia. It is driven by confusion and conflation of pre-conceived notions of an individual's gender identity, gender expression and/or physical sex (bodily) characteristics.

There are of course mechanisms and instruments available to counter these abuses both at the national and international level. For us the resolution of the African Commission on Human and Peoples' Rights on protection against violence in cases of this kind, merits special attention. And the recent appointment by the United Nations Human Rights Council of Vitit Munarbhorn as an independent expert on sexual orientation and gender identity is of considerable significance.

Scientific studies on LGBTI are not only helpful but essential to improve our understanding of humanity. But the purpose of scientific investigation into and evaluation of LGBTI individuals should not be to pathologise or justify LGBTI. As a natural part of the world's biological diversity and variety, different gender identity should need no justification. 'If you hear someone say that homosexuality is unnatural, you can be pretty sure you are not listening to a scientist' (Dr Marc Breedlove, Rosenberg Professor of Neuroscience, Michigan State University).

Attempts to pathologise have perhaps led to pseudoscience that could result in undue bias and discrimination. For example, subjecting people like our Olympic athlete Caster Semenya to all sorts of abuse and discrimination is a poignant reminder of this brutal truth. Similarly, we should reject the idea (of making the apologetic justificatory argument) that LGBTI people deserve equal rights because they are 'born this way'. In a free and democratic society founded on the prevalence of human rights, it should not matter whether one's sexual orientation is inborn or a choice. The right to dignity, equality and freedom embrace the right to make that choice.

On 24 February 2014, Uganda's president, Yoweri Museveni, signed a draconian Anti-Homosexuality Bill into law after 2 months of declining to do so. Misusing science to justify the Ugandan anti-gay law, he said he changed his mind following the findings of a special scientific committee his Health Ministry had appointed earlier in the month: 'Their unanimous conclusion was that homosexuality, contrary to my earlier thinking, was behavioural and not genetic.' However, one of the members of this committee stated that the report had not made that finding.

'They misquoted our report', said Paul Bangirana, a clinical psychologist at Makerere University in Kampala. 'The report does not state anywhere that homosexuality is not genetic, and we did not say that it could be unlearnt.' The point though is that even if scientists had concluded that the LGBTI element of humanity was behavioural, there was no rational basis to prohibit it.

Gender diversity and bodily diversity (particularly intersex variations) are still misunderstood and often confused with sexual orientation in South Africa and many other countries.

As a result of cisnormative conceptions of gender, many transgender women are still classified as 'gay men' and transgender men as 'lesbian women'. Violence against transgender men is often unreported and unpunished, or conflated and misreported in 'corrective rape' statistics, often framed as an issue solely in the lesbian community, particularly in townships.

Human rights violations against intersex persons in medical settings continues because the diagnostic nomenclature and language used in medical and public discourse to name, classify, describe and understand intersex bodies are stigmatising and pathologising.

The Constitution of our country and those of many others prohibit discrimination on numerous grounds including race, gender, sexual orientation, disability and age. It is necessary to consider whether societal approaches and attitudes in South Africa are consistent with what our Constitution says. The evaluation is important because legal instruments in other countries in and outside Africa embrace these protections in their constitutions or other legal instruments. The question inevitably arises whether legal enactments or provisions, although undoubtedly necessary and positive, are enough to create a truly egalitarian society in which all human beings are equal. South Africa can be used as an example.

Our Constitution says in effect that men and women are equal. Yet, experience tells that a vast majority of men regard themselves as superior to women. It is even more significant that most women would look up to men on the basis of an implicit understanding that men are stronger and more powerful than they are. We are sometimes very much our own oppressors.

As far as race equality is concerned, it is worth noting that we have talked for many years about the transformation of the legal profession. One of the ways to achieve this is to ensure that black lawyers get more paying work. This has not yet happened in a substantial sense. A black political leader has said that he thinks that white lawyers are better than black ones, and that he has a white lawyer. So does the overwhelming majority of black people in relatively large wealthy corporates. Self-oppression again?

We can now look at the position of the LGBTI community. Our Constitution prohibits any unfair discrimination against any person by the state or another person on the grounds, amongst others, of sexual orientation. The Constitutional Court of our country has held that gay and lesbian people should be allowed to marry and that consensual homosexual conduct is not a crime. Yet it cannot be refuted that many in our country still believe that gay and lesbian people who relate sexually to each other do so in sin.

In South Africa, as well as in the rest of Africa and much of the world, there is a disjuncture or disconnect between what constitutions or other legal instruments provide and what the courts say on the one hand, and the status of the hearts and minds of people on the other. It has been suggested that our Constitution is out of step with the hearts and minds of our people. Does this mean that our Constitution is worth nothing? Certainly not.

The Constitution is a vision for the kind of society that all of us need to work together to create and nurture. Indeed, the Constitution imposes on all of us a duty to actively contribute to the kind of society contemplated therein. This has implications for the conduct of all human beings, lawyers as well as social and natural scientists alike. All of us must know the Bill of Rights contained in our Constitution. Each of us must understand, internalise, embrace and ultimately practise its values. And then we need to propagate those values so that in time more and more people embrace them. We cannot assume that these values are understood and known by all, not even by lawyers who might profess to do so.

And then there is work for us to do as scientists. There is need for multi-disciplinary research to establish the exact causes or reasons, or the factors that contribute towards them, for the prejudices in relation to sex and gender of the people of Africa and others. We can then devise a scientific way of changing the hearts and minds of all of us so that our thinking accords with the values of our Constitution.

Although it is impossible to say with precision where we are on the road to sex and gender transformation in Africa, it would seem that we have not gone very far, and that the road ahead is long and arduous. Lawyers and the law cannot achieve this change alone. All of us need to contribute to a social revolution to achieve this result.

AUTHOR:
Hannah R. Simon[1,2]

AFFILIATIONS:
[1]Department of Biological Sciences, University of Cape Town, Cape Town, South Africa

[2]Department of Environmental and Geographical Sciences, University of Cape Town, Cape Town, South Africa

CORRESPONDENCE TO:
Hannah Simon

EMAIL:
hannah.r.simon@gmail.com

KEYWORDS:
conservation; social activism; environmental justice; poverty; green culture

Understanding the polarisation of environmental and social activism in South Africa

Heralded as the decade of social awareness, the start of the 21st century has brought with it a growing cognisance of what is good and right. Through time, the world has birthed different schools of justice, which have all added their flavours to the hypothetical cocktail of human well-being. Two of these schools – *social justice* and *environmental justice* – have attracted the most attention.

The United Kingdom's Department of Forestry has deemed social justice the 'shar[ing] of social, environmental and economic benefits' and environmental justice the 'distribution of environmental benefits'[1]. When either of these justices are not met, waves of resistance – so-called social and environmental activism, respectively – invariably arise. Social activism effectively caters directly to the needs of people, whilst environmental activism prioritises the state of the environment.

Theoretically, a country should be able to protect a society along with its environment. In fact, in many countries, governments, citizens and policymakers have been able to do just that: for example, Bolivia's 'Law of Mother Earth' or Stuttgart's 'Cool City Framework'[2]. Hence, in many cases globally, the two are not seen as separate entities. However, attempts to rectify South Africa's turbulent socio-economic past have resulted in social justice being wedged at the forefront of post-apartheid plans and policies. Consequently, saving the environment has fallen subsidiary to some of the more pressing social issues – such as housing, cultural expression and employment – in South Africa. The result of this sidelining has been the polarisation of environmental and social activism, with the subsequent prioritisation of the latter. Here are three key reasons why this phenomenon may be evident globally, and may be especially true in South Africa.

Swift social change versus steady environmental change

In many ways, environmental activism and social justice cannot be equated because they occur over very different scales in space and time. Drastic and definite changes in global climate operate over hundreds to millions of years; the life cycle of a local socio-economic issue, however, can complete within a single year. For example, if we culled carbon dioxide emissions today, it is likely, although not certain, to take millennia for it to leave the atmosphere as gas and calcify to rock. In the past, social transformation in South Africa has also taken some time. Tracing the history of colonialism and white dominance from the 1800s all the way to the white/non-white economic dichotomy today helps to illustrate this. Irrespective of the numerous post-apartheid policy changes, over 20 years into democracy, the average white person still earns approximately six times as much as a person of colour.[3]

However, more and more evidence is suggesting that social reform operates, on average, over much shorter time frames than environmental intervention can. For example, within a single month in late 2015, South Africa witnessed one of its biggest student uprisings since Soweto 1976. The #FeesMustFall and #RhodesMustFall movements, active across most of the country's universities, made socio-economic demands that were recognised and addressed almost immediately.

As we broaden our scope to encompass larger and longer spatio-temporal scales – so characteristic of most environmental issues – we lose more and more certainty of future projections. In light of this, and of the poverty so characteristic of South African society, how can one reasonably expect the country's urban poor to curtail electricity usage today, for example, when the long-term future of earth's energy stocks are so unreliable? Should we be modifying cause without certainty of effect in a country in which resource demand is far exceeding supply? Hence, environmental activism is often not a priority in South Africa because it requires an invaluable investment of patience and risk before actual change can be seen. In a political economy such as South Africa's, economic desperation does not allow for risk taking and money wasting. Support for social activism, on the other hand, has grown because over reasonable time frames an issue can often be identified and solved, with some degree of certainty.

Obscure environmental issues versus tangible social solutions

Environmental activism prioritises something that is relatively esoteric whilst social justice offers tangible solutions. Worldwide, climate change scientists sometimes fail at gaining support because their numbers and reasoning cannot be understood by the general public. Still, scientists often argue that their work *is* accessible, explaining simply that, for example, sea levels will rise along with global surface temperatures. The Intergovernmental Panel on Climate Change has estimated that at worst-case scenario global surface temperatures will rise by approximately 4 °C by 2100.[4] This is simple enough, but how well does it represent the lived experience of the average South African? To what extent does it help us to understand how climate change will actually affect us in our or in our children's lifetimes? If we cannot wholly understand the potential repercussions of global change, then it is unlikely that we will garner enough support to mitigate them.

In some respects, social issues such as racism and gender inequality have also been heralded as extremely esoteric. Based almost entirely on personal experience, many social issues cannot be reduced to neat mathematical formulae and solved. Interestingly, these same issues have most often been accepted as direct, accessible and solvable. For example, at the University of Cape Town, tangible 'solutions' – such as the renaming of buildings, the eradication of fee increases and the removal of colonial memorabilia – that directly impacted social well-being have, especially since late 2015, been implemented. Such solutions have provided concrete, quantitative and relatively simple evidence that socio-economic problems can, in theory, be compensated for. Whether or not

these solutions have been effective in executing true change is another topic of contention. Environmental problems, on the other hand, require rigorous data analyses before any steps can be taken to curb global change. This explains why it is so difficult to provide viable solutions that can benefit both people and the environment at the same time. Further, it emphasises why social change is often prioritised over environmental intervention in South Africa.

A green culture versus a cultural tradition

Environmental justice and social activism can also be at odds in South Africa because of a clash in world views. Aspects of South African society present a battle between culture and the environment, whereby many cultural practices in this country are extremely environmentally unfriendly. Rearing livestock, for example, has been heralded as one of the largest sources of greenhouse gases, the most prominent water pollutant and the key driver of biodiversity loss. It has been estimated that approximately 10 000 L of water is required to produce a single 450-g meat patty.[5] With this in mind, environmentalists worldwide are likely to encourage a shift from meat-eating towards a plant-based diet.

However, South Africa's cultural and culinary history is likely to stall or to even resist this shift. From the beginning of the Upper Palaeolithic, San hunter–gatherers have been killing and eating large game.[6] This practice is deeply embedded in South African culture and, although it has been deemed environmentally sustainable, it represents the mechanisms on which modern meat production and consumption – practices that are invariably not sustainable – are based. Today, 24 September has been set aside annually as a public holiday (National Heritage Day) and is devoted to celebrating our heritage through the making of a *braai*. National Heritage Day has been labelled a 'noble cause that will contribute to strengthening South Africa as a nation through social cohesion'[7]. This is representative of a bigger meat-eating creed that characterises South African society, across cultures, and without proper consideration of the potential environmental impacts.

In light of the turbulent socio-economic past of this country, social cohesion and nation-building is indeed an urgent issue that needs addressing, but unfortunately is often pursued at the expense of the environment. Conservationists have delved deeper into the environment–society clash by exploring other cultural practices that impede the health of South Africa's ecosystems. One such example is through the utilisation of leopard skins in traditional religious rituals conducted by members of the Shembe Church, a hybrid of Zulu and Old Testament Christian religions. During the 1970s, about 50 000 leopards were killed annually to sustain these practices.[8] In theory, the leopard skin trade could, according to environmentalists, quite easily be obliterated. However, who decides whether biodiversity or an entire cultural practice should be jeopardised? Such a clash in world views makes it difficult to simultaneously balance the needs of the environment with that of South African society.

Amongst others, three key variables – time scale, tangibility and culture – illustrate the contrasts between environmental justice and social activism. Debates continue as to whether these variables can be dealt with together. On one hand, a nexus of governance, environment and people can and should be formed, ensuring that each issue is tackled relative to the other. Scientists Gordon McGranahan and David Satterthwaite have coined this phenomenon 'bridging the green and brown agendas'[9]. However, South Africa's case is complicated. This country has urgent socio-economic matters which need to be addressed before finding too many causes and fixing none. Given the current socio-economic standing of this country, environmental cooperation is unlikely to ever be achieved until people have been granted true and democratic social justice.

The opinions expressed herein do not necessarily reflect the views of the University of Cape Town.

References

1. Lucas K, Walker G, Eames M, Fay H, Poustie M. Environment and social justice: Rapid research and evidence review. London: Policy Studies Institute; 2004.

2. Rehan R. Cool city as a sustainable example of heat island management case study of the coolest city in the world. HBRC J. 2014;12(2):191–204. http://dx.doi.org/10.1016/j.hbrcj.2014.10.002

3. Wilkinson K. Race, poverty and inequality: Black first land claims fact-checked. Africa Check. 2015 27; FACTSHEET. Available online from: https://africacheck.org/reports/race-poverty-and-inequality-black-first-land-first-claims-fact-checked/

4. Intergovernmental Panel on Climate Change (IPCC). Final draft of the special report on emission scenarios. Technical Support Unit of IPCC Working Group III. Bilthoven: Netherlands Environmental Assessment Agency; 2000.

5. Oppenlander RA. Food choice and sustainability: Why buying local, eating less meat and taking baby steps won't work. Minneapolis, MN: Langdon Street Press; 2013.

6. Barnard A. Hunters and herders of southern Africa: A comparative ethnography of the Khoisan peoples. Cambridge: Cambridge University Press; 1992. http://dx.doi.org/10.1017/CBO9781139166508

7. Scannell J. National Braai Day: Our mission [homepage on the Internet]. c2007 [cited 2016 Sep 01]. Available from: http://braai.com/national-braai-day-mission/

8. Ramsay S. #ShockWildlifeTruths: Will the Zulu traditional dress wipe out the SA leopard? News24. 2016 February 29; Traveller24 [cited 2016 Sep 01]. Available from: http://traveller24.news24.com/Explore/Green/shockwildlifetruths-will-the-zulu-skin-trade-wipe-out-the-leopard-20160229

9. McGranahan G, Satterthwaite D. Environmental health or ecological sustainability: Reconciling the brown and green agendas in urban development. In: Pugh C, editor. Sustainable cities in developing countries. London: Earthscan; 2013. p. 73–90.

AUTHOR:
Eben Maré[1]

AFFILIATION:
[1]Department of Mathematics and Applied Mathematics, University of Pretoria, Pretoria, South Africa

CORRESPONDENCE TO:
Eben Maré

EMAIL:
Eben.mare@up.ac.za

A note on aspects of risk and return for South African bond investors

Gentlemen prefer bonds – Andrew Mellon (1855–1937)

South African bond returns typically increase with duration. On a risk-adjusted basis, I demonstrate that bond returns appear most attractive in the short to medium maturity-term area.

Bond risk and returns

Bonds represent debt instruments that allow issuers to finance capital needs. The bond represents an obligation of the issuer to pay the holder principal as well as regular interest amounts. Typical bond issuers are sovereign governments. According to a National Treasury report[1], the size of the South African debt capital markets measured in excess of ZAR2 trillion as at the end of March 2015.

Using returns from the 2016 version of the Credit Suisse Global Investment Sourcebook[2], I observe that over the period 1900 to 2015, the annualised real return for South African bonds was 1.8% compared with 1.0% for shorter-term bills. Inflation measured 4.9% per year over this period (i.e. the nominal yield on bonds was 6.7% compared with 5.9% for bills). In comparison, US annual real returns were 2.0% for bonds and 0.8% for bills and inflation was 2.9% over the same period.

Bond returns have varied considerably. Over the period 2000 to 2015, for example, South African annual real bond returns amounted to 4.9% compared to real returns of 2.2% on shorter bills.[2] (In comparison the US return for the same period was 5.4% for bonds and -0.4% for bills.)

Bonds are inherently more risky than bills. Bonds mostly represent longer-term financial commitments than bills. It would therefore seem 'intuitively obvious' that bonds should yield higher returns (on average) than bills.[3-6]

I am specifically interested in the question of whether the excess return of bonds over bills represents an adequate investment case in the South African market, measured in terms of risk versus return. In essence I am trying to answer the question: Which bonds would I invest in to achieve the best risk/reward? Traditionally bond risk is measured in terms of so-called duration[7] and volatility of returns. I shall, however, use VaR and CVaR as defined below[8] to measure risk.

The All Bond Index (ALBI)[9] is a composite index of the top 20 bonds in South Africa, dually ranked on the basis of market capitalisation as well as liquidity (bonds with a maturity of less than 1 year are excluded). The ALBI01, for example, represents a sub-index of the ALBI which includes only bonds with a maturity of 1–3 years.

For the purposes of this study, I collected monthly total return data over the period 31 December 1999 to 31 December 2015, considering the ALBI and its sub-indices. Annualised, average monthly return and risk statistics are shown in Table 1. VaR denotes the so-called value-at-risk at 98% confidence, i.e. a loss level one would anticipate experiencing with a 2% frequency. The conditional VaR (CVaR) is the loss-expectation of realisations below the 98%-VaR level (see Alexander[8]).

Table 1: All Bond Index (ALBI) and sub-index risk and return statistics

	ALBI01	ALBI02	ALBI03	ALBI04	ALBI
ALBI sub-split	1 to 3 years	3 to 7 years	7 to 12 years	12 years plus	
Return	9.85%	11.33%	12.39%	12.60%	11.65%
Standard deviation	2.79%	5.69%	8.08%	10.95%	7.21%
VaR	1.09%	2.65%	4.60%	6.37%	3.59%
CVaR	1.31%	3.26%	5.23%	7.43%	4.72%
Kurtosis	2.66	2.56	1.48	2.05	1.92
Skewness	0.49	0.44	0.13	0.34	0.14

The reader will notice the increase in returns between the ALBI01 to ALBI04 buckets, which is accompanied by a rise in the risk levels as measured by the standard deviation of returns, VaR and CVaR measures.

I treat the ALBI sub-indices as investable portfolios. I solve for various portfolios (combinations of ALBI01 to ALBI04) using the monthly returns data, which yield specific excess return levels at the lowest risk measured in terms of CVaR. Figure 1 denotes the efficient frontier so obtained. Figure 2 shows the allocation of assets across ALBI01 to ALBI04 (higher excess returns require more allocation to higher yielding assets and corresponding levels of risk) based on the efficient frontier. In Figure 3, I calculate the ratio of the VaR corresponding to a certain excess return portfolio – this figure shows that investors are (on average) not compensated for bearing extra risk in terms of longer duration. Allocation to longer duration portfolios should therefore be based on a tactical view pertaining to the business cycle (see Ilmanen[10]).
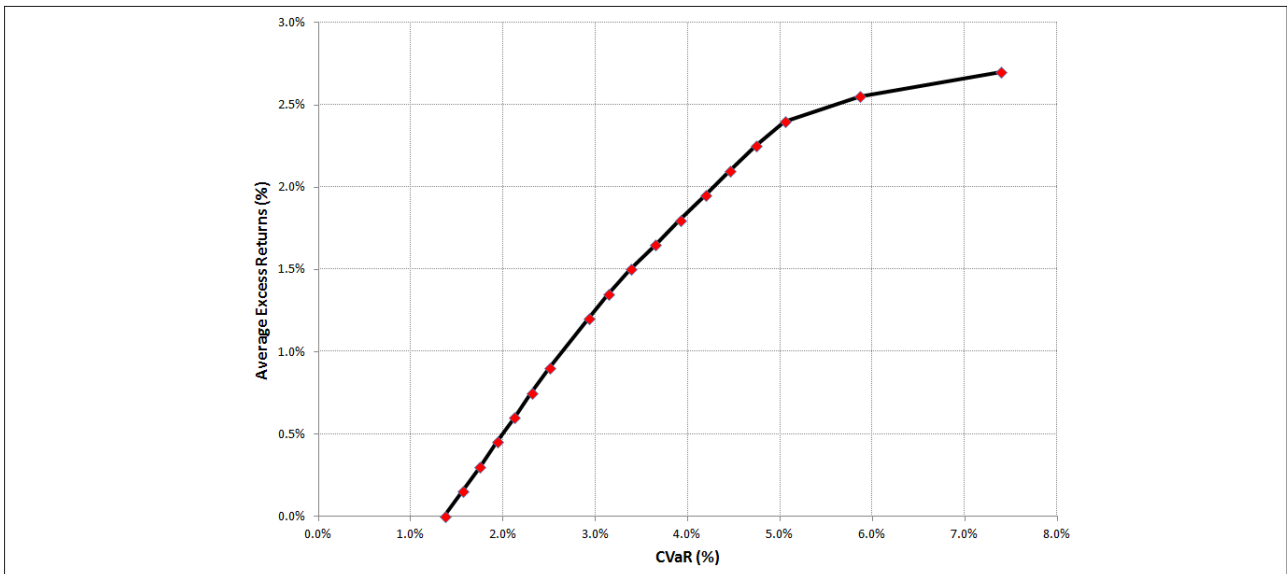
Figure 1: Efficient frontier.
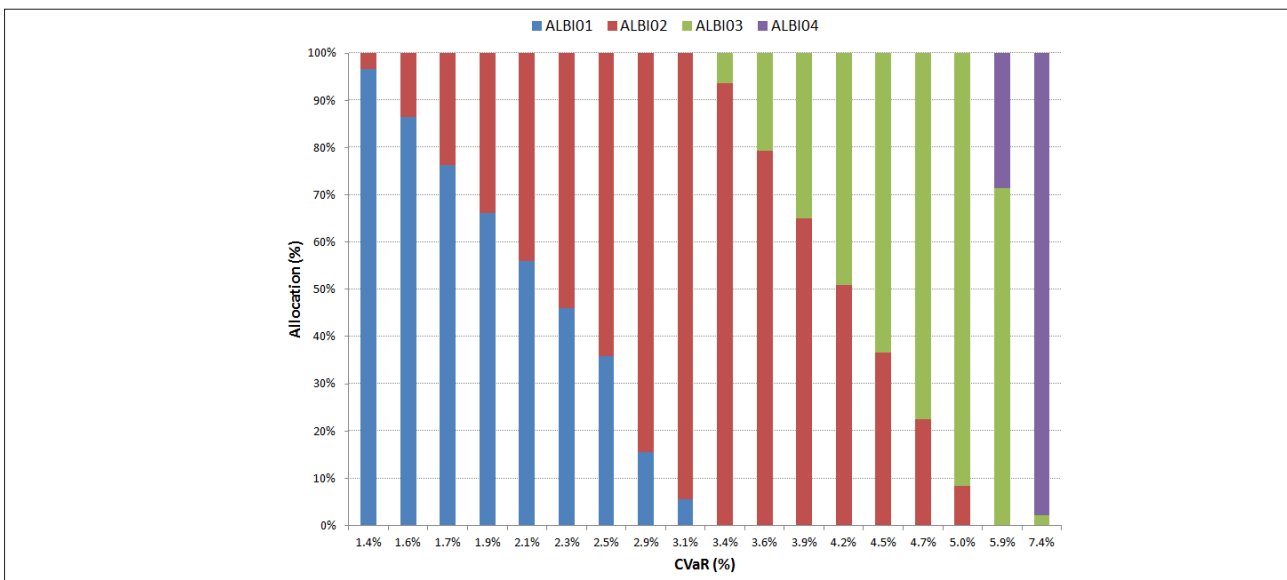
Figure 2: Efficient frontier allocation.
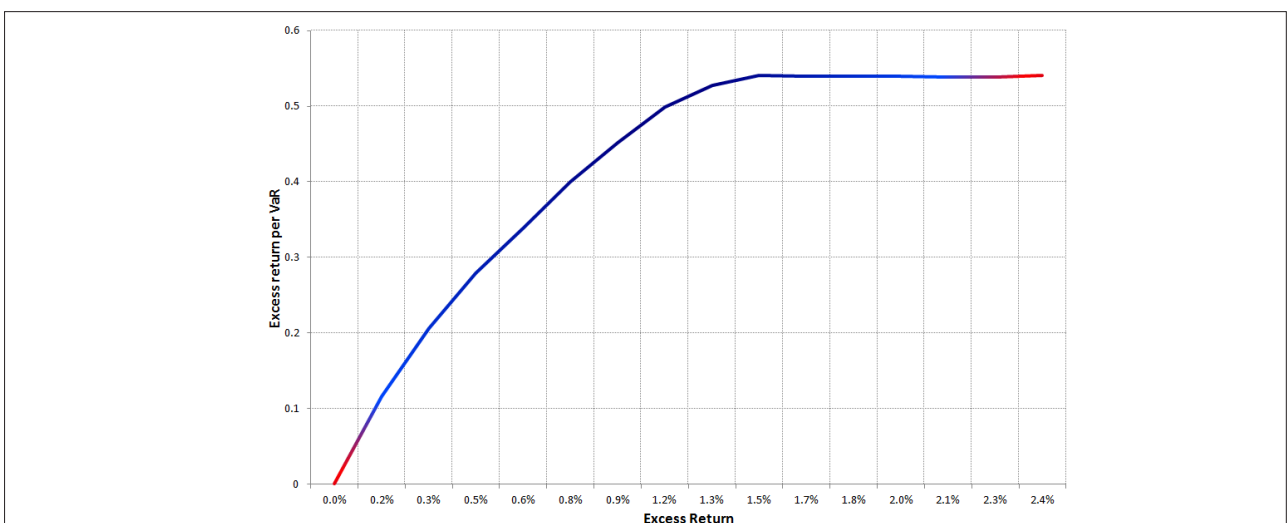
Figure 3: Excess return per unit of VaR (98% confidence level).

Conclusion

On the basis of average monthly returns of investable bond indices, I observe that excess bond returns tend to increase as the corresponding duration increases. On a risk-adjusted basis, however, it would appear that South African bond investors achieve optimal results using short-duration portfolios. Short-duration bond portfolios also outperform bills. Long-duration extension could enhance returns based on tactical views such as the business cycle and inflationary expectations.

References

1. South African National Treasury. Debt management report 2014/2015. Pretoria: National Treasury; 2015. Available from: http://www.treasury.gov.za/publications/other/Debt%20Management%20Report%202014-15.pdf

2. Dimson E, Marsh P, Staunton M. Credit Suisse Global Investment Returns Sourcebook 2016. Zurich: Credit Suisse; 2016.

3. Ilmanen A. Does duration extensions enhance long-term expected returns? J Fixed Income. 1996;6(2):23–36. http://dx.doi.org/10.3905/jfi.1996.408172

4. Fama EF. Term premiums in bond returns. J Financ Econ. 1984;13(4):529–546. http://dx.doi.org/10.1016/0304-405X(84)90014-X

5. Durham JB. Another view on US Treasury term premiums. FRB of New York Staff Report 658. 2013 December 01. http://dx.doi.org/10.2139/ssrn.2377405

6. Grieves R, Mann SV, Marcus AJ, Ramanlal P. Riding the bill curve. J Portfolio Manage. 1999;25(3):74–82. http://dx.doi.org/10.3905/jpm.1999.319717

7. Ilmanen A. How well does duration measure interest rate risk? J Fixed Income. 1992;1(4):43–51. http://dx.doi.org/10.3905/jfi.1992.408032

8. Alexander C. Market risk analysis. Volume IV: Value at risk models. Chichester: John Wiley & Sons; 2009.

9. Johannesburg Stock Exchange. All Bond Indices [homepage on the Internet]. No date [cited 2016 Oct 24]. Available from: https://www.jse.co.za/services/market-data/indices/fixed-income-index-series/all-bond-indices

10. Ilmanen A. Forecasting US bond returns. J Fixed Income. 1997;7(1):22–37. http://dx.doi.org/10.3905/jfi.1997.408197

Science to policy – Reflections on the South African reality

AUTHORS:
Sophie von der Heyden[1] (iD)
Peter Lukey[2]
Louis Celliers[3]
Kim Prochazka[4]
Amanda T. Lombard[5]

AFFILIATIONS:
[1]Evolutionary Genomics Group, Department of Botany and Zoology, Stellenbosch University, Stellenbosch, South Africa

[2]Strategic Environmental Intelligence Unit, Department of Environmental Affairs, Pretoria, South Africa

[3]Coastal Systems Research Group, Natural Resources and the Environment, Council for Scientific and Industrial Research, Pietermaritzburg, South Africa

[4]Fisheries Management, Department of Agriculture, Forestry and Fisheries, Roggebaai, South Africa

[5]Coastal and Marine Research Unit, Department of Botany, Nelson Mandela Metropolitan University, Port Elizabeth, South Africa

CORRESPONDENCE TO:
Sophie von der Heyden

EMAIL:
svdh@sun.ac.za

Research is a key resource in a knowledge economy and governance system. In order to enable research to benefit the nation and to contribute to growing the knowledge-based economy (the aims of the Global Change Grand Challenge, and specifically the Society and Sustainability Research Programme), the gap between research, knowledge production and policy and management (i.e. the knowing-doing gap[1]) needs to be closed, yet closing this gap remains a complex challenge[2]. This year's annual SANCOR (South African Network for Coastal and Oceanic Research) Forum meeting addressed this gap through consultation with a variety of stakeholders from the coastal and marine science community. Our brief was to provide for reflection and discussion on aspects of the science–policy–management interface within South Africa and this commentary provides a summary of the Forum discussions. We detail some current challenges of integrating coastal and marine science into policy and decision-making in South Africa, highlight 'success stories' and provide some thoughts on maximising overlap and building a sound science–policy interface. Although couched in the context of marine and coastal sciences, our findings will resonate with other scientific disciplines. Similarly, the challenges in and opportunities for creating constructive dialogue for evidence-based decision-making are not specific to South Africa, so we draw on national, international and collective experience to provide an avenue for doing so. In this commentary we highlight current examples of mismatch between science and policy by focusing on barriers resulting from legislation, politics and a general lack of process for better integration. In particular, we focus on the complexities of evidence-based decision-making at different scales, and how international scientific engagement has helped shape policy in South Africa. We finish by providing some perspectives, directions and examples to help narrow the gap and foster better science–policy integration into the future.

The gap between policy support and science: How messy is it really?

The theory behind the gap

With increasing globalisation, the impacts of climate change, the focus on growth economies, food security and increasing (or maintaining) human well-being and livelihoods, there have never been as many challenges in and opportunities for science to inform, contribute to and support decision-making. Science has made meaningful contributions to decision-making processes and several publications have focused on the importance of this relationship. These contributions have impacted fields such as conservation[3,4], fisheries[5] and medicine[6], yet barriers still exist that prevent the ready integration of research into policy. In South Africa, some notable initiatives do exist that try to actively engage across the divide, including the Programme to Support Pro-poor Policy Development under the auspices of the Department of Planning, Monitoring and Evaluation – yet such targeted programmes remain few. All too often the science–policy interface is limited to opportunities during public participation events, rather than allowing for the constructive 'co-creation' or consultative approaches (Figure 1) that lead to overall benefits in the design and implementation of policies.
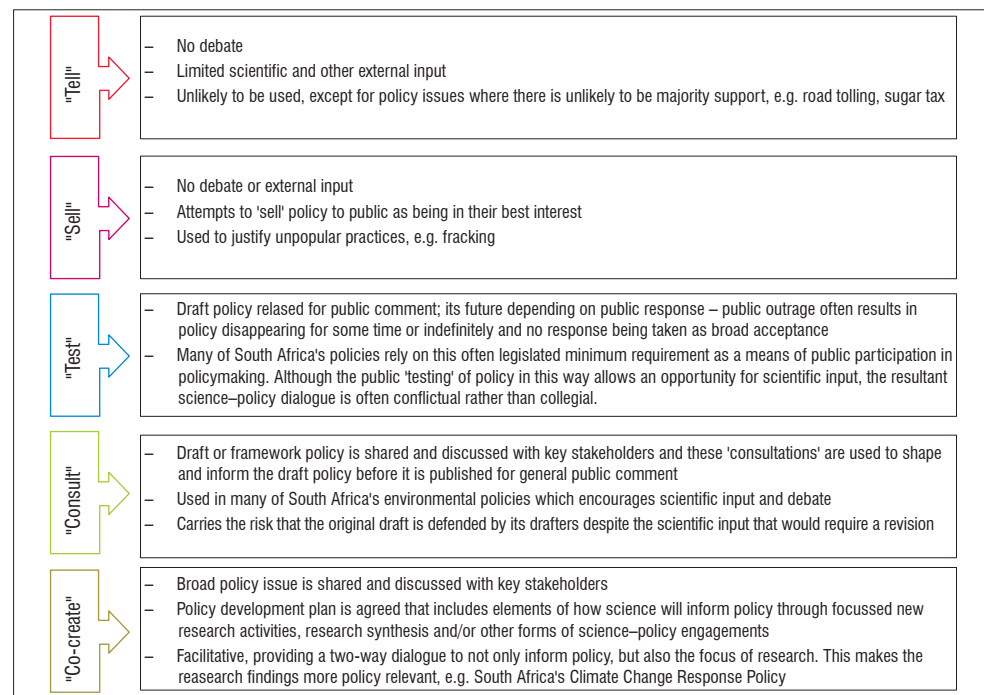
Figure 1: Senge's[11] five decision-making strategies may be used as a framework to describe policymaking in South Africa and the opportunities for evidence-based policymaking associated with each strategy.

Box 1:   Common barriers to successful integration of science into evidence-based decision-making and policy

1. A disjunct in timing – a research project from grant application to completion can take several years, whereas policymakers work on much shorter time scales.

2. A lack of training for scientists in navigating the decision-making processes that lead to formulation of policy, as well as in transdisciplinary research.

3. A lack of mutual understanding of the scientific process and the policy cycle, which includes the time scale and characteristics of achieving a successful outcome. Success in science is predominantly measured by academic impact whereas managers aim to achieve positive societal impact. Academic impact, generally speaking, often falls short of providing a timely and appropriate contribution to societal change. Conversely, enduring societal impact is not generally possible without timely access to appropriate knowledge.

4. A lack of understanding of the limitations of scientific findings and equally important the lack of understanding of the multidimensional decision-making required for real-world problems faced by managers and politicians.

5. Presented scientific data being ignored because policymakers have not been convinced of its policy importance, significance, relevance or implications.

6. A lack of opportunities that allow researchers and managers space to engage constructively around the gaps in knowledge production, focusing on policy-relevant data, and to build relationships based on mutual trust and understanding. This includes lack of consultation between policymakers and scientists when policies are being formulated, and can result in scientists being unwilling to engage in science to enable or inform what they consider to be poorly formulated policies.

7. Mismatch in the language and processes used by scientists and policymakers. There is a distinct issue that arises from the difference between the language of probability and significance (as used by scientists), and the language of risk, resource availability and perception, used by managers and politicians.

8. Scientists are often unable or unwilling to provide the certainty and absoluteness required and often demanded by managers and policymakers. On the flip side, there is often a lack of clear policy objectives that would enable policy-relevant science to be undertaken.

9. Lack of data accessibility and capacity to find and evaluate the relevant data and its originators.

10. Poor feedback of scientific impact on policy, which prevents researchers from gauging the actual impact of their work.

11. Lack of accountability (perceived or otherwise) of government and policymakers.

Box 1 provides an overview of some of the common challenges. These challenges are not mutually exclusive, but rather act as multiple, additive obstacles. Surprisingly, however, little research has gone into understanding the contribution that each of these factors makes to science–policy overlap. The question 'What is it that can be done to increase impacts that are transparent, easily applied and consistent?' therefore resonates with both the research community and policy developers. Clearly, evidence-based decision-making will not only benefit policy processes, but also contribute to more effective implementation as it provides a mechanism for ownership and commitment from stakeholders. Put differently, how can the scientific process and the policy cycle be better integrated to create conditions for decision-making based on appropriate and quality scientific evidence, and the needs and desires of communities? Below we focus on four examples.

Example 1: 'Too much, too soon'

The *Integrated Coastal Management (ICM) Act* provides one good example of the science–policy disconnect within the context of the science and policy gap. South Africa is among only a handful of countries globally, particularly in Africa, that has a formalised and institutionalised ICM,[7] which includes legal instruments, policy documents as well as guidelines and strategies and the use of state of the coast reporting, among others. The ICM is also strongly and formally institutionalised within the governance system in South Africa. The ICM Act compels the creation of 'coastal committees' at national and provincial levels, as well as a host of other technical and political fora forming part of the coastal management landscape down to local government level. Theoretically, the substantial effort to establish a system of integrated coastal management provides ready and usable science to the policy pathway. It provides feedback mechanisms between spheres of government and also between civil society sectors. However, underlying the 2008 ICM Act is the enormous complexity of its implementation, which has reduced the impact and reach of the strong legislation. Moreover, determining its effectiveness is hampered by the availability of data; the first 5-year policy cycle remains to be evaluated before improving and expanding the initial processes.

Example 2: Individual decisions versus a formal process

Other issues at the messy science–policy interface pertain to the legislative pathways leading to evidence-based decision-making which, in some countries, is regulated by national legal instruments. An extreme example is the *Magnusson-Stevens Fishery Conservation and Management Act* of the United States of America[8], which is highly prescriptive. In contrast, South Africa's *Marine Living Resources Act*[9] is virtually silent on the requirement for integrating research. Nevertheless, the overarching United Nations Convention on Law of the Sea[10], together with various voluntary instruments of the Food and Agriculture Organization, empower the relationship between science and decision-making in fisheries, even in the absence of strong national legislation. Thus, a close connection between science, policy and management has been a long-standing feature of South African fisheries management. The primary weakness of this arrangement is, however, that it can be heavily influenced by the preferences of individual decision-makers and the extent to which scientific information is counter-weighed by less rigorous information relating to social well-being, food security and local economic drivers. Thus the relationship between science, policy and management in legislation, while enabling, may not necessarily impact the decision-making processes effectively.

Example 3: Exclusion of science during policy formulation

Scientists are often not privy to the formulation of policy although ultimately, they are significantly affected by the legislation. For example, the current drafting of the new Marine Spatial Planning Bill did not include input from scientists not affiliated with government in its formulation, although South Africa has a strong background in marine spatial planning with extensive data from multiple disciplines that could richly inform policy; in fact, South Africa is at the forefront of marine spatial planning globally. However, the only opportunity for scientists to comment was during the public participation period which did not allow for the science, knowledge and process to be reviewed and integrated into the bill. This situation falls under Senge's[11] 'test' scenario (Figure 1), in which important stakeholders can only participate after the drafting of the initial legislation. Ironically, once enacted, scientists will be expected to perform policy-relevant science that they had no meaningful input in developing.

Example 4: Political decisions trump 50 years of science

Fisheries around the world provide some of the best-known and clear examples of the 'trumping' of scientific advice by political opinion or expedience. Throughout the history of South Africa's fisheries, such overriding of scientific advice has been a significant contributing factor

leading to the current poor state of many of our nearshore fisheries resources. A pertinent example is Tsitsikamma, South Africa's oldest Marine Protected Area, declared in 1964. Since its declaration there have been several attempts (frequently in periods leading up to elections), to gain access for local communities to fish within the Marine Protected Area. Many years of research have provided solid evidence on the positive impacts of this closed area in protecting certain linefish species not only within the confines of the Marine Protected Area itself, but also over a far wider geographic area where far-reaching positive spillover effects from the closed area contribute to the recovery of depleted fish stocks across their entire range. Research has further demonstrated that associated benefits to coastal communities more broadly (i.e. spanning several hundreds of kilometres on either side of the closed area) outweigh the benefits that could accrue from small and highly localised communities catching fish within the closed area. Despite the availability of such strong scientific evidence to the contrary, access to fish in key areas of the Marine Protected Area was recently granted in response to political pressure from a small group of local residents. Eventually the matter reached the courts, at which time the scientific evidence, and the administrative process used by government to grant the access, was fully considered, and the decision was reversed. This is a clear example of how political influence can be used to override 50 years of scientific evidence.

Converting scientific findings into policy: Differences in scale

Approaches for integrating research into practice and policy differ at local, regional, national and global levels, requiring different workflows and interactions, which adds an additional level of complexity to implementation. In general, at national scales, policies tend to focus on principle-based guidance, providing frameworks for decision-making and broad measures for successful implementation. In a developing nation, the burden is, however, on the state to provide scientific support for decision-making within provinces and local government. Therefore the science to policy pathway differs for the three spheres of government in South Africa and depends on the policy in question. For example, there is virtually no distribution of responsibility outside the national sphere of government when dealing with 'ocean matters'. Fisheries, oil and gas, offshore mining and shipping industries (i.e. big industries with high-value resources) all have a distinctly top-down management design. This allows for a shorter (but potentially heavily politicised) science to policy pathway, with direct relationships between national stakeholders and regulators, as well as the science that supports decision-making at that scale. In these instances, research has a clear and direct pathway to policy. Science supporting these large industries often deals with large, complex and dynamic ecological systems, and science products related to some aspects may be incomplete and rely on expert opinion. So, while the science to policy pathway may be short, the potential impact of decision-making based on partial or incomplete science exists. Conversely, this also provides greater opportunity for the introduction of unfounded opinion and poor science as evidence, as well as opportunities for discrediting good science although it is highly relevant.

In contrast, at finer scales (such as the local government level), science that can contribute to more informed decision-making requires a much longer value-chain, resulting in very specific solutions for equally specific problems. This places greater burden on scientists to understand the flow of evidence-based information through a highly complex stakeholder universe, e.g. municipalities or provincial government. The relationships between stakeholders are complex, often competing and, in many cases, result in the creation of 'wicked problems' that never satisfy all parties. Therefore, while it is possible to design processes that broadly address common issues, the local context (environmental, institutional, etc.) makes it extremely challenging to find a one-size-fits-all solution. For example, climate change adaptation, local resource management and local economic development all require very specific, place-based conditions for local implementation which may not be met with a single solution.

Taking success in informing international policy back to South Africa

South Africa has played important roles at the international level in promulgating global legislation. For example, South Africa's involvement in the Intergovernmental Panel on Climate Change (IPCC) highlights how well South Africans actively engage in a dedicated science–policy dialogue when an efficient and effective platform for this kind of dialogue exists. This international engagement can also have knock-on effects nationally. For example, the Minister of Finance's 2008 budget speech to Parliament provides evidence of how this international engagement impacted on local policy when the Minister noted that '…the United Nations International Panel on Climate Change, in which a South African team led by the Department of Environmental Affairs and Tourism played an active role, has added impetus to the need for policy change'[12]. This high level of engagement has continued with a South African scientist co-chairing IPCC's Working Group II for the compilation of its 6th Assessment Report and provides continuing support for the possibilities of engagement at the national level.

A further example of the inclusion of science in policy and management decisions in South Africa is highlighted by the interactions of fisheries scientists and managers. Fisheries management globally has a long-standing tradition of directly linking science to decision-making and this same close relationship has, for over 100 years, been a feature of South African fisheries management. The existence of this close relationship is confirmed by feedback from international fisheries science experts who participate in the annual international peer review of aspects of South African fisheries science. They frequently indicate that South Africa is on par with respect to international best practice. Similarly, the rigorous audits of the science, policy, management and enforcement landscape that are regular features of fisheries eco-labelling or certification schemes, to which some South African fisheries belong, further confirm the existence of this close relationship.

South Africa has also played a leading international role in integrating science-based biodiversity planning into policy[13], and attracting significant international funding into programmes such as the Cape Action Plan for People and the Environment (CAPE)[14]. CAPE has subsequently underpinned many national and regional biodiversity plans and forms of legislation, and has defined a methodological framework of best practice within a well-networked community of practice (which includes scientists, managers and policymakers who meet annually at the National Biodiversity Planning Forum). Many publications[15,16] have been produced from this work and highlight the important role of stakeholder engagement and user-useful products. This initiative, along with examples from climate change and fisheries management, illustrate how international workflows have positively impacted national agendas, highlighting the continued need for the inclusion of South African researchers in high-level programmes, in roles that will ultimately support local endeavours at all scales.

Better integration of science into policy: Some perspectives

Research shows that policy- and decision-makers are cognisant of the importance of science and knowledge production. Indeed, in the South African context, although Cronin and Sadan[17] found a relatively poor use of scientific evidence by senior government officials, they also found that these officials desired the use of more scientific evidence in decision-making. Recognising and negating the obstacles (Box 1; Figure 1) is a good start to more integrated decision-making, but this will differ within the context of the management questions asked, as well as the sphere of government in which the integration of data is required. However, the opportunities for evidence-based decision-making can only be fully realised by consultative or co-created approaches, with active engagement at all levels of the policy process by a wide variety of stakeholders. Here we identify some factors that can help smooth the transition between research, policy and management and the successful implementation of evidence-based policymaking.

Box 2: The National Climate Change Response Policy: A case study of doing it differently using co-production

Broadly, 'co-production' is an approach that increases knowledge exchange among scientists and decision-makers. In this approach, managers actively participate in scientific research programmes from the onset, collaborating with researchers throughout every aspect of the study including design, implementation and analysis (Figure 2). This approach fosters a stronger understanding of the research and also increases the ownership and its subsequent communication.[19] Although the development of South Africa's National Climate Change Response Policy may be regarded as a product of co-production, the approach differed from that described by Cvitanovic et al.[19] in that it was the scientists who were actively engaged in the policy development process from the onset, collaborating with policymakers throughout the policymaking process, rather than vice versa.

The initiation of a dedicated climate change response policy development process took place at the National Climate Change Conference held in 2005. At this conference, 'over 600 representatives from government, business, the scientific and academic communities, and civil society considered the science relating to climate change and key responses to the potential social and economic impacts associated with the compelling scientific evidence of climate change'[23]. The conference consisted of two parallel and overlapping sessions, with a dedicated Climate Change and Science Conference hosting African and international scientists that primarily focused on climate change science. Overlapping with this session, was the National Consultative Conference on Climate Change within the context of testing and informing South Africa's policies, strategies and action plans; directing South Africa's international negotiations on climate change; charting the way forward on future commitments; generating inputs for the Second National Communication on Climate Change; revising policies to take into account new scientific developments; and more closely coordinating South Africa's environmental approach with the national energy strategy.[24] Every day the Consultative Conference started with a briefing from the Science Conference based on their previous day's proceedings and this briefing informed and directed the policy discussions of the day.

This strong science–policy dialogue was maintained by having the same official being responsible for the management of the policy development process and as the 'client' for the compilation of South Africa's Second National Communication to the United Nations Framework Convention on Climate Change – what Cvitanovic et al.[19] refer to as the 'embedding' approach to increasing knowledge exchange among scientists and decision-makers.

Furthermore, throughout the policy development process, any issue that was seen as requiring further research to better inform the evolving policy was flagged and, following a second national policy conference in 2009 and the public commenting period on the climate change response Green Paper, specific research was commissioned to inform the final White Paper. That the policy and Second National Communication were approved for publication at the same Cabinet meeting in October 2011 is evidence of how these science and policy processes were closely linked and 'co-produced'.

These factors include better coordination and cooperation between the different levels of government, the use of legislated opportunities to appeal poorly informed decisions, using a 'co-creation' approach to policy development (see Box 2 for an example from the national climate change agenda), and, importantly, provision of opportunities for project planning and research formulation that include multiple stakeholders to jointly identify and plan science (Figure 1).

One other notable achievement in South Africa is the National Biodiversity Assessment (NBA)[18], led by the South African National Biodiversity Institute. The NBA comprises a series of technical reports that review, synthesise and make accessible, data on biodiversity and ecosystems, anthropogenic pressures on these, as well as the status of knowledge on natural systems. The success of the NBA to a large part is a result of its participatory approach, with the leaders of each NBA component actively engaging with networks of researchers and scientists who provide data and analysis. The NBA has made significant impacts at various levels (Figure 2), notably feeding into evidence-based decision-making through ventures such as the National Biodiversity Strategy and Action Plan. Although these impacts are not direct, they do provide a process of strengthening the uptake of science not only to the highest level of government in South Africa, but beyond into international projects such as the Convention on Biological Diversity.

Solutions of scale, legislation and easily transferred knowledge

Scalable solutions, that cross different spheres of management, are important in narrowing the science–policy gap, particularly in the multiscale coastal and marine domain. For example, the current separation of coastal (terrestrial) and marine planning domains should be reconsidered to include more diverse stakeholder communities, and to recognise land–sea ecological linkages. Currently, municipalities are not included in large offshore industry decision-making, although production and/or benefits flow over municipal domains (as is the case for fisheries, oil and gas, etc.) all with infrastructure, labour and local economic development implications.

Improving the existing legislation can help bridge the gap between science, policy and management goals, particularly in areas in which such scientific information and advice is routinely and predictably available, and where its use is considered international best practice.

This improvement will foster links between stakeholders and procedures, thus facilitating the direct flow of information to policy development, rather than these links occuring on an ad-hoc basis.

Further, successful links between science and decision-making are to a large extent enabled by structures and procedures which are in place to ensure a direct flow of scientific information into policy development and management. Examples of these structures are the scientific working groups of the fisheries management agency (currently the Branch Fisheries Management of the Department of Agriculture, Forestry and Fisheries), which generate targeted, management-relevant advice, with well-established procedures for transfer of the scientific information and advice to decision-makers. The role of these groups in facilitating the transfer of scientific advice and information cannot be overemphasised.

Finally, whether taking a path of 'embedding', 'co-production' or through knowledge brokering in an enabling environment[19], the inclusion of multiple stakeholders from the outset will go some way in uniting disparate entities that have common goals. It is clear that differences exist in the extent to which science is translated in decision-making across different 'disciplines' in South Africa. We need a range of options that can be tailored to assist in bridging the gap between science, policy and management and ultimately to facilitate evidence-based decision-making.

A knowledge base built on transdisciplinary research

One of the approaches for reducing the length of the science to policy pathway is the evolving definition of what is known as transdisciplinary research – the highest form of integrated research.[20-22] Transdisciplinarity involves not only the natural, social and health sciences in a humanities context, but also incorporates participants from outside of scientific fields (e.g. land managers, user groups and the general public) and it is this level of integration, combined with participatory approaches, that transcends traditional disciplinary boundaries. Unlike multidisciplinary research that shares knowledge across disciplines in thematically based investigations with multiple goals (but does not generate new integrative knowledge), trans- or interdisciplinary research synthesises and harmonises links between disciplines in a coordinated and coherent whole that focuses on 'real-world' system problems. The difference between multi- and transdisciplinarity is the level of integration and cooperation with the added imperative of bridging disciplinary viewpoints.
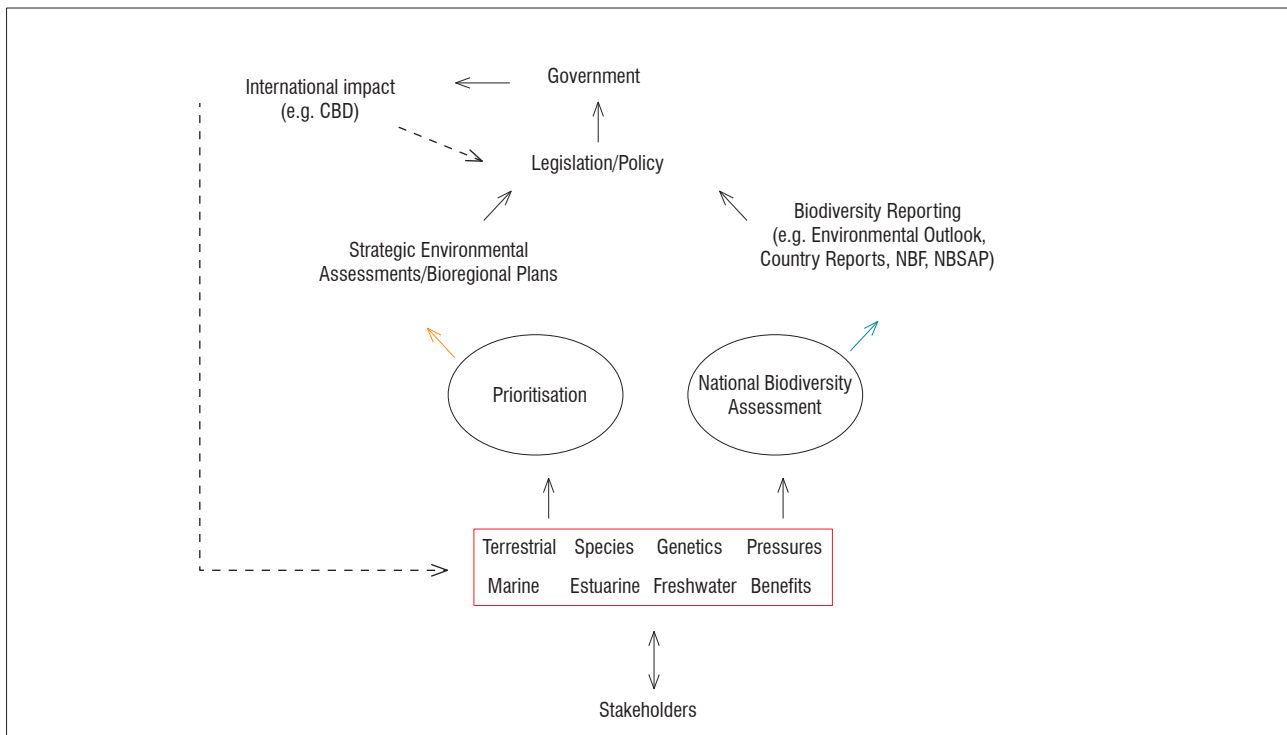
CBD, Convention on Biodiversity; NBSAP, National Biodiversity Strategy and Action Plans; NBF, National Biodiversity Framework

Figure 2: The National Biodiversity Assessment is a multistakeholder report[18] and assesses the state of South Africa's biodiversity, across terrestrial and aquatic environments, emphasising spatial (mapped) information for both ecosystems and species. It has direct and indirect impacts, with the latter feeding into both national and international policy through various other strategic and reporting mechanisms.

This is important for several reasons, particularly because science and the knowledge industry is increasingly being challenged by funders and stakeholders to demonstrate impact beyond academic excellence, with the concept of societal impact firmly embedded in the language of funders, stakeholders and civil society. Transdisciplinary processes will help improve understanding between those generating the scientific information and advice, and those using it. On the one hand, decision-makers require an understanding of the scientific information and how it has been generated. On the other hand, scientists require an understanding of the type of scientific information that is useful to decision-makers, of the time frames on which scientific information and advice is required, and of the nature of inputs (other than the scientific inputs) that may also inform the decision-making process. If the end-user forms part of the project team (and both receives science products and makes substantive conceptual input), the science to policy pathway is shortened. For example, resource groups in which scientists and managers work well together tend to be subject to the least arbitrary decision-making, resulting in the best-managed, most optimally exploited fisheries, and thereby have the most impact.

As a community, SANCOR is cognisant of the potential of transdisciplinary research, but the cost and complexity of such projects and programmes will require new methods and attitudes before transdisciplinarity will become a mainstream research approach. Importantly, in whichever disciplines South African researchers engage, there needs to be more intentional planning of pathways from science activities to outputs, outcomes and impacts.

Acknowledgements

References

1. Pfeffer J, Sutton RI. Knowing "what" to do is not enough: Turning knowledge into action. Calif Manage Rev. 1999;42:83–107. http://dx.doi.org/10.2307/41166020

2. Knight AT, Cowling RM, Rouget M, Balmford A, Lombard AT, Campbell BM. Knowing but not doing: Selecting priority conservation areas and the research–implementation gap. Conserv Biol. 2008;22:610–617. http://dx.doi.org/10.1111/j.1523-1739.2008.00914.x

3. Sutherland WJ, Pullin AS, Dolman PM, Knight TM. The need for evidence-based conservation. Trends Ecol Evol. 2004;19:305–308. http://dx.doi.org/10.1016/j.tree.2004.03.018

4. Pullin AS, Knight TM, Stone DA, Charman K. Do conservation managers use scientific evidence to support their decision-making? Biol Conserv. 2004;19:245–252. http://dx.doi.org/10.1016/j.biocon.2003.11.007

5. Food and Agriculture Organization of the United Nations (FAO). FAO code of conduct for responsible fisheries. Rome: FAO; 1995. Available from: http://www.fao.org/3/a-v9878e.pdf

6. Sackett DL, Rosenberg WM, Gray JM, Haynes RB, Richardson WS. Evidence based medicine: What it is and what it isn't. BMJ. 1996;312:71–72.

7. Celliers L, Rosendo S, Coetzee I, Daniels G. Pathways of integrated coastal management from national policy to local implementation: Enabling climate change adaptation. Mar Policy. 2013;39:72–86. http://dx.doi.org/10.1016/j.marpol.2012.10.005

8. Magnuson-Stevens Fishery Conservation and Management Reauthorization Act of 2006. Available from: http://www.nmfs.noaa.gov/sfa/magact/

9. Marine Living Resources Act no 18 of 1998. Available from: http://www.gov.za/sites/www.gov.za/files/a18-98.pdf

10. UN General Assembly. Convention on the Law of the Sea [document on the Internet]. 1982 December 10 [cited 2016 Nov 07]. Available from: http://www.refworld.org/docid/3dd8fd1b4.html

11. Senge PM. The fifth discipline: The art and practice of the learning organization. New York: Doubleday/Currency;1990.

12. Manuel TA. Budget speech 2008 [document on the Internet]. 2008 February 20 [cited 2016 Nov 07]. Available from: http://www.treasury.gov.za/documents/national%20budget/2008/speech/speech.pdf

13. Balmford A. Conservation planning in the real world: South Africa shows the way. Trends Ecol Evol. 2003;18:435–438. http://dx.doi.org/10.1016/S0169-5347(03)00217-9

14. Cowling RM, Pressey RL, Rouget M, Lombard AT. A conservation plan for a global biodiversity hotspot – the Cape Floristic Region, South Africa. Biol Conserv. 2003;112:191–216. http://dx.doi.org/10.1016/S0006-3207(02)00425-1

15. Pierce SM, Cowling RM, Knight AT, Lombard AT, Rouget M, Wolf T. Systematic conservation planning products for land-use planning: Interpretation for implementation. Biol Conserv. 2005;125(4):441–458. http://dx.doi.org/10.1016/j.biocon.2005.04.019

16. Knight AT, Cowling RM, Campbell BM. An operational model for implementing conservation action. Conserv Biol. 2006;20(2):408–419. http://dx.doi.org/10.1111/j.1523-1739.2006.00305.x

17. Cronin GP, Sadan M. Use of evidence in policy making in South Africa: An exploratory study of attitudes of senior government officials. Afr Eval J. 2015;3(1), Art. #145, 10 pages. http://dx.doi.org/10.4102/aej.v3i1.145

18. Driver A, Sin KJ, Nel NJ, Holness S, Van Niekerk L, Daniels F, et al. National Biodiversity Assessment 2011: An assessment of South Africa's biodiversity and ecosystems – Synthesis report. Pretoria: South African National Biodiversity Institute and Department of Environmental Affairs; 2012. Available from: http://www.sanbi.org/nba

19. Cvitanovic C, Hobday AJ, Van Kerkhoff L, Wilson SK, Dobbs K, Marshall NA. Improving knowledge exchange among scientists and decision-makers to facilitate the adaptive governance of marine resources: A review of knowledge and research needs. Ocean Coast Manage. 2015;112:25–35. http://dx.doi.org/10.1016/j.ocecoaman.2015.05.002

20. Choi BCK, Pak AWP. Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: 1. Definitions, objectives, and evidence of effectiveness. Clin Invest Med. 2006;29:351–364.

21. Max-Neef MA. Foundations of transdisciplinarity. Ecol Econ. 2005;53:5–16. http://dx.doi.org/10.1016/j.ecolecon.2005.01.014

22. Stock P, Burton RJF. Defining terms for integrated (multi-inter-trans-disciplinary) sustainability research. Sustainability. 2011;3:1090–1113. http://dx.doi.org/10.3390/su3081090

23. South African Department of Environmental Affairs (DEA). Action For Climate Change, Conference Statement, 20 October 2005. Pretoria: DEA; 2005.

24. South African Department of Environmental Affairs (DEA). Bojanala, Daily Journal of the National Climate Change Conference. Issue 1: 2005 October 17. Pretoria: DEA; 2005.

REVIEWER:
Menán du Plessis

AFFILIATION:
Department of General Linguistics, Stellenbosch University, Stellenbosch, South Africa

EMAIL:
Menan.du.Plessis@gmail.com

Towards the study of South African literature as an integrated corpus

The title of this new work by Helize van Vuuren refers to a dancing rattle made from dried springbok ears laced together by a thong, and is an appropriate metaphor for her set of eight gracefully interlinked academic essays in which she explores the possibility of a more inclusive approach to South African literature, looking in particular at intersections between literary productions in various indigenous Khoesan languages – but mainly the !Ui language, IXam – and Afrikaans. Along the way she gives thoughtful consideration to many of the most pressing questions in current South African discourse studies, such as the issue of possible cultural appropriation, how to meet the expression of historical trauma with adequate sensitivity, how to traverse the gap between oral and written literature, how to assess material available only in translation, and in particular, how to further an appreciation of South African literature as a multilingual and multimodal but nevertheless integrated corpus.

For many South Africans of a particular social group and class, South African literature was often narrowly imagined during the 1960s to be written almost entirely in either English or Afrikaans, by people of colonial descent. It was only slowly that we discovered giants like Wally Serote and Dennis Brutus – and no doubt at least partly because they wrote in English – so that when the landmark *Penguin Book of South African Verse* appeared[1] in 1968 with its inclusion of poetry translated from Sotho, Xhosa and Zulu, and even a few fragments of translated material intended to represent Khoi and San lyrics, it was for many a moment of epiphany. It was at this same time that a particular kind of neo-Romanticism was emerging in the West, with manifestations in movements such as the hippy counter-culture, the experimentation with heightened consciousness and mind-altered states, the 'Back to the Land' movement, the beginnings of environmental awareness, and the turning to non-Western cultures in quest of wisdoms perceived to be more ancient and more authentically in harmony with nature. This zeitgeist was part of the context in which the members of the famous Marshall family worked over many years to produce their documentaries in the form of extensive film footage and books about the Juǀhoan of Nyae Nyae, while it was at this time also that Bleek and Lloyd's *Specimens of Bushman Folklore* was re-introduced to the South African public, also in 1968, in a facsimile reprint of the original 1911 edition.[2]

With the rediscovery of the Bleek and Lloyd manuscript notebooks by Roger Hewitt in the 1970s, there was an intensification of interest in the IXam material, and while some scholars continued in the romantic vein of the earlier period – even to the extent of perpetuating the fascination with shamanism and states of trance, or invoking universal archetypes – in other cases, new insights came to the fore. As a result of this later work, it is better appreciated today that the narratives were of course dictated to linguists, rather than delivered in their usual manner and in their more usual context with a spontaneously reacting and participating live audience. The process of transcription may have been laborious, and has almost certainly led to a loss of vividness, including details of gesture, speech mimicry and song that would have been part of a living performance. It is also better understood today that the IXam speakers were not channelling disembodied voices from some timeless lost world, but that both the speakers and the linguists who worked with them in the later part of the 19th century were inevitably embedded within a particular historical context and social milieu.[3] The recognition of such factors has prompted literary scholars to search for new ways of approaching the material, with Michael Wessels, for example, proposing and demonstrating an 'intertextual' method, in which he considers the IXam corpus as a whole – although always in its translated form only – and where, in order to make sense of a particular term in a particular text, he makes comparative reference to the uses of that term throughout the entire available 'network of signifiers'[4].

Van Vuuren writes that she began her own academic engagements with the IXam material during the 1990s, when she was struck by the continued lack of its proportionally appropriate incorporation into anthologies of South African literature. As a core part of her own project, Van Vuuren here devotes two extended and ground-breaking chapters to the neglected body of work by Gideon von Wielligh – in particular his early 20th-century volumes of *Boesmanstories* – which, she writes, 'could feasibly be seen as the missing link between IXam orality and the beginning of one stream of Afrikaans literature'.

The IXam stories developed by Von Wielligh were based (like the Khoekhoe and Swazi material he also published) on stories told to him by elderly members of communities encountered by him throughout his childhood and adult working life, chiefly in the vicinity of the Cederberg and Calvinia, but also in regions now known as the Northern Cape, Mpumalanga and the Free State. The essential point about these interactions is that they took place through the shared medium of Afrikaans, at a time when Afrikaans itself was hardly a written language, so that aspects of Von Wielligh's style have their origin, as Van Vuuren notes, 'in its own Afrikaans orality'. While Von Wielligh's 'Bushman stories' have generally been relegated to the category of children's literature, and have often been dismissed as derivative or even concocted, Van Vuuren makes a convincing case for their legitimacy, while their value, she suggests, is twofold. On one hand, the IXam (and Khoekhoe) stories collected by Von Wielligh are notable for 'the poetic sparks they lit in many instances, and from which Afrikaans poetry grew in the early 20th century.' On the other hand, Von Wielligh's IXam archive may be seen as a valuable supplement to the Bleek and Lloyd archive, and can be drawn on in the same way for purposes of the intertextual method of elucidation.

Van Vuuren moves on to describe an approach she terms 'contextual', which complements the intertextual one and is essentially multidisciplinary, involving cross-reference to fields such as archaeology, anthropology and botany, while also acknowledging the historically rooted nature of the material. She demonstrates such a contextualised reading in Chapter 4, in which she focuses on the English translation of a small and almost inscrutable IXam lyric from the published *Specimens*, 'A song sung by the star !Gãunǔ and especially the Bushman women', which

asks, 'Does the ‖garraken flower open?', and twice replies that 'The ‡kŭ-kx'am is the one which opens'. The ‖garraken were identified by Lloyd as edible bulbs, while the second flower was identified by her as the rain daisy (or ox-eye daisy), *Dimorphotheca annua* (now *pluvialis*), which is a large cream-coloured daisy notable for its dark centre. (The name given by Dia!kwain may have been a borrowed Kora expression meaning 'black mouth'.) In an attempt to make sense of the four lines, Van Vuuren makes reference to aspects of the hunter–gatherer lifestyle, and discusses the importance of the bulbs (*uintjies*) that were traditionally harvested from the wild by the women. With this wider context considered, she concludes that 'the questions in the song are probably either mistranslated or intended to mean, "Are the *uintjie* flowers open already?"' The interest of the women, she suggests, 'is not in the daisy per se, but rather in the hidden edible corm and its as-yet unopened flower'. She proposes then that the last line might better be translated as 'The daisy is open'. (This suggested replacement for Lloyd's originally given version will be revisited below.)

Perhaps the most powerful essay in the collection is the book's third chapter, which discusses the personal testimonies of the ǀXam speaker, ‖Kabbo, and a Swazi man identified only as 'Rooizak' who was condemned to death in 1874, and who conversed at length with a German missionary before his death. (The German records of this agony-filled encounter were rediscovered a century later and translated into English by Peter Delius.) Her study of these two texts stems, Van Vuuren explains, from 'an increase of interest in holocaust literature', which, in the context of contemporary discourse studies in South Africa, takes the form of a growing interest in the expression of historical trauma. 'Both testimonies', she says, 'have a prison experience as the starting point and cause of trauma'. She later adds:

> In the intense dialogue between the Swazi prisoner and the evangelists nothing is more striking than the conflict between their different cultures, different justice systems, different customs and the absolutely powerless situation in which Rooizak finds himself. In spite of the heavily mediated nature of this text, it is still one of the most striking South African testimonies of one man's trauma and spiritual torture.

Van Vuuren does not explicitly make the point herself, but it is indeed one of the very manifestations of such trauma that the South African story is still all too often told in the languages and louder voices of those from formerly dominant classes.

In the second part of the book, which begins with a philosophical essay (Chapter 5) on 'The Bushman in our Consciousness', Van Vuuren moves on to examine such issues as the question of possible 'cultural appropriation', through essays on the novels of Piet van Rooyen (Chapter 6), and the poetry of Antjie Krog, Stephen Watson and Eugène Marais (Chapter 7), and D.J. Opperman (Chapter 8). The 'Namibian oeuvre' of the Afrikaans writer Piet van Rooyen spans three novels that appeared between 1994 and 2001, and which arose out of the author's personal experience as a development worker among the Juǀʼhoan at Nyae Nyae. Van Vuuren seems to be somewhat ambivalent in her assessment of this work, although it is clear that she has reservations about the author's romantic nostalgia for some imagined lost age, and the stereotypical projection of the San as 'children of nature'.

The essay on the three poets revisits the bitter saga that began when Stephen Watson notoriously accused Antjie Krog of plagiarising his own poeticised versions of some of the ǀXam material (in fact, of Bleek and Lloyd's English *translations* of that material). As a way of throwing more light on issues surrounding the Watson–Krog spat ('a long, distasteful and unnecessary controversy'), Van Vuuren offers us a fresh consideration of Eugène Marais' *Dwaalstories*, which – much like the *Boesmanstories* and *Dierestories* of Von Wielligh – are supposed to have been obtained, through the medium of Afrikaans, from elderly storytellers of 'Bushman' and Korana origin. Marais' stories were obtained in the Waterberg region of the province now known as Limpopo, and it is this material that was the spark and source material

for his own *Die Lied van die Reën* ('The dance of the rain'). Van Vuuren highlights the complexity of the notion of 'cultural appropriation' in such a context, where it is material from an oral tradition that has been the source of inspiration. Once again, Van Vuuren seems ambivalent in her final assessment, suggesting that any 'migration of the oral tradition' of the San is on one hand a form of tribute, and may also have a function of preservation, while on the other there is an undeniable benefit that generally accrues in one direction only, in the form of heightened stature for the writer who draws on the material.

Van Vuuren's collection of essays is without question a valuable addition to the changing field of South African literary studies. Having said this, and at the risk of seeming to carp, there are a few issues that cannot be left unaddressed. For one thing, the lay user should exercise considerable care when consulting Dorothea Bleek's *Bushman Dictionary*[5], which lumps together data from a range of different sources of greatly varying quality, for languages that belong not only to the !Ui and Taa divisions of the TUU family (Bleek's 'Southern Bushman'), but also to the JU family ('Northern Bushman'); and which even covers languages ('Central Bushman') that in reality belong to the KHOE family, which includes not only Kalahari varieties such as Naro and Khwe, but also Khoekhoe varieties such as Nama, Dama and Kora (or !Ora). It is only data labelled 'S1' in the *Dictionary* that illustrates ǀXam itself, while as a general rule it is only safe to expand a search for meanings of ǀXam words to entries for related !Ui languages, such as those labelled 'S2' and 'S3'.

The greatest issue, however, and one not by any means unique to Van Vuuren's work, concerns a persistent belief that the ǀXam material cannot be directly accessed in the original language. While both intertextual and contextual approaches have much to offer, as Van Vuuren herself demonstrates, in the end it is surely nothing less than the text itself that we should take as our starting point. It is a poor excuse to protest that we are powerless to do so simply because there are no speakers left: linguists have been able, after all, to establish the syntax of extinct languages such as Ancient Egyptian and Hittite, where they have had far less to work with! It is true that we do not yet have a comprehensive reference grammar for ǀXam (although linguists based in Germany are currently working on one) – but we have ample resources to proceed with in the meantime. Bleek and Lloyd left us a wealth of carefully transcribed material, in many cases accompanied by detailed translation or at least partial glosses, while it is hardly as though these scholars undertook *no* linguistic analysis at all: Lloyd's detailed notes on the morphology and syntax of the language are reflected in the numerous entries that were later incorporated into the *Dictionary* by her niece, who also collated them into a sketch grammar.[6,7] (Tom Güldemann has provided an updated version of this sketch in which he uses a more contemporary linguistic terminology.[8])

We already know enough about the ǀXam language to realise, for example, that the revised reading Van Vuuren proposes (p.100) for the *uintjieblom* song is untenable. The fourth line of the song, which she would like to re-interpret as 'the daisy is open', is:

> ‡kŭ-kx'am kang ‖kuang ā ‖khou beri-sing

This line contains two aspectual markers, *kang* and ‖kuang, which both signify a continuous action or process, while the auxiliary ‖khou, which seems to have been associated with verbs of motion or process, gives the implication 'become'. It is difficult to find an idiomatic English equivalent for all of this, but the sense of the original line can perhaps clumsily be conveyed by 'the ox-eye daisy is indeed being the one that goes open', or as it was given by Lloyd in the first place, 'the ‡kŭ-kx'am is the one that opens'. One way to make sense of this might be to treat the little song as a rubric for distinguishing the more useful geophytes from mere daisies (which do not have bulbs), as the flowers of a daisy habitually close at night and open up in the morning, which is the time when the women would have gone out in search of food.

These minor notes aside, Van Vuuren's string of sensitive and resonant essays unequivocally makes a contribution to current South African literary studies, and provides a significant impetus to the quest for a more integrated concept of South African literature. Someday perhaps,

when our universities start to transform their academic programmes, studies of this kind may even take place within the ideal context of a holistic school of languages, linguistics and literature.

References

1. Cope J, Krige U, editors. The Penguin book of South African verse. Harmondsworth: Penguin Books; 1968.

2. Bleek WHI, Lloyd LC. Specimens of Bushman folklore. London: George Allen & Co.Ltd.; 1911; facs. repr. Cape Town: Struik; 1968.

3. Bank A. Bushmen in a Victorian world: The remarkable story of the Bleek–Lloyd collection of Bushman folklore. Cape Town: Double Storey; 2006.

4. Wessels M. Bushman letters: Interpreting IXam narrative. Johannesburg: Wits University Press; 2010.

5. Bleek D. A Bushman dictionary. New Haven, CT: American Oriental Society; 1956.

6. Bleek D. Bushman grammar. Z Eingeborenen-Sprachen. 1928/29;19:81–98.

7. Bleek D. Bushman grammar. Z Eingeborenen-Sprachen. 1929/30;30:161–174.

8. Güldemann T. Southern Khoesan (Tuu). In: Vossen R, editor. The Khoesan languages. Milton Park: Routledge; 2013. p. 241–249, 419–431.

Military psychologists as scientists and practitioners

Military psychology is a sub-discipline of psychology that encompasses a wide range of academic and professional activities such as assessment and measurement (monitoring and selection), diagnosis, intervention (prevention, support, treatment and rehabilitation), training and development, human engineering, organisational development, and research. Because of its comprehensive scope, psychologists from different registration categories (such as clinical, counselling, industrial and research) contribute to different areas of military psychology. Military psychology has been practised in South Africa for almost five decades. In the rest of Africa, military psychology is not as formalised and some may view it as still in its infancy.

The book under review, *Military Psychology for Africa,* focuses on the application of psychological principles and methods in military environments. This beautiful book will be welcomed by psychologists and mental health-care practitioners whose clients include members of the security forces and their dependants. The publication of the book is a landmark in the history of psychology in Africa. Now, for the first time, African researchers and practitioners have access to a comprehensive book that acknowledges the unique social, cultural, economic and political dynamics that influence military activities on the continent.

The editor of the book, Gielie van Dyk, is a clinical psychologist and professor of psychology at the South African Military Academy (Stellenbosch University) in Saldanha, South Africa. The Military Academy offers professional military higher education and attracts students from several countries on the African continent.

Van Dyk invited a substantial team of experts – comprising 16 authors and 11 reviewers – to contribute to the book. The authors include academics and practitioners (both military psychologists and private practitioners) and most (10) are members of the South African National Defence Force. They are qualified in industrial (11), clinical (4) or research (1) psychology. A chapter that is likely to attract considerable attention to the role of traditional healing practices in military psychology was written by a traditional healer who also holds a doctorate in psychology. With the exception of one Ugandan contributor, all the authors are from South Africa. The review panel comprised eight South African academics as well as one reviewer each from Australia, Nigeria and Uganda.

The aim of the contributors was to produce a text that psychologists, researchers, students, military trainers and mental health-care practitioners would find useful. In my view, they achieved this aim in a commendable way. The 20 chapters combine theory, research findings, case studies, as well as guidelines from military reports, field and training manuals. The book is likely to stimulate further theorising, application and research.

The book comprises 479 pages and is divided into seven sections comprising between two and six chapters. The chapters focus on the role and contributions of military psychology to:

(1) conventional military operations, (2) peacekeeping, (3) the development and maintenance of morale, (4) the selection of military personnel and civil military coordination officers, (5) the management of trauma, (6) the prevalence and management of HIV/AIDS, (7) the well-being of soldiers (e.g. trauma, HIV/AIDS and traditional healing), (8) the improvement of combat readiness and the identification of critical factors, (9) the challenges facing military families, (10) the psychological well-being and resiliency of military families, (11) the role of personality in military psychology, (12) military leadership (including leadership in Africa), (13) military work readiness, and (14) factors involved in subjective career success of soldiers in Africa.

Two chapters also review the need for military psychology in Africa, with specific reference to the Central African Republic and Uganda.

I am particularly impressed with four aspects of the book: the comprehensive coverage of a range of topical issues, the generic and reader-friendly structure of the chapters, the informative and interesting content, and the inclusion of many recent references. In my view, all the chapters make a constructive contribution to the general theme of the book. However, I believe that five chapters stand out amongst the others. These are Chapter 1 (the role of military psychology in conventional operations), Chapter 4 (the selection of soldiers and military personnel), Chapters 6 and 8 (the management of trauma) as well as Chapter 12 (the factors relating to combat readiness). In my opinion each of these chapters makes a noteworthy contribution to knowledge production in military psychology.

Some features of the publication detract somewhat from its superb standard. Firstly, I acknowledge the challenges faced by second-language authors. Still, I believe that this beautiful and informative book deserves a higher standard with regard to linguistic presentation. For example, the Preface is an important section and one of the first to be read. Regrettably, the editorial standard of the Preface does not match the standard of the publication as a whole. Secondly, the excessive use of bullets in some chapters undermines the cohesiveness and weakens the underlying arguments (e.g. pages 16–24, 103–108, 140–143, 301–308, 315–318, 337–341, 361–367). Thirdly, the contents of the chapters overlap in places, which is not surprising considering that 16 authors contributed to the publication, but in most cases I found the overlap unnecessary and difficult to justify.

In conclusion, this publication represents a ground-breaking and courageous accomplishment. The book is a valuable, practical and comprehensive guide that offers insights and ideas to psychologists, mental health-care practitioners and military leaders alike. Although the book is primarily aimed at an African audience, it addresses universal military issues that psychologists all over the world would be interested in. The authors drew from more than 1100 bibliographic sources (of which approximately 25% were produced by African authors). The publication is a first for the continent and I expect that it will achieve classic status in years to come. Another reviewer considered the book 'long overdue' and I concur. I recommend the book to researchers, practitioners and students who share an interest in military psychology. I hope that the book will receive the attention it deserves, and that readers will use it as a springboard to take military psychology in Africa to the next level. Lastly, I hope that the second edition will be based on a higher percentage of African bibliographic sources, research and content.

From biological control to controlling biology

As an African zoologist, the first part of the title (*The Serengeti Rules*) did not fascinate me – my first thought was that this book was yet another travel diary of an American tourist about the wonderful African wilderness. However, the subtext (*The Quest to Discover How Life Works and Why it Matters*) intrigued me, creating an anticipation of understanding life at a deeper level. The author, a scientist, is a gifted writer with a special passion for telling science stories that bridges the space between the worlds of scientists and non-scientists. But the fact that Carroll is a biochemist/molecular biologist, occupying himself with sub-cellular research far removed from life on the Serengeti plains, may create some apprehension. He does not see the understanding and mastery of control systems as a negative but instead views this paradigm shift in the words of Paul Simon – '*These are the days of miracles*'. In the Introduction, Carroll lays the foundation for what is to follow: the importance of understanding biological control, i.e. the rules of regulation not only at molecular level but also at ecological level. Carroll poses the question: Why are we humans not applying the same vigour we apply to understanding control at the molecular level to the understanding of rules at higher levels? This is a book about the stories of pioneering scientists who relentlessly worked to understand biological control systems and how this knowledge has helped – and should continue to help – us to manage life on the planet.

Part I of the book, titled 'Everything is regulated', starts with a stand-off between the author and an elephant bull in Tarangire National Park in northern Tanzania. Carroll recaps what happened, the 'fight-or-flight' response, and outlines the research from Harvard physiologist Walter Cannon and his students a century ago when they discovered the 'stress response'. Working on physiological shock related deaths during World War I, Cannon discovered the importance of the buffering capacity of bicarbonate, in 'snatching' several wounded soldiers from death. The 'wisdom of the body', referring to the regulation of vital parameters within a narrow range, intrigued him. He coined the process 'homeostasis', which later became recognised as one of the most fundamental principles in physiology. The 'Cannon story' sets the scene and at the end of Chapter 1, Carroll makes the homeostasis link to numbers regulation in nature and Charles Darwin's remarks about elephant numbers. The next story though is not about Darwin, but about Charles Elton, one of the founders of modern ecology, who was also fascinated by the regulation of animal numbers. From early on, Elton figured out and published on food chains and complex food webs. His fascination with variation in population numbers started when he read about linked fluctuations in species in the food chain. Lemming populations showed these oscillations, and a similar phenomenon was seen in Canadian rabbit populations – called the Lynx relationship. At age 26, Elton wrote his 200-page classic *Animal Ecology* in just 85 days! In this book, he established several patterns and ecological principles, and emphasised the fundamental importance of regulating animal numbers in a food chain context, with linkage to Cannon's homeostatic control. Carroll skillfully draws a parallel between the contributions to understanding regulation processes in physiology by Cannon and in animal numbers by Elton, thereby converging thoughts towards understanding biological control systems.

In Part II, titled 'The logic of life', Carroll uses a quote from Monod and Jacob: 'Anything that is found to be true of *E. coli* must be true of elephants'. Carroll reiterates that although ecological and physiological regulation had been established by Elton and Cannon, they were only at the start of untangling the 'rules' of regulation. He proceeds to tell the stories of how the well-known rules of physiological regulation were discovered, specifically through the work of two French biologists, Jacques Monod and Francois Jacob, who described the 'logic' of life. At this point, Carroll reveals the drive and 'logic' behind writing this book about 'analogous rules of regulation and logic' operating not only at the molecular level but also at the ecological level. He convincingly shows that these principles should be taken into account when considering human and environmental health issues, especially when the rules of regulation are broken. Not surprisingly, Chapter 3 forms the foundation of the book, with the Monod and Jacob model surfacing in the chapters to follow. Monod was one of the co-founders of the field of molecular biology and won a Nobel Prize for his role in unravelling the regulation rules at the molecular level using bacteria as model. It was only after Monod discovered a new player in the enzyme regulation game, and managed to 'flip the logic around', that he found that the 'inducer' actually controls a 'repressor', creating a so-called 'double-negative logic'. In order to make progress in understanding how the 'repressor' worked, Monod teamed up with Francois Jacob. Carroll highlights the keystone discovery of negative feedback systems and the importance of Monod and Jacob's discovery of enzyme control reinforcing the generality of these 'rules of regulation'.

In Chapter 4, Carroll continues the historical expedition, telling the reader about the unravelling of the regulation of cholesterol. Carroll describes the courage of scientists, and the importance of serendipity, when telling the tale of the US scientist Ancel Keys. Keys put his own body on the line in a high altitude human experiment to study how the body responds and adjusts to very high altitudes. The high altitude camping got Keys interested in survival diets during the war of 1944, and after the war contrasting statistics regarding deaths from heart disease got his attention. After discovering that poor people generally had lower cholesterol levels in their blood and suffered fewer cardiovascular-related deaths than richer American people, he launched several large-scale studies, including an international study in 1958 with 12 000 men on different diets. Keys raised awareness by stating that 'what people were eating was making them sick'. Against this background, Carroll tells the story of heart disease risk and cholesterol regulation and treatment, starting with two adolescent patients of Goldstein who suffered heart attacks at an early age. He explains how US scientists Goldstein and Brown discovered the presence and function of membrane-bound LDL receptors and their absence in patients suffering from hypercholesterolemia. Carroll then asks whether these 'insights into the rules of cholesterol regulation' would ever find their way into treating the disease? Although the answer is yes, it took a rollercoaster ride to get to this point. He tells the story of a Japanese scientist, Akira Endo, who independently worked with fungi and discovered statin therapy, 'a penicillin for cholesterol' – a potent inhibitor of the reductase enzyme Goldstein and Brown identified as an important early

regulator in the cholesterol control pathway. Although Goldstein and Brown won a Nobel Prize, Endo did not get much recognition (or money) for his discovery and inventions with fungal extracts. In 2003, this oversight was rectified with a conference held in Endo's honour in which Goldstein and Brown acknowledged his contribution to the significant decline in coronary deaths since Ancel Keys raised the warning flag.

The last chapter in Part II, dealing with stories about regulation at cellular level, tells the tale of cancer cells. The title of Chapter 5 – 'Stuck accelerators and broken brakes' – says it all. Carroll introduces the US scientist, Janet Rowley, who made breakthroughs in our understanding of the genetic basis of cancer, starting with leukemia, and described specific chromosomal translocations. From this first evidence for the existence of viral oncogenes, Carroll takes the reader along the path towards the point where it becomes clear that cancer is also a disease of regulation, metaphorically described by Carrol 'like a stuck accelerator in a car' situation. Here the Monod and Jacob model surfaces again, including the possibility of interference with the negative feedback, so 'foot slipping off the (genetic) brake' phenomenon. The next challenge was to find a way to repair the 'genetic brakes' in cancer cells.

To this point in the book, Carroll skillfully introduces the reader to the rules and logic that regulate intracellular compounds and cell numbers. But finally, in Part III, he sets out to take the reader on field expeditions to confirm that the ecological rules are analogous to the molecular rules of regulation. However, this all happens at a different level with limited opportunities for controlled experimentation. As Carroll put it on the first page, 'The anticipation in our Land Cruiser rises', and so it is for the reader. As a zoologist, I enjoyed Chapter 6, titled 'Some animals are more equal than others'. Carroll tells the story of Robert Paine, his life as a biology student and his interaction with the ecologist Fred Smith. Paine realised that to understand nature, one needs to understand the regulation of numbers, and to do this, he needed situations in which he could break the rules (intervene) and conduct so-called 'kick it and see' ecology experiments. He found such an opportunity at Mukkaw Bay, on the Pacific coast, where he removed starfish (predators) and tossed them into the ocean – according to Carroll, 'one of the most important experiments in the history of ecology' – thereby inventing the concept of a 'keystone species'. Carroll discusses more case studies – from otters controlling sea urchins (the latter controlling kelp) to fish–herbivore–algae cascades – to establish another concept, trophic cascades, characterised by a top-down control system but with the same principle of a double negative logic described by Monod and Jacob. As Carroll remarks: 'What began as tossing of starfish, has led to two fundamental insights' (the first two Serengeti rules). Firstly, keystones exert significant effects on the stability and diversity of their communities (hence not all species are equal); secondly, some members of food chains have strong (top-down) indirect effects on lower trophic levels. In Chapter 7, Carroll takes us back into the Serengeti to introduce Tony Sinclair, an Oxford scholar who decided to spend his life in the Serengeti, starting with the understanding of the rules of regulation in an exponentially increasing buffalo population in the 1960s and 1970s. Carroll tells the story of the rinderpest virus and the control of buffalo and wildebeest, showing that a pathogen could also have a keystone role. The control of the virus unleashed amazing changes waiting for Sinclair to discover,

such as the outbreak of trees and an increase in giraffe numbers. Carroll yet again connects the dots and draws attention to multiple levels of negative regulation, indeed triple negative logic, while conservationists initially blamed elephants for the loss of trees in the Serengeti! The role of wildebeest as competitors for resources becomes evident when following the numbers of grasshoppers, gazelles and others. Carrol now asks: What controls wildebeest in the absence of rinderpest? Could it be the body size predation issue or poaching? And what controls the larger species if predation does not? Density-dependent regulation or negative feedback as in molecular regulation? But then Sinclair observes mass deaths of wildebeest during a severe drought of 1993. There are thus two major ways of numbers control: predation and availability of food. But, Carroll asks, is there a way for larger antelope, like wildebeest, to circumvent these limiting factors? Migration? Carroll summarises Sinclair's Serengeti story by stating that our knowledge about regulation explains and answers so many 'Why' questions, but more importantly that the general rules of regulation turn out to be remarkably similar. Understanding these rules of regulation allows us to diagnose where they are broken and hopefully take the effort to cure them.

In Chapter 8, titled 'Another kind of cancer', Carroll introduces the reader to well-selected ecosystem examples in which 'regulatory rules' were broken (ecological cancers). Relevant here in South Africa, there is a discussion on the blooms of so-called blue-green algae, *Microcystis* (correctly known as cyanobacteria), known to produce lethal toxins (microcystin) in water associated with pollutants like phosphate compounds and glyphosate herbicides. Through several more examples (insect plagues, baboon and cownose ray population explosions), Carroll shows that common to all, just like tumour suppressors, understanding the rules of regulation is key to healing. Carroll also adds several more ecological cancers, such as brown planthopper insects plaguing the rice fields in Cambodia, explosions of olive baboon populations in Ghana and cownose rays in the USA. The common denominator among these three, very different, ecological cancers, is a decrease in natural predators (spiders, lions and sharks), just like tumour suppressors that act as proliferation brakes. Carroll recalls a comment by Sinclair that all problems of wildlife populations can be grouped into three categories: 'too many, too few and too much'. In the next chapter, Carroll presents a selection of examples of well-thought-out bio-manipulations. Striking is the story of the wolves of Yellowstone National Park. This brings us to one of the most ambitious restoration projects in Africa – that of the Gorongosa National Park in central Mozambique. In this instance, scientists had to deal with the communities surrounding the park and with poaching – humans as the top predator.

In the Afterword, Carroll tells one last story of the smallpox virus and the man who took the trouble to save the 'house on fire' – very similar to the 'rinderpest story', another example of how humans can control biology when human health is at stake.

This book by Sean Carroll is a life changer! A must read for anyone vaguely interested in science history to appreciate how miracles in science discovery happen and to understand the universality of biological regulation – 'Serengeti rules' that could be applied in therapy on a global scale.

The phenomenon of skin lightening: Is it right to be light?

AUTHORS:
Lester M. Davids[1] (iD)
Jennifer van Wyk[2]
Nonhlanhla P. Khumalo[2]
Nina G. Jablonski[3]

AFFILIATIONS:
[1]Department of Human Biology, University of Cape Town, Cape Town, South Africa

[2]Division of Dermatology, Groote Schuur Hospital, Cape Town, South Africa

[3]Department of Anthropology, Pennsylvania State University, State College, Pennsylvania, USA

CORRESPONDENCE TO:
Lester Davids

EMAIL:
Lester.davids@uct.ac.za

ARTICLE INCLUDES:
× Supplementary material
× Data set

Chemicals capable of lightening the skin – variously known as skin-bleaching, skin-lightening, depigmenting, skin-evening and skin-brightening agents – are among the most commonly used skin preparations in the world. Globally, Africa reportedly exhibits a high prevalence of skin lightener use. In this review, we provide both clinical and social perspectives on skin lightener use in Africa, with particular emphasis on South Africa. We narratively explore the timeline associated with skin lightener use in South Africa and attempt to interweave the social rhetoric of this specific paradigm. Despite the risks associated with exposing the skin to known constituents of these formulations, such as hydroquinone and mercury, chronic use continues. In spite of legislation banning hydroquinone and mercury in cosmetics in South Africa, these ingredients are present in widely available products. We recommend better implementation of policies and greater ethical responsibility of multinational cosmetic companies in addition to the initiation of a system of random product testing and penalties that could improve industry compliance.

Significance:

- There is a high prevalence of skin lightener use in Africa.

- Despite legislation banning harmful compounds, these compounds are still used in skin lightening formulations.

- There is an urgent need to implement policies and recommendations for preventing the influx and illicit sale and use of untested skin lighteners.

Introduction

The global use of skin lighteners is not new. Although skin lighteners have been used for centuries, only in the last century has the production of these materials become commercialised and global.[1] Currently, Africa has the highest number of studies reporting on the global prevalence of skin lightener use (see Table 1). Despite toxic systemic effects, application of topical skin lighteners remains popular throughout the African continent. The market for commercial skin lighteners has grown in the Caribbean, Asia and the Far East. The chief reasons for this growth remain varied, but, undoubtedly, are strongly linked to historical racism, the perceived social benefits of lighter skin and the marketing expertise of the multinational cosmetics companies now involved in their production.[2-4] More recently, incentives for skin lightening use by pigmented people arise in part, and mostly subconsciously, from colourism – that is, the preference for lighter skin tone because of its association with positive social outcomes.[1]

Sub-Saharan Africa

The prevalence of skin lightener use across the African continent (Table 1) has become a common part of life in African communities. In the late 1960s, 60% of urban African women reported using skin lightener formulations, making these formulations the fourth most commonly used household product (after soap, tea and tinned milk).[5] Further evidence of these practices is derived from local vernaculars – in Mali and Senegal, the terms '*caco*' and '*xeesal*' are used to describe the practice[6], and in Ghana, the term '*nensoebenis*' describes ochronosis (the hyperpigmentation and damage to the skin as a result of chronic skin lightener use)[7]. In South Africa, in various ethnic languages related to indigenous tribes, the practice is described as '*ukutsheyisa*' (isiXhosa for 'to chase beauty') and '*ukucreamer*' (isiZulu for 'applying creams on the skin'). Around Johannesburg, the word '*mashubaba*' is used in urban township slang to describe ochronosis. The motivation driving the practice is often the desire to lighten one's skin because of a perceived notion of increased privileges, higher social standing, better employment and increased marital prospects associated with lighter skin. This perception, coupled with influential marketing strategies from transnational cosmetic houses using iconic celebrities, increases the allure for women primarily, but also increasingly, men.[8,9] Unfortunately, the main fear is that the presence of these legally available products could potentially cloud the distinction of the consumer between products that are tested and those that are damaging and illegal. This distinction needs to be pertinently highlighted.[9]

Use of skin lightening agents and their biological consequences

A diverse range of compounds is currently used in skin lightener preparations. However, the 'big four' – mercury-containing compounds (known as mercurials), hydroquinone and its derivatives (monobenzyl ether and monomethyl ether), potent corticosteroids and retinoids – are still considered the primary sources of damage. The biological consequence of each of these is considered below.

Mercurials

Mercurials (mercury and its derivatives) are the oldest known skin lighteners.[10] They act to lighten the skin by replacing the copper required for tyrosinase activity and antagonistically inhibiting melanin in melanocytes.[11] Although banned by the US Food and Drug Administration in 1973 and by the European Union (EU) in 1976, mercurials continue to be produced and exported to countries in Africa and the Caribbean.[1] Moreover, mercury poisoning through the use of skin lightener preparations continues to be reported from Africa[12]; Europe[13,14]; USA[15,16]; Australia[17] and Hong Kong[18].

Table 1: Summary of the global prevalence of skin lightening practices in Africa and the most commonly used skin formulations

Country and study type	Prevalence of skin lightening (*n*)	Most commonly used skin lightener	Reference
Nigeria, community study	72.4% females (*n*=450)	Hydroquinone; mercury; corticosteroids; locally concocted soaps/creams	Adebajo[54]
Nigeria, hospital study	40% of female and 2% of male patients (*n* not given)	Hydroquinone; class I and II steroids; plant derivatives	Ajose[55]
Nigeria, hospital study	58.7% females (*n*=414)	Hydroquinone; steroids; mercurials; kojic acid; alpha hydroxyl acids	Nnoruka et al.[56]
Senegal, maternity centre survey	68.7% females (*n*=99)	Hydroquinone; corticosteroids; mercurials	Mahé et al.[48]
Senegal, hospital study	67% females (*n*=65)	Not specified	Diongue et al. [57]
Mali, hospital study	Not specified in study	Hydroquinone; steroids; mercurials	Faye et al.[6]
Togo, community study	58.9% females (*n*=910)	Hydroquinone; mercurials; steroids	Pitcheé et al.[49]
Burkina Faso, community study	44.3% females (*n*=447)	Phenolics; steroids; mercurials; combination of unknown agents	Traore et al.[58]
Togo, hospital study	40.7% females (*n*=119)	Hydroquinone; topical corticosteroids	Kombaté et al.[59]
Somali, community study	Not specified in study	Mercurials	Adawe and Oberg[60]
South Africa, hospital study	32.6% females (*n*=600)	Hydroquinone; corticosteroids; niacinamide; vitamin C; traditional plant combinations	Dlova et al.[31]
Rwanda, survey	35% females (*n* not given)	Not specified	Kamagaju et al.[61]
Cameroon, hospital study	43.6% females (*n*=658)	Hydroquinone; mercurials; steroids; plant derivatives; vitamin C	Kouotou et al.[62]

Hydroquinone

Hydroquinone, chemically known as 1,4 dihydroxybenzene, is an effective skin lightener which acts by preventing the synthesis of melanin through competing with its natural substrate, the enzyme tyrosinase, in a reversible reaction.[19-22] In addition, it inhibits both DNA and RNA synthesis in epidermal melanocytes[22] and becomes toxic to melanosomes[11]. Banned by South Africa in 1980, followed by the EU and Japan in 2001, hydroquinone continues to be available in the USA in dermatologically based preparations[1] and its use for treating conditions such as melasma, pregnancy-induced hyperpigmentation, lentigines and other skin depigmentary disorders such as vitiligo[23] continues to raise opposition by dermatologists.

Chronic, unsupervised use of hydroquinone could result in a hyperpigmented, skin-altering condition called exogenous ochronosis (Figure 1). Characterised by progressive asymptomatic hyperpigmentation, the lower layers of the skin show signs of collagen and elastic fibre degeneration.

Corticosteroids

Most often used in combination with hydroquinone and/or mercurials in skin lightener preparations, corticosteroids are categorised according to their potencies from class I (most potent) to class VII (least potent).[24] Clobetasol propionate, licensed in 1973, is a class I fluorinated corticosteroid (generally prescribed at a concentration of 0.05%) which causes vasoconstriction and immunosuppression and results in hypopigmentation or skin lightening.[10,24] As clobetasol propionate is very stable and able to penetrate the epidermal and dermal skin layers, its inclusion in skin lightening preparations was immediate and its use was manifested in numerous African and Afro-Caribbean communities.[25,26]

Photos: N. Khumalo; published with permission from the individuals.

Figure 1: Evidence of exogenous ochronosis resulting from chronic use of skin lightening formulations.

Unfortunately, chronic use of this highly potent corticosteroid leads to severe consequences such as cutaneous atrophy, with clinical manifestations of skin thinning and fragility, telangiectasia and striae.[24]

Retinoids

Today, the vast majority of skin lightening agents include a form of retinoid. Topical retinoids include all trans-retinoic acid, 13-cis-retinoic acid (isotretinoin), retinol, retinaldehyde, tazarotene and adapalene.[27] Well known for their anti-ageing effects[28], retinoids function to lighten hyperpigmented skin by decreasing melanosome transfer from epidermal melanocytes, inhibiting both tyrosinase transcription and melanin synthesis[29,30].

With respect to management of skin lightener use, it is – undoubtedly – the current undivided global opinion of most clinicians and scientists that the inclusion of inorganic and other forms of compounds within cosmetic formulations should be strictly prohibited.[18] However, difficulties ensue with compliance and relate to each country's policies and the enforcement thereof. This situation reinforces the pertinent need for a global standardisation of cosmetic policies and recommendations.

South Africa and depigmenting agents

South Africa, in the context of its political overtones, has a distinctive history within the globalisation of skin lighteners.[31] The period 1960–1970 saw a tremendous increase in the popularity of skin lightener use in South Africa with the market peaking in the 1970s.[5] Whether this peak was because of aggressive marketing and intelligent advertisement strategies by emerging cosmetic companies or politically motivated cannot be quantified, but the broad appeal of these products continued to rely on the popular ideologies that linked lighter skin to power and beauty.[32] An early dermatological study reported on a 6-year trial consisting of 840 volunteers who were drawn from various race groups with skins varying from very fair to very dark. The volunteers were subjected to open skin tests, 'normal usage' tests and standard 48-h closed-patch tests. In all, over 7000 skin test areas were examined. The results showed that concentrations of hydroquinone of 3% or less produced negligible adverse effects, irrespective of the base or the colour of the user's skin. This study stressed that any confusion of hydroquinone with the hazardous monobenzyl ether of hydroquinone (monobenzone) should be avoided.[33] Another study[34] conducted on 347 patients seen over a 6-month period at the Chris Hani Baragwanath Hospital in Johannesburg further reported on patients presenting with depigmented patches over their skins. The authors found that this effect was a result of the repeated use of an overnight 'depigmenting' cream containing hydroquinone at a concentration greater than 3%.[34] By 1975, the South African government banned the inclusion of all forms of mercury in any skin formulations.

Also in 1975, the South African skin lightener cosmetic industry was estimated at ZAR12.8 million per annum (about USD1.6 million at the current exchange rate). Despite warnings from renowned dermatologists, skin lightening creams continued to be illicitly sold and used for 'facial improvement and the elevation of social standing'[33]. It was claimed at the time that hydroquinone (at a 3% concentration) remained the only substance to reasonably depigment darker skin.[33,34] Findlay et al.[35] quoted patients as saying that if, on application, a product 'stings or burns the skin' that it is 'working well'. This perception of patients has led to the continued, unregulated use of even stronger formulations with consequential skin damage – a condition called exogenous ochronosis (Figure 1).[35] By 1979, there was a growing concern regarding the high occurrence of ochronosis cases and a study by Dogliotti and Leibowitz[36] showed that 6% of 5463 patients seen over a 2-year period presented with skin damage related to the use of marketed cosmetic preparations. These products included greasy paraffin derivatives, bleaching creams, lanoline mixtures and surface moisturising agents.[36] Individual or multiple products were applied in the hope of 'clearing undesirable patchy hyperpigmentation, curing pimples and changing a dark complexion to one which was bright, lovely and seductive'[36]. This statement may have implied the perception that black skin was the antithesis of these descriptions, highlighting to young, impressionable black South Africans the 'need' for ongoing use.

By 1980, the South African cosmetic market industry had swelled to an annual turnover of about ZAR25 million (~USD3.1 million), with the market described as 'phenomenal' and 'ripe for the picking'[37]. Findlay et al.[35] further reported that evidence of ochronosis related to the chronic use of a skin lightener containing 8–15% hydroquinone. This cream was subsequently withdrawn from the market and replaced with the monobenzyl ether of hydroquinone known as monobenzone.[38] These polyphenolics, traditionally used in the rubber and photographic trades[39], work effectively as skin lighteners as they inhibit the enzyme (tyrosinase) primarily involved in producing melanin in the epidermis of human skin[40]. Unfortunately, chronic use leads to ochronosis[35] (Figure 1).

Clinical reports[35,37,41] increased in frequency and their description of the severity of the adverse effects of skin lighteners motivated the government to regulate the amount of hydroquinone in over-the-counter formulations to less than 2% by 1980[32]. Despite this regulation, during the early 1980s no fewer than 58 brands of skin lighteners were still found to be available in pharmacies and supermarkets in the Johannesburg area, and patients in the area were presenting with severe ochronosis.[42]

BOX 1:   Early evidence of skin lightening in South Africa

> It would be remiss in this review involving South Africa and the use of skin lighteners to not mention the influence of the Krok family. Twin sons Abie and Solly became managing directors of Twins Pharmaceuticals Holdings and with their acute business acumen, soon established a cosmetic empire with their 'Super Rose' skin-lightening cream and lotion. 'Super Rose' was so successful that, along with another popular brand, 'Ambi', the range was extended to men, in the form of a stronger version suitable for male skins, called 'He-Man'. These formulations (which were not fully scientifically or pharmacologically tested) made the Krok brothers extremely wealthy. With impetus from the Dermatology Society of South Africa, resistance to these formulations also came from the Black Consumer Union, the Foundation for African Business and Consumer Services as well as the Black Taxi Association, culminating in 1990 with the complete ban in South Africa of all cosmetics containing hydroquinone.[32]

In comparison with the late 1970s and testament to the need for a 'fairer skin', by 1986 in South Africa, the total revenue of skin lighteners was estimated at GBP30 million (about ZAR390 million).[41] The consequence of this market was becoming more evident. A study reported the prevalence of ochronosis in black individuals attending two South African hospitals at 15% in male patients and 42% in female patients; the prevalence was 69% amongst users of skin lighteners.[41] The authors also found an inverse relationship between ochronosis and level of education of the individual. It was noteworthy that even products containing 2% hydroquinone used with a sunscreen were reported to cause ochronosis.[41,42]

Owing in part to the Black Consciousness movement, laws emanating from the US Federal Drug Administration and EU's Drug Administration RAPEX (Rapid Alert System for Non-Food Products), as well as a significant change in the country's political profile, by 1990, South African products containing hydroquinone were completely banned.[32] Moreover, South Africa became one of the first countries in the world to restrict cosmetic advertisements from claiming the ability to bleach, 'lighten' or 'whiten' skin.[5] Despite this restriction, the skin lightener market did not experience the huge downturn expected. In contrast, there was an increase in use which could be explained by a massive rise in local, informal settlements and their own associated markets possibly further influenced by the opening of South African borders to neighbouring countries after 1994.

As the late 1990s approached, consumerism – the ability of advertisements and media to affect the purchasing of goods – became a major influential factor in the sale of skin lightening formulations. Coupled to consumerism, the influence of social media, mobile telephones with roaming applications (apps) and influential celebrity icons emanating from the West (Hollywood in the USA), the Far East (China and Japan) and Asia (Bollywood in India), accelerated skin lightener use beyond what could be globally imagined.[43] In addition, market expansion of large, multinational cosmetic houses as well as the rapid expansion of start-up, entrepreneurial companies, fuelled the global expansion.[9]

Today, skin lightening cosmetics are widespread and new marketing campaigns describing the 'success' of those with fairer skin, have included products attractive to both young men and young women across Africa.[31] However, not all people who use these products do so to lighten their complexion. In a recent South African study of 600 women of African and Indian ancestry, it was reported that 32.7% used skin lightening products, but the main reason cited was treatment of skin problems (66.7%), with skin lightening accounting for only 33.3% of use.[44] This point emphasises the need to inform populations that early and prompt treatment of conditions such as acne vulgaris reduces the risk of scarring.

Policies and recommendations

Despite numerous countries in Africa making a concerted effort to stop the chronic use of skin lightening products through national bans of constituent compounds such as hydroquinone and mercury, there still remains an inconsistent level of regulation within the sector. With many people still engaging in this practice, discovering new means to curb the use of skin-bleaching products and prevent dangerous health consequences remains a pressing area of inquiry. The sustained use of these products is suggested to be primarily a consequence of classification of these products as cosmetic rather than pharmaceutical and a lack of regulation means there is no consistent requirement for ingredient labelling. Williams[45] suggests that had skin lighteners been classified as drugs and not cosmetics, the product licence would have probably long been withdrawn. Moreover, many product labels do not list all the ingredients, and in some developing countries there is evidence of misbranding.[45] Most African countries have regulatory organisations. In South Africa, the watchdog organisation is the Cosmetic, Toiletries and Fragrance Association (CTFA). This association, as in other African countries, controls the policies relating to labelling and regulation and should work closely with governmental sectors relating to importation and availability of products. Unfortunately, the current status quo seems to be a lack of enforcement of existing regulation – a topic that needs to be addressed at the governmental level. However change is evident, and examples include the governments of Nigeria, South Africa, Kenya, Zimbabwe and, more recently, the Ivory Coast, banning the import and sale of skin-bleaching products that contain mercury and hydroquinone.[46] Unfortunately governments cannot address the issue through policy changes alone.[46-49] Hall[50] suggested that in order to initiate real change, destigmatisation of dark skin needs to be advocated, and subsequent studies acknowledging the realities of these motivations are needed. Others suggest that more pressure on corporations to change their concept of beauty may be influential in discouraging skin bleaching.[51,52] The concept of beauty is a tentative and sensitive issue. For example, Eagle et al.[53] posed the question of ethical responsibility of multinational cosmetic companies. They suggest that these multinational companies may be inadvertently (or indeed consciously) sustaining prejudices and stereotypes. Of particular concern is that these companies have the financial resources to reinforce the desirability of these products by using marketing techniques glamorising the use of their products.[53]

Conclusions

The use of illegal ingredients in skin lightener products continues to prevail throughout many African countries. A strong argument for safety exists to justify an interim ban (in all countries) on over-the-counter creams containing hydroquinone, pending long-term trials on the safety of these products. Moreover, the onus should be on manufacturers rather than consumers to establish such safety. Many countries in Africa and the African Diaspora, including Uganda, Kenya, South Africa and Gambia, have banned skin lightener products, while others – such as Ghana, Zambia, Jamaica and, most recently, Ivory Coast, have promoted public health education to dissuade people from using bleaching creams.[9] Overall, Africa, and particularly South Africa, needs to continue research into the motivation for skin lightening. Given the inability of government bans to effectively prevent skin lightening, current studies should investigate other strategies aimed at improving industry compliance,[8,49,52] such as random testing and penalties.

References

1. Jablonski NG. Living color: The biological and social meaning of skin color. Berkeley, CA: University of California Press; 2012.

2. Glenn EN. Yearning for lightness: Transnational circuits in the marketing and consumption of skin lighteners. Gender Soc. 2008;22(3):281–302. http://dx.doi.org/10.1177/0891243208316089

3. Glenn EN. Consuming lightness: Segmented markets and global capital in the skin-whitening trade. In: Glenn EN, editor. Shades of difference: Why skin color matters. Stanford, CA: Stanford University Press; 2009. p. 166–187.

4. Rondilla JL, Spickard P. Is lighter better? Skin-tone discrimination among Asian Americans. Lanham, MD: Rowman and Littlefield Publishers Inc.; 2007.

5. Thomas LM. Skin lighteners, black consumers and Jewish entrepreneurs in South Africa. Hist Workshop J. 2012;73(1):259–283. http://dx.doi.org/10.1093/hwj/dbr017

6. Faye O, Keita S, Diakité FS, Konaré HD, Ndiaye HT. Side effects of depigmenting products in Bamako, Mali. Int J Dermatol. 2005;44(S1):35–36. http://dx.doi.org/10.1111/j.1365-4632.2005.02809.x

7. Dadzie O, Petit A. Skin bleaching: Highlighting the misuse of cutaneous depigmenting agents. J Eur Acad Dermatol. 2009;23(7):741–750. http://dx.doi.org/10.1111/j.1468-3083.2009.03150.x

8. Hunter ML. Buying racial capital: Skin-bleaching and cosmetic surgery in a globalized world. J Pan Afr Stud. 2011;4(4):142–164.

9. Bird D, Caldwell H, DeFanti M. The quest for beauty: Asia's fascination with pale skin. In: Oglesby RA, LeBlanc HP III, Adams MG, editors. Global business perspectives. Beltsville, MD: International Academy of Business Disciplines; 2010. p. 26–32.

10. Olumide YM, Akinkugbe AO, Altraide D, Mohammed T, Ahamefule N, Ayanlowo S, et al. Complications of chronic use of skin lightening cosmetics. Int J Dermatol. 2008;47(4):344–353. http://dx.doi.org/10.1111/j.1365-4632.2008.02719.x

11. Denton CR, Lerner AB, Fitzpatrick TB. Inhibition of melanin formation by chemical agents. J Invest Dermatol. 1952;18(2):119–135. http://dx.doi.org/10.1038/jid.1952.16

12. Saffer D, Tayob H, Bill PL, Baily P. Continued marketing of skin-lightening preparations containing mercury. S Afr Med J. 1976;50(39):1499.

13. Summa JD. Chronic mercury poisoning from cosmetic creams. Munch Med Wochenschr. 1975;117(26):1121–1124.

14. Otto M, Ahlemeyer C, Tasche H, Von Mühlendahl KE. Endemic mercury burden caused by a bleaching ointment in Balken refugees. Gesundheitswesen. 1994;56(12):686–689.

15. McRill C, Boyer LV, Flood TJ, Ortega L. Mercury toxicity due to use of a cosmetic cream. J Occup Environ Med. 2000;42(1):4. http://dx.doi.org/10.1097/00043764-200001000-00004

16. Weldon MM, Smolinski MS, Maroufi A, et al. Mercury poisoning associated with a Mexican beauty cream. Western J Med. 2000;173(1):15–18. http://dx.doi.org/10.1136/ewjm.173.1.15

17. Dyall-Smith DJ, Scurry JP. Mercury pigmentation and high mercury levels from the use of a cosmetic cream. Med J Aust. 1990;153(7):409–410, 414–415.

18. Tang HL, Chu HK, Mak YF, Lee W, Cheuk A, Yim KF, et al. Minimal change disease following exposure to mercury-containing skin lightening cream. Hong Kong Med J. 2006;12(4):316–318.

19. Chan TYK. Inorganic mercury poisoning associated with skin-lightening cosmetic products. Clin Toxicol. 2011;49(10):886–891. http://dx.doi.org/10.3109/15563650.2011.626425

20. Gupta AK, Gover MD, Nouri K, Taylor S. The treatment of melasma: A review of clinical trials. J Am Acad Dermatol. 2006;55(6):1048–1065. http://dx.doi.org/10.1016/j.jaad.2006.02.009

21. Parvez S, Kang M, Chung H-S, Cho C, Hong M-C, Shin M-K, et al. Survey and mechanism of skin depigmenting and lightening agents. Phytother Res. 2006;20(11):921–934. http://dx.doi.org/10.1002/ptr.1954

22. Parvez S, Kang M, Chung H-S, Bae H. Naturally occurring tyrosinase inhibitors: Mechanism and applications in skin health, cosmetics and agriculture industries. Phytothery Res. 2007;21(9):805–816. http://dx.doi.org/10.1002/ptr.2184

23. Jimbow K. Vitiligo: Therapeutic advances. Dermatol Clin. 1998;16(2):399–407. http://dx.doi.org/10.1016/S0733-8635(05)70021-8

24. Pels R, Sterry W, Lademann J. Clobetasol propionate – Where, when, why? Drugs Today. 2008;44(7):547–557. http://dx.doi.org/10.1358/dot.2008.44.7.1122221

25. Olumide YM, Elesha SO. Hydroquinone induced exogenous ochronosis. Niger Med Pract. 1986;11:103–106.

26. Keane FM, Munn SE, Taylor NF, Du Vivier AWP. Unregulated use of clobetasol propionate. Brit J Dermatol. 2001;144(5):1095–1096. http://dx.doi.org/10.1046/j.1365-2133.2001.04213.x

27. Sorg O, Kasraee B, Salomon D, et al. The potential depigmenting activity of retinaldehyde. Dermatology. 2013;227(3):231–237. http://dx.doi.org/10.1159/000354294

28. Ortonne J-P, Marks R. Photodamaged skin: Clinical signs, causes and management. London: Martin Dunitz Ltd.; 1999.

29. Ortonne J-P. Retinoid therapy of pigmentary disorders. Dermatol Ther. 2006;19(5):280–288. http://dx.doi.org/10.1111/j.1529-8019.2006.00085.x

30. Roméro C, Aberdam E, Larnier C, Ortonne JP. Retinoic acid as modulator of UVB-induced melanocyte differentiation. Involvement of the melanogenic enzymes expression. J Cell Sci. 1994;107(4):1095–1103.

31. Dlova N, Hamed SH, Tsoka-Gwegweni J, Grobler A, Hift R. Women's perceptions of the benefits and risks of skin-lightening creams in two South African communities. J Cosmet Dermatol. 2014;13(3):236–241. http://dx.doi.org/10.1111/jocd.12104

32. Thomas LM. Skin lighteners in South Africa: Transnational entanglements and technologies of the self. In: Glenn EN, editor. Shades of difference: Why skin color matters. Stanford, CA: Stanford University Press; 2009. p. 188–210.

33. Bentley-Phillips B, Bayles MAH. Cutaneous reactions to topical application of hydroquinone. Results of a 6-year investigation. S Afr Med J. 1975;49(34):1391–1395.

34. Dogliotti M. Survey of skin disorders in the urban black population of South Africa. Brit J Dermatol. 1975;93(3):259–270. http://dx.doi.org/10.1111/j.1365-2133.1975.tb06491.x

35. Findlay GH, Morrison J, Simson I. Exogenous ochronosis and pigmented colloid milium from hydroquinone bleaching creams. Brit J Dermatol. 1975;93(6):613–622. http://dx.doi.org/10.1111/j.1365-2133.1975.tb05110.x

36. Dogliotti M, Leibowitz M. Granulomatous ochronosis: A cosmetic-induced skin disorder in blacks. S Afr Med J. 1979;56(19):757–760.

37. Findlay G, De Beer H. Chronic hydroquinone poisoning of the skin from skin-lightening cosmetics. S Afr Med J. 1980;57:187–190.

38. Westerhof W, Kooyers T. Hydroquinone and its analogues in dermatology – A potential health risk. J Cosmet Dermatol. 2005;4(2):55–59. http://dx.doi.org/10.1111/j.1473-2165.2005.40202.x

39. Anderson F, Kessler J. Quinone continuous recycle process for electrolytic conversion of benzene to quinone. US Patent US3758392 A; 1973.

40. Parvez S, Kang M, Chung HS, Bae H. Naturally occurring tyrosinase inhibitors: Mechanism and applications in skin health, cosmetics and agriculture industries. Phytother Res. 2007;21(9):805–816. http://dx.doi.org/10.1002/ptr.2184

41. Hardwick N, Van Gelder LW, Van der Merwe CA, Van der Merwe MP. Exogenous ochronosis: An epidemiological study. Brit J Dermatol. 1989;120(2):229–238. http://dx.doi.org/10.1111/j.1365-2133.1989.tb07787.x

42. Phillips JI, Isaacson C, Carman H. Ochronosis in black South Africans who used skin lighteners. Am J Dermatopath. 1986;8(1):14–21. http://dx.doi.org/10.1097/00000372-198602000-00003

43. Hunter ML. "If you're light you're alright": Light skin color as social capital for women of color. Gender Soc. 2002;16(2):175–193. http://dx.doi.org/10.1177/0891243022104895

44. Dlova NC, Hamed SH, Tsoka-Gwegweni J, Grobler A. Skin lightening practices: An epidemiological study of South African women of African and Indian ancestries. Br J Dermatol. 2015;173(Suppl 2):2–9. http://dx.doi.org/10.1111/bjd.13556

45. Williams H. Skin lightening creams containing hydroquinone: A case for a temporary ban. Brit Med J. 1992;305:903–904. http://dx.doi.org/10.1136/bmj.305.6859.903

46. Olumide YM. Use of skin lightening creams. Brit Med J. 2010;341:345–346. http://dx.doi.org/10.1136/bmj.c6102

47. Lewis KM, Robkin N, Gaska K, Njoki LC. Investigating motivations for women's skin bleaching in Tanzania. Psychol Women Quart. 2011;35(1):29–37. http://dx.doi.org/10.1177/0361684310392356

48. Mahé A, Perret J, Ly F, Fall F, Rault J, Dumont A. The cosmetic use of skin-lightening products during pregnancy in Dakar, Senegal: A common and potentially hazardous practice. T Roy Soc Trop Med H. 2007;101(2):183–187. http://dx.doi.org/10.1016/j.trstmh.2006.06.007

49. Pitché P, Kombaté K, Tchangai-Walla K. Cosmetic use of skin-bleaching products and associated complications. Int J Dermatol. 2005;44(s1):39–40. http://dx.doi.org/10.1111/j.1365-4632.2005.02811.x

50. Hall RE. The bleaching syndrome in the context of somatic norm image among women of color: A qualitative analysis of skin color. Eur J Soc Sci. 2010;17(2):180–185.

51. Lewis KM, Gaska K, Robkin N, Martin A, Andrews E, Williams J. The need for interventions to prevent skin bleaching: A look at Tanzania. J Black Stud. 2012;43(7):787–805. http://dx.doi.org/10.1177/0021934712446701

52. Shimo A. The quest for a lighter shade of pale. Maclean's. 2008 November 14;121(46):150–152.

53. Eagle L, Dahl S, Low DR. Ethical issues in the marketing of skin lightening products. In: ANZMAC 2014 Proceedings; 2014 December 01–03; Brisbane, Australia. Brisbane: ANZMAC; 2014. p. 75–81. Available from: http://www.anzmac.org/_files/1423706830anzmac%202014%20proceedings.compressed.pdf

54. Adebajo SB. An epidemiological survey of the use of cosmetic skin lightening cosmetics among traders in Lagos, Nigeria. W Afr J Med. 2002;21(1):51–55.

55. Ajose FOA. Consequences of skin bleaching in Nigerian men and women. Int J Dermatol. 2005;44(S1):41–43. http://dx.doi.org/10.1111/j.1365-4632.2005.02812.x

56. Nnoruka E, Okoye O. Topical steroid abuse: Its use as a depigmenting agent. J Natl Med Assoc. 2006;98(6):934.

57. Diongue M, Ndiaye P, Douzima P-M, Seck M, Seck I, Faye A, et al. Economic impact of skin-lightening products on household income in sub-Saharan Africa: The case of Senegal. Med Sante Trop. 2013;23(3):308–312.

58. Traore A, Kadeba JC, Niamba P, Barro F, Ouedraogo L. Use of cutaneous depigmenting products by women in two towns in Burkina Faso: Epidemiologic data, motivations, products and side effects. Int J Dermatol. 2005;44(s1):30–32. http://dx.doi.org/10.1111/j.1365-4632.2005.02807.x

59. Kombaté K, Mouhari-Toure A, Saka B, Akakpo AS, Maboudou A, Pitché P, et al. Acne and skin bleaching in Lome, Togo. Int J Dermatol. 2012;51(s1):27–29. http://dx.doi.org/10.1111/j.1365-4632.2012.05560.x

60. Adawe A, Oberg C. Skin-lightening practices and mercury exposure in the Somali community. Minn Med. 2013;96(7):48–49.

61. Kamagaju L, Morandini R, Gahongayire F, Stévigny C, Ghanem G, Pirotte G, et al. Survey on skin-lightening practices and cosmetics in Kigali, Rwanda. Int J Dermatol. 2016;55(1):45–51. http://dx.doi.org/10.1111/ijd.12833

62. Kouotou E, Bissek A, Nouind C, Defo D, Sieleunou I, Ndam E, editors. Dépigmentation volontaire: Pratiques et dermatoses associées chez les commerçantes de Yaoundé (Cameroun) [The practice of skin bleaching and associated skin diseases among female traders in Yaounde (Cameroon)]. Ann Dermatol Venereol. 2015;142(6–7):443–445. French. http://dx.doi.org/10.1016/j.annder.2015.02.023

AUTHORS:
Hugo G. Nami[1,2]
Paloma de la Peña[3,4] iD
Carlos A. Vásquez[1,5]
James Feathers[6]
Sarah Wurz[3,4] iD

AFFILIATIONS:
[1]CONICET-IGEBA, 'Daniel A. Valencio' Palaeomagnetism Laboratory, Department of Geological Sciences, University of Buenos Aires, Buenos Aires, Argentina

[2]Department of Anthropology, National Museum of Natural History, Smithsonian Institution, Washington DC, Washington, USA

[3]Evolutionary Studies Institute, School of Geosciences, University of the Witwatersrand, Johannesburg, South Africa

[4]School of Geography, Archaeology and Environmental Sciences, University of the Witwatersrand, Johannesburg, South Africa

[5]Ciclo Básico Común, University of Buenos Aires, Buenos Aires, Argentina

[6]Luminescence Laboratory, University of Washington, Seattle, Washington, USA

CORRESPONDENCE TO:
Paloma de la Peña

EMAIL:
paloma.delapenya@gmail.com

Palaeomagnetic results and new dates of sedimentary deposits from Klasies River Cave 1, South Africa

Palaeomagnetic data from Klasies River main site Cave 1 (Eastern Cape Province, South Africa) are reported. Natural remanent magnetisation directions obtained from 77 oriented samples were determined by progressive alternating field demagnetisation methodology. Three palaeomagnetic samplings from the Witness Baulk from the Middle Stone Age (MSA) Late Pleistocene White Sand member and the Holocene Later Stone Age (LSA) middens in Cave 1 were dated and analysed to obtain the palaeomagnetic directions recorded in the sediments. Here we provide new optically stimulated luminescence (OSL) dates for the White Sand Member, and new accelerator mass spectrometry (AMS) radiocarbon dates for the LSA midden of areas not previously dated. The palaeomagnetic analysis took into account rock magnetism and directional analysis. The former reveals that the main magnetic carrier was magnetite; the latter shows that characteristic remanent magnetisation of normal and anomalous directions were observed in the lower portion of the White Sand Member and LSA midden. Normal directions correspond to the palaeosecular variation record for South Africa during the Late Pleistocene. On the other hand, the anomalous directions recorded in the LSA midden might represent the likely Sterno-Etrussia geomagnetic field excursion which occurred during the Late Holocene and is observed in other places on the planet. Finally, the directional data obtained are a potential tool for discussing the age of deposits corresponding to those periods.

Significance:
- New dates confirm and extend previous age determinations for the LSA and White Sand Member from Klasies River

Introduction

In earth's history, the Late Pleistocene and Holocene cover the last ca 126 000 years. This period has seen major cultural developments as recorded in the Middle Stone Age (MSA) and Later Stone Age (LSA). A number of techno-complexes are used to describe such changes on a broad level; for example, Still Bay and Howiesons Poort in marine isotope stage (MIS) 4 (75 000–58 000 years ago)[1-3] of the MSA and Wilton and post-classic Wilton of the LSA that occur within marine isotope stage 1[4]. Refining chronologies are vital to understand cultural change. For example, while the technological characteristics of the MSA Still Bay and Howiesons Poort are comparatively well described, its chronostratigraphic boundaries are debated as different dating methods indicate longer or shorter chronologies.[5-7] For the LSA, discussions revolve around nomenclature more than temporal boundaries,[4] but here too there could be a closer understanding of cultural change through time. There is thus an opportunity to expand the methodological range to obtain more perspectives on chronological resolution in the South African MSA and LSA. Here we focus on the use of palaeomagnetic features as distinguishing and chronological markers at Klasies River main site for a Late Pleistocene MSA and a post-classic Wilton occurrence.

During the rock and sedimentary sequences formation process, certain minerals lock in a record of the direction and intensity of the geomagnetic field (GMF). Diverse kinds of palaeomagnetic investigations are carried out on sedimentary deposits, because at the moment of its formation, magnetic minerals were magnetised parallel to the GMF. In this way, the sediments provide data on different GMF features, such as palaeosecular variations, excursions and reversals occurring through the history of the earth.[8] Actually, the GMF has alternated between periods of normal polarity, in which its direction was the same as it is at present, and reverse polarity, in which the field was the opposite. These periods are called chrons, and their durations vary from thousands to millions of years. If compared with high magnitude modifications such as reversals, secular variations are the small changes occurring slowly and progressively in all parts of the GMF with time scales in the order of decades to millennia. The strength and direction of the total field vary as a result of changes in strength and direction of the dipole and non-dipole components. Hence, there is a global signature to the secular variation but also significant differences.[9] Disturbances spanning a short time period that do not result in reversal are called geomagnetic excursions. Detailed knowledge of these excursions has important geomagnetic and stratigraphic implications, becoming useful as a magnetostratigraphic geochronological tool.[10-15]

Spanning the last 780 000 years, during the Brunhes Chron, the GMF polarity has been 'normal' as it is now. However, there has been a number of occasions when the GMF either briefly reversed or behaved anomalously. This fact indicates that this normal polarity has been interrupted by significant departures from the dipole field configuration.[10,16-18] These departures are considerably larger than those seen in secular variations observed during historical times, and sometimes even attain opposite polarity, originating in GMF excursions. By definition, the excursions are short intervals of anomalous field directions that occur within a broader period of 'stable' normal or reversed magnetic polarity.[18-22] However, while certain excursions are known to be global, others

may have been on a continental scale only. During the Brunhes Chron, several anomalous records were observed in different materials, times and places.[16,17,19,23]

In South Africa, palaeomagnetic research related to palaeoanthropological and archaeological investigations was employed to date sites and localities that yielded evidence of the oldest hominin remains going back to the terminal Miocene and Pliocene.[24-26] However, its use in sequences spanning the Pleistocene and Holocene is scarce. As part of a research programme aimed to deepen the knowledge of the GMF behaviour during these geological epochs in different places of the world,[27-32] a number of archaeological and geological sedimentary sections in South Africa were sampled. As a result of this investigation, here we report the preliminary results of the detailed research performed at the Klasies River (KR) main site on the southern coast of South Africa.

Site description and sampling

KR is situated on the Tsitsikamma coast between the mouth of the Klasies River and Druipkelder Point in the Eastern Cape Province (Figure 1). Main site (34°06.30'S; 24°23.26'E) comprises two caves – Caves 1 and 2 – and two overhangs (termed Caves 1A and 1B) – features within and against a Table Mountain sandstone cliff that faces the Indian Ocean. At KR main site the MSA consists of a well-stratified 20-m sequence[33] whereas the LSA was deposited only in Cave 1 between ~4800 BP and 2500 BP[34]. Main site (Figure 2) was excavated by Wymer in the 1960s[34], by Deacon between 1984 and 2010[33,35,36], and, since 2013, by Wurz. Deacon divided the stratigraphy into a series of members formed by sediments of similar lithology, including the Light Brown Sand (LBS), Shell and Sand (SAS), Rockfall (RF), and Upper and White Sand (WS) members. Above the WS member in Cave 1 a LSA midden occurs that has been excavated and numbered from Layers 12 to 1 by Singer and Wymer[34]. An area of the LSA midden was excavated by Binneman[37]. The samples analysed for this paper are from the Witness Baulk in Cave 1 (Figure 2), in particular the WS member, termed Layer 13 by Singer and Wymer[34], and the LSA deposits that overlie this member (Supplementary figure 1).

The Witness Baulk is located along the central part of the cave in a north–south direction (Figure 2 and Supplementary figure 1) and contains the LBS, SAS, WS members and two LSA middens. The LBS and lower part of the SAS member consists of in-situ deposits, but the upper part (SASW sub-member) was formed by talus material from Cave 1A deposits. These talus deposits blocked the entrance of Cave 1, and during much of MIS 5, 4 and 3, the cave was not open for habitation. During a phase when the cave was open, the WS member was formed on top of the SAS member by clean, light-coloured aeolian sands with silt lenses.[34(p.27)] The sand is typical of a regressional fine-grained aeolianite, that may indicate a retreating sea level which was initially near the cave but eventually quite distant.[38(p.147)] Calcareous suspended material that formed in pools of standing water near the rear of the cave associated with this member is also present. The WS member (Layer 13) was presumably deposited after a period of major erosion of the MSA III deposits in Cave 1A. The main reason for this inference is that there is no trace of similar sands in the stratigraphy beneath Cave 1A.[34(p.114)] Before excavation, the WS member had a maximum thickness of close to 1 m in the western part of the cave[34(p.9)] and it thinned out towards the east section. The WS member contains what is known as the MSA IV industry with predominantly quartzite artefacts in the size range of those of Howiesons Poort. The MSA IV has proportionally fewer blades than points, with the latter described as 'unusually small'[34(p.114)]. There are no archaeological features such as hearths supporting a geological origin for the bulk of this member, and archaeological material occurs mostly near the base.[38] Bada and Deems[39] provided an aspartic acid racemisation date of 65 ka for a sample taken 1.5 m below the surface of the Witness Baulk in the WS member. An optically stimulated luminescence (OSL) date of 70.7±7.4 ka was obtained by Feathers[40] (Supplementary table 1). Here we report on new single-grain OSL dating and the palaeomagnetic orientation of samples from this member. New radiocarbon dating of the LSA and WS member in the areas sampled for palaeomagnetic investigation was undertaken to ensure the accurate contextualisation of the palaeomagnetic data. The dates obtained by Singer and Wymer[34] and Binneman[37] as presented in Table 1 originate from other areas of these deposits. A hiatus, associated with grey flowstones and stalagmitic structures, occurred after the deposition of the WS member, implying that the cave must have been essentially closed off at this time.[38] Wave erosion associated with the rise of the sea level in the mid-Holocene removed blocking aeolianite or shell midden deposits and re-opened the cave.
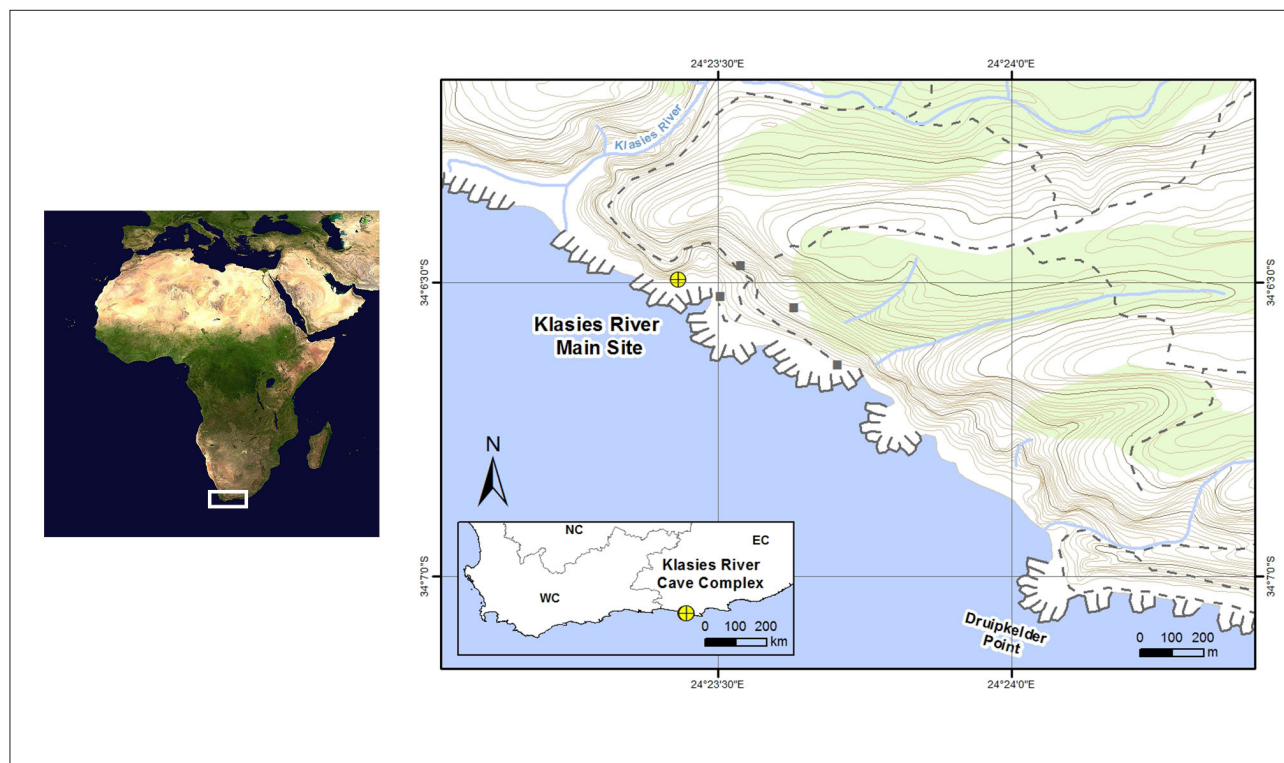
Figure 1:   Location of Klasies River sites in South Africa (34°06.30'S; 24°23.26'E).
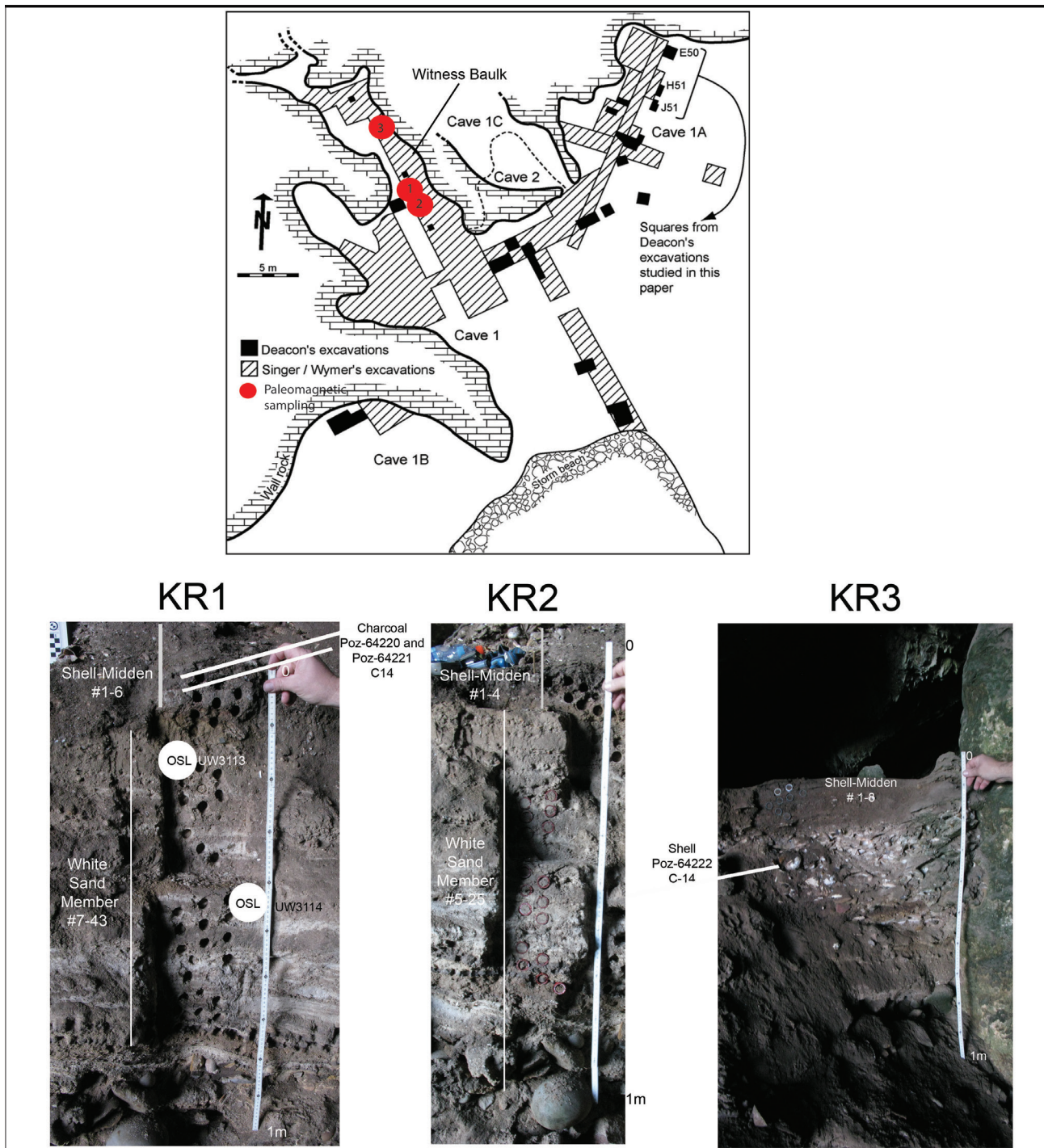
Source (upper panel): Modified after Deacon and Geleijnse[33]

Figure 2: (upper) Map of Cave 1 showing the location of the three palaeomagnetic samplings in the Witness Baulk. (lower) Close-ups of the three palaeomagnetic and dating samplings (OSL and C14).

The LSA shell midden debris formed in two major episodes, documented in Layers 1 to 12.[34] Singer and Wymer[34] excavated a significant portion of the LSA middens, and subsequently Binneman[37] undertook a square-metre excavation of the middens in the eastern profile of the Witness Baulk, approximately at Square B of Singer and Wymer's Main Cutting[34(Figure 3.2)] (Supplementary figure 1). The lower midden, LSA I (Layers 7–12), consists of abundant marine shell with thin ash layers[34,38] (also see Supplementary figure 1). The lower midden is more compact than the upper midden. The top of Layer 7 contains crushed shell and lime cementation that had occurred on the surface, indicating a break in deposition. Carbon-14 (C14) dating undertaken by Singer and Wymer[34]

for Layers 7 and 10, and a further date by Binneman[37] for the base of the midden, indicate occupation at between 4800 BP and 3700 BP (Table 1).

The upper LSA midden – LSA II (Layers 1–6) – originally had a depth of about 1.5 m, but a part of this deposit has been eroded away from the Witness Baulk (Supplementary figure 1). A temporary high sea level (+4 m) that occurred not too long after 2450 BP between Layers 4 and 3 destroyed much of the MSA and LSA deposits as a result of storm-wave erosion.[38(p.146)] The top midden dates to between 2500 BP and 2800 BP (see below Table 1). As explained above, new accelerator mass spectrometry (AMS) dates for charcoal and shell samples associated with the palaeomagnetic studies have been obtained for the LSA middens, and are reported on below.

Table 1: Uncalibrated radiocarbon dates from Later Stone Age (LSA) Phase I and II in Klasies River Cave 1[34,37] and new radiocarbon dates for the LSA midden (bolded)

Cultural attribution	Depth	Type of sample	Laboratory reference	BP	Calibration BP
LSA I	Layer 7	Charcoal	GX-0973	4695±180	X
LSA I	Layer 10	Charcoal	GX-0970	4755±95	X
LSA I	Binneman base	Charcoal	Pta-3905	3780±60	X
	-120 mm from current surface	**Charcoal**	**Poz-64220**	**3425±35**	**3812–3800, 3721–3550 (94%), 3534–3492 (5%) (curve SHCal13)**
	-120 mm from current surface	**Charcoal**	**Poz-64221**	**3605±35**	**3974–3810 (77%), 3802–3720 (23%) (curve SHCal13)**
LSA II	Layer 1	Charcoal	GX-0969	2525±85	X
	-150 mm from current surface	**Shell**	**Poz-64222**	**2655±30**	**2451–2280 (mean prob. = 2340) (marine 13 curve - Delta R = 400)** **2276–1932 (mean prob. = 2084) (marine 13 curve - Delta R = 615)**
LSA II	Surface of Layer 7	Shell	GX-0971	2795±85	x

The three new C14 samples were calibrated with Calib radiocarbon package. Poz-64220 and Poz-64221 samples (both charcoal samples) were calibrated with SHCal13 curve. The multiple age ranges for Poz-64220 and Poz-64221 refer to different intersections of the dates in one of the SHcal13 curve plateaus. The shell sample Poz-64222 was calibrated with the default Delta R=400 and from the data from south of Hout Bay (the nearest reference for KR) where the Delta R is 615.

Table 2: New dates based on optically stimulated luminescence

Laboratory reference	Depth below top of White Sand Member (mm)	Number of grains	Equivalent dose[†]	Over-dispersion from central age model (%)	Dose rate (Gy/ka)	Date (ka)
UW3113	150	146	60.2±2.6	45±4	0.90±0.05	66.5±4.8
UW3114	450	106	55.7±2.9	36±5	0.99±0.05	56.3±4.6

[†]Determined from central age model for UW3114, and by the largest component of the finite mixture model for UW3113.
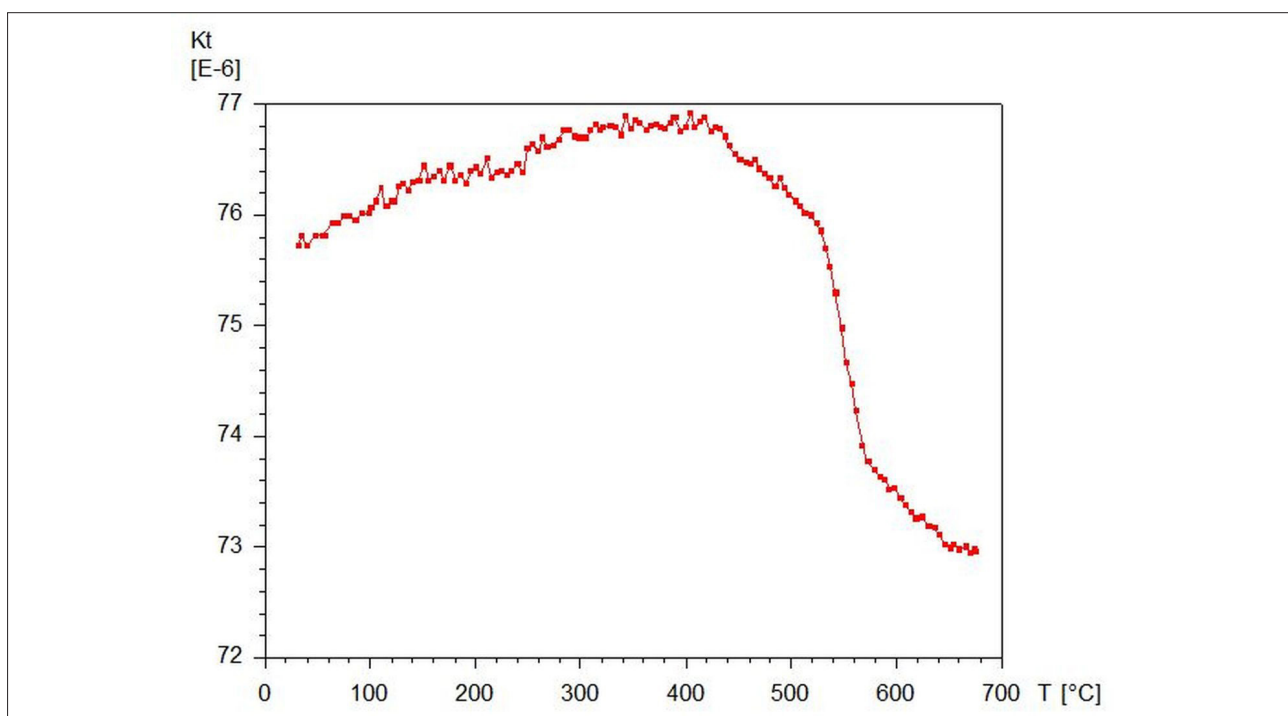
Figure 3: Thermomagnetic curve showing the variation of magnetic susceptibility with temperature for sample RM2.

The LSA middens are associated with typical 'informal' southern Cape coastal stone tool assemblages characterised by low levels of curation and 'unstandardised' artefacts and described as the 'Kabeljous' industry. This industry sometimes occurs contemporaneously with the Wilton industry on the coast.[37(p.151)] In the Kabeljous industry, formally retouched tools are few, but the thick-backed scraper-knives, sometimes termed giant crescents[34(p.127)], are quite distinctive.[37] In addition to stone artefacts, Singer and Wymer[34] recorded and described querns, pounders, a bored stone, a grooved stone, a slate palette, ostrich egg shell beads, a perforated cowry and sinkers. Bone tools occur in significant proportions in both the upper and lower middens.

Dating

The palaeomagnetic research performed in KR Cave 1 provided an opportunity to obtain additional AMS and single-grain OSL dates for the LSA middens and the upper part of the WS member of the Witness Baulk relating to the sampling. However, the main purpose of the new dating samples was to accurately contextualise all the palaeomagnetic results that will be discussed below.

Both LSA middens and the WS member have been dated previously. The previous and new C14 and OSL dating results are synthesised in Tables 1 and 2 and Supplementary table 1. For the new AMS dates – Poz-64220 and Poz-64221 – a similar age should be expected, as the charcoal samples taken were from a similar height (Figure 2). Indeed, this is the case; in Table 1 it is shown that both samples date to ~3500 BP. The date for Square D Poz-64222 is more recent (~2600 BP). The LSA midden thinned out towards the back of the cave (Supplementary figure 1). What is preserved at present is an uneven eroded surface of the deposit that Singer and Wymer excavated (Supplementary figure 1). Poz-64222 therefore represents a more recent layer of the LSA midden, even if it is slightly lower than those of the other two layers.

The new C14 and OSL dates largely confirm previous chronological assessments. Our C14 dates for phase I are somewhat younger than those obtained by Singer and Wymer[34], but similar to that of Binneman[37]. Poz-64220 and Poz-64221 date to ~3500 BP, the Binneman[37] date is 3780±60 BP, whereas the Singer and Wymer date is about 4800 BP. It must be taken into account that the dates by Singer and Wymer and Binneman result from conventional C14 dating whereas the new dates were obtained with the AMS technique. For LSA II, our date for Poz-64222 is 2655±30 BP – also slightly younger than that obtained by Singer and Wymer (2795±85 BP) for the surface of Layer 7. These dates may indicate that the LSA I at Klasies River occurred until 3500 BP, and that LSA II lasted until closer to 2000 BP than previously estimated.

For the Witness Baulk, two samples for OSL dating were taken at different depths adjacent to palaeomagnetic sampling KR1 and KR2 (Table 2, Supplementary tables 2–7 and Supplementary figure 2). The two OSL dates, UW3113 and UW3114, were derived from 180–212-μm single-grain quartz. The equivalent dose was determined on more than 100 grains for each sample. Using an over-dispersion of 30% (from dose recovery) as typical for a single-aged sample, a finite mixture model showed that all grains for UW3114, and 94% of grains for UW3113, were consistent with a single component (Supplementary figure 2). This argues for well-bleached, unmixed samples. Using a weighted average (central age model) for the equivalent dose and dividing through by the bulk dose rate provides the ages given in Table 2. While nearly within 1σ error terms, the two dates are inverted stratigraphically. Both, however, are within 1σ of the date reported by Feathers[40]. This inversion might be partially caused by the dose rate. The dose rate of the sample was assumed to be characteristic of the samples' entire radioactive environment, which may not be fully true, especially for UW3113 which is close to the shell midden. Details of the dating method are given in the supplementary material.

Palaeomagnetic study

Sampling procedures

The palaeomagnetic samples were taken from the fine sediments of the eastern profile exposed by Singer and Wymer's main cutting (see their Figure 2.1[34]; see also Figure 2 and Supplementary figure 1). Most of the deposit is formed by fine-grained sand with lenses of silt and clay, which are interbedded by thin lenses of fine gravel. The sampling was performed in the upper layers of the WS member and the back slope area of the overlying LSA midden relating to Singer and Wymer's Squares C and D.

Three palaeomagnetic samplings named KR1 ($n=43$), KR2 ($n=28$) and KR3 ($n=8$) were taken (Figure 2). KR1 and KR2 were collected in Singer and Wymer's Square C, and KR3 in a well-defined, and apparently undisturbed, stratum from Square D close to the stalagmite (Figures 2 and 3). The LSA midden was sampled in KR1 (#1–9), KR2 (#1–4) and KR3 (#1–8). The WS member was sampled in the same areas in KR1 (#10–43) and KR2 (#5–28) below the LSA. In the WS member, the interbedded levels of gravels were not sampled, and cores (KR1 #25–27, #37–42 and KR2 #26–30) were taken from the silt and clay levels. The cores were taken vertically using 25-mm long and 20-mm diameter cylindrical PVC plastic containers. In KR1 and KR3, the cylinders were carefully pushed into the sediments, overlapping each other by about 50%. In KR2, each core was sampled continuously without overlapping. The sample's orientation was measured using a Brunton compass. Samples were consolidated with sodium silicate once removed and they were numbered from the top to the bottom. The depth of each sample is depicted in Figure 2. One sampling – named RM – was taken nearby to KR1 for rock magnetic analysis. The samples ($n=12$) were taken with an interval of ~50 mm.

Rock magnetic study

This analysis was performed using the variation of magnetic suscep-tibility with frequency, which is a tool to identify small magnetic particles in the superparamagnetic state (SP). Because the grain size distribution follows a lognormal distribution, the presence of these SP particles is indicative of bigger particles whose grain size is in the stable single domain state (SSD),[41,42] and which are very good magnetic carriers of remanence.

Samples were measured in the susceptibilimeter AGICO model MK1-FA. Several amplitude fields ranging from 5 A/m to 700 A/m in f1 (=1000 Hz), 5 A/m to 350 A/M in f2 (=4000 Hz) to 200 A/m in f3 (=16 000 Hz) were used. The parameter χ_{fd}% (percentage difference of susceptibility with frequency) is calculated according to:

$$\chi_{fd}\% = 100 \times [(\chi_{lf} - \chi_{hf}) / \chi_{lf}],$$

where χ_{lf} is the magnetic susceptibility in low frequency and χ_{hf} is the susceptibility in high frequency. The resulting measurements are shown in Supplementary table 8.

For magnetite, the parameter χ_{fd}% was correlated with grain size in the limit SP/SSD when the frequencies ranged between 1000 Hz and 16 000 Hz at ambient temperatures. Figure 3 shows the variation of susceptibility with temperature for the RM2 sample; the drop near 580 °C is an indicator of magnetite as the main magnetic carrier. When SSD concentration is high, a peak near 580 °C is apparent – this is named Hopkinson peak. The fact that the Hopkinson peak is suppressed can be related to SP grains, because this sample has the higher χ_{fd}%. Most of the samples show values higher than 10% in f1f3 frequencies and values higher than 6% in f1f3 frequencies; only two samples – RM10 and RM12 – exhibited lower values for χ_{fd}% in both differences in frequency. In these samples it can be argued that the concentration of SP fraction is lower; therefore the concentration of viscous grains is also lower.

Remanence directions analysis

All samples were subjected to detailed stepwise alternating field (AF) demagnetisation in progressive steps of 3, 6, 9, 12, 15, 20, 25, 30, 40 and 60 mT with a three-axis static degausser attached to a 2G cryogenic magnetometer (755 R). Additional steps of 80 mT, 100 mT and 120 mT were used for some specimens. The KR samples showed a common pattern with similar reliable magnetic behaviour. Most samples from the LSA midden had a gradual decrease of magnetisation with almost all remanence erased at 30 mT (KR1 1; KR2 1; KR3 3, 5; Figure 4a,k,o–p); from the WS member, a few samples at 30–40 mT (KR1 27, 31, 33, 42; Figure 4f–g,i–j), and most samples between 50 mT and 60 mT (KR1 13, 25; Figure 4d–e), while a few indicate a drop at 80–120 mT (KR1 32; KR2 9, 10; KR3 7; Figure 4h,m,n,q).
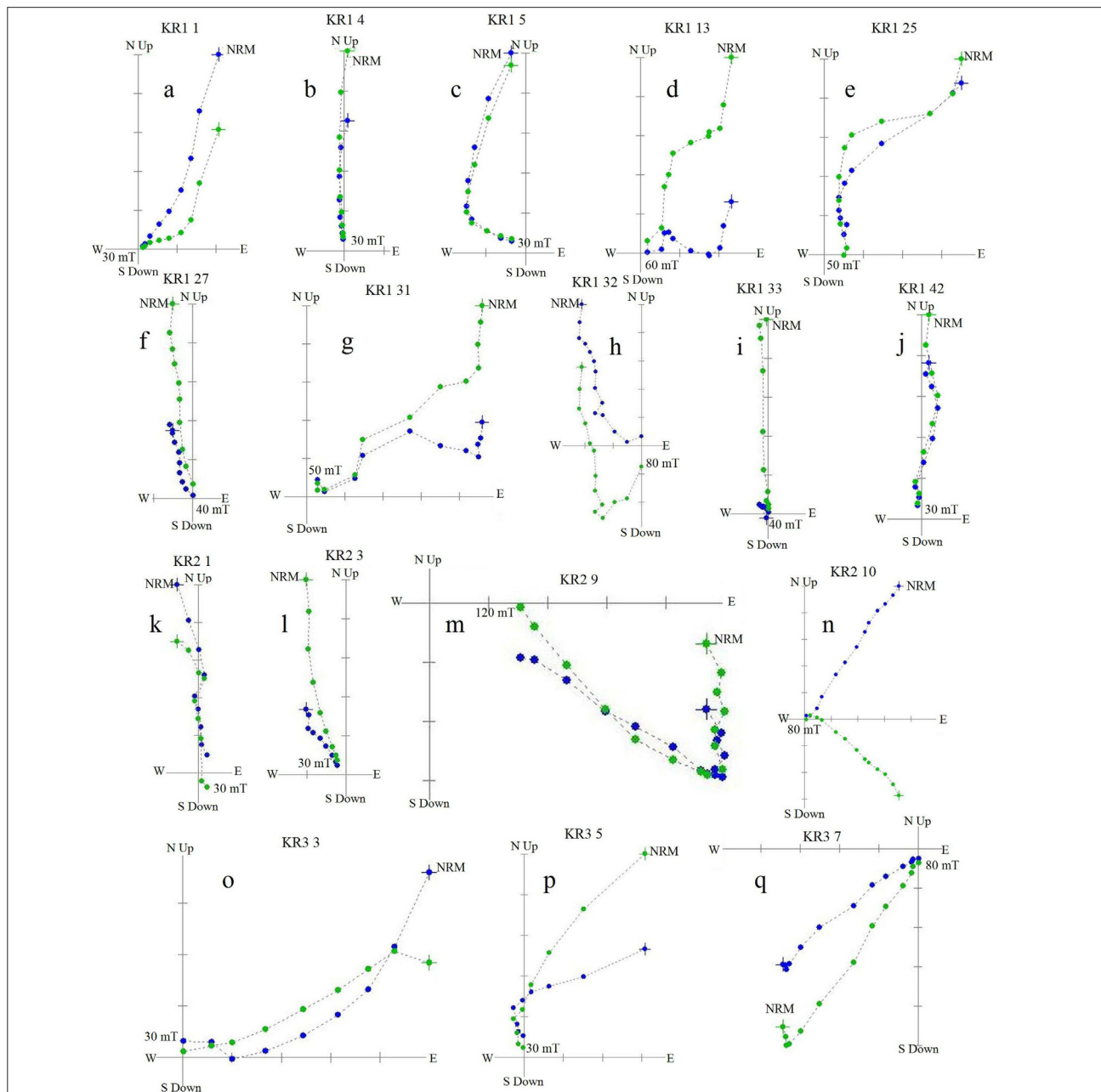
NRM, natural remanent magnetisation

Figure 4: Vector component diagrams of alternating-field demagnetisation behaviour for representative samples from KR Cave 1: (a–j) KR1, (k–n) KR2 and (o–q) KR3. The totality of the vector projection diagrams illustrated in the figures is directional data with corrected field. Blue and green symbols correspond to projection on the horizontal and vertical planes, respectively.

Palaeomagnetic directions were determined by the 'Remasoft 3.0 palaeomagnetic data browser and analyzer' computer program.[43] Some specimens were not processed because their orientation marks were lost (KR2 1 and 2), or because they were too unstable to isolate directions (KR1 30; KR3 4; Figure 5). The characteristic remanent magnetisation (ChRM) directions were calculated using principal components analysis.[44] In most cases, a ChRM could be defined trending in the Zijderweld diagrams towards the coordinate's origin (e.g. KR1 1, 5, 13; KR2 10; KR3 5, 7; Figure 4a–c,n,p–q). The samples shown in Figure 4 were highly reliable, displaying similar patterns of univectorial behaviour in the vector diagrams projection (KR1 4, 27, 33; KR2 10; Figure 4b,f,i,n). Others had two magnetic components with one decaying to the origin in the vector diagrams projection (KR1 1, 5, 32; KR3 5; Figure 4a,c,h,p). Four samples (namely KR1 5, 32; KR3 5; Figure 4c,p) displayed two magnetisation components with different directions. As described

in Supplementary table 9, the secondary ('soft') component in three samples recorded a normal GMF position while the 'hard' magnetisation is interpreted as an early remanence acquired during formation of the sedimentary deposit. A viscous secondary component easily removed in a few samples (KR1 1; KR2 1; KR3 7; Figure 4a,k,q) between 3 mT and 12 mT was not considered further. There was one sample with three magnetic components; one of which (KR1 13; Figure 4d) appears to decay to the origin. Finally, a few specimens exhibited residual directions not erased at demagnetisation fields of 40–50 mT, causing some deviation away from the origin. Nevertheless, they do not affect the isolated ChRM (e.g. KR1 25, 31; Figure 4e,g). Despite the rejection, it is important to point out that the KR1 30 stereoplot illustrated in the upper panel of Figure 5 shows more than one component contained in a very low inclination plane. In this regard, the two magnetisations observed in this rejected sample are in agreement with the two components recorded

in a nearby specimen (i.e. KR1 32; Figure 4h) and the anomalous directions observed until KR1 33. Many samples showed either a high (e.g. KR2 10; Figure 4n) or a low negative (e.g. KR1 1, 27, 33; Figure 4a,f,i), and positive inclinations (e.g. KR1 32; KR3 7; Figure 4h,q). Most samples yielded a normal direction (KR1 1, 4, 13, 27; KR2 1, 3; Figure 4a–b,d,f,k–l), a few samples showed northwesterly (KR1 5; Figure 4c), northeasterly or easterly directions (KR1 1, 31; KR2 10; Figure 4a,g,n), several of which had steep inclinations (KR1 1, 27, 33; KR2 1; Figure 4a,f,i,k). Oblique reverse and reversed or 'anomalous' southward directions were found in various cores from KR3 (7; Figure 4q) and the upper portions of KR2. The maximum angular deviations were generally within low values, ranging from 0° to 5° (KR1=71.4%, KR2=64.3% and KR3=57.1%) and from 5.1° to 10° (KR1=28.6%, KR2=39.7% and KR3=43.9%).

The number and intervals of demagnetisation steps used to isolate the ChRM of each sampling are given in Supplementary table 10. This analysis shows that the KR samples display normal, intermediate and reverse magnetic remanence of low negative and positive inclination values, mostly in KR2. Figure 5 (lower panel) illustrates the stereographic projection of ChRM for the three KR samplings. They show a cluster of normal directions and intermediate and reversed directions in KR1 to KR3. Magnetograms of the declination and inclination profiles from each sampled section are exhibited in Figure 6. KR3 and the upper part of the KR1 and KR2 sites yielded records with anomalous directions. Between samples 3 to 15, the upper portion of KR2 varies widely with transitional positions between samples 9 to 13 gradually changing to normal positions; KRM3 mostly shows anomalous directions far from the present GMF; KR1 and KR2 exhibit wide amplitude pulses both in declination and inclination. Remarkably consistent is the declination agreement on specific specimens such as KR2 7 and KR3 7. From KR1 13 and KR2 16, a significant but gentle eastward shift in the declination over 60° is observed. Most of the specimens of the WS member display normal GMF positions; however, there are also anomalous directions far from the present GMF. Normal directions in the central part of KR1 (#6–31) and KR2 (#14–29) show the record of the palaeosecular variation of the normal GMF. In the lowermost part of KR1, a wide pulse with transitional positions both in declination and inclination between normal samples KR1 29 and 35 is observed. Differences in KR1 and KR2 logs might be for several reasons. On the one hand, samples KR2 1 and 2 were not processed. On the other hand, the sampling interval in KR2 is different from that in KR1; these facts might explain in part the differences between the upper and lower portions. The record of inclination in the WS member presents a certain similar behaviour ranging between ~-40° and 70°. The two anomalous and the normal features observed in the logs are shown respectively by the Greek letters α, β and γ in Figure 6. After checking that the directional data from KR was useful to assess a Fisherian distribution, the site mean directions of KR1 and KR2 were computed using Fisher's[45] statistic. Additionally, a mean direction was calculated using the normal directions from both samplings. From this study, it is observed that KR mean directions show a consistent agreement; they are also located close to the International Geomagnetic Reference Field (IGRF) direction (D=332.86°, I=-64.55°) for 2014 – the year in which the sampling was performed (Supplementary table 10; Figure 7). Supplementary tables 8 and 9 depict the virtual geomagnetic pole positions (VGP) calculated from the directions of the sites reported here. When plotted on a present world map, VGPs show intermediate and reverse positions from the rotation axis of the earth (Figure 8).
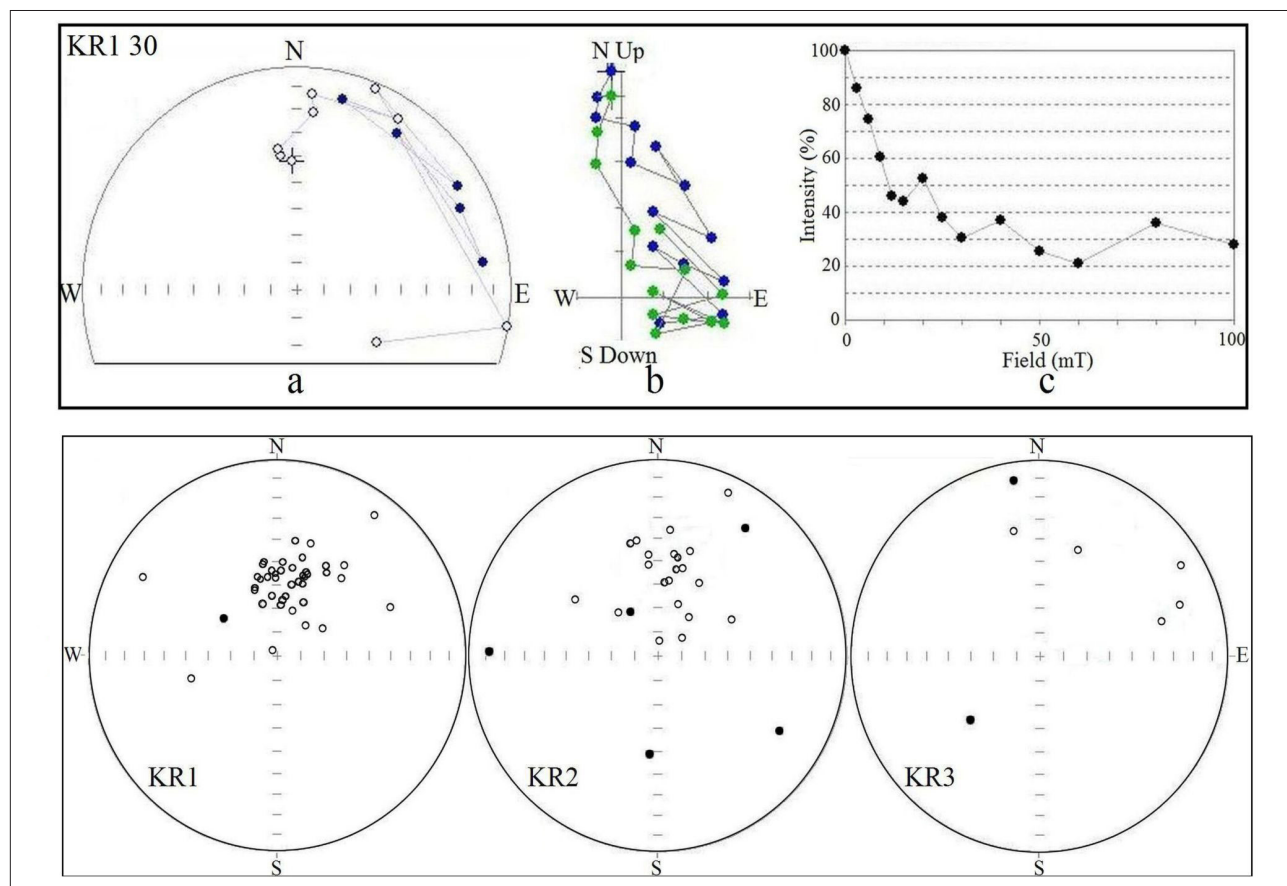
Figure 5: (upper) Example of an unreliable sample to isolate directions from KR1 (#30) showing more than one component, contained in a very low inclination plane: (a) stereographic projection, (b) Zijderveld diagram and (c) demagnetisation curve. (lower) Stereographic projections of directional data with field correction of characteristic remanent magnetisations of each sample for the sections reported in this paper. Negative inclination is indicated by open circles and positive inclination by solid circles.

The virtual poles in the northern hemisphere are located in northern North America, Greenland, and north, east and southeastern Asia. VGPs in the southern hemisphere are situated in South America, west of central Africa, Australia and Antarctica. As seen in Figure 8d, VGPs from KR show a consistency between the apparent systematic in VGP location observed in other places during the Pleistocene–Holocene transition and Holocene.[46](Figure 12b) Interestingly, in earth's history, longitudinal bands across Australasia and the Americas have been observed in transitional records since the early Jurassic.[47]

Discussion and conclusion

Palaeomagnetic data obtained in the fine-grained sediments have shown anomalous and normal directions. Normal directions were mostly recorded in the WS member, and anomalous directions in the lowermost portion of KR1. The intermediate, oblique reverse and reverse polarity far from the present GMF were recorded in the upper part of KR1 and KR2, representing LSA I, and the totality of KR3, relating to LSA II.
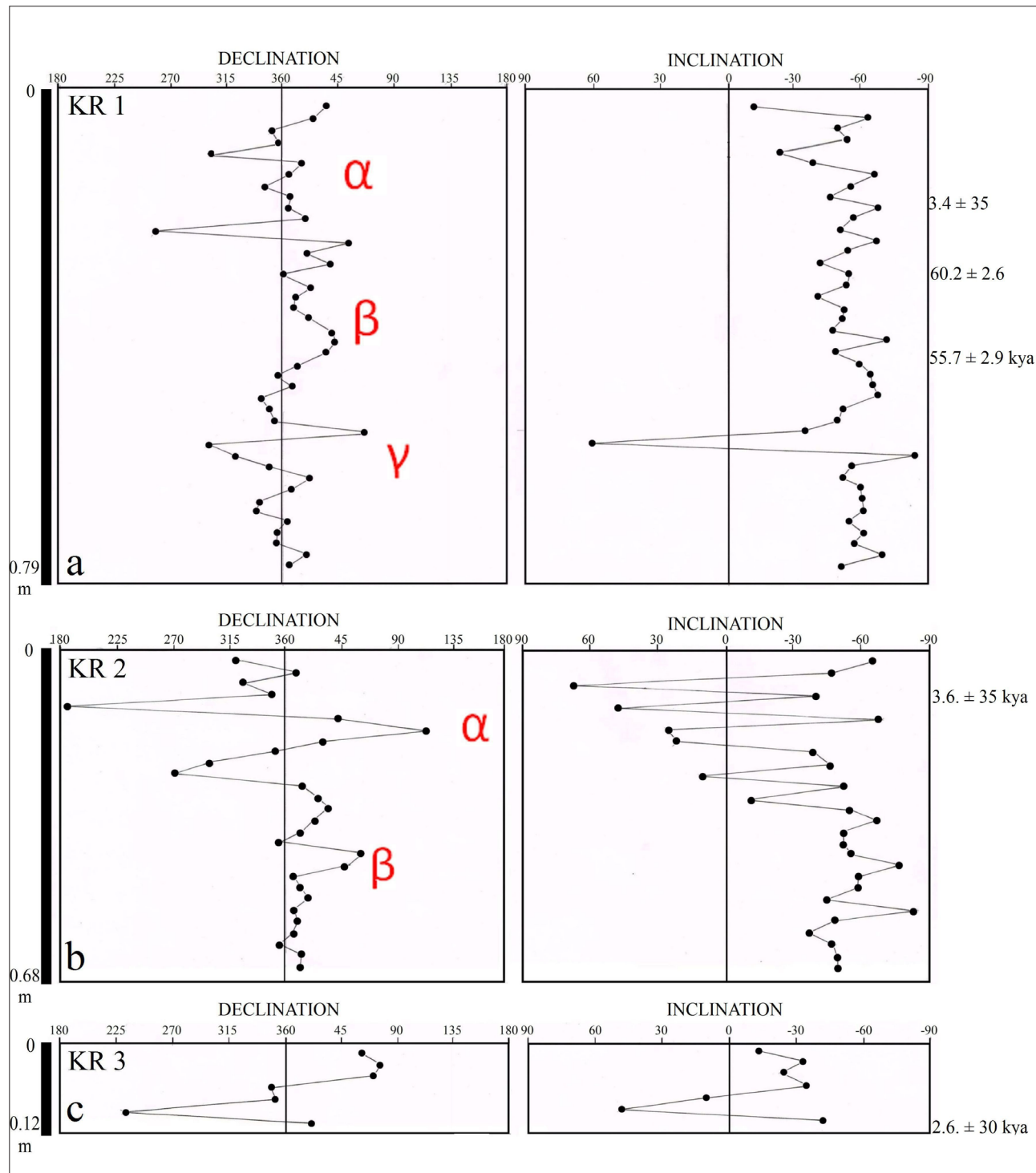
Figure 6: Declination and inclination logs according to depth and the absolute dates determined through accelerator mass spectrometry and optically stimulated luminescence from (a) KR1, (b) KR2 and (c) KR3. Greek letters denote the features with more conspicuous long direction departures from its upper (*α*) and lower (*γ*) portions, and the normal directions recorded in the middle part (*β*).
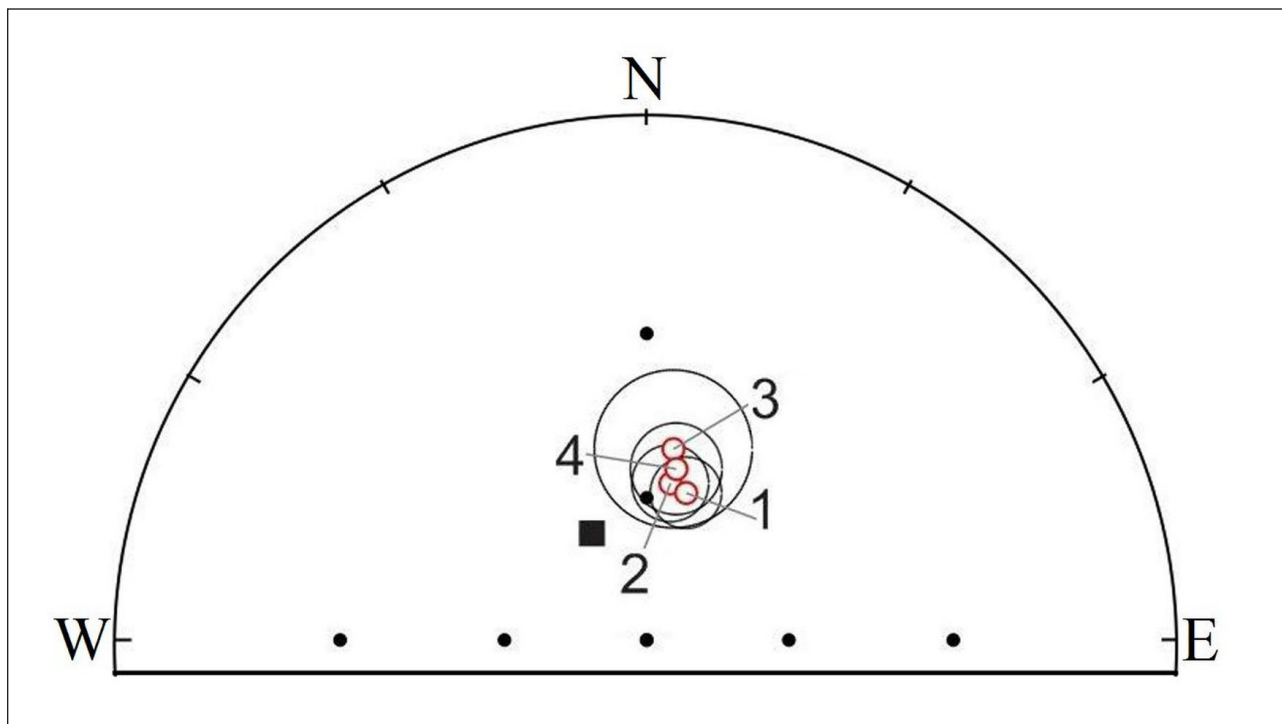
Figure 7: Site mean magnetisation directions for (1) KR1, (2) KR1 #6–31, (3) KR2 and (4) KR2 #14–30. The International Geomagnetic Reference Field direction is indicated by the square.
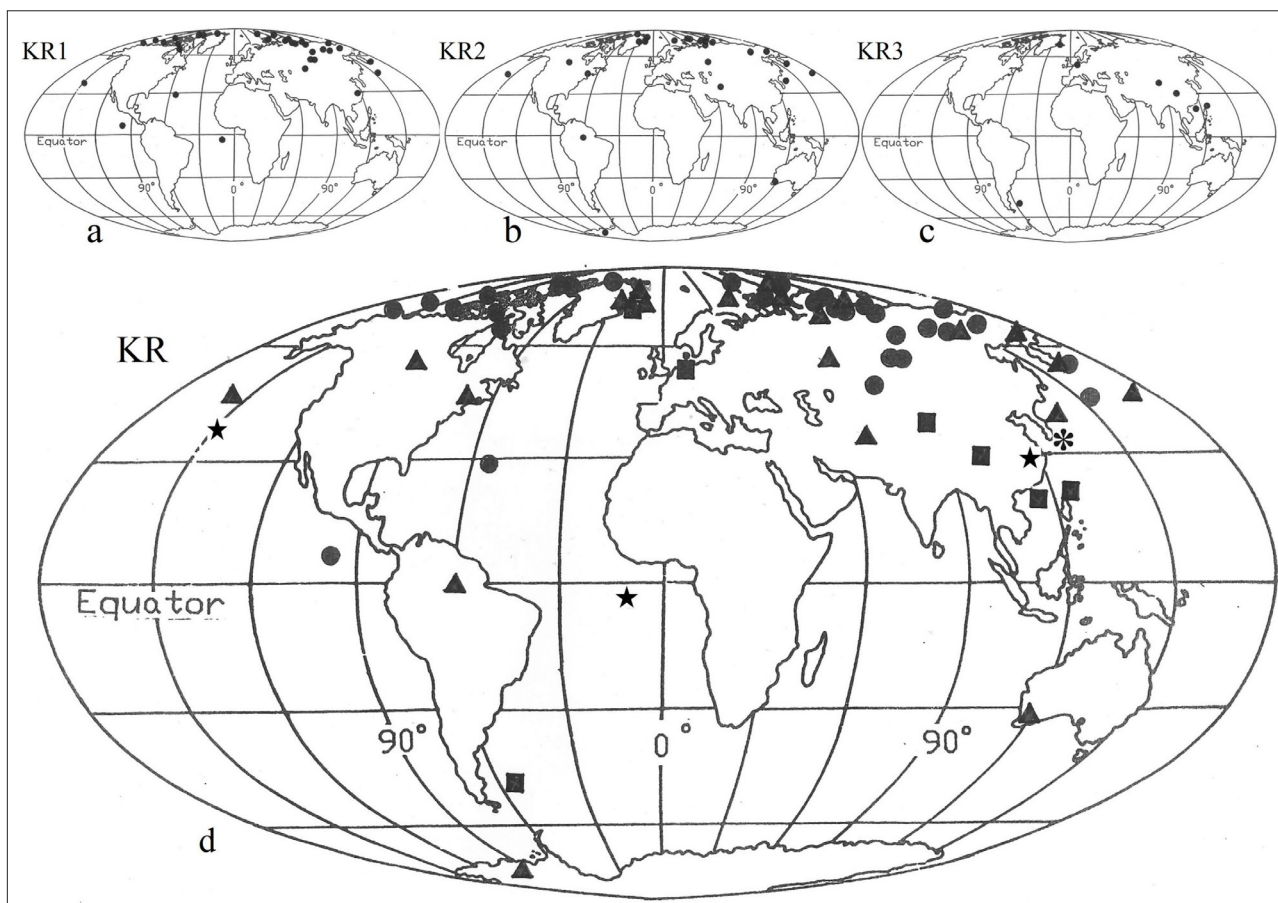
●, KR1; ▲, KR2; ■, KR3. The stars represent samples KR31 to 33 and the asterisk shows the VGPs calculated from the secondary direction of KR3 #5.

Figure 8: Virtual geomagnetic pole (VGP) paths plotted on a world map for the sections mentioned in the text: (a) KR1, (b) KR2, (c) KR3 and (d) all VGPs from Klasies River.

Anomalous palaeomagnetic records may occur for a number of reasons. Actually, various causes can give rise to the measurement of anomalous directions of remanent magnetism which does not reflect the true GMF behaviour, such as diverse deposition processes, chemical alterations as well as sedimentary physical disturbances.[22,48] Also, the anomalous directions may reflect true GMF excursions, which are defined when VGPs differ by more than 45° from the geographic pole during normal or reverse polarity, signifying that they may be considered a major deviation in GMF behaviour.[14,18,22] Hence, if the anomalous directional data observed at KR are not a result of a sedimentary artefact and/or perturbations, they correspond to the Late Holocene GMF record at southern Africa during the $\sim \leq 3.5$ ka to 2.5 ka interval; on the other hand, the normal directions occurring at the middle portion of KR1 and most of KR2 of the WS member logs represent the Late Pleistocene palaeosecular variation record during $\sim$60–65 ka. Regardless of the relative scarcity of samples with anomalous directions in the lower portion of KR1, these samples might presumably represent an anomalous Late Pleistocene GMF record with a similar age.

The directions recorded in the LSA middens may suggest that the likely GMF excursion observed in other parts of the earth during the Holocene[31,49] might also be present in southern Africa. Particularly at $\sim$2.5–3.0 ka, a number of localities in the world yielded records of the excursion named Etrussia or Sterno-Etrussia.[50-53] Additionally, sea cores from Tahiti provided data on this recent event.[54] In the southern hemisphere, Late Holocene deposits also produced records with reverse positions.[49,55,56] However, the above-mentioned dates obtained at KR Cave 1 allow inquiry into the possibility that the likely Late Holocene GMF excursion in South Africa might have started earlier, at $\sim \leq 3.5$ ka. At a similar latitude to KR, but in the western hemisphere, a Holocene declination and inclination log with wide swings similar to KR are present in the Misiones Province, northeastern Argentina.[31] Even though this event is not dated, it may be significant. Therefore, the polarities recorded in the LSA midden sections of KR1 and KR2 as well KR3 might be a manifestation of the above-mentioned geomagnetic anomaly occurring during the Late Holocene. In the lower portion of KR1, marked with the letter γ in Figure 6, there are some samples that may also be a manifestation of possibly anomalous GMF behaviour observed in other places of the world during the Late Pleistocene. At that time, a well-known excursion is the Laschamp geomagnetic event, dated at $\sim$40 ka.[18,57,58] However, considering the new OSL dates obtained in the WS member, the aforementioned anomalous record may correspond to an older GMF event. Actually, Nowaczyk and colleagues[25(Table 5)] informed a number of records with dates ranging from 65 ka to 86 ka corresponding to the Norwegian-Greenland Sea event. Interestingly, Lund and colleagues[16(Table 1)] estimated an age of 61 ± 2 ka for this excursion. Also, resulting from Ocean Drilling Program Leg 172, Lund and associates[16(Figure F1)] reported 14 plausible excursions within the Brunhes Chron recorded at sites 1060 to 1063, separated by more than 1200 km. During the Late Pleistocene, they yielded evidence of an undated excursion labelled as 5a, which is placed between two anomalies dated at $\sim$40 kya (3b) and $\sim$125 kya (5b). The anomalous directions observed in the lower portion of KR1 may likely correspond to the above-mentioned Late Pleistocene events.

Because anomalous directions may be a result of sedimentological or rock magnetic effects, and because of the scarcity of similar records around the world, most scholars do not accept as real GMF behaviour the anomalies reported for the last $\sim$11 ka BP. However, there is growing evidence suggesting the hypothesis of the global excursional state of the GMF with no contemporaneous intermediate or reverse directions during this time span[31,32], and the possible occurrences of other excursions during the Pleistocene.[16,17,54] Hence, if the KR records represent true GMF behaviour, they have provided valuable insights.

Palaeomagnetic directions recorded in sedimentary deposits may be used as chronostratigraphic tools[11,13,14], which might be helpful for archaeological relative and absolute dating[10,11,59,60]. In the case of the record of the LSA midden overlying the WS member, it may be useful to correlate LSA deposits from South Africa. The anomalous directions of the lower portion of KR1, in the WS member, may be instrumental to relate MSA deposits. The palaeosecular variation observed in the middle part of KR1 and KR2 magnetograms, relating to the MSA, may also be used with a similar aim.[10,61,62] Hence, the observed likely palaeomagnetic features have much potential to be useful as a marker for deposits containing LSA and MSA remains, respectively, in the region.

Acknowledgements

Authors' contributions

H.G.N. and P.d.I.P. conceptualised the project; H.G.N. performed the palaeomagnetic analyses; J.F. performed the OSL analysis; C.A.V. performed the rock magnetism analysis; H.G.N., S.W. and P.d.I.P. performed the field work; and H.G.N., C.A.V., P.d.I.P., S.W. and J.F. wrote the manuscript.

References

1. Wurz S. Technological trends in the Middle Stone Age of South Africa between MIS 7 and MIS 3. Curr Anthropol. 2013;54(S8):305–319. http://dx.doi.org/10.1086/673283

2. Mackay A, Stewart BA, Chase BM. Coalescence and fragmentation in the late Pleistocene archaeology of southernmost Africa. J Hum Evol. 2014;72:26–51. http://dx.doi.org/10.1016/j.jhevol.2014.03.003

3. Wadley L. Those marvellous millennia: the Middle Stone Age of Southern Africa. Azania. 2015;50(2):155–226. http://dx.doi.org/10.1080/0067270X.2015.1039236

4. Lombard M, Wadley L, Deacon J, Wurz S, Parsons I, Mohapi M, et al. South African and Lesotho Stone Age sequence updated (I). S Afr Archaeol Bull. 2012;67(195):120–144.

5. Jacobs Z, Roberts RG, Galbraith RF, Deacon HJ, Gran R, Mackay A, et al. Ages for the Middle Stone Age of southern Africa: Implications for human behavior and dispersal. Science. 2008;322(5902):733–735. http://dx.doi.org/10.1126/science.1162219

6. Jacobs Z, Roberts RG. An improved single grain OSL chronology for the sedimentary deposits from Diepkloof Rockshelter, Western Cape, South Africa. J Archaeol Sci. 2015;63:165–192. http://dx.doi.org/10.1016/j.jas.2015.01.023

7. Tribolo C, Mercier N, Douville E, Joron JL, Reyss JL, Rufer D, et al. OSL and TL dating of the Middle Stone Age sequence at Diepkloof Rock Shelter (South Africa): A clarification. J Archaeol Sci. 2013;40(9):3401–3411. http://dx.doi.org/10.1016/j.jas.2012.12.001

8. Kodama KP. Palaeomagnetism of sedimentary rocks: Process and interpretation. Chichester: Wiley-Blackwell; 2012. http://dx.doi.org/10.1002/9781118384138

9. Stenberg RS The geophysical basis of archaeomagnetic dating. In: Eighmy JL, Sternberg RS, editors. Archaeomagnetic dating. Tucson, AZ: University of Arizona Press; 1990. p. 5–28.

10. Tarling DH. Palaeomagnetism: Principles and applications in geology, geophysics, and archaeology. New York: Chapman and Hall; 1983. http://dx.doi.org/10.1007/978-94-009-5955-2

11. Barendrest RW. Using palaeomagnetic remanence and magnetic susceptibility data for the differentation, relative correlation and absolute dating of quaternary sediments. In: Mahaney WC, editor. Quaternary dating methods. Amsterdam: Elsevier; 1984. p. 101–140. http://dx.doi.org/10.1016/S0920-5446(08)70067-0

12. Parkes PA. Current scientific techniques in archaeology. New York: St. Martin's Press; 1986.

13. Herz N, Garrison EG. Geological methods for archaeology. New York: Oxford University Press; 1998.

14. Merrill RT, McFadden PL. The use of magnetic field excursions in stratigraphy. Quaternary Res. 2005;63:232–237. http://dx.doi.org/10.1016/j.yqres.2005.02.007

15. Westaway R. The effect of changes in the earth's moment of inertia during glaciation on geomagnetic polarity excursions and reversals: Implications for Quaternary chronology. Curr Sci. 2003;84:1105–1115.

16. Lund SP, Williams T, Acton G, Clement B, Okada M. Brunhes epoch magnetic field excursions recorded in ODP Leg 172 sediments. In: Keigwin V, Rio D, Acton G, editors. Proceedings of the Ocean Drilling Project, Scientific Results. 2001;172, Art. #10, 18 pages. http://dx.doi.org/10.2973/odp.proc.sr.172.216.2001

17. Lund SP, Stoner JS, Channell JET, Acton G. A summary of Brunhes paleomagnetic field variability recorded in Ocean Drilling Program cores. Phys Earth Planet In. 2006;156:194–204. http://dx.doi.org/10.1016/j.pepi.2005.10.009

18. Laj C, Channel JET. Geomagnetic excursions. In: Kono M, editor. Treatise on geophysics. Amsterdam: Elsevier; 2007. p. 373–416. http://dx.doi.org/10.1016/b978-044452748-6/00095-x

19. Singer B. A Quaternary geomagnetic instability time scale. Quat Geochronol. 2014;29–52. http://dx.doi.org/10.1016/j.quageo.2013.10.003

20. Watkins N. Polarity group sets up guidelines. Geotimes. 1976;21:18–20.

21. Thouveny N, Creer KM. Geomagnetic excursions in the past 60 ka: Ephemeral secular variation features. Geology. 1992;20:399–402. http://dx.doi.org/10.1130/0091-7613(1992)020<0399:GEITPK>2.3.CO;2

22. Verosub K, Banerjee SK. Geomagnetic excursions and their paleomagnetic record. Rev Geophys. 1977;15:145–155. http://dx.doi.org/10.1029/RG015i002p00145

23. Gubbins D. The distinction between geomagnetic excursions and reversals. Geophys J Int. 1999;137:F1–F4. http://dx.doi.org/10.1046/j.1365-246x.1999.00810.x

24. Jacobs JA. Geomagnetic excursion. In: Gubbins D, Herrero-Bervera E, editors. Encyclopedia of geomagnetism and paleomagnetism. Dordrecht: Springer; 2007. p. 311–313. http://dx.doi.org/10.1007/978-1-4020-4423-6_115

25. Nowaczyk NR, Frederichs TW, Eisenhauer A, Gard G. Magnetostratigraphic data from late Quaternary sediments from the Yermak Plateau, Arctic Ocean: Evidence for four geomagnetic polarity events in the last 170 ka of the Brunhes Chron. Geophys J Int. 1994;117:453–471. http://dx.doi.org/10.1111/j.1365-246X.1994.tb03944.x

26. Thackeray JF, Kirschvink JL, Raub TD. Palaeomagnetic analyses of calcified deposits from the Plio-Pleistocene hominid site of Kromdraai, South Africa. S Afr J Sci. 2002;98:537–540.

27. Herries AIR, Adams JW, Kuykendall KL, Shaw J. Speleology and magnetobiostratigraphic chronology of the GD 2 locality of the Gondolin hominin-bearing paleocave deposits, North West Province, South Africa. J Hum Evol. 2006;51:617–631. http://dx.doi.org/10.1016/j.jhevol.2006.07.007

28. Herries AIR, Shaw J. Palaeomagnetic analysis of the Sterkfontein palaeocave deposits: Implications for the age of the hominin fossils and stone tool industries. J Hum Evol. 2011;60:523–539. http://dx.doi.org/10.1016/j.jhevol.2010.09.001

29. Nami HG. Possible Holocene excursion of the earth's magnetic field in southern South America: New records from archaeological sites in Argentina. Earth Planet Space. 1999;51:175–191. http://dx.doi.org/10.1186/BF03352222

30. Nami HG. Probable Middle Holocene geomagnetic excursion at the Red Rock Archaeological Site, California. Geofís Int. 1999;18:239–250.

31. Nami HG. New detailed Holocene paleomagnetic records with anomalous geomagnetic field behavior in Argentina. Geoacta. 2012;37(2):83–116.

32. Nami HG. New paleomagnetic results and evidence for a geomagnetic field excursion during the Pleistocene-Holocene transition at Pichincha Province, Ecuador. Geofís Int. 2015;52:127–148. http://dx.doi.org/10.1016/j.gi.2015.04.009

33. Deacon HJ, Geleijnse VB. The stratigraphy and sedimentology of the main site sequence, Klasies River, South Africa. S Afr Archaeol Bull. 1988;43:5–14. http://dx.doi.org/10.2307/3887608

34. Singer R, Wymer J. The Middle Stone Age at Klasies River Mouth in South Africa. Chicago, IL: University of Chicago Press; 1982.

35. Deacon HJ. Two Late Pleistocene-Holocene archaeological depositories from the southern Cape, South Africa. S Afr Archaeol Bull. 1995;50:121–131. http://dx.doi.org/10.2307/3889061

36. Deacon HJ, Wurz S. Middle Pleistocene populations and the emergence of modern behaviour. In: Barham L, Robson Brown K, editors. Human roots: Africa and Asia in the Middle Pleistocene. Bristol: Western Academic Press; 2001. p. 55–63.

37. Binneman JNF. Symbolic construction of communities during the Holocene Later Stone Age in the southeastern Cape [Phd thesis]. Johannesburg: University of the Witwatersrand; 1995.

38. Butzer KW. Sediment stratigraphy of the Middle Stone Age sequence at Klasies River Mouth, Tsitsikama Coast, South Africa. S Afr Archaeol Bull. 1978;33(128):141. http://dx.doi.org/10.2307/3888151

39. Bada JL, Deems L. Accuracy of dates beyond ^{14}C dating limit using the aspartic acid racemisation reaction. Nature. 1975;255:218–219. http://dx.doi.org/10.1038/255218a0

40. Feathers JK. Luminescence dating in less than ideal conditions: Case studies from Klasies River main site and Duinefontein, South Africa. J Archaeol Sci. 2002;29:177–194 http://dx.doi.org/10.1006/jasc.2001.0685

41. Hrouda F. Models of frequency-dependent susceptibility of rocks and soils revisited and broadened. Geophys J Int. 2011;187:1259–1269. http://dx.doi.org/10.1111/j.1365-246X_2011.05227.x

42. Shcherbakov VP, Fabian K. On the determination of magnetic grain-size distributions of superparamagnetic particle ensembles using the frequency dependence of susceptibility at different temperatures. Geophys J Int. 2005;162:736–746. http://dx.doi.org/10.1111/j.1365-246X_2005.02603.x

43. Chadima M, Hrouda F. Ramasoft 3.0. Paleomagnetic data browser and analyzer. Brno: Agico Inc.; 2007.

44. Kirschvink JL. The least-squares line and plane and the analysis of palaeomagnetic data. Geophys J R Astr Soc. 1980;62:699–718. http://dx.doi.org/10.1111/j.1365-246X.1980.tb02601.x

45. Fisher RA. Dispersion on a sphere. Proc R Soc Ser A. 1953;217:295–305. http://dx.doi.org/10.1098/rspa.1953.0064

46. Nami HG. New detailed paleosecular variation record at Santa Lucía archaeological site (Corrientes Province, northeastern Argentina). Geofís Int. 2011;50(2):163–175.

47. Vizán H, Van Zele MA. Jurassic–Early Cretaceous intermediate virtual geomagnetic poles and Pangaean subduction zones. Earth Planet Sci Lett. 2008;266(1–2):1–13. http://dx.doi.org/10.1016/j.epsl.2007.09.044

48. Langereis CG, Van Hoof AAM, Rochette P. Longitudinal confinement of geomagnetic reversal paths as a possible sedimentary artifact. Nature. 1992;358:228–230. http://dx.doi.org/10.1038/358226a0

49. Nami HG. Paleomagnetic results from Argentinean Patagonia: New evidence for the Holocene geomagnetic excursion in southern South America. In: Fernandez LD, editor. The Quaternary period. In press 2015.

50. Petrova GN, Pospelova GA. Excursions of the magnetic field during the Brunhes Chron. Phys Earth Planet In. 1990;63:135–143. http://dx.doi.org/10.1016/0031-9201(90)90067-8

51. Wiegank F, Petrova GN, Pospelova GA. Magneto-chronostratigraphic scale model, Brunhes Chron. Geomag Field Quat. 1990;62:169–177.

52. Dergachev VA, Raspopov OM, Van Geel B, Zaitseva GI. The ´Sterno-Etrussia´ geomagnetic excursion around 2700 BP and changes of solar activity, cosmic ray intensity, and climate. Radiocarbon. 2004;46:661–681.

53. Guskova EG, Raspopov OM, Piskarev AL, Dergachev VA. Magnetism and paleomagnetism of the Russian Arctic marine sediments. In: Proceedings of the 7th International Conference 'Problems of Geocosmos'; 2008 May 26–30; St. Petersburg, Russia. St. Petersburg: St. Petersburg State University; 2008. p. 380–385.

54. Lund SP, Platzman E, Thouveny N, Camoin G. Evidence for two new paleomagnetic field excursions ~2,500 and ~12,500 years ago from the South Pacific Ocean Region (Tahiti). American Geophysical Union Fall Meeting; 2007 December 10–14; San Francisco, CA, USA. Washington DC: American Geophysical Union; 2007. Abstract #GP42A-05. Available from: http://adsabs.harvard.edu/abs/2007AGUFMGP42A..05L

55. Nami HG, Sinito AM. Evidence of a possible excursion of the geomagnetic field registered during the Late Holocene in the Province of Chubut, Argentina. Geoacta. 1993;20:19–26.

56. Sinito AM, Nami HG, Gogorza C. Analysis of palaeomagnetic results from Holocene sediments sampled at archaeological excavations in South America. Quaternary S Am Antarc Peninsula. 1997;10:31–44.

57. Guillou H, Singer BS, Laj C, Kissel C, Scaillet S, Jicha BR. On the age of the Laschamp geomagnetic excursion. Earth Planet Sci Lett. 2004;227:331–343. http://dx.doi.org/10.1016/j.epsl.2004.09.018

58. Nowaczyk NR, Arz HW, Frank U, Kind J, Plessen B. Dynamics of the Laschamp geomagnetic excursion from Black Sea sediments. Earth Planet Sci Lett. 2012;351–352:54–69. http://dx.doi.org/10.1016/j.epsl.2012.06.050

59. Thompson R. Paleomagnetic dating. In: Smart PL, Frances PD, editors. Quaternary dating methods — A user's guide. Technical guide 4. Cambridge: Quaternary Research Association; 1983. p. 177–197.

60. Wolfman D. Geomagnetic dating methods in archaeology. Adv Arch Method Theory. 1984;7:363–458. http://dx.doi.org/10.1016/B978-0-12-003107-8.50011-6

61. Opdyke ND, Channell JET. Magnetic stratigraphy. San Diego, CA: Academic Press; 1996. http://dx.doi.org/10.1016/S0074-6142(06)80010-0

62. Stoner JS, St-Onge G. Magnetic stratigraphy in paleoceanography: Reversals, excursions, paleointensity, and secular variation. Developments Mar Geol. 2007;1:99–138. http://dx.doi.org/10.1016/S1572-5480(07)01008-1

AUTHORS:
Rabelani Mudzielwana[1]
Mugera W. Gitari[1] (iD)
Titus A.M. Msagati[2]

AFFILIATIONS:
[1]Environmental Remediation and Pollution Chemistry Research Group, Department of Ecology and Resource Management, University of Venda, Thohoyandou, South Africa

[2]College of Science Engineering and Technology, University of South Africa, Johannesburg, South Africa

CORRESPONDENCE TO:
Mugera Gitari

EMAIL:
mugera.gitari@univen.ac.za

Characterisation of smectite-rich clay soil: Implication for groundwater defluoridation

Groundwater is a widely used and affordable source of drinking water in most of the rural areas of South Africa. Several studies have indicated that groundwater in some boreholes in South Africa has a fluoride concentration above the level recommended by the World Health Organization (1.5 mg/L). Fluoride concentrations above the permissible limit (>1.5 mg/L) lead to dental fluorosis, with even higher concentrations leading to skeletal fluorosis. In the present work, we evaluate the application of smectite-rich clay soil from Mukondeni (Limpopo Province, South Africa) in defluoridation of groundwater. The clay soil was characterised by mineralogy using X-ray diffraction, by elemental composition using X-ray fluorescence and by morphology using scanning electron microscopy. Surface area and pore volume was determined by the Brunauer–Emmett–Teller surface analysis method. Cation exchange capacity and pH_{pzc} of the soil were also evaluated using standard laboratory methods. Batch experiments were conducted to evaluate and optimise various operational parameters such as contact time, adsorbent dose, pH and initial adsorbate concentration. It was observed that 0.8 g/100 mL of smectite-rich clay soil removed up to 92% of fluoride from the initial concentration of 3 mg/L at a pH of 2 with a contact time of 30 min. The experimental data fitted well to a Langmuir adsorption isotherm and followed pseudo second order reaction kinetics. Smectite-rich clay soil showed 52% fluoride removal from field groundwater with an initial fluoride concentration of 5.4 mg/L at an initial pH of 2 and 44% removal at a natural pH of 7.8. Therefore smectite-rich clay soil from Mukondeni has potential for application in defluoridation of groundwater. Chemical modification is recommended to improve the defluoridation capacity.

Significance:

- Physicochemical and mineralogical characterisation of smectite-rich clay soil

- Defluoridation of groundwater using smectite-rich clay soil

- Adsorption modelling using adsorption isotherms and kinetic models

Introduction

Groundwater is the most affordable and widely used natural resource for many of the rural areas in South Africa. Its chemical nature is one of the most important criteria that determine its usefulness for a specific need and as such not all groundwater is fit for drinking.[1] Although the presence of fluoride in drinking water, within permissible limits, is beneficial for the production and maintenance of healthy bones and teeth, excessive intake of fluoride can cause dental or even skeletal fluorosis.[2] The permissible limit set by the World Health Organization (WHO) is 1.5 mg/L.[3] The presence of fluoride in high concentration in drinking water is therefore a serious health concern.

Fluorosis is an incurable disease caused by drinking water with fluoride concentrations of greater than 1.5 mg/L for extended periods of time and therefore efforts must be made to prevent it by providing drinking water with fluoride concentrations below 1.5 mg/L. Ncube and Schutte[4] indicated that groundwater which has fluoride concentrations beyond the recommended limits for drinking water will require defluoridation. Meenakshi and Maheshwari[5] reported various methods for defluoridation, including adsorption methods, ion exchange, membrane techniques and precipitation methods. Among all the methods of defluoridation which have been reported, adsorption has been considered an effective method for defluoridation of groundwater in rural areas because it uses materials which are readily available and inexpensive. Different materials such as clay soils[6], activated alumina[7], activated carbon[8] and other low-cost materials have been tested for defluoridation of groundwater. Among the tested adsorbents, activated alumina is widely used because it is inexpensive and it has a high defluoridation capacity.[5] Recently, more studies have focused on the use of clay soils for defluoridation because of their good adsorptive properties such as a large specific surface area, chemical and mechanical stability, layered structure and high cation exchange capacity. Several studies that have reported the use of clay soils for defluoridation have been published, including a study on illite-goethite soil in China[9], a study on selected South African clays[6], the use of chemically modified bentonite clay[10], magnesium incorporated bentonite clay[11], Fe^{3+}-modified bentonite in South Africa[12] and Al^{3+}-modified bentonite in South Africa[13]. Smectite-rich clay from Mukondeni (a village in the Limpopo Province of South Africa) has been used for a very long time to make ceramic water filters and has been found to be effective in improving water quality.[14] Its use in the adsorption of inorganic pollutants has never been evaluated, although it is known to be rich in smectite.[15] Smectite-rich clay soil is available naturally in large abundance at no cost.

We evaluated the physicochemical properties of the black Mukondeni smectite-rich clay soil and its potential application in defluoridation of groundwater using batch experiments. Optimum operating conditions such as contact time, adsorbent dosage, initial adsorbate concentration and initial pH were determined. The regeneration and re-usability potential of smectite-rich clay soil as well as the effects of co-existing ions on defluoridation efficiency was also evaluated. Lastly, the mechanism of fluoride adsorption on the clay soil was elucidated.

Material and methods

Sample collection and preparation

Smectite-rich clay soil was collected from Mukondeni Village, Vhembe District in the Limpopo Province of South Africa. Field water was collected from the community borehole in Siloam, in the Vhembe District. All reagents and total ionic strength adjustment buffer (TISAB-III) were obtained from Rochelle Chemicals & Lab Equipment CC, South Africa Ltd and were of analytical grade. A stock solution containing 1000 mg/L fluoride was prepared by dissolving 2.21 g NaF in 1 L of Milli-Q water (18.2 MΩ/cm) and fluoride solutions for batch experiments were prepared from fresh stock fluoride solution by appropriate dilution.

Preparation of clay soil

The soil was washed by mixing with Milli-Q water (18.2 MΩ/cm) at a ratio of 1:5 in a 1-L beaker; the mixture was stirred for 5 min, and the procedure was repeated twice. After stirring, mixtures were agitated for 15 min using a Stuart reciprocating shaker and then centrifuged for 10 min at 5000 rpm. Samples were then dried in an oven for 12 h at 105 °C. Soil samples were then milled to pass through a 250-μm sieve.

Physicochemical and mineralogical characterisation

Mineralogical and chemical composition of the clay soil were determined using X-ray diffraction (XRD) (PANalytical X'Pert pro power) and X-ray fluorescence (XRF) (Thermo Fisher ARL9400 XP+ sequential XRF with WinXRF software), respectively. Surface morphology was determined using scanning electron microscopy (SEM) (Leo1450 SEM, voltage 10 kV, working distance 14 mm). Surface area and pore volume were determined by the Brunauer–Emmett–Teller method using micrometrics TriStar II. Functional groups and surface chemistry were determined using Fourier transform infrared spectroscopy (FTIR). Cation exchange capacity (CEC) was determined using ammonium acetate buffers at pH 5.4 and 7.4. The concentration of exchangeable cations was determined using flame atomic absorption spectra (600 PerkinElmer). Point of zero charge (pH_{pzc}) was determined using titration at 0.1 M, 0.01 M and 0.001 M KCl ionic strength.

Batch defluoridation experiments

Batch experiments were carried out to evaluate the effect of contact time, initial adsorbate concentration, adsorbent dosage and pH on fluoride adsorption. A 100-mL volume of a known fluoride concentration was pipetted into a 250-mL plastic bottle and a known mass of the smectite-rich clay was suspended in the mixtures and then agitated for 30 min in a table shaker to attain equilibrium. After agitation, samples were filtered using 0.45-μm pore membrane filters. Filtrates were analysed for residual fluoride using an ion selective electrode calibrated with four standards containing 1 mL of TISAB III per 10 mL of solution. The same ratio was maintained for the sample. TISAB III was added to decomplex fluoride ions (F^-) from Al^{3+}, Fe^{3+} and Si^{4+} complexes and to maintain pH at between 5.2 and 5.5, optimum for the F^--selective electrode. To evaluate the effect of contact time, the contact time was varied from 5 min to 270 min. The effect of adsorbent dosage was evaluated by varying the adsorbent dosage from 0.1 g to 2 g. To evaluate the effect of adsorbate concentration, the initial concentration was varied from 3 mg/L to 15 mg/L and to evaluate the effect of initial pH, the pH of the solution was adjusted from 2 to 12 using 0.1 M NaOH and 0.1 M HCl. The effect of competing ions was investigated by spiking a 3 mg/L fluoride solution with 5 mg/L of SO_4^{2-}, Cl^-, CO_3^{2-} and NO_3^-, separately; a blank experiment was set for control with initial fluoride concentration of 3 mg/L. Mixtures were agitated for 30 min. All experiments were carried out at room temperature. Equations 1 and 2 were used to calculate the percentage of removal and adsorption capacity Q, respectively.

$$\% \text{ removal} = \left(\frac{C_o\text{-}C_e}{C_o} \right) \times 100, \qquad \text{Equation 1}$$

where C_o is the initial fluoride ion concentration and C_e is the equilibrium fluoride ion concentration in mg/L.

$$Q = \left(\frac{C_o\text{-}C_e}{m} \right) \times V, \qquad \text{Equation 2}$$

where C_o is the initial fluoride concentration (mg/L), C_e is the fluoride concentration at equilibrium (mg/L), V is the volume of solution (L) and m is the mass of the adsorbent (g).

Regeneration of smectite-rich clay soil

Regeneration of the adsorbent was carried out as follows: 0.8 g of fluoride-loaded clay was agitated with 100 mL of 0.1 M NaOH for 30 min on a table shaker. After agitation, the adsorbent was filtered through a 0.45-μm pore membrane filter and the filtrate was diluted to 100 mL and then analysed for desorbed fluoride. The collected adsorbent on filter paper was washed with Milli-Q water and then dried at 110 ºC for 3 h. Regenerated adsorbent was then re-used for defluoridation up to five times.

Method of analysis

The fluoride concentration in the treated water sample was measured using an ion selective electrode (9609 BNWP Orion, USA) attached to a Thermo Scientific Orion Star A215 ISE pH meter. A similar ion meter coupled with a pH electrode was used for measuring the pH of treated samples. All experiments were conducted in triplicate for better accuracy and the mean values were reported.

Results and discussion

Physicochemical characterisation

X-ray diffraction analysis

XRD analyses were carried out to identify the mineral phase of the adsorbent. As shown in Figure 1, the clay soil is mainly composed of montmorillonite, quartz, albite and anthophyllite. The quantitative results further confirm the presence of smectite (60.34%) as the major mineral and the presence of quartz (20.21%) and plagioclase (20.4%) as minor minerals.

M-Montmorillonite
Q-Quartz
A-Albite
An-Anthophyllite
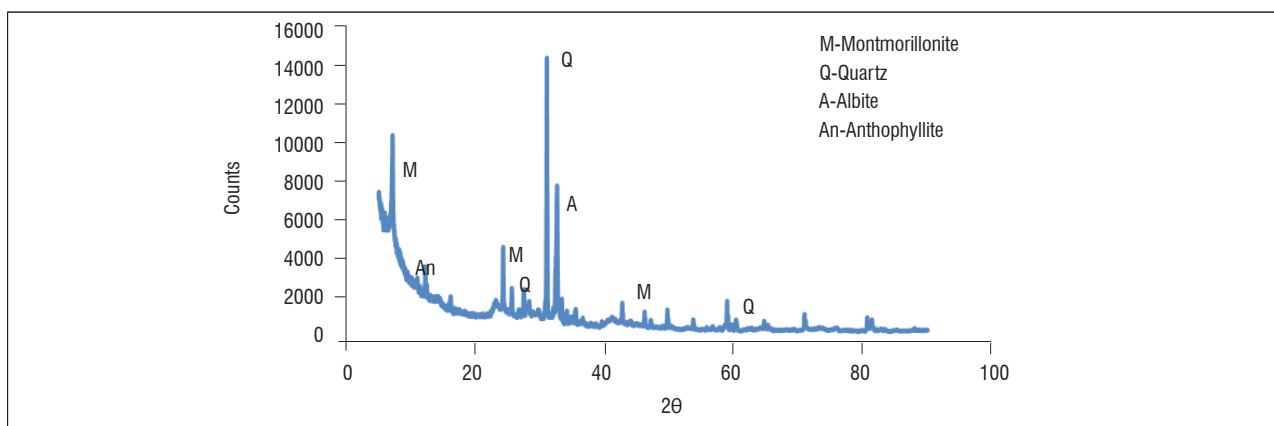
Figure 1: X-ray diffraction spectrum of the smectite-rich clay soil.
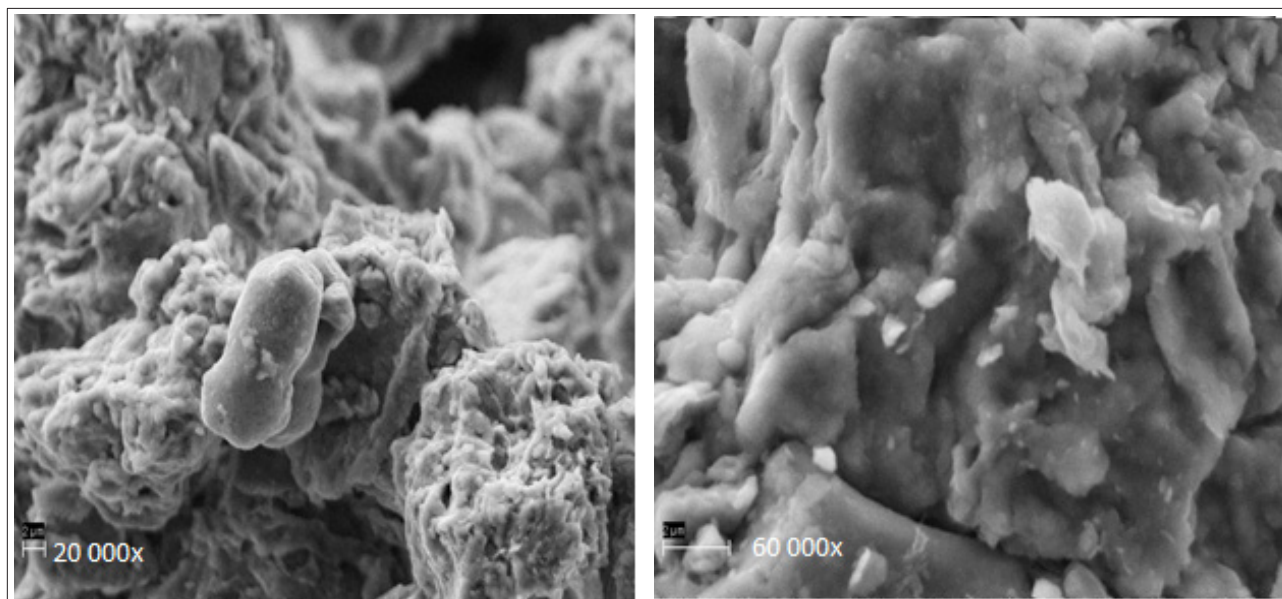
Figure 2: Scanning electron micrograph of the surface morphology of smectite-rich clay soil.

X-ray fluorescence analysis

Table 1 presents the major elemental composition of smectite-rich clay soil. The analysis reveals that silica (SiO_2) is the main component at 62.9%, followed by Al_2O_3 at 14.35%. High concentrations of SiO_2 and Al_2O_3 reveal that the clay soil is an aluminosilicate material.

Table 1: Physicochemical properties of the smectite-rich clay soil

Parameter		Value
Al_2O_3 (%w/w)		14.35
CaO (%w/w)		1.79
Fe_2O_3 (%w/w)		5.45
K_2O (%w/w)		0.93
MgO (%w/w)		2.79
MnO (%w/w)		0.8
Na_2O (%w/w)		2.52
P_2O_5 (%w/w)		0.04
SiO_2 (%w/w)		62.9
TiO_2 (%w/w)		0.75
Loss on ignition		7.71
Cation exchange capacity (meq/100 g)	pH 5.4 pH 7.4	74.23 79.93
pH_{pzc}		6.0
Pore diameter (nm)		7.04
Surface area (m²/g)		20.35

Scanning electron microscope analysis

Surface morphology of smectite-rich clay is shown in Figure 2. At lower magnifications, smectite-rich clay soil showed an irregular porous structure, whereas images at higher magnifications revealed that smectite-rich clay soil has a smooth surface.

Fourier transform infrared spectroscopy analysis

FTIR analysis was carried out in order to understand the surface chemistry and functional groups of smectite-rich clay soil and also to help understand the adsorption mechanism. Figure 3 shows the FTIR spectra of raw and fluoride-loaded smectite-rich clay soils. The analysis shows three main absorption regions of smectite-rich clay soil, i.e. 3000–3800 cm⁻¹, 1300–1800 cm⁻¹ and 500–1200 cm⁻¹. The same absorbance bands were observed by Toor et al.[16] in montmorillonite clay. A notable difference was observed in those regions in raw smectite-rich clay soil and F⁻-loaded smectite-rich clay soil. The absorption band at 3694.38 cm⁻¹ may be a result of stretching and vibration of the structural OH⁻ group and the broad band at 3620.51 cm⁻¹ may be attributed to the stretching and vibration of structural hydroxyl groups and water. In the lower frequencies, a strong band at 996.06 was observed which may be caused by the stretching and vibration of Si-O groups in the clay mineral. The IR peak at 910 cm⁻¹ may be attributed to stretching and vibration of Al-OH-Al. After fluoride sorption there was a decrease in the intensity of transmittance in all the peaks. This decrease confirms that the fluoride adsorption in smectite-rich clay soil was mainly through interaction with OH groups and also direct interaction with the metal surface.
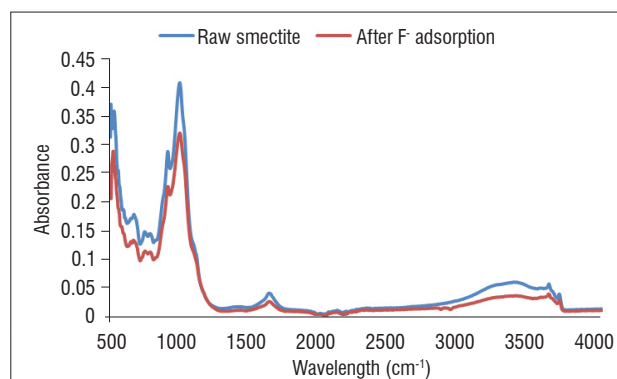
Figure 3: Fourier transform infrared spectroscopy spectra for raw and fluoride-loaded smectite-rich clay soil.

CEC, surface area, pore volume and diameter and pH_{pzc}

CEC refers to the total quantity of exchangeable cations that the soil can retain on their surfaces by electrostatic force at given pH. The CEC for smectite-rich clay was identified by measuring the concentrations of Na^+, K^+, Mg^{2+} and Ca^{2+} in ppm released in solution and the concentrations were then converted to milliequivalent per 100 g

(meq/100 g). CEC values of smectite-rich clay soil at pH 5.4 and 7.4 are presented in Table 1. The CEC for smectite-rich clay soil was found to be 74.23 (meq/100 g) at pH 5.4 and it increased to 79.93 (meq/100 g) at pH 7.4. From the results it can be said that CEC is not dependent on pH as there was no significant difference in the concentration of cations at different pH. The results correlate with the results by Gitari et al.[12] who also concluded that CEC is independent of pH. It can also be concluded that the smectite-rich clay soil has moderate CEC. The surface area and pore size of the clay are the significant properties that influence the sorption of ions by the clay. Surface area, pore volume and pore size of smectite-rich clay soil were determined using the Brunauer–Emmett–Teller method and the results are summarised in Table 1. The smectite clay has an average pore diameter of 7.04 nm, which suggests that it is mesoporous in nature. pH_{pzc} of the smectite-rich clay soil evaluated in this study was found to be 6.0, which is lower than that for bentonite clay soil reported by Gitari et al.[12] which was found to be 8.2. pH_{pzc} refers to the suspension pH at which the clay soil has zero net charges on the surface[17]. Above the pH_{pzc} the clay is negatively charged and below the pH_{pzc} the clay is positively charged.[18] Therefore, adsorption of anions is likely to be high at a pH below 6.

Batch adsorption experiments

Batch experiments were carried out to evaluate the effects of contact time, adsorbent dosage, initial concentration, pH and co-existing ions in the adsorption of fluoride onto smectite-rich clay soil.

Effect of contact time

The effect of contact time in adsorption of fluoride was evaluated at various initial concentrations ranging from 3 mg/L to 10 mg/L and adsorbent dosage of 0.3 g/100 mL. Contact time was varied from 5 min to 270 min at a 250-rpm shaking speed. Results are shown in Figure 4a which illustrates the plot of percentage fluoride removal with time. It was observed that percentage of fluoride removal increased gradually from 5 min up to 30 min, with a slight decrease at 60 min, after which the system seemed to approach equilibrium. This pattern was observed at all initial concentrations. The increase might be attributed to the availability of active sites for fluoride sorption which result in more exchange of ions and the decrease might be an indication that the active sites were getting saturated as the interaction continued, with the system finally reaching equilibrium. The percentage fluoride removal was reduced with an increase in initial fluoride concentration. Optimum fluoride uptake was observed at 30 min and this contact time was applied in subsequent experiments.

Effect of adsorbent dosage

The effect of adsorbent dosage was evaluated by varying the amount of clay from 0.1 g/100 mL to 2.0 g/100 mL. An initial fluoride concentration of 3 mg/L was used. The mixtures were agitated for 30 min at 250 rpm shaking speed. The results are reported as percentage fluoride removal in Figure 4b. It was observed that fluoride removal increased gradually from 46% at an adsorbent dosage of 0.1 g/100 mL to 62.2% at 0.8 g/100 mL. The increase is likely a result of an enhancement in the number of adsorption active sites available for adsorption of F- as the adsorbent mass increased.[11] A slight change in the percentage F- removal was observed for an adsorbent dosage of 0.8 g/100 mL, which may be an indication that the system had reached equilibrium. Similar results have been reported by Kamble et al.[10] with chemically modified bentonite for which they cited two reasons: (1) better utilisation of the available active sites at low adsorbent dose in comparison to high adsorbent dose at which too many sites are available for limited quantity of adsorbate and (2) reduced driving force for adsorption as high adsorbent dose causes lower equilibrium fluoride concentration. A dosage of 0.8 g/100 mL was then chosen as the optimum adsorbent dosage and was applied in subsequent experiments.

Effect of initial adsorbate concentration

The effect of adsorbate concentration was evaluated by varying the initial concentration from 3 mg/L to 15 mg/L at 30 min, 60 min and 120 min contact time at a shaking speed of 250 rpm and an adsorbent dosage of 0.8 mg/100 mL. Results are reported in terms of percentage removal in Figure 4c. It is observed that the percentage F- removal decreased with an increase in initial concentration. The same trend was observed for all three contact times. According to Thakre et al.[11], the decrease in fluoride adsorption is because of the availability of more fluoride ions in solution at higher fluoride concentration, which also indicates that the fluoride-binding capacity of smectite-rich clay soil was approaching exhaustion. An initial adsorbate concentration of 3 mg/L was chosen as the optimum concentration for subsequent experiments.

Effect of initial pH

The effect of initial pH on percentage fluoride removal was evaluated by varying the initial pH of the solution from 2 to 12 using 0.1 M NaOH and 0.1 M HCl. The results are presented in Figure 4d. pH of the medium is one of the important parameters that influence fluoride removal efficiency significantly and help to understand the fluoride-uptake mechanism of the adsorbent. From the results presented in Figure 4d it is evident that smectite-rich clay showed significantly high fluoride removal efficiency over a wide range of pH, from 2 to 10. The optimal pH – with a fluoride removal of about 92% – was an acidic pH of 2. A decrease to 31.33% was observed at pH > 10. At pH 2 the surface of the clay is positively charged and F- would be electrostatically attracted to the clay surface, which explains the high F- adsorption at pH 2. The trend of change in pH during fluoride adsorption showed that at a pH below 6, the final pH was higher than the initial pH while at a pH above 6 the solution pH decreases. The increase may be attributed to the release of OH- ions from the adsorbent surface during ion exchange and the decrease in final pH at initial pH above the pH_{pzc} may be attributed to the release of H+ ions from the adsorbent surface.
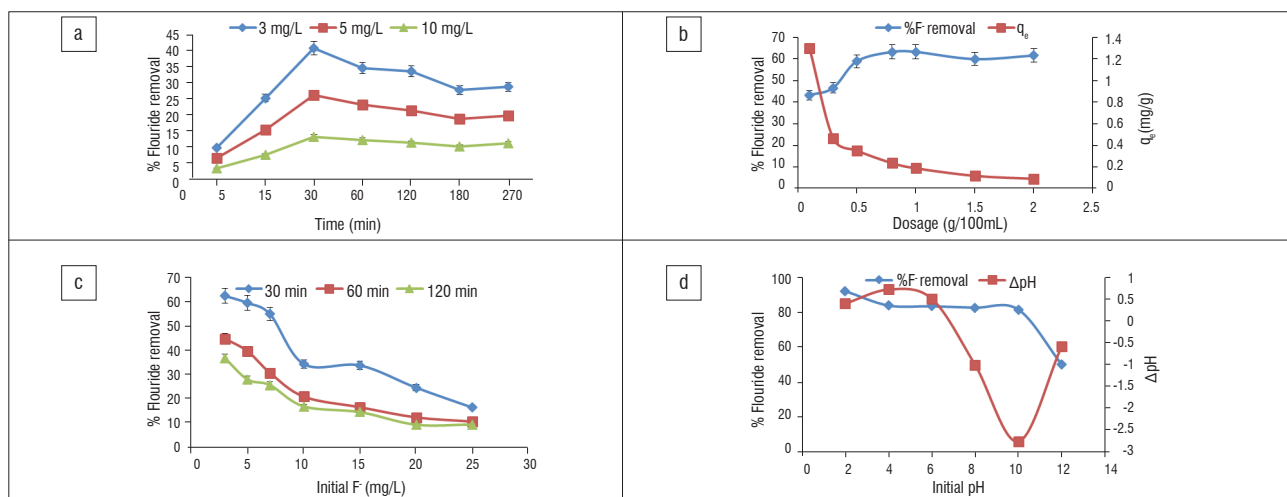
Figure 4: Optimisation of fluoride removal by smectite-rich clay soil: (a) effect of contact time, (b) effect of adsorbent dosage, (c) effect of initial concentration and (d) effect of initial pH.

Effect of co-existing anion

Considering the complexity of natural water sources, many anions such as sulfates, chlorides, bicarbonates and nitrates might be present and might affect the efficiency of an adsorbent to adsorb fluoride. In order to evaluate the effects of co-existing anions on adsorption of fluoride by smectite-rich clay soil, defluoridation was conducted in the presence of each of 5 mg/L SO_4^{2-} Cl^-, CO_3^{2-} and NO_3^-, separately, at an initial fluoride concentration of 3 mg/L; a blank experiment represented the control with an initial fluoride concentration of 3 mg/L. The results are presented in Figure 5.
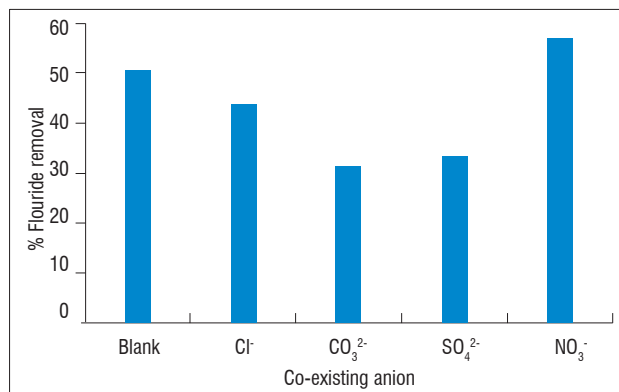
Figure 5: Effects of co-existing anions on fluoride removal by smectite-rich clay soil (at 3 mg/L fluoride, 0.8 g/100 mL adsorbent dosage and 30 min contact time at 250 rpm).

From the results in Figure 5 it is observed that the presence of NO_3^- increased the percentage of fluoride removal to 57%. This increase may be a result of an increase in the ionic strength of the solution or a weakening of lateral repulsion between adsorbed fluoride ions. However, the presence of Cl^-, SO_4^{2-} and CO_3^{2-} decreased the fluoride adsorption to 44%, 33.33% and 31.66%, respectively, possibly because these ions compete with fluoride ions for the surface functional groups on the adsorbent surface, thereby decreasing the fluoride removal. Fluoride removal in the presence of anions increased in the order of $CO_3^{2-} > SO_4^{2-} > Cl^- > NO_3^-$.

Regeneration of adsorbent

To evaluate the regeneration and recyclability of smectite-rich clay soil, five successive adsorption and desorption cycles were performed at an initial concentration of 3 mg/L fluoride, pH of 5.54 and contact time of 30 min. The results are presented in Figure 6. It is observed that the percentage fluoride removal after the first cycle decreased slightly from 56% to 54.6% and continued to decrease after each cycle. The same trend was reported in Zhang et al.[19] and Jia et al.[20] and it was attributed to inadequate regeneration of the adsorbent.
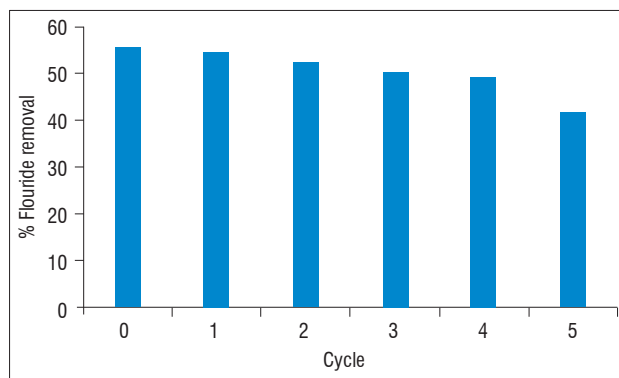
Figure 6: Percentage fluoride removal by smectite-rich clay soil in successive cycles (at 3 mg/L fluoride, 30 min contact time at 250 rpm and pH 5.54).

Defluoridation of field water

The efficiency of smectite-rich clay soil for fluoride removal was tested by using field groundwater collected from the Siloam borehole at the established optimum conditions of pH, contact time and adsorbent dosage. The adsorption of fluoride from groundwater by smectite-rich clay soil was conducted at optimum pH (pH 2) and natural pH (pH 7.8). Table 2 shows the physicochemical parameters of the field water before and after treatment.

Table 2: Physicochemical parameters of field water before and after treatment

Parameter	Field groundwater	Natural pH (7.8)	Optimum pH (2)
pH	7.8	7.3	2.68
Conductivity (μS/cm)	192	155	140
Total dissolved solids (mg/L)	203	168	135
F^- (mg/L)	5.4	3	2.8
Cl^- (mg/L)	32	39	37
SO_4^{2-} (mg/L)	12	11	8
NO_3^- (mg/L)	1.13	ND	2
Br^- (mg/L)	2.1	ND	ND
PO_4^{3-} (mg/L)	3	ND	2
Ca (mg/L)	2.1	3.23	4
Na (mg/L)	72	75.61	80
Al (μg/L)	11.1	20.1	103
Fe (μg/L)	16.5	1.1	<0.001

ND, not detected

The percentage F^- removal was found to be lower at natural pH (44%) than that at optimum pH (52%), as shown in Table 2, which may be a result of the abundance of OH^- ions – which compete with F^- ions – at natural pH (7.8). The percentage F^- removal achieved from field groundwater is lower than the percentage F^- removal achieved with synthetic fluoride solution (92%). This finding may be a result of the presence of competing anions such as Br^-, PO_4^{3-} and SO_4^{2-} in the field water. From Table 2 it can be seen that the concentration of these co-existing ions decreased – indicating that they were adsorbed during the adsorption process. The concentration of Al and Fe chemical species in treated water were found to be within the WHO recommended limits of 0.2 mg/L and 0.3 mg/L, respectively, for potable water. The results show that smectite-rich clay soil from Mukondeni can be used for defluoridation of groundwater containing fluoride at concentrations below 3 mg/L in order to achieve the recommended limit of 1.5 mg/L. However, the fact that the soil removed a higher percentage at lower pH may be the limitation in its application, as the water will need to be acidified before treatment.

Adsorption modelling

Langmuir and Freundlich equations are generally used to describe the equilibrium relationship between adsorbate concentration and the amount of adsorbate adsorbed on the surface of the adsorbent. The Langmuir adsorption isotherm is the most applicable isotherm commonly applied in solid/liquid systems to describe saturated monolayers sorption.[21] It assumes that the adsorbent surface is uniform, i.e. all the adsorption sites are equivalent and adsorption molecules do not interact.[19]

The linear equation of Langmuir adsorption is expressed by Equation 3 as follows:

$$\frac{C_e}{Q_e} = \frac{1}{Q_m b} + \frac{C_e}{Q_m},$$

Equation 3

where C_e is the equilibrium concentration (mg/L), Q_e is the adsorption capacity (mg/g), Q_m is the theoretical maximum adsorption capacity (mg/g) and b is the Langmuir constant related to enthalpy of adsorption (L/mg). Q_m and b are determined from the slope and intercept of the plot of $\frac{C_e}{Q_e}$ vs. c_e. Figure 7a shows the plot of the Langmuir isotherm. The values of Langmuir parameters Q_m and b are reported in Table 3. In order to predict the adsorption efficiency of the process, the dimensionless equilibrium parameter (R_L) was calculated using Equation 4:

$$R_L = \frac{1}{1+bC_o},$$

Equation 4

where C_o is the initial concentration and b is the Langmuir constant. A value of R_L less than 1 generally indicates favourable adsorption and a value greater than 1 indicates unfavourable adsorption. Calculated R_L values at various contact times in Figure 7b lie within the range 0–1, indicating that the adsorption process was favourable at room temperature for all the adsorbate concentrations tested.

The Freundlich adsorption model is an empirical model which is indicative of the surface heterogeneity of the adsorbent and considers multilayer adsorption.[10] It is expressed by Equation 5:

$$\log Q_e = \log K_f + \frac{1}{n} \log C_e,$$

Equation 5

where C_e is the equilibrium concentration (mg/L), Q_e is the amount adsorbed at equilibrium (mg/g), K_f is the Freundlich constant related to adsorption capacity and represents the strength of adsorptive bonds and $1/n$ is the adsorption intensity. When $0<1/n<1$, the adsorption is favourable; when $1/n=1$, the adsorption is irreversible; and when $1/n>1$, the adsorption is unfavourable. The values of K_f and $1/n$ are obtained from the slope and intercepts of a linear plot of log Q_e vs log C_e. Figure 7c shows the Freundlich isotherm. Values of K_f and $1/n$ are reported in Table 3. Based on the correlation coefficient values for both contact times, the adsorption data fit better to the Langmuir adsorption isotherm than the Freundlich adsorption isotherm. Better fit of the Langmuir isotherm model suggests monolayer uniform adsorption on the surface of the adsorbent. Calculated model constants for Langmuir and Freundlich isotherms are shown in Table 3. The value of R_L, a dimensionless equilibrium parameter, and the value of $1/n$ – the Freundlich adsorption intensity – were both between 0 and 1, which indicates that the adsorption of F⁻ onto smectite-rich clay soil was favourable.

Adsorption kinetics

Theoretically, the adsorption of fluoride onto a solid particle is controlled by different mechanisms. These mechanisms involve diffusion or transport of fluoride from bulk solution to the exterior surface of the adsorbent particle, adsorption of F⁻ onto the particle surface, movement of the solute within the pores of the particle, attachment of the solute at sites on the interior surface of the adsorbent via sorption, complexation and intra-particle diffusion phenomena.[22,23] Pseudo first and second order rate reactions and the intra-particle diffusion models are widely used to elucidate the adsorption mechanisms and the rate-limiting factors.

A pseudo first order equation is represented by Equation 6. The model is widely used to describe liquid–solid phase adsorption systems, and it is the earliest known kinetic model describing the adsorption rate based on the adsorption capacity.[24]

$$\log(q_e\text{-}q_t) = \log(q_e)-\frac{k_{ad}t}{2.303}$$

Equation 6

A pseudo second order equation is represented by linear Equation 7 and is used to describe chemisorption, as well as cation exchange reactions:

$$\frac{t}{q_t} = \frac{1}{K_{2ads}q_e^2} + \frac{1}{q_e}t,$$

Equation 7

where q_e and q_t (both in mg/g) are the amounts adsorbed per unit mass at a time, t (in min). K_{ad} (min⁻¹) and K_{2ads} (g/mg.min) are first and second order rate constants (g/mg.min). The value of K_{ad} is determined from the slope and intercepts of log (q_e-q_t) vs t (min) and the value of K_{2ads} is determined from the slope and intercepts of t/q_t vs t. Figure 7d and 7e show plots of pseudo first and second order, respectively. The plots of pseudo second order yielded straight lines and higher correlation coefficients at both concentrations. The plots of pseudo first order did not give straight lines. Therefore adsorption of fluoride onto smectite-rich clay soil followed pseudo second order and occurred through chemisorption. Model constants for pseudo first and second order are presented in Table 3. In order to predict the rate-limiting step in the adsorption of F⁻ onto smectite-rich clay soil, the possibility of intra-particle diffusion was evaluated using the Weber–Morris model of intra-particle diffusion expressed by Equation 8[23]:

$$qt = K_i t^{0.5}$$

Equation 8

where q_t is the amount adsorbed (mg/g) at a given time t (min) and K_i (mg/g.min) is the intra-particle diffusion rate constant determined from the slope of t⁰·⁵ vs q_t. Figure 7f shows the plot of the amount of fluoride adsorbed and t⁰·⁵. The intercept did not pass through the origin at each of the evaluated concentrations and the data show a bilinear plot, with the initial portion (phase 1) indicating the boundary layer diffusion and the other represents the intra-particle diffusion (phase 2) and the values of K_{i1} and K_{i2} (intra-particle diffusion co-efficient rate constant for phase 1 and phase 2, respectively) obtained from the plot are shown in Table 3. At both initial concentrations, the K_{i1} value is higher than the K_{i2} value, suggesting that the initial sorption step is more rapid than the final step (phase 2), which may be because of differences in the rate of mass transfer in the initial and final stages of adsorption.[20] This observation is connected to a mechanism consisting of an external mass transfer followed by diffusion into micro- and mesoporous surfaces. This means adsorption of F⁻ onto smectite-rich clay soil is a complex process involving both boundary and intra-particle diffusion adsorption processes.

Mechanistic aspects of adsorption of F⁻ onto smectite-rich clay soil

The mechanism of fluoride adsorption onto smectite-rich clay was evaluated by comparing the FTIR spectra of smectite-rich clay soil before and after adsorption (Figure 3). FTIR spectra showed that the clay soil surface is mainly characterised by Si-OH and Al-OH groups which may be easily modified by changing the pH of the medium. pH$_{pzc}$ determination showed that at pH 5.8 the surface has neutral charge, below 5.8 the surface is positively charged and above 5.8 it is negatively charged. Figure 4d shows that the adsorption of F⁻ onto smectite-rich clay soil was optimum at low pH and decreases as the pH increases, which may have been a result of the increase in OH⁻ which competes with F⁻ for adsorption sites. After F⁻ adsorption, the FTIR spectra showed a decreased peak intensity of transmittance indicating that during adsorption there was ion exchange between OH⁻ in the surface and F⁻ in bulk solution, leading to the formation of new bonds such as SiF and AlF. Different mechanisms can be suggested at various pH levels. At pH below 5.8, the surface of the clay is positively charged (Equation 9) and therefore F⁻ ions will be electrostatically adsorbed to the surface of the clay (Equation 10). Furthermore, low pH causes protonisation of surface –OH groups to –OH₂⁺ which facilitates the ligand exchange mechanism because of stronger attractive forces between fluoride and the adsorbent surface and the presence of more hydroxylated sites for exchange of F⁻ than at high pH.[8] At moderate pH, fluoride adsorption occurs via ion exchange (Equation 11). At a pH above the pH$_{pzc}$ at which the surface of the clay is negatively charged, fluoride adsorption also occurs via ion exchange (Equation 12).
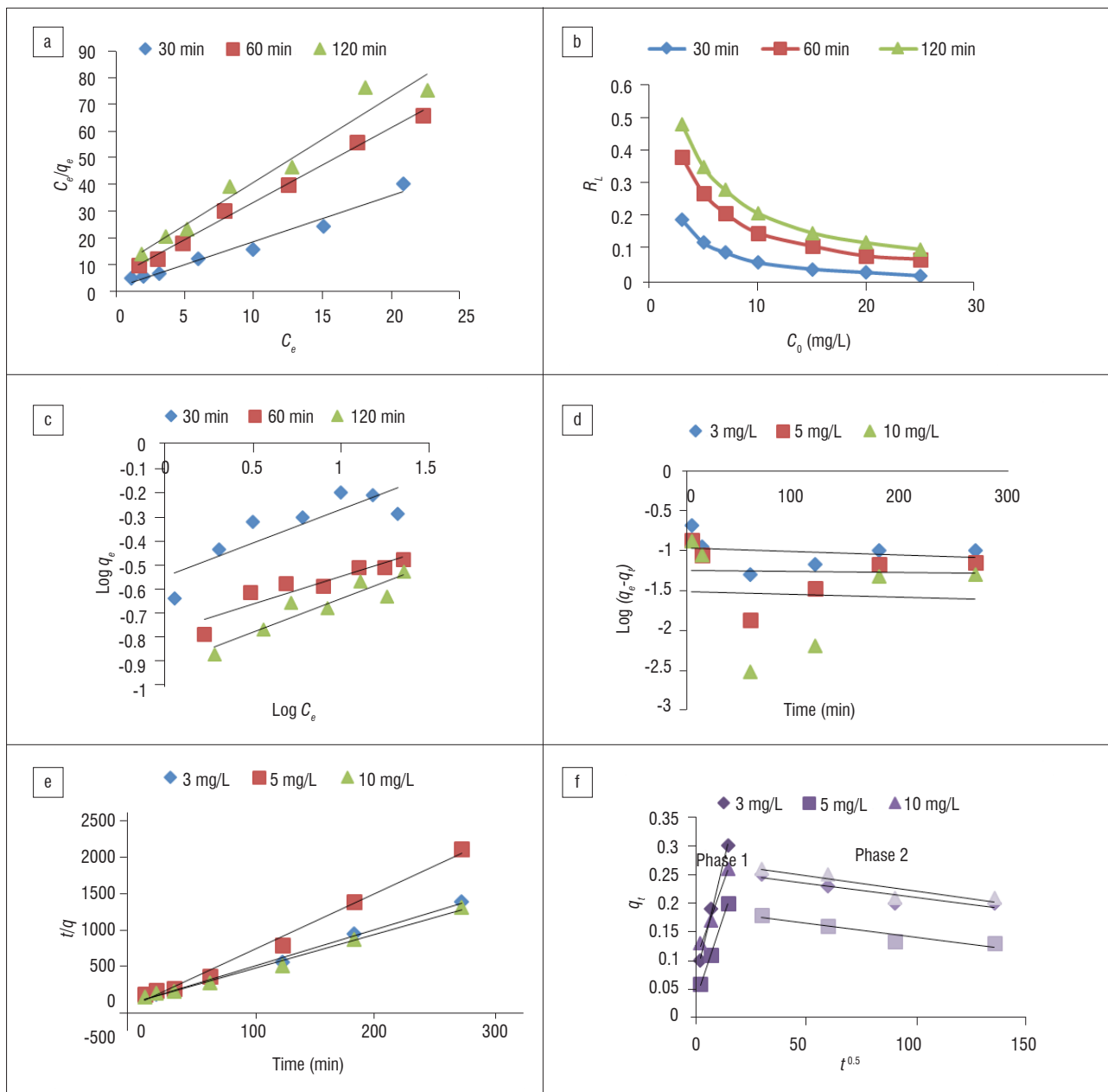
Figure 7: (a) Langmuir isotherm, (b) equilibrium dimensionless parameter, (c) Freundlich isotherm, (d) pseudo first order, (e) pseudo second order and (f) intra-particle diffusion of the adsorption of fluoride by smectite-rich clay soil.

Table 3: Constant values for adsorption isotherms and adsorption kinetics model

	Langmuir isotherm			Freundlich isotherm		
	Q_m (mg/g)	b (L/mg)	R^2	K_f (mg/g)	$1/n$	R^2
30 min	0.58	1.36	0.97	0.28	0.27	0.73
60 min	0.35	0.53	0.99	0.2	0.22	0.85
120 min	0.31	0.36	0.99	0.12	0.27	0.63

	Pseudo first order			Pseudo second order			Intra-particle diffusion	
	$q_{e(exp)}$ (mg/g)	K_{ad} (min^{-1})	R^2	K_{2ad} (g/mg.min)	$q_{e(cal)}$ (mg/g)	R^2	K_{i1}	K_{i2}
3 mg/L	0.19	1.15x10^{-3}	0.05	0.98	0.23	0.99	0.05	-0.005
5 mg/L	0.13	2.3x10^{-4}	0.001	1.13	0.2	0.98	0.01	-0.005
10 mg/L	0.21	6.9x10^{-4}	0.003	1.12	0.21	0.99	0.01	-0.005

$$\equiv M(OH)_2 + H^+ \leftrightarrow \equiv MOH_3^+ \qquad \text{Equation 9}$$

$$\equiv MOH + H_3O^+ + F^- \leftrightarrow \equiv MOH_2^+\text{-}F^- + H_2O \qquad \text{Equation 10}$$

$$\equiv MOH + F^- \leftrightarrow MF + OH^- \qquad \text{Equation 11}$$

$$\equiv M(OH)_2^- + 2F^- \leftrightarrow MF_2 + 2OH^- \qquad \text{Equation 12}$$
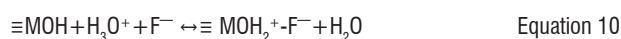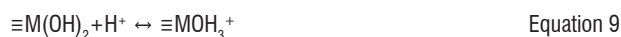
where $\equiv M$ represents Si and Al in the adsorbent surface.

Conclusions

In the assessment of the use of smectite-rich clay soil from Mukondeni for the defluoridation of groundwater, the following major conclusions were made:

- Optimum conditions were found to be 30 min contact time at a 250 rpm shaking speed on a table shaker, with an adsorbent dosage of 0.8 g/100 mL, a 3 mg/L adsorbate concentration and a pH of 2. Fluoride removal at optimum conditions was $\approx$92% for an initial concentration of 3 mg/L.

- Smectite-rich clay soil can be successfully regenerated using 0.1 M NaOH.

- Adsorption isotherms showed that the data fit better to the Langmuir isotherm than to the Freundlich isotherm, suggesting that adsorption of fluoride occurred on a monolayer surface.

- Adsorption kinetic studies showed that the data fit better to pseudo second order than to pseudo first order suggesting that the adsorption occurs via chemisorption.

- The data did not fit well to the intra-particle diffusion model of Weber–Morris and showed a bilinear plot which indicated that the adsorption of fluoride onto smectite-rich clay soil is a complex process.

- Ligand exchange and ion exchange mechanisms of F⁻ adsorption are suggested to be the main mechanism taking place during the adsorption process.

- Field water defluoridation experimental results showed that smectite-rich clay soil is a potential candidate material for defluoridation.

- We recommend further research to improve the adsorptive capacity of the smectite-rich clay soils.

Acknowledgements

Authors' contributions

M.W.G. was the project leader; M.W.G. and R.M. were responsible for the experimental and project design; R.M. carried out the experiments and performed the calculations; T.A.M.M. made conceptual contributions; and R.M. and M.W.G. wrote the manuscript.

References

1. Nagendra CR. Fluoride and environment: A review. In: Martin JB, Madha SV, Vasantha KT, editors. Proceedings of the Third International Conference on Environment and Health; 2003 December 15–17; Chennai, India. Chennai / Toronto: Department of Geography, University of Madras / Faculty of Environmental Studies, York University; 2003. p. 386–399.

2. Susheela AK. Treatise on fluorosis. New Delhi: Fluorosis Research and Rural Development Foundation; 2001.

3. World Health Organization. Guideline for drinking water. Geneva: World Health Organization; 2011.

4. Ncube EJ, Schutte CF. The occurrence of fluoride in South African groundwater: A water quality and health problem. Water SA. 2005;31(1):35–40. http://dx.doi.org/10.4314/wsa.v31i1.5118

5. Meenakshi SK, Maheshwari RC. Fluoride in drinking water and its removal. J Hazard Mater. 2006;137(1):456–463. http://dx.doi.org/10.1016/j.jhazmat.2006.02.024

6. Coetzee PP, Coetzee LL, Mubenga S. Characterization of selected South African clays for defluoridation of natural water. Water SA. 2003;29(3):331–338.

7. Ghorai S, Pant KK. Investigations on the column performance of fluoride adsorption by activated alumina in a fixed-bed. Chem Eng J. 2004;98:165–173. http://dx.doi.org/10.1016/j.cej.2003.07.003

8. Ma Y, Wang SG, Fan M, Gong WX, Gao BY. Characteristics and defluoridation performance of granular activated carbons coated with manganese oxides. J Hazard Mater. 2012;168:1140–1146. http://dx.doi.org/10.1016/j.jhazmat.2009.02.145

9. Wang Y, Reardon EJ. Activation and regeneration of a soil sorbent for defluoridation of drinking water. Appl Geochem. 2001;16:531–539. http://dx.doi.org/10.1016/S0883-2927(00)00050-0

10. Kamble S, Dixit P, Rayalu SS, Labhsetwar NK. Magnesium incorporated bentonite clay for defluoridation of drinking water. Desalination Water Treat. 2009;249:687–693.

11. Thakre D, Rayalu S, Kawade R, Meshram S, Subrt J, Labhsetwar N. Magnesium incorporated bentonite clay for defluoridation of drinking water. J Hazard Mater. 2012;180:122–130. http://dx.doi.org/10.1016/j.jhazmat.2010.04.001

12. Gitari WM, Ngulube T, Masindi V, Gumbo JR. Defluoridation of groundwater using Fe³⁺-modified bentonite clay: Optimization of adsorption condition. Desalination Water Treat. 2015:53(6):1578–1590. http://dx.doi.org/10.1080/19443994.2013.855669

13. Masindi V, Gitari, MW, Ngulube T. Defluoridation of drinking water using Al³⁺-modified bentonite clay: Optimization of fluoride adsorption conditions. Toxicol Environ Chem. 2014;96(9):1294–1309. http://dx.doi.org/10.1080/02772248.2014.977289

14. Tyeryar M, Hackett C, Harsh D, Hackett T. Establishing a ceramic water filter factory in Limpopo Province, South Africa. Jefferson Public Citizens. 2012;3:104–110.

15. Dacosta FA, Muzerengi C, Mhlongo SE, Mukwevho GF. Characterization of clays for making ceramic pots and water filters at Mukondeni Village, Limpopo Province, South Africa. J Eng Appl Sci. 2013;8(11):927–932.

16. Toor M, Jin B, Dai S, Vimonoses V. Activating natural bentonite as a cost effective adsorbent for removal of Congo red in waste water. J Ind Eng Chem. 2014;1979:1–9.

17. Spark DL. Environmental soil chemistry. 2nd ed. Amsterdam: Academic Press; 1997.

18. Bohn HL, Mcneal BL, O'Connor G. Soil chemistry. New York: John Wiley and Sons; 1975.

19. Zhang Z, Tan Y, Zhong M. Defluoridation of wastewater by calcium chloride modified natural zeolite. Desalination. 2011;276:246–252. http://dx.doi.org/10.1016/j.desal.2011.03.057

20. Jia Y, Zhu BS, Zhang KS, Jin Z, Luo T, Yu XY, et al. Porous 2-line ferrihydrite/byerite composite (LFBC): Fluoride removal performance and mechanism. Chem Eng J. 2015;268:325–336. http://dx.doi.org/10.1016/j.cej.2015.01.080

21. Sun Y, Fang Q, Dong J, Cheng X, Xu J. Removal of fluoride from drinking water by natural stilbite zeolite modified with Fe(III). Desalination. 2011;277:121–127. http://dx.doi.org/10.1016/j.desal.2011.04.013

22. Maliyekkal SM, Sharma AK, Philip L. Manganese-oxide-coated alumina: A promising sorbent for defluoridation of water. Water Res. 2006;40:3497–3506. http://dx.doi.org/10.1016/j.watres.2006.08.007

23. Chen N, Zhang Z, Feng C, Li M, Zhu D, Sugiura N. Studies on fluoride adsorption of iron-impregnated granular ceramics from aqueous solution. Mater Chem Phys. 2011;125:293–298. http://dx.doi.org/10.1016/j.matchemphys.2010.09.037

24. Oladoja NA, Helmreich B. Batch defluoridation appraisal of aluminium oxide infused diatomaceous earth. Chem Eng J. 2014;258:51–61. http://dx.doi.org/10.1016/j.cej.2014.07.070

Stirred cell ultrafiltration of lignin from black liquor generated from South African kraft mills

AUTHORS:
Paul Kekana[1,2]
Bruce Sithole[1,2]
Deresh Ramjugernath[2]

AFFILIATIONS:
[1]Forestry and Forest Products Research Centre, Natural Resources and the Environment, Council for Scientific and Industrial Research, Durban, South Africa

[2]Discipline of Chemical Engineering, University of KwaZulu-Natal, Durban, South Africa

CORRESPONDENCE TO:
Bruce Sithole

EMAIL:
bsithole@csir.co.za

ARTICLE INCLUDES:
✓ Supplementary material
✗ Data set

Ultrafiltration of lignin from black liquor was carried out in a stirred batch cell using polyethersulfone membranes. Parameters such as operating pressure, feed concentration, stirring rate and membrane cut-off size were varied and their effects on lignin retention and permeate flux were investigated. The operating pressure, feed concentration and stirring rate were varied in the ranges 150–350 kPa, 3–9% and 200–400 rpm, respectively. The membranes used had cut-off sizes of 5 kDa, 10 kDa and 20 kDa. A one-factor-at-a-time experimental design approach was applied in this study. Retention of lignin increased with increases in operating pressure, feed concentration and stirring rate, but decreased with an increase in molecular cut-off size of the membrane. Permeate flux on the other hand increased with increases in pressure, stirring rate and molecular cut-off size of the membrane but decreased with an increase in feed concentration. The extraction of lignin from black liquor was successfully carried out and extraction efficiencies as high as 86% could be achieved depending on the experimental conditions. The study was concluded with the recommendation of conducting additional experiments using a pilot plant in a continuous mode.

Significance:
- The extraction of lignin from black liquor was successfully carried out and extraction efficiencies as high as 86% were obtained. The results can be used to extend the ultrafiltration of black liquor to an industrial scale.

Introduction

The extraction of lignin from black liquor can be achieved by means of acid precipitation, ultrafiltration and the application of selective solvents.[1,2] The extraction method employed should effectively separate the lignin from the black liquor, without causing an imbalance of the cooking chemicals, Na_2S and NaOH. Acid precipitation is the most commonly used method and it has reached the most advanced state of development and implementation.[3,4] Drawbacks of using precipitation include the use of H_2SO_4 which upsets the liquor cycle chemical balance with excess sulfur and the process of filtering and separating the lignin precipitate is hampered by the formation of colloids.[4] The application of selective solvents such as ionic liquids allows for lignin structures of low molecular weight to be obtained, which are used in the synthesis of high added value products, but with very high production costs.[5] Ultrafiltration is therefore seen as a promising way of separating lignin from kraft black liquors. The qualities that make membrane processes the ideal technique for separation purposes in biorefineries include their excellent fractionation capability and comparatively low energy requirement.[6,7]

The fractionation of lignin by sequential precipitation from softwood industrial black liquor samples collected from different chemical pulping stages was investigated by Alekhina et al.[8] The authors studied the isolation efficiency as well as the impact of fractionation and pulping severity on lignin structure. It was observed that the spectroscopic characterisation of the isolated lignin revealed significant alteration in its structure and functionalities as a function of the pH, while the alteration in structure and properties of the samples was only marginal. Lignin precipitates isolated at pH 10.5 exhibited the highest purity (lowest content of polysaccharides) while the samples precipitated at a pH of 2.5 revealed the highest carbohydrate content (low purity). These results indicated that up to 85% of the lignin solubilised in softwood black liquor can be recovered in high purity by lowering the pH to 5. Jin et al.[9] developed a novel membrane-assisted electrochemical approach for the precipitation of lignin from black liquor. In the process, the pH in the black liquor solution was lowered to 4.7 as a result of water electrolysis, without the addition of acid or CO_2, leading to pH-dependent lignin precipitation. In the study it was observed that the electrochemical cell performance was significantly influenced by the condition of the electrolytes. The solution conductivity and the applied current density were also found to play important roles in altering the pH of the system. In addition, more than 70% of the chemical oxygen demand value in the black liquor solution was decreased at the set pH, which was comparatively higher than that of the conventional acidic precipitation. This difference was attributed to lignin precipitation and oxidation. The authors concluded that this novel membrane-assisted electrochemical approach may serve as a promising and cost-effective technique for the extraction of lignin and recovery of caustic from black liquor. Velez and Thies[10] studied the precipitation of liquefied-lignin fractions from a softwood kraft black liquor by acidification with CO_2 in a 2-L vessel with a 45°-angle conical bottom to facilitate the collection of liquid-lignin fractions. The process took place at 115 °C and 6.2 bar over a pH range of 13.6–9.5. In their study, fresh kraft black liquor was used as the feed only for the first fractionation and the partially spent black liquor from the first then served as the feed to the second fractionation, and so on. Seven of these liquefied-lignin fractions were produced from the study and they were found to be highly hydrated phases, containing 32.3–48.2 wt% water. The authors also concluded that the metal contents in the lignin fractions were significantly reduced compared to the original black-liquor feed.

Several studies have been undertaken to investigate the extraction of lignin from black liquor using ultrafiltration.[1-2,6,11-14] A comparative study on the ultrafiltration of kraft black liquor using different flow modules was carried out by Satyanarayana et al.[15] The different modules investigated included radial cross flow, rectangular cross

flow and stirred cell. The authors concluded that the observed rejection and permeate flux were higher in the stirred cell module than with the radial and rectangular cells. Ultrafiltration of black liquor in a stirred batch cell using a cellulose acetate membrane of 5 kDa was carried out by Bhattacharjee and Bhattacharya[16]. The membrane parameters such as solute permeability and reflection coefficient were determined to characterise the membrane. It was found that the flux for black liquor was comparatively lower than other solutes such as polyethylene glycol which could be attributed to the increased thickness of the polarised layer. Bhattacharjee and Bhattacharya[17] studied the performance of ultrafiltration of black liquor by using a laboratory fabricated stirred and rotating disk batch ultrafiltration cell. The study was carried out as an attempt to minimise flux decline in order to obtain enhanced flux for the treatment of black liquor obtained from sulfite-pulping industries. The authors concluded that membrane rotating was more efficient in reducing concentration polarisation compared to stirring action. However, it is worth mentioning that the lignin dissolved from such alkaline sulfite pulping processes differs from kraft lignin.

The present work was undertaken to investigate the possibility of extracting lignin from black liquor generated from South African kraft mill black liquors. To our knowledge, a study of this nature has not been conducted in the South African context, so it is expedient to study the feasibility of ultrafiltration of lignin in black liquors from South African mills. In this study, the effects of operating pressure, membrane cut-off size, stirring rate and feed concentration on the extent of lignin retention and permeate flux were investigated. It has to be emphasised that only the effects of these parameters were investigated: the study did not include the optimisation of the parameters as the results from a stirred cell study cannot be applied on an industrial scale. The results only assist in the scale up of the process to a continuous pilot plant scale which is industrially relevant. The role of fouling is also monitored in the study and the results reported.

Materials and methods

Materials

Black liquor was procured from a South African eucalyptus kraft mill. The black liquor was diluted with deionised water to get the desired concentrations. Hydrophilic polyethersulfone membranes with molecular cut-off sizes of 5 kDa, 10 kDa and 20 kDa were procured from Memcon (Pty) Ltd, South Africa. The membranes are usable in the pH range 0–14 and are resistant to temperatures up to 95 °C. The composition of the black liquor used is reported in Table 1.

Table 1: Concentrations of selected fractions in black liquor

Analyte	% (m/m)
Total solids content	14.53
Lignin	4.53
Ash content (dry basis)	9.02
Water content	84.97
Hemicelluloses (sugars)	
Xylose	0.03
Arabinose	0.02
Galactose	0.01

Apparatus

A stirred dead-end filtration configuration as shown in Figure 1 was employed for the ultrafiltration experiments. The stirred cell (Amicon 8400, Merck, South Africa) had a volume of 400 mL and a diameter of 76 mm. The cell was connected to a compressed air cylinder; the pressure in the cell was controlled by adjusting the pressure regulator

on the outlet side of the cylinder and was monitored by a calibrated pressure gauge on the inlet of the cell. The effective filtration area of the cell was 41.8 cm^2 and the clearance between the flat stirrer hanging from the top inside the cell and the membrane was 1.5 mm. Stirring inside the cell was accomplished by using a magnetic stirrer with a digital display (MS-H280-Pro, Scilogex, USA). A constant temperature in the cell during operation was achieved by means of a heating copper coil inserted around the stirred cell body. Water from a heating bath was circulated in the heating coil using a peristaltic pump (323, Watson Marlow, South Africa).
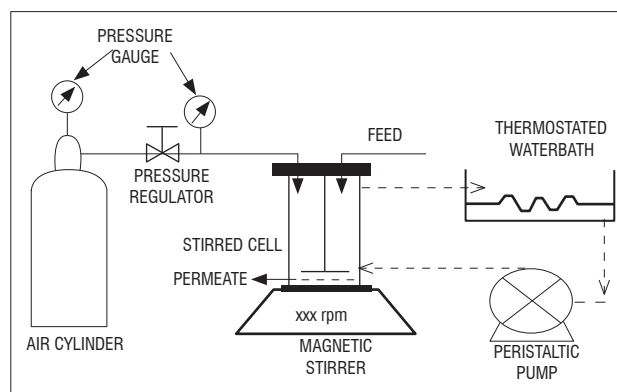
Figure 1: Schematic of the batch ultrafiltration set-up.

Analysis

The total dissolved solids concentration was determined gravimetrically by evaporating a known volume of black liquor sample at 105 °C for 24 h and determining the weight of the residue. The ash content was measured by heating the residue from the total dissolved solids measurement to 950 °C and weighing the sample afterwards. The ash content is indicative of the concentration of inorganic matter in the sample. Lignin can be measured from the light absorption at a wavelength of 280 nm as it contains phenolic groups which absorb light. A UV-visible spectrophotometer (Varian CARY 50 CONC) was used to measure the UV light absorption of the black liquor samples. Where necessary, the samples were diluted with 0.1 M NaOH and an absorption constant of 24.6 L/g.cm was used. The hemicellulose (sugar) content of the black liquor was determined using high-performance liquid chromatography, comprising an autosampler (Perkin Elmer, Series 200), column oven (Perkin Elmer, Series 200), chromatography interface (Perkin Elmer, 600 Series), pulsed amperometric detector (Dionex) and analytical column (CarboPac PA-1, Dionex). The black liquor was first adjusted to a pH of 5 to 6 using 6 mol/L HCl and then the sugars were hydrolysed via heating with 4% H_2SO_4 at 121 °C for 1 h. The mass balance of selected analytes indicated that the error from analysis did not exceed 10% (see Supplementary tables 1 and 2 of the supplementary material).

Ultrafiltration procedure

Ultrafiltration experiments were carried out to investigate the effect on lignin retention and transient flux decline of three variables: transmembrane pressure (150, 250 and 350 kPa), stirring rate (200, 300 and 400 rpm) and feed concentration (3, 6 and 9 wt%). One parameter was varied as the others were held constant to get an exact picture on dependence. All experiments were conducted at 60 °C by circulating a heating water stream at 95 °C and 15 rpm in the coil around the stirred cell body. The disk membrane was placed on the porous support and the cell was assembled. Pure water flux at different transmembrane pressures was measured and plotted against the transmembrane pressure. The membrane resistance for each membrane was calculated from the slope of this plot. This calculation was followed by the actual experiment by charging the cell with 200 mL black liquor solution. The transmembrane pressure and stirrer speed were adjusted to desired levels using a pressure regulator and the speed controller on the magnetic stirrer, respectively. The duration of each experiment was 1 h and permeate at different time intervals was measured by collecting

10 mL of permeate in a measuring cylinder and recording the time of this collection. The retentate was collected by opening the stirred cell after each run. After each run, the cell was dismantled and the membrane thoroughly washed with deionised water to remove any deposition. The membrane was soaked in distilled water overnight and pure water flux was checked again to observe any variation in its hydraulic resistance before its reuse. This procedure was repeated after every experiment.

Results and discussion

Membrane hydraulic resistance

The hydraulic resistance of each membrane investigated was determined by recording the pure water flux as a function of pressure in the stirred cell and obtaining a linear relationship between the two parameters for each membrane. The membrane resistance R_M was calculated from Equation 1[18] and the results are shown in Table 2.

$$R_M = \frac{\Delta P}{\mu_w J_w}$$

Equation 1

where ΔP is the transmembrane pressure, J_w is the pure water flux and μ_w is the viscosity of water with the effect of temperature taken into consideration.

Table 2: Membrane hydraulic resistances (R_M) of all investigated membrane sizes

Membrane cut-off size (kDa)	$(10^{12} m^{-1})$
5	14.5
10	5.08
20	3.39

From Table 2 it is observed that the membrane resistance increased with decreasing cut-off size of the membrane. This finding was expected, because for membranes with lower molecular weight cut-off, the pore sizes become constricted and a reduction in flux is expected.

Stirred cell ultrafiltration of black liquor

Stirred cell ultrafiltration tests were conducted to investigate the effects of operating pressure, stirring rate and feed concentration on the extent of retention and flux. Polyethersulfone membranes were used and the pressure, feed concentration and stirring rate were varied in the range 150–350 kPa, 3–9% and 200–400 rpm, respectively. The lignin retention in the stirred cell was calculated as follows[19]:

$$R = \frac{\ln(c_R/c_F)}{\ln(VCF)} = \frac{\ln[VCF-(c_P/c_F)(VCF-1)]}{\ln(VCF)}$$

Equation 2

where c_F, c_P and c_R are the feed, permeate and retentate concentrations (g/L), respectively, and VCF is the volume concentration factor (dimensionless). Lignin retention values calculated in this manner

consider either the retentate or the permeate concentrations, and the corresponding VCF ratio, allowing comparison of retention values throughout the process. VCF is dimensionless and given as[19]:

$$VCF = \frac{V_F}{V_R}$$

Equation 3

where V_F and V_R are the feed and retentate volumes, respectively. The VCF can be useful in the calculation of observed retention values of a solute during batch processes.[19]

Typical lignin analysis results obtained from the study are tabulated in Table 3. These are for pressure variation experiments in which the stirring speed and feed concentration were fixed at 200 rpm and 9%, respectively. From Table 3, it can be seen that the lignin fraction was concentrated in the retentate stream (for example at an operating pressure of 350 kPa); the lignin concentration increased to 81.2 g/L from 36.4 g/L in the raw sample. As is also shown in Table 3, the error associated with the analysis was below 10%. All results from other experiments showed a similar trend (data not shown).

Membrane selection

Ultrafiltration was carried out in a stirred cell with three membranes that had cut-off sizes of 5 kDa, 10 kDa and 20 kDa, because black liquor is a polydispersed solution containing solutes within a wide distribution of molecular weights (and sizes) ranging from 100 to 100 000. The selection of a membrane cut-off size is therefore important for the efficiency of the process.[15] The variation of retention and flux with pressure are shown in Figures 2 and 3, respectively. For this study, the feed concentration and stirring rate were fixed at 9% and 250 rpm, respectively. In Figure 2, it is shown that retention increased when a membrane with a lower cut-off size was used. For instance, at a transmembrane pressure of 150 kPa, retention was 83.9% for the 5-kDa membrane, whereas under the same experimental conditions, the retention of the 20-kDa membrane was 56.1%. The same trend was observed for the entire pressure range tested. The results show that membranes with higher cut-off sizes allow more solutes with sizes less than the pore size of that particular membrane to pass through.

The initial permeate flux, as shown in Figure 3, was found to be comparatively higher for membranes with higher cut-off sizes for an operating pressure of 150 kPa. The initial permeate flux for the 5-kDa membrane was comparatively lower at 15.1 L/m²h and it remained fairly constant throughout the course of the run.

For the 10-kDa membrane it was 33.1 L/m²h, which was a twofold increase in flux compared to that of the 5-kDa membrane. This flux was maintained at 33.1 L/m²h for 30 min and started to decrease gradually thereafter. For the 20-kDa membrane, the flux was 47.9 L/m²h (an almost 45% increase compared to that of the 10-kDa membrane) and it could only be maintained for 20 min, after which it started to decrease rapidly during the course of the run. The decline in permeate flux for the 10-kDa and 20-kDa membranes can be attributed to the deposition of the solute particles in the membrane pores as well as build-up of the deposited layer on the membrane surface. The combined effect of these resulted in a reduction in permeate flux during the course of operation.[15]

Table 3: Lignin mass balance (at a stirring rate of 200 rpm and feed concentration of 9%)

Experimental conditions	Lignin concentration (g/L)		Sample volume after run (mL)		Calculated mass of lignin in sample (g)		Lignin (% in feed)		Analysis error (%)
	Permeate	Retentate	Permeate	Retentate	Permeate	Retentate	Permeate	Retentate	
150 kPa	15.87	61.6	106	94	1.71	5.67	23.54	77.86	-1.41
250 kPa	13.19	62.5	108	92	1.39	5.87	19.20	80.73	0.07
350 kPa	13.54	81.17	131	69	1.77	5.60	24.36	76.95	-1.31
Raw sample	36.39		200		7.28				
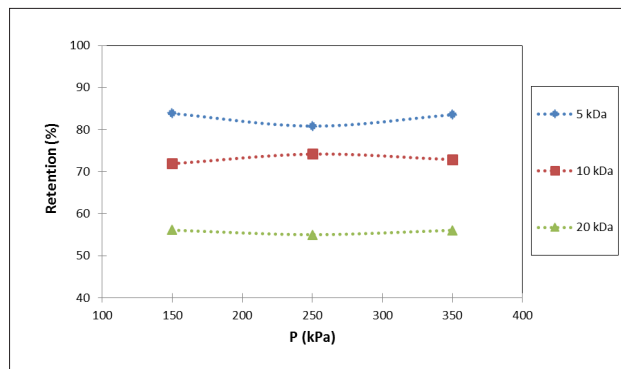
Figure 2: Variation in lignin retention with pressure for different membranes in a stirred cell (at a stirring speed of 250 rpm and a feed concentration of 9%).

Figure 3: Variation in permeate flux as a function of time for different membranes in a stirred cell (at a stirring speed of 300 rpm, feed concentration of 9% and pressure of 150 kPa).

Figure 4: Variation in lignin retention with pressure at different feed concentrations in a stirred cell (with a membrane size of 10 kDa and a stirring speed of 300 rpm).
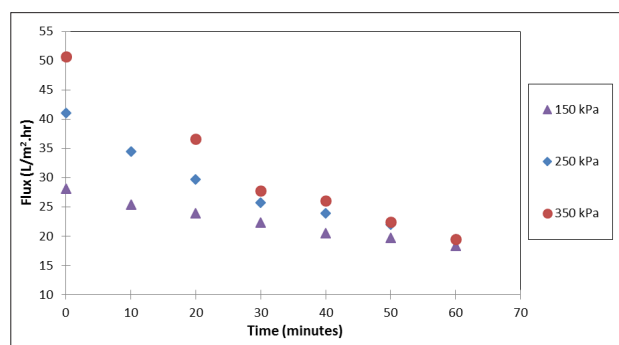
Figure 5: Variation in permeate flux as a function of time at different pressures in a stirred cell (at a feed concentration of 6% and stirring speed of 200 rpm with a membrane size of 10 kDa).

Effect of pressure on lignin retention and permeate flux

Figures 4 and 5 represent the effect of pressure on retention and permeate flux, respectively. For this study, a 10-kDa membrane was used and Figure 4 is plotted for feed concentrations 3%, 6% and 9%. An increase in pressure from 150 kPa to 250 kPa, as shown in Figure 4, resulted in only a marginal increase in the retention of lignin of less than 2% for each concentration tested. However, for the 6% and 9% solutions, as the pressure was increased further from 250 kPa to 350 kPa, the retention increased from 60.7% to 67.3% and from 73.8% to 79.5%, respectively. This increase was attributed to the fact that, as the pressure is increased, the convective diffusion of solutes to the membrane surface is increased, which in turn results in severe concentration polarisation. This polarisation leads to an increase in viscosity of the solution near the membrane surface as the higher molecular weight organics are retained by the membrane and this viscous layer is responsible for pre-sieving of the organic solutes.[12]

The high initial permeate flux values at higher pressures are as a result of an increase in driving force of the solvent through the membrane, as shown in Figure 5. However, the decrease in permeate flux during the course of operation is rapid for higher pressures. For ultrafiltration at 350 kPa, after 20 min, the flux declined by 45% from 50.66 L/m²h to 27.7 L/m²h whereas that for ultrafiltration at 250 kPa declined by 28% from 41.0 L/m²h to 29.7 L/m²h in the same time interval. For ultrafiltration at 150 kPa, after 20 min, only a 15% decline in permeate flux (from 28.1 to 23.9L/m²h) was observed. This difference is because the rate of deposition of solutes on the membrane surface occurs at a faster rate as the pressure is increased, which reduces the net driving force of the solvent transport through the membrane as a result of the increase in osmotic pressure of the solution at the surface. Furthermore, the deposited layer becomes compact at higher pressures, which increases the resistance against the solvent.[12,15,20]

Effect of feed concentration on lignin retention and permeate flux

The effect of black liquor feed concentration on the extent of retention and flux are described in Figures 6 and 7. For this study, a 10-kDa membrane was used and the stirring rate was fixed at 300 rpm for all experiments. The retention of lignin in the stirred cell increased with an increase in feed concentration in the entire range tested, as shown in Figure 6. For instance, at an operating pressure of 350 kPa, an increase in feed concentration from 3% to 9% resulted in an increase in retention from 54.8% to 72.3%. For an operating pressure of 250 kPa, the extent of retention increased from 30.1% to 66.9% as the feed concentration was increased in the same range. This difference can be explained by the fact that as the ultrafiltration process proceeds, the viscosity of the solution near the membrane surface increases as a result of organic solutes that are rejected by the membrane.[12] This layer serves to sieve organic solutes. Because the viscosity of black liquor is dependent on its concentration, increasing the feed concentration of the solution results in an increase in the rate of formation of the viscous layer near the membrane surface as well as its viscosity. Black liquor components have a molecular weight (and size) distribution ranging from 100 Da to 100 000 Da; therefore, when a viscous layer is formed near the membrane surface, the retention of low molecular weight lignin fractions that would, under normal circumstances, permeate through the membrane are retained, thus increasing the retention.[15] The same trend was observed for other operating pressures tested but it has to be highlighted that, at a constant feed concentration, higher operating pressures resulted in higher retention values. For example, at a fixed feed concentration of 6%, increasing the pressure from 250 kPa to 350 kPa resulted in an increase in retention from 41.9% to 57.9%. This increase is because at a higher pressure the viscous layer formed at the membrane surface becomes more compact, which enhances its ability to sieve low molecular weight lignin fractions.[12,15]
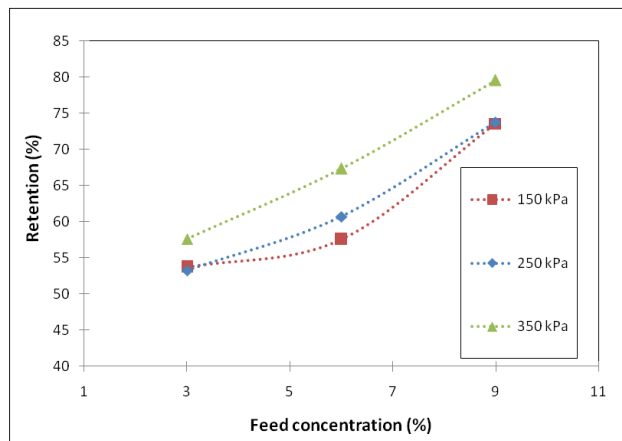
Figure 6: Variation in lignin retention with feed concentration at different pressures in a stirred cell (at a stirring speed of 200 rpm and a membrane size of 10 kDa).

The variation of permeate flux as a function of time for different feed concentrations is displayed in Figure 7. It was observed that the initial permeate flux, as shown in Figure 7, was comparatively higher for less concentrated solutions. The initial flux readings were 95.7 L/m²h, 62.9 L/m²h and 44.2 L/m²h for the 3%, 6% and 9% solutions, respectively. The variation with initial flux as a function of feed concentration is as a result of osmotic pressure resistance which is established within seconds of operation and this resistance is strongly dependent on feed concentration.[11] The reduction in flux with an increase in feed concentration is because the bulk concentration in the stirred cell increases as the solutes are rejected by the membrane, which in turn increases the extent of solute deposition near the membrane surface. The flux is reduced as the resistance to solvent transport through the membrane is increased in the polarised layer.[21] It is worth noting from Figure 7 that, after a filtration duration of 20 min, the permeate flux for the 6% and 9% solutions reached the same value and decreased at the same rate until the end of the filtration run. Also after 40 min, the same flux reading was recorded for all solutions tested, which indicates the same magnitude of total resistance to flow. For the 3% solution, the rate of depletion of solvent in the stirred cell was comparatively higher, which resulted in the accelerated build-up of a deposited layer, and hence the flux reached the same value as that of the 6% and 9% solutions. For the 6% and 9% solutions, the osmotic pressure was comparatively higher because of an increased concentration which retards the rapid depletion of the solvent from the stirred cell, and hence the flux decreased gradually with time.
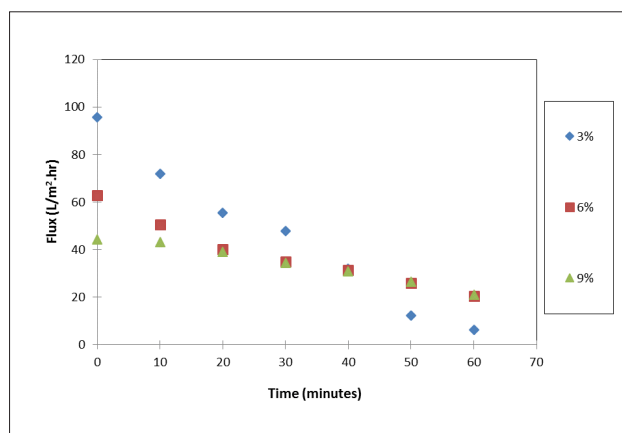
Effect of stirring rate on lignin retention and permeate flux

Figures 8 and 9 show the variation of lignin retention and permeate flux, respectively, with stirring rate in the stirred cell. In Figure 8 it is shown that as the stirring rate is increased, there is a corresponding increase in lignin retention in the entire concentration range investigated. As the stirring rate was increased from 200 rpm to 400 rpm, there was a corresponding increase in retention from 59.9% to 73.6% for the 3% solution. For the 6% and 9% solutions, the extent of retention increased from 61.5% to 79.1% and from 67.8% to 86.6%, respectively. A possible explanation for this observation is that at a low stirring speed, the rate of diffusion of solutes from the membrane surface to the bulk solution is comparatively lower as a result of the significant reduction in the influence of external forced convection that is induced by stirring. Because black liquor is a polydispersed solution, there is a possibility that lignin fractions with molecular weight sizes that are less than the cut-off size of the membrane permeate through the membrane, thus reducing the extent of retention. The backward diffusion of solute from the membrane surface to the bulk is enhanced as the stirring rate in the stirred cell is increased, and forced convection of solutes prevents them from being deposited on the membrane surface. This prevention in turn minimises the possibility of the low molecular weight lignin fractions from permeating through, thus increasing the extent of retention.

The variation of flux as a function of time at different stirring rates is represented by Figure 9. This study was conducted with a solution of 3% and operating pressure of 150 kPa. As the stirring rate increased from 200 rpm to 300 rpm, as shown in Figure 9, the initial permeate flux increased from 55.2 L/m²h to 81.3 L/m²h. This increase results because as the rate of backward diffusion of solutes from the membrane surface to the bulk is enhanced by an increase in stirring rate, there is less build-up of osmotic pressure and the greater effective pressure difference results in an increase in flux.[15] It is worth noting that in the 20–40 min time interval, the fluxes for 300 rpm and 400 rpm were approximately equal to each other, indicating that the extent of minimisation of osmotic pressure at the membrane surface was approximately the same as the stirring rate was increased above 300 rpm. This means the resistance to permeation of the solvent was approximately the same in that specific time interval. The rapid decline in permeate flux for stirring rates of 300 rpm and 400 rpm was as a result of an increase in the bulk solution viscosity which increased the osmotic pressure (the extent of the increase is minimised by stirring) as the solvent permeated through the membrane at a faster rate than at 200 rpm. This supposition is substantiated by the significant deviation in flux values for 300 rpm and 400 rpm in the 50–60 min time interval. This deviation indicates that the resistance to solvent flow at a stirring rate of 300 rpm increases as a result of the slower back diffusion of solute from the membrane surface as the viscosity and concentration of the solution increases. At 400 rpm, this effect is offset by the faster rate of back diffusion of solute from the surface, which minimises the increase in osmotic pressure/resistance.
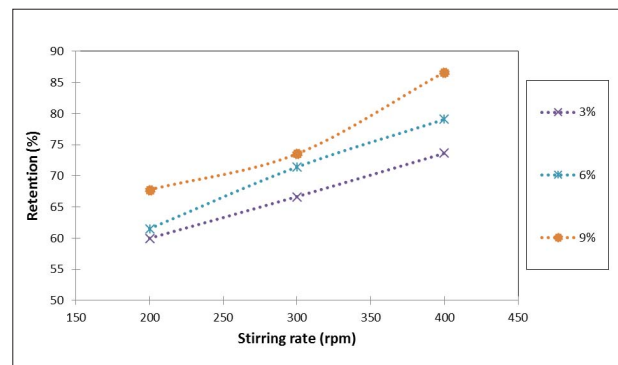

Figure 7: Variation in permeate flux as a function of time at different feed concentrations in a stirred cell (at a pressure of 250 kPa and a stirring speed of 300 rpm, with a membrane size of 10 kDa).
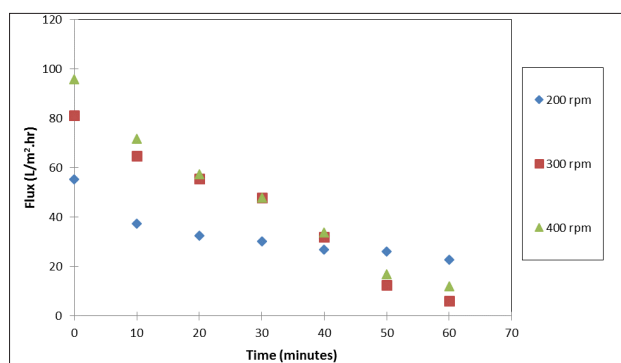

Figure 8: Variation of lignin retention with stirring rate at different feed concentrations in a stirred cell (at a pressure of 150 kPa with a membrane size of 10 kDa).

Figure 9: Variation of permeate flux as a function of time at different stirring rates in a stirred cell (at a pressure of 150 kPa and a feed concentration of 3% with a membrane size of 10 kDa).

Solids content analysis

Table 4 represents the ash content analysis results of the fractions collected at the end of the ultrafiltration runs. An optimum membrane cut-off size should result in minimal retention of cooking chemicals (i.e. sodium and sulfur). In addition to the fact that these chemicals have to be reused in the cooking process, it is necessary to avoid forming ash and sulfur dioxide when these inorganic elements are combusted in a normal furnace.[22] The increase in ash content in the retentate fractions is as a result of the retention of multivalent ions that are associated with the retained organic matter. For example, for the 5-kDa membrane and an operating transmembrane pressure of 150 kPa, as observed in Table 4, the ash content increased from 5.8% (raw) to 7.1% (retentate stream). The same trend was observed for all different combinations of membrane cut-off size and process conditions (data not shown). This result was also observed by Wallberg et al.[22] when they used polymeric and ceramic membranes to fractionate and concentrate black liquor. The ash that is observed in the permeate streams is attributed to monovalent ions as their retention is insignificant.[22]

Table 5 shows the solids content analysis results of the fractions collected at the end of the ultrafiltration runs. The increase in solids content in the retentate streams as the pressure is increased, with the other variables (membrane cut-off size, stirring speed, feed concentration) maintained at constant values, is attributed to the increase in driving force which accelerated the rate of solvent permeation out of the stirred cell, resulting in the increased concentration of the retentate stream.[15] For

example, for the 10-kDa membrane, as recorded in Table 5, the solids content in the retentate stream increased from 11.2% to 15.4% as the pressure was increased from 150 kPa to 350 kPa. The same trend (i.e. the concentration of solids in the retentate stream) was observed for all different combinations of membrane cut-off size and process conditions (data not shown).

Conclusions

A bench-scale study was conducted to investigate the feasibility of extracting lignin, by means of ultrafiltration, from black liquor generated using hardwoods sourced from the southern African region. It was found that amongst the membranes investigated, the highest retention of lignin was achieved using the 5-kDa membrane. Although the 5-kDa size membrane had the highest retention, its flux was too low for a comparative study. Therefore, to investigate the effect of operating parameters on both the flux and retention, a 10-kDa size membrane was used. The parameters investigated (operating pressure, feed concentration and stirring rate) had significant effects on the retention of lignin and flux and it would be worthwhile to investigate them further on a pilot plant scale. Membrane fouling was observed to have a profound negative effect on flux values in the study and the mechanisms that contributed to this were combinations of osmotic pressure build up and the development of a viscous gel layer on the membrane during operation. This study should be augmented by studying the ultrafiltration of lignin in a pilot plant in a continuous mode to (1) overcome the effect of fouling associated with batch filtration and (2) to generate a set of data that can be applicable to ultrafiltration of black liquor on an industrial scale.

Acknowledgements

Authors' contributions

This work forms part of P.K.'s MSc dissertation, thus he was responsible for performing the experiments, for laboratory analysis and for writing the manuscript. B.S. is the corresponding author and project leader; he supervised and reviewed the work and gave comments on the manuscript. D.R. acted as co-supervisor during the MSc study, assisted with the experimental design and analysis of results and provided comments on the manuscript.

Table 4: Ash content analysis (%) of ultrafiltration fractions from the membrane selection study

Pressure (kPa)	Membrane cut-off size (kDa)					
	5		10		20	
	Permeate	Retentate	Permeate	Retentate	Permeate	Retentate
150	4.14	7.09	4.61	8.85	5.5	9.49
250	4.27	8.83	4.7	8.7	4.52	11.22
350	3.9	8.75	4.61	10.46	4.82	12.07

Table 5: Solids content analysis (%) of ultrafiltration fractions from the membrane selection study

Pressure (kPa)	Membrane cut-off size (kDa)					
	5		10		20	
	Permeate	Retentate	Permeate	Retentate	Permeate	Retentate
150	5.55	9.54	5.37	11.18	6.55	13.78
250	4.73	12.72	5.45	12.91	6.17	16.12
350	4.94	14.54	5.6	15.38	6.21	17.72

References

1. Wallberg O, Linde M, Jönsson A. Extraction of lignin and hemicelluloses from kraft black liquor. Desalination. 2006;199:413–414. http://dx.doi.org/10.1016/j.desal.2006.03.094

2. Dafinov A, Font J, Garcia-Valls R. Processing of black liquors by UF/NF ceramic membranes. Desalination. 2005;173(1):83–90. http://dx.doi.org/10.1016/j.desal.2004.07.044

3. Moshkelani M, Marinova M, Perrier M, Paris J. The forest biorefinery and its implementation in the pulp and paper industry: Energy overview. Appl Therm Eng. 2013;50(2):1427–1436. http://dx.doi.org/10.1016/j.applthermaleng.2011.12.038

4. Toledano A, García A, Mondragon I, Labidi J. Lignin separation and fractionation by ultrafiltration. Sep Purif Technol. 2010;71(1):38–43. http://dx.doi.org/10.1016/j.seppur.2009.10.024

5. Chatel G, Rogers RD. Review: Oxidation of lignin using ionic liquids – An innovative strategy to produce renewable chemicals. ACS Sustainable Chem Eng. 2014;2(3):322–339. http://dx.doi.org/10.1021/sc4004086

6. Jönsson A, Wallberg O. Cost estimates of kraft lignin recovery by ultrafiltration. Desalination. 2009;237(1–3):254–267. http://dx.doi.org/10.1016/j.desal.2007.11.061

7. Banerjee S, De S. An analytical solution of Sherwood number in a stirred continuous cell during steady state ultrafiltration. J Membr Sci. 2012;389:188–196. http://dx.doi.org/10.1016/j.memsci.2011.10.029

8. Alekhina M, Ershova O, Ebert A, Heikkinen S, Sixta H. Softwood kraft lignin for value-added applications: Fractionation and structural characterization. Ind Crop Prod. 2015;66:220–228. http://dx.doi.org/10.1016/j.indcrop.2014.12.021

9. Jin W, Tolba R, Wen J, Li K, Chen A. Efficient extraction of lignin from black liquor via a novel membrane-assisted electrochemical approach. Electrochim Acta. 2013;107:611–618. http://dx.doi.org/10.1016/j.electacta.2013.06.031

10. Velez J, Thies MC. Solvated liquid-lignin fractions from a kraft black liquor. Bioresour Technol. 2013;148:586–590. http://dx.doi.org/10.1016/j.biortech.2013.08.097

11. Doherty WOS, Mousavioun P, Fellows CM. Value-adding to cellulosic ethanol: Lignin polymers. Ind Crop Prod. 2011;33(2):259–276. http://dx.doi.org/10.1016/j.indcrop.2010.10.022

12. De S, Bhattacharya PK. Flux prediction of black liquor in cross flow ultrafiltration using low and high rejecting membranes. J Membr Sci. 1996;109(1):109–123. http://dx.doi.org/10.1016/0376-7388(95)00188-3

13. Bhattacharjee C, Sarkar P, Datta S, Gupta BB, Bhattacharya PK. Parameter estimation and performance study during ultrafiltration of kraft black liquor. Sep Purif Technol. 2006;51(3):247–257. http://dx.doi.org/10.1016/j.seppur.2006.02.005

14. Jönsson A, Nordin A, Wallberg O. Concentration and purification of lignin in hardwood kraft pulping liquor by ultrafiltration and nanofiltration. Chem Eng Res Des. 2008;86(11):1271–1280. http://dx.doi.org/10.1016/j.cherd.2008.06.003

15. Satyanarayana SV, Bhattacharya PK, De S. Flux decline during ultrafiltration of kraft black liquor using different flow modules: A comparative study. Sep Purif Technol. 2000;20(2–3):155–167. http://dx.doi.org/10.1016/S1383-5866(00)00086-1

16. Bhattacharjee C, Bhattacharya PK. Flux decline analysis in ultrafiltration of kraft black liquor. J Membr Sci. 1993;82:1–14. http://dx.doi.org/10.1016/0376-7388(93)85088-E

17. Bhattacharjee C, Bhattacharya PK. Ultrafiltration of black liquor using rotating disk membrane module. Sep Purif Technol. 2006;49(3):281–290. http://dx.doi.org/10.1016/j.seppur.2005.10.011

18. Schwarze M, Le DK, Wille S, Drews A, Arlt W, Schomäcker R. Stirred cell ultrafiltration of aqueous micellar TX-100 solutions. Sep Purif Technol. 2010;74(1):21–27. http://dx.doi.org/10.1016/j.seppur.2010.05.003

19. Britz TJ, Robinson RK, editors. Advanced dairy science and technology. Oxford: Blackwell; 2008. http://dx.doi.org/10.1002/9780470697634

20. Noordman TR, Ketelaar TH, Donkers F, Wesselingh JA. Concentration and desalination of protein solutions by ultrafiltration. Chem Eng Sci. 2008;57(4):693–703. http://dx.doi.org/10.1016/S0009-2509(01)00403-1

21. Bhattacharjee C, Datta S. A mass transfer model for the prediction of rejection and flux during ultrafiltration of PEG-6000. J Membr Sci. 1997;125(2):303–310. http://dx.doi.org/10.1016/S0376-7388(96)00247-5

22. Wallberg O, Jönsson A, Wimmerstedt R. Fractionation and concentration of kraft black liquor lignin with ultrafiltration. Desalination. 2003;154:187–199. http://dx.doi.org/10.1016/S0011-9164(03)80019-X

Impact of mother tongue on construction of notes and first-year academic performance

AUTHORS:
Shalini Dukhan[1] (iD)
Ann Cameron[2]
Elisabeth Brenner[3]

AFFILIATIONS:
[1]School of Animal, Plant and Environmental Sciences, University of the Witwatersrand, Johannesburg, South Africa

[2]Science Teaching and Learning Centre, University of the Witwatersrand, Johannesburg, South Africa

[3]School of Molecular and Cell Biology, University of the Witwatersrand, Johannesburg, South Africa

CORRESPONDENCE TO:
Shalini Dukhan

EMAIL:
Shalini.dukhan2@wits.ac.za

The purpose of this study was to identify whether there are any differences in the quality of the notes constructed in English between students for whom English is a first language and those for whom it is a second language. Subsequently we assessed whether this difference, if any, affected their grades. Unsurprisingly, the first-language students produced better structured and more detailed notes; they also performed better academically than their second-language peers. However, when students were provided with training that focused on using writing as a means to promote critical thinking, there was an improvement in the personalisation of their notes. The improvement in grades was significant for second-language students. Thus the university has a pivotal role to play in preparing students for academic success by providing them with supportive measures to aid their transition into first year.

Significance:

- The work illustrates that writing can be used as a tool for students to improve their learning and their academic performance.

- Second-language students' grades improve when writing interventions are provided early in the year.

- Students need to take on the responsibility for their learning; lecturers also have a responsibility in scaffolding learning.

Introduction

In 1994, educational opportunities were opened to previously disadvantaged populations, which resulted in increasing numbers of black students gaining entry into universities.[1] Teachers at previously 'black' schools often teach by means of code-switching.[2,3] (Code-switching refers to the alternation between learners' home language and English in order to make complex concepts more understandable to learners.) When students from these schools enter a university at which the medium of instruction is English, they may not have developed the English competency which is required and which is fundamental to their academic success. In this study, we focused on first- and second-language students' experiences of note-taking and note-making at school and in their first year of university, and the impact of their writing strategies on their grades at an English-medium university.

There are a number of compounding factors which influence second-language students' proficiency in English, and thus their academic success. Firstly, when students enter the university academic environment they navigate their approach to learning and their student identity, i.e. in terms of their cognitive and linguistic experiences.[4] This means that undergraduates navigate the ways in which they are 'behaving, interacting, valuing, thinking, believing, speaking…reading and writing'[4(p.111)]. Thus, as they navigate their identity they begin to identify with the discourse and develop their identities within the discourse. Schools in South Africa are mostly teacher driven, and rarely provide the opportunity for students to develop an identity within the discourse.[4]

Secondly, any underperformance of second-language undergraduates is usually attributed to limited language proficiency, but underdeveloped reading skill also plays a part. While Weideman and van Rensburg[5] support the argument that language proficiency is related to academic success, Pretorius[6-8] has shown that even though there is a relationship with language, reading proficiency has a stronger connection to the students' grades. Reading ability is based on decoding and comprehension of texts. At primary school the focus is on decoding.[6] As students progress through their schooling it is generally assumed that they will independently learn to deconstruct and make meaning of texts. The scaffolding of relational thinking receives little attention at school, with little done to assist students to move towards deep levels of comprehension.[6]

Previous studies[5,7] have shown that students who develop fluent reading ability in their mother tongue are able to transfer their reading skills when they learn in a second language. Therefore, emphasis should lie in the development of reading ability in the mother tongue. Within the South African context, however, English is seen as a language of status compared to African languages[5], therefore parents usually desire their children to learn English early in their schooling[3,7]. The level of skill in reading ability is associated with the students' skills in expressing themselves in their writing.[8,9] Thus the development in relational thought during the process of learning to read is believed to be related to the level of skill in writing. Most academic assessment at university is based on a student's ability to communicate their understanding via their writing; it is therefore critical to place a spotlight on the students' writing skills.

Thirdly, research has been published on whether the conception of learning that students develop is cognitively appropriate for the university academic landscape. Because of their prior schooling experience,[5] second-language English students usually have not developed the necessary level of skill in their reading[6,7] and writing abilities[8,10]. Reading and writing abilities hinge on the development of inferential processing (i.e. the ability to link ideas and concepts in a text to other areas of the text or to relevant concepts in one's memory)[11] – this phenomenon is also known as anaphoric resolution[12]. Students who can resolve anaphoric ties satisfactorily have a better comprehension of their texts, and therefore are likely to perform better on assessments. The level of student

engagement with anaphoric resolution is likely to be evident in the notes that they make. Undergraduates who paraphrase, and whose notes contain accurate content not provided by the lecturer, are more likely to consider links and connections with different content areas in the text, and are more likely to relate their new knowledge to their current knowledge structure. Poor inferential reading can be evidenced in poor inferential writing; thus a student's writing can give clues to their cognitive level of engagement. More importantly, the process of writing can assist students identify gaps in their understanding.

Carstens and Fletcher[10] noted that writing interventions which focus on the development of understanding and handling of content, structure and style, and on language development, can enrich learning experiences and students' grades. We hypothesise that – as writing is an individual activity which students are expected to master before they enter university, and is central to university assessment – by focusing on the development of writing abilities students would be able to self-identify gaps in their learning. The process of compiling study notes could provide the mechanism for students to develop a deep approach to learning, and thus improve their grades. Training disadvantaged students in the skill of writing has been shown to improve their academic literacy[13], but there is a dearth of literature on whether a focus on the development of the students' skill in constructing their notes is likely to influence first-year students' writing abilities, *and* the depth of their study notes, *as well as* their academic performance.

Undergraduates need to comprehend a lecture taught in English, including making sense of the terminology of the discourse and being able to capture sufficient content in their notes.[14] The students' understanding of the vocabulary within science is important because it influences their comprehension of the material presented during fast-paced lectures. Biology is particularly rich in specialist terminology, so the issue of language support is important in this context. Students who have not had sufficient experience using English to converse may require additional guidance to be able to understand scientific terms.[15]

When students have difficulty with the vocabulary of the discipline, the quality of the notes they capture in the lecture will be affected. Most students take notes to aid lecture recall, to improve their understanding, and to capture information that is required for assessments.[16] But the cognitive overload that second-language students face in lectures could compromise the opportunity for learning. Concerns about the limited support that second-language students have received in developing academic language skills provided the motivation for investigating the impact of language on note-making and hence on academic performance. Because the literature has shown that the quality of students' notes is linked to grades,[17] the question that guided this study was whether the practice of writing could be used to enrich the second-language students' learning experience and improve their grades.

The quality of notes that students capture in class (note-taking) affects the way in which they engage with and revise their notes after class (note-making). The note-making phase is more closely associated with generative learning than the note-taking phase.[18] Generative learning refers to the students' creation of links between prior knowledge and new knowledge.[19,20] This means that during note revision, if second-language students spend more effort in understanding the language in which the concepts are taught than on enhancing their depth of understanding, the quality of their notes will be compromised and they will miss generative learning opportunities. The lack of generative learning will then impact on the grades that students achieve.

It was surmised that by analysing and comparing the notes prepared by first- and second-language students, in comparison to the content lecturers provide, insight should be gained regarding the differences in the ways in which these students approached their note-making and learning. The results were then viewed in relation to the grades that students received in tests and examinations in their first year.

Methods

This study forms part of a larger one which focused on the relationship between the notes that first-year students constructed and their grades.

The study was conducted in a biology course at a South African university, and extended from Semester 2 in 2009 to Semester 1 in 2011. Two lectures per semester were identified for this research. A 45-minute lecture was observed and video-recorded for each of four lecturers per year. The video-recorded lectures were transcribed for analysis and compared to data collected from students. In each year of the study, between 43% and 60% of the participants were second-language students.

Students voluntarily completed the questionnaires. The return rate of completed questionnaires for each year was about 25%. The questionnaire probed note-taking and note-making experience at school and university. The students' responses to the questionnaire were analysed according to:

- prior schooling experience in constructing notes
- use of notes for assessment preparation at school
- their expectations of university lecturers
- additional information provided on the lecturers' slides
- how students captured notes
- the reasons for note-taking and note-making

In each year, 30 students per year (i.e. 90 students in total) were randomly selected from those who completed the questionnaire. These students participated in one-on-one, face-to-face interviews. The first interview was conducted 2 months into Semester 1, the second interview 1 month into Semester 2, and the third at the beginning of the following year. Interviews were used to gather insight on how students' views, expectations and experiences of their note-taking and note-making practices changed as they gained experience at university.

The notes which participating students used for study purposes were photocopied, and analysed in terms of the quantity and quality of content. A total of 240 sets of notes (from eight lectures for 30 students each) was analysed. A rubric developed by us (Appendix 1 in the supplementary material) was the starting point for the qualitative analysis of students' notes; this rubric was based on SOLO taxonomy. SOLO taxonomy is used to evaluate a student's cognitive ability on certain tasks, and defines different cognitive levels and knowledge dimensions.[21] Cognitive abilities are defined at the structural, conceptual and procedural level, and an increase in mental abilities is observed as one moves up each of these cognitive levels. Other researchers (e.g. Reed et al.[22] and Granville and Dison[23]) have modified SOLO taxonomy so that it could be used within a disciplinary context or generically. Our study outlines how SOLO taxonomy can be used in nuanced ways. In this study, the application of the different cognitive levels can be gauged by analysing students' notes. The more links to prior knowledge and greater incorporation of the students' input, the higher the cognitive level of the student in that piece of work. The qualitative results from the analysis using the rubric are referred to as the notes score.

Students' notes were also analysed quantitatively according to a comprehensive set of notes made by one of the authors (S.D.). The number of information units provided in students' notes was calculated. Information units are blocks of information or whole ideas, and comprise a sentence, clause or stand-alone phrase.[24] Numerous other authors[24-27] have used the measure of information units as a method of analysis of notes. By looking at the number of information units in the students' notes, a comparison was made of the amount of content students noted from the lecturers' verbal explanations and from the visuals provided in class. Additional information that students captured from other resources was also considered. Two peers in the education field independently repeated this analysis, and evaluated 10% of the notes, with more than 95% agreement in terms of analysis.

Students' grades were obtained from departmental records. Data were quantitatively analysed using a single-factor analysis of variance and KyPlot version 2.0.

Ethics clearance was granted by the university's Human Research Ethics Committee (HREC Non-Medical Protocol number: 2009ECE114). Informed consent was obtained from all participants, and the principles of the Declaration of Helsinki were adhered to.

Results and discussion

Undergraduates at the university at which this study was conducted had experienced some communication in English at school. In this study, first language refers to students whose mother tongue is English, while second language refers to students whose mother tongue is not English.

Data obtained from the interviews support previous findings (e.g. Rollnick[2] and Kapp[3]) that code-switching continues into secondary school. A comment from a second-language first-year student was that because code-switching was not practised at university, language posed a barrier to her understanding of the content presented during lectures:

> *I used to go and ask my teacher; she used to explain in my language; but here* [at university] *everything is done in English; I'm far from home; there's no one to ask.*

A total of 10% of the students commented that their level of competence in English affected the depth of meaning that they were able to construct, and thus affected their quality of notes. Students with low levels of vocabulary associated with the language of instruction and with the discourse struggle with anaphoric resolution.[12,13] Moreover, when second-language students do not have opportunities to practise anaphoric resolution, there is an impact on the depth of understanding that they are likely to access and thus the time that it takes for them to form a deep understanding of lecture content.[13]

Two international studies have reported that South African learners are ranked amongst the lowest in international standards in reading, and that the language in which learners were taught at school and the language of assessment most likely contributes to the students' poor academic performance.[28,29] As students progress through school, it is essential that they learn to formulate a cohesive representation of their texts[8]; thus the focus should lie in the development of language proficiency[3,6,7,30] and reading ability.[5-7] Because the expectation is that students would have developed the required level of skill at school, lecturers at university do not generally focus on developing reading skills.[3,9,13,31]

School experience influences transition into first year

Compared with first-language students, on average 11% more of the second-language students reported that, unlike the situation at university, they had been provided with comprehensive notes from their teachers at school (Table 1). These students reported that they usually did not engage further with the material other than to memorise the content. This approach results in surface learning because students are not compelled to resolve anaphoric ties within topics, and to form a global understanding of topics.

Table 1: First- and second-language students' note-making experiences at school

Year	First-language English students		Second-language English students	
	Received notes from the teacher	Some experience in note-making	Received notes from the teacher	Some experience in note-making
2009	5	7	11	7
2010	4	13	5	8
2011	5	9	8	8
Combined	14	29	24	23

This practice on the part of the teacher results in a lack of generative learning opportunities on the part of the students. Teachers who provide comprehensive notes are taking on the responsibility for student learning because they are, in essence, providing material which the learners rote learn. Students are not then encouraged to engage with their work at a deeper cognitive level. However, when teachers scaffold inferential processing activities in their classroom then students are more likely to engage in higher cognitive processes and in constructing meaning.[11] This also means that practice in reading and writing would enable more efficient anaphoric resolution.[12]

Mother tongue influences the quality of students' notes

Data from this study show that when students entered university, those who were accustomed to receiving notes from the teacher at school used the lecture slides as study notes, because their previous school experience had led them to conclude that the slides provided all the information that they were required to know. But at university, lecturers usually use slides as an aid to their teaching, and to provide key points so that students can use them to focus and extend their learning after class.

Based on data from the analysis of the video-recorded lectures, the detail that lecturers provided verbally (i.e. material that was not on the slides) ranged from 8% to 76% of the total information provided during the lecture. This proportion makes it essential that students should know how to use the lecturers' slides to enhance their learning. However, from the interviews it became clear that second-language students did not seem to understand that the lecture slides were a base on which they needed to expand. They stated that they struggled with copying content from the slides and noting material from the lecturers' verbal explanations while also decoding biology and English vocabulary, which impacted on the quality of their notes in class and the depth of the notes they made after class. These findings support Wildsmith's[30] assertion that, although they are able to read and understand, students struggle with writing in the formal academic context. It is likely that because students' learning experiences have not usually involved higher cognitive processing (i.e. relational thinking[23]), they do not have the know-how or inclination to formulate extended notes at university.

In Semester 1, the second-language students captured half the quantity of notes from the verbal and visual aspects of the lectures that their peers did (Table 2). The first-language students also noted more content from resources other than that provided in the lecture (Table 2). And the notes that second-language students made were not as coherent in terms of the development of ideas and organisational structure, as reflected by the notes score (Table 2). Although these observations were made when quantitatively and qualitatively analysing the students' notes (Tables 2 and 3), there were no statistically significant differences.

In Semester 2, the situation changed: second-language students captured more facts for Lecturer 3's sections in 2009 and 2010 than their first-language peers (Table 3). Lecturer 3 provided keywords, and probably because of their comprehension of English, the first-language students did not capture as much content because they better understood the concepts that the lecturer explained. In contrast, second-language students were more likely to feel compelled to capture in entirety what the lecturer was saying. For Lecturer 4's section, when students had to note key points (as opposed to keywords only) from the slides, second-language students noted fewer points than did first-language students (Table 3). Lecturer 4 based the lecture content on personal research in the field and lectured in a conversational manner; the second-language students may have had difficulty in deciphering what content needed to be noted and therefore captured less material than their first-language peers.

By Semester 2, second-language students had likely realised the importance of the lecturers' verbal explanations in relation to the content that was tested, and therefore knew that the material provided on the lecturers' slides was not sufficient for their studying. They learned that they needed to listen to – and understand – the lecturers' verbal explanations, as shown in a comment from a second-language student:

For the first semester it was difficult; I had a problem in taking down notes, so now [Semester 2] I've realised that what the lecturer is saying is more important than going through the notes [slides], they would go through the slides but they add a little bit and I've realised that those are the things that they ask you in the test.

However, first-language students were aware, from the beginning of the year, that the verbal aspect of the lectures contained the detailed material that would be examined, and consequently acted to capture some of this information in their notes. In 2010 and 2011, the first-language students noted more content from the verbal lecture than did the second-language students, and made more coherent notes, as shown by the notes score (Table 2 and 3). Storch[14] reported that students should write more coherent and personalised notes as they gain experience and a better grasp of English within university; our data show that language of instruction can impact on the quality and quantity of the notes that are made.

Compared with the second-language students, the first-language students noted almost twice the amount of information from the textbook (Table 2). During interviews it emerged that second-language students lacked experience in using textbooks at school, and therefore found it difficult to know how to deepen their learning and class notes using the textbook at university. Students are expected to read their texts, class notes and library books so that they can extend the knowledge which is provided in the lecture. But students are not reading at levels that show mastery.[5-7,12] Students are able to decode but are not able to comprehend texts at the required levels.[5,6] This inability stems from issues around literacy prior to the students entering tertiary education. These issues include limited access to books at schools, lack of encouragement to read for pleasure, and limited printed resources in African languages.[6]

Using textbooks develops the students' ability to engage with material in a manner that allows for the construction of cognitive learning.[32] But a major barrier to learning science is the grammar of science.[33] Not having a good command of scientific vocabulary not only influences a student's ability to read and write in a scientific manner but also their ability to understand the scientific content.[34] The range of vocabulary that the students had developed prior to their first year at university influenced how much they understood when they read lecture slides and their textbooks:

Most of the time if I'm using the textbook, I write it in the way the textbook gives it but then I would make myself understand it better, [because] I realised that there was this vocabulary gap, so to familiarise myself with the vocabulary I have to somehow incorporate the vocabulary the textbook uses, which is university vocabulary, into my work because if I don't then I will be stuck in an exam just not knowing what they are asking.

Students who have limited vocabulary struggle with anaphoric resolution, and this impacts on their ability to paraphrase.[13] Students need to gain experience within the academic environment of the university in order to understand the level of self-regulation that is required, and this understanding enables them to have realistic expectations of their role in first year. Additionally, to improve academic performance, there should be adequate exposure to written texts at school.[13] As the schooling system currently stands, the majority of students has not been groomed to use writing to critique texts or as a vehicle to explore their understanding, yet at university they are expected to have developed the required skill to use writing as a way of communication *and* to evidence higher-order cognitive skills.[3] Our findings show that as much as students need to take charge of their learning, the

Table 2: Comparison of Semester 1 note-making averages between first- and second-language students

| | L1 2010 | | L2 2010 | | L1 2011 | | L2 2011 | | Combined | | | |
| | | | | | | | | | L1 | | L2 | |
	E1	E2	E1	E2	E1	E2	E1	E2	E1	E2	E1	E2
Number of facts	69	36	87	44	52	43	72	36	62	41	80	42
Number of verbal aspects	66	28	84	39	40	39	70	38	55	35	78	38
Number of visual aspects (content on slide/board/overheads)	59	28	82	39	40	39	69	39	51	34	76	38
Additional information from other resources (e.g. textbook)	17	7	5	1	11	6	3	2	18	7	5	1
Notes score	18	15	21	14	19	16	21	17	19	15	21	14

L, lecturer; E1, English first-language students; E2, English second-language students

Table 3: Comparison of Semester 2 note-making averages between first- and second-language students

| | L3 2009 | | L4 2009 | | L3 2010 | | L4 2010 | | Combined | | | |
| | | | | | | | | | L3 | | L4 | |
	E1	E2	E1	E2	E1	E2	E1	E2	E1	E2	E1	E2
Number of facts	83	90	33	24	139	144	64	39	115	112	35	18
Number of verbal aspects	83	90	31	22	127	140	57	39	69	57	31	17
Number of visual aspects (content on slide/board/overheads)	83	90	31	22	133	140	62	39	111	110	47	30
Additional information from other resources (e.g. textbook)	1	0	2	0	6	4	1	0	6	4	1	0
Notes score	10	9	9	9	16	12	17	9	16	12	17	9

L, lecturer; E1, English first-language students; E2, English second-language students

university also has an obligation to support these students become academically successful.

A series of workshops that focused on developing the students' abilities to critically read, write and develop an argument[34] was offered to all first-year students in 2010 and 2011. These workshops aimed to show students how to make meaning of content when they constructed notes. When second-language students saw the value of using their notes as a vehicle for the process of learning and knowledge transformation, their academic performance improved (Table 4).

Language affects the quality of notes and academic performance

There was a narrower gap between the academic performance of first- and second-language students in 2010 and 2011 compared with those in 2009. Compared with their second-language peers in 2009, the first-language students performed better on all assessments. However, in 2010, the gap in performance between the groups decreased. There was a further decrease in this gap in 2011, and in some assessments the second-language students performed slightly better, on average, than the first-language students (Table 4).

Table 4: Assessment averages (%) for first- and second-language students

Assessment	2009		2010		2011		Combined	
	E1	E2	E1	E2	E1	E2	E1	E2
Test 1	52	41	68	66	57	58	68	66
Test 2	73	65	50	44	49	52	51	44
Exam 1	50	39	60	56	56	52	60	56
Exam 2	45	35	48	39	45	48	50	39
Test 3	52	44	49	44	47	44	50	44
Test 4	62	52	59	55	40	36	60	55
Exam 3	59	43	50	47	44	45	50	47
Exam 4	59	52	62	53	48	48	64	53

While it is possible that the university changed their selection criteria and accepted more academically capable students in 2010 and 2011 than those in 2009 – or that there may have been changes in the teaching and learning environment at school which facilitated the second-language students to develop a higher level of skills relevant to the university academic environment – it is also likely that the improved performance of the first-year second-language students in 2010 and 2011 was a result of the implementation of the workshops. The workshops enabled students to see that the lecturers' slides were not the product that needed to be learnt, but rather that the slides represented a guideline. The students became aware that they needed to read further and personalise their notes with more details. This is evident in the notes scores that the 2010 and 2011 cohorts received compared with the scores of the 2009 cohort (Tables 2 and 3). This finding corroborates that of Pretorius'[12] that there is a strong link between an increase in resolution of anaphoric references and proficiency in English. Moreover, in agreement with Carstens and Fletcher[9,10], writing interventions can improve academic performance. This study provides some insight into how the use of writing can scaffold second-language learning (and enhance academic performance) within an English-medium university.

During the transformation of notes there was the likelihood of a trans-formation in the students' knowledge and understanding as well, which then resulted in better comprehension of lecture content. This is because when students revise and personalise their notes they engage more deeply in determining how their understanding of content aligns with what is presented on the lecture slides, i.e. the content that the lecturer expects students to comprehend. Thus, during the process of revising their notes, students identify gaps in their knowledge and can then seek information from other resources, such as textbooks, to better their understanding. If students are taught to critically analyse texts and arguments, then the process of understanding and critiquing the scientific discourse will be more accessible to them.[33]

Conclusion

First- and second-language students' experience of, and competence in, English influenced the depth of meaning they were able to achieve in class and when reading textbooks; this influence had an impact on the quality of notes that they constructed. The grades that students achieved on assessments were found to be related to the medium of instruction. However, when the students were provided with workshops that focused on writing as a means to promote critical thinking, and thus when note-making was used as a means to deepen comprehension, there was a general improvement in the standard of notes that second-language students made, and hence in their grades.

By being cognisant of the challenges that second-language students face in both English and biology, lecturers could scaffold their lectures to be more inclusive of these students. It is essential that training be given at the beginning of the year if second-language students are going to gain epistemological access to subjects like biology when they enter university.

Authors' contributions

S.D. conceptualised and designed the project as part of her PhD research; she collected and analysed the data and wrote the manuscript. A.C. and E.B. were supervisors and provided guidance on the study design, data collection, data analysis and write-up.

References

1. Fiske EB, Ladd HF. Elusive equity: Education reform in post-apartheid South Africa. Cape Town: HSRC Press; 2004.

2. Rollnick M. Current issues and perspectives on second language learning in science. Stud Sci Educ. 2000;35:93–122. http://dx.doi.org/10.1080/03057260008560156

3. Kapp R. Students' negotiations of English and literacy in a time of social change. Econ Writ. 2012:591–614.

4. Kapp R, Bangeni B. 'I was just never exposed to this argument thing': Using a genre approach to teach academic writing to ESL students in the Humanities. In: Herrington A, Moran C, editors. Genre across the curriculum. Logan, UT: Utah State University Press; 2005. p. 109–127.

5. Weideman A, Van Rensburg C. Language proficiency: Current strategies, future remedies. J Lang Teaching. 2002;36:152–164. http://dx.doi.org/10.4314/jlt.v36i1-2.6010

6. Pretorius EJ. Reading ability and academic performance in South Africa: Are we fiddling while Rome is burning? Lang Matters. 2002;33:169–196. http://dx.doi.org/10.1080/10228190208566183

7. Pretorius EJ, Mampuri DM. Playing football without a ball: Langauge, reading and academic performance in a high-poverty school. J Res Reading. 2007;30:38–58. http://dx.doi.org/10.1111/j.1467-9817.2006.00333.x

8. Pretorius EJ. The comprehension of logical relations in expository texts by students who study through the medium of ESL. System. 2006;34:432–450. http://dx.doi.org/10.1016/j.system.2006.02.003

9. Carstens A. Generic versus discipline-specific writing interventions: Report on a quasi-experiment. South Afr Linguist Appl Lang Stud. 2011;29:149–165. http://dx.doi.org/10.2989/16073614.2011.633363

10. Carstens A, Fletcher L. Quantitative evaluation of a subject-specific essay-writing intervention. South Afr Linguist Appl Lang Stud. 2009;27:319–332. http://dx.doi.org/10.2989/SALALS.2009.27.3.9.943

11. Pretorius EJ. Reading and the Unisa student: Is academic performance related to reading ability? Progressio. 2000;2:35–48.

12. Pretorius EJ. English as a second language learner differences in anaphoric resolution: Reading to learn in the academic context. Appl Psycholinguist. 2005;26:521–539. http://dx.doi.org/10.1017/S0142716405050289

13. Wildsmith-Cromarty R, Steinke K. The write approach: Can R2L help at tertiary level? Per Linguam. 2014;30:38–54. http://dx.doi.org/10.5785/30-1-570

14. Storch N. The impact of studying in a second language (L2) medium university on the development of L2 writing. J Second Lang Writ. 2009;18:103–118. http://dx.doi.org/10.1016/j.jslw.2009.02.003

15. Parkinson J. Acquiring scientific literacy through content and genre: A theme-based language course for science students. Engl Specif Purp. 2000;19:369–387. http://dx.doi.org/10.1016/S0889-4906(99)00012-5

16. Badger R, White G, Sutherland P, Haggis T. Note perfect: An investigation of how students view taking notes in lectures. System. 2001;29:405–417. http://dx.doi.org/10.1016/S0346-251X(01)00028-8

17. Titsworth BS, Kiewra KA. Spoken organisational lecture cues and student note-taking as facilitators of student learning. Contemp Educ Psychol. 2004;29:447–461. http://dx.doi.org/10.1016/j.cedpsych.2003.12.001

18. Kiewra KA, DuBois NF, Christian D, McShane A, Meyerhoffer M, Roskelley D. Note-taking functions and techniques. J Educ Psychol. 1991;83:240–245. http://dx.doi.org/10.1037/0022-0663.83.2.240

19. Peper RJ, Mayer RE. Note taking as a generative activity. J Educ Psychol. 1978;70:514–522. http://dx.doi.org/10.1037/0022-0663.70.4.514

20. Peper RJ, Mayer RE. Generative effects of note taking during science lectures. J Educ Psychol. 1986;78:34–38. http://dx.doi.org/10.1037/0022-0663.78.1.34

21. Biggs JB, Collis KF. Evaluating the quality of learning: The SOLO taxonomy. New York: Academic Press; 1982. http://dx.doi.org/10.1016/B978-0-12-097552-5.50002-8

22. Reed Y, Grandville S, Janks H, Makao P, Stein P, Van Zyl S, et al. (Un)reliable assessment: A case study. Perspect Educ. 2003;21:15–28.

23. Granville S, Dison L. Thinking about thinking: Integrating self-reflection into an academic literacy course. J Engl Acad Purposes. 2005;4:99–118. http://dx.doi.org/10.1016/j.jeap.2004.07.009

24. Hughes C, Suritsky SK. Note-taking skills of university students with and without learning disabilities. J Learn Disabil.1994;27:20–24. http://dx.doi.org/10.1177/002221949402700105

25. Di Vesta F, Gray SG. Listening and note-taking. J Educ Psychol. 1972;63:8–11. http://dx.doi.org/10.1037/h0032243

26. Dunkel P. The content of L1 and L2 students' lecture notes and its relation to test performance. TESOL Q. 1988;22:68–90. http://dx.doi.org/10.2307/3586936

27. Kiewra KA, Benton S, Kim S, Risch N, Christensen M. Effects of note-taking format and study technique on recall and relational performance. Contemp Educ Psychol. 1995;20:172–187. http://dx.doi.org/10.1006/ceps.1995.1011

28. Howie S, Venter E, Van Staden S, Zimmerman L, Long C, Du Toit C, et al. PIRLS 2006 South African children's reading literacy achievement. Pretoria: Centre for Evaluation and Assessment, University of Pretoria; 2008.

29. Howie S, Van Staden S, Tshele M, Dowse C, Zimmerman L. PIRLS 2011 South African children's reading literacy achievement. Pretoria: Centre for Evaluation and Assessment, University of Pretoria; 2012.

30. Wildsmith R. The African languages in South African education 2009-2011. Lang Tech. 2013;46:120–124. http://dx.doi.org/10.1017/S0261444812000420

31. Haggis T. Pedagogies for diversity: Retaining critical challenge amidst fears of 'dumbing down'. Stud High Educ. 2006;31:521–535. http://dx.doi.org/10.1080/03075070600922709

32. Fang Z. Scientific literacy: A systemic functional linguistics perspective. Sci Educ. 2004;89:335–347. http://dx.doi.org/10.1002/sce.20050

33. Wellington J, Osborne J. Language and literacy in science education. Buckingham: Open University Press; 2001.

34. Brenner E, Nichols P. Critical engagement through writing (CEW): Using writing to promote critical thinking in large classes. In: Hornsby DJ, Osman R, De Matos Ala J, editors. Teaching large classes: Interdisciplinary perspectives for quality tertiary education. Stellenbosch: Sun Press; 2013. p. 97–112. http://dx.doi.org/10.18820/9780992180690

Student-perceived criteria for assessing university relevance in community development

AUTHORS:
Joseph Francis[1] iD
Beata Kilonzo[1]
Pertina Nyamukondiwa[1] iD

AFFILIATION:
[1]Institute for Rural Development, University of Venda, Thohoyandou, South Africa

CORRESPONDENCE TO:
Pertina Nyamukondiwa

EMAIL:
pertinam@gmail.com

In sub-Saharan Africa, universities are increasingly being called upon to contribute more towards combating poverty and promoting development in rural areas. Yet, it is still argued that universities are ivory towers, and as a result, their contribution to finding sustainable solutions to issues hampering the realisation of improved quality of life of people in rural areas remains unsatisfactory. This perception emanates from the universities' apparent failure to articulate and demonstrate how they can achieve the desired goal stated above. Moreover, there are no universally embraced criteria for assessing the relevance of a rural area based university to the community it serves. This study was therefore carried out to determine the perceptions of University of Venda undergraduate students on what they believed were appropriate criteria for assessing the relevance of a rural area based university in community development in South Africa. Reflection circles, anchored on participatory research techniques, were used to engage the students. The results of the engagement were organised into sub-themes. The most prominent perceptions were: 'A university has active long-term community-based development initiatives'; 'A university is continuously addressing the real needs of the communities in question'; 'University initiatives are creating jobs for its graduates and community members'; and 'Continuous community requests for university assistance in solving the challenges militating against development'. The wide range of perceptions of students observed in this study is a useful input into initiatives seeking to develop an objective tool for assessing the relevance of a rural area based university in community development.

Significance:

- A set of criteria that students believe should be used to assess the relevance of universities in community development were generated.

- The criteria can be used to develop an index that might serve as a tool for ranking university relevance to their constituencies.

- The criteria can also be used to sharpen the business of community engagement directorates in rural-based universities.

Introduction

In sub-Saharan Africa, universities are increasingly being called upon to contribute more meaningfully towards combating poverty and promoting development in rural areas.[1-3] However, most universities continue to operate like ivory towers whose contribution to the improvement of the quality of life of people in rural areas is unsatisfactory.[4,5] Boyer[4] shares his frustration with the ivory tower persona and profession-orientation of universities, which tends to alienate them from the larger community and lead to the marginalisation of the most pressing civic, social, economic, environmental and moral challenges. Olowu[6] extends this argument by noting that most rural universities are disconnected from the community's needs and rarely take advantage of the strengths of their immediate societies. MacGregor and Makoni[7] recommend that universities should be 'citadels not silos, defending communities around them rather than being inward-looking…if they are to actively advance their development goals'. This situation highlights the need for universities to contribute to implementation of the transformation goals defined in the White Paper on Higher Education of 2014 and the National Plan for Higher Education 2015/16–2019/20, in particular in rural areas, so that universities take a lead in building a more just and progressive society.

Another view that is gaining prominence in the higher education sector is the need for universities to become effective engines for preparing students to apply their expertise in facilitating the development of grassroots communities, countries and regions.[8,9] In support of this view, Fullerton[9] believes in a system that benefits university researchers, students and the community. The author makes reference to benefits from community engagement through improving universities' core intellectual and academic work, in part, through giving students and faculty real-world experience which can positively impact both research and teaching. Fullerton[9] contends that through engaging communities when tackling the problems they face and also involving students in such processes, rural planning and development scholars can bring about valuable experiences for all role players. Anderson[10] expands this contention by pointing out that better insight and implications from the research are gained when they are developed in collaboration with those who live the life.

Both urban and rural areas face considerable developmental challenges, among which are social exclusion, citizen protests because of a lack of basic needs or services, inadequate income to meet household demands, inappropriate disposal of toxic waste, unsustainable farm and non-farm enterprises, and lack of sustainable industries, most often within the reach of local universities.[8,11,12] Despite the existence of this wide range of challenges, rural-based higher education institutions, particularly in developing countries, continue to operate in ways that do not reflect the aspirations and expectations of their constituencies.[8,9,13] Thus, interventions seeking to redefine their roles in community development and societal transformation in general are justified.

Some scholars share the belief that quite often universities descend on communities to carry out needs assessments[14,15] and then abandon them soon after writing research reports and or assignments that merely personally advance the investigators[8,14,16]. Bodorkós and Pataki[16] claim that lack of credibility of science today stems from the way it is currently institutionalised. Rarely do university researchers share the results of their work with the communities in which the research is conducted. Nor are the research outcomes implemented. It seems this lack of credibility is rooted in the universities' failure to articulate how they can effectively contribute to societal transformation and development.

One way of making universities relevant is to strengthen their interaction with local communities, civic organisations, businesses and industry partners for local development.[17,18] According to the then South African Department of Education (DoE)[17(p.4-5)], this would be achieved by 'contributing to the development of communities through application and extension of its knowledge and expertise'. Students who participate in such actions are exposed to practical and relevant training which equip them with the skills needed to improve local and national development.[1,9,19] Initiatives or programmes of that nature must reflect the voices of the students involved and community members concerned, in line with the adage, 'Nothing about us without us'[20]. Thus, taking this situation into consideration, an exploratory case study was carried out to 'harvest' the perceptions of students at the University of Venda (Univen) in South Africa on what they regarded as suitable criteria that could be used to assess the relevance of a rural area based university in community development. Before describing the study area, the context of a rural area based university in South Africa is clarified in order to build a better understanding of what necessitated the current study. Also, community development and engagement are briefly explained.

Conceptualising rural-based universities in a South African context

Since the first university (University of Bologna in Italy) was established in 1088, promotion of teaching, research and knowledge transfer have remained as the major roles of these institutions.[21-23] However, universities have continued to evolve in response to new challenges facing nations and availability of improved technologies. According to Etzkowitz et al.[24], in order to address the new challenges, universities are expected to improve regional and national economic performance. At the same time, they should also address their own financial challenges. In Africa, universities were promoted and established to assist decolonised nations to 'build the capacities of emerging governments to develop, manage resources, alleviate poverty of their people and close the gap between them and the developed world'[1(p.4-5)]. In South Africa, the potential of rural-based universities to play a significant role in facilitating sustainable development, especially in the communities within their vicinity, is not in doubt. This argument takes cognisance of: the universities' strategic locations within the rural communities which they are supposed to serve[25]; availability of relevant physical infrastructure for training and empowering of communities; and potential to promote teaching and learning and build strong partnerships[21]. However, poverty and underdevelopment remain rife in rural areas and residents often struggle to feed themselves.[26-28] These conditions present opportunities for action research and learning platforms for students as well as academics. Apparently, the Integrated Sustainable Rural Development Strategy of 2000 was designed in a way that enabled rural-based universities to direct their work towards building the capacity of individuals and local communities to champion their own development. It was assumed that they were strategically located to play a catalytic development role and would be easily linked to designated local industrial development nodal points. Community engagement activities were expected to emerge in the form of synergistic partnerships with local businesses, civic bodies, industry, and local and provincial governments, leading to meaningful socio-economic growth and development.

A conducive climate for mounting effective community engagement initiatives in the country does exist. Successive post-1994 govern-ments formulated various strategies and polices meant to redefine, restructure and transform the higher education system and society in general.[28-30] Among these were the Constitution of the Republic of South Africa in which Chapter 2 is dedicated to the Bill of Rights, *Higher Education Act No. 101 of 1997* and White Paper No. 3 on the Transformation of Higher Education notice 1196 of 1997. The White Paper identified the need for societal transformation and ways to address the development challenges facing the broader society.[30,31] Moreover, the Reconstruction and Development Programme of 1994 had earlier outlined this challenge. The new mandate that the DoE[17], Bunting[28] and Badat[29] espoused provided an opportunity for rural-based universities to play a more meaningful role in promoting sustainable rural development.[28,30] However, as Nkomo and Sehoole[21] reveal, there is a dearth of research focusing on the contribution of rural-based universities to sustainable development. Clearly, this situation presents an opportunity for the universities to co-originate with the affected constituencies, practical solutions to existing challenges. This is likely to be a central pillar of relevance of a rural-based university to rural communities in South Africa.

Community development in perspective

Community development is a process in which residents of an area come together to take collective action and generate solutions to common problems.[32] According to Zadeh and Ahmad[33], it often takes place when community members genuinely participate in the development process. Rural communities have particular strengths and assets that can be harnessed in order to improve their quality of life. They might benefit from educational institutions such as universities through financial and administrative support.[9] Higher education institutions should strive to fill this gap so that they facilitate community-led planning and implementation of local development initiatives. It is imperative to note that participation of local people in their own social change is a fundamental tenet of community development. Therefore, any outside intervention must fall within the structures and norms of the community in question.

Community engagement clarified

Many definitions of community engagement exist. Community engagement is grounded in the principles of organising people in a particular locality and adheres to fairness, justice, empowerment, participation and self-determination.[34-37] A common denominator of the myriad of definitions is the role that university students and staff must play. For purposes of this paper, we have adopted the Maddison and Laing[38] view of community engagement – which is that it 'takes a particular form, and is context-dependent – arising for institutions from their individual histories and locations, and from their view about their strategic position'. There is also a need to highlight that although the engaged activities are in the local environment in relation to the campus, globalisation has virtually extended the sphere of influence of universities. Globalisation thus justifies the need to ensure that community engagement embraces stakeholders well beyond the campus.

In South Africa, community engagement is defined from the perspective of the legislations and policies on the higher education system, which recognise it together with teaching and learning, and research as the key pillars of university business. South Africa's White Paper on Higher Education of 1997 calls upon universities to 'demonstrate social responsibility…and their commitment to the common good by making available expertise and infrastructure for community service programmes'. One of its key objectives is to 'promote and develop social responsibility and awareness amongst students of the role of higher education in social and economic development through community engagement'.

Description of the study area

Univen was established in 1982 and is a historically disadvantaged higher education institution located in Thohoyandou. Thohoyandou is the administrative centre of the predominantly rural Thulamela Municipality within the Vhembe District Municipality of the Limpopo Province

of South Africa. Other municipalities that make up this District are Makhado, Musina and Mutale. Several villages surround Thohoyandou and the university. A considerably high poverty rate, which Statistics South Africa[39] estimates to be 40%, and backlogs in service delivery (the worst being electricity, potable water and sanitation) characterise the area. The poverty evident in the Vhembe District and other areas in the rural areas of South Africa is rooted firmly in the discriminatory policies of the apartheid system implemented prior to 1994, when the country attained democracy. During the apartheid era, the entire Vhembe District was part of the Venda Homeland or Bantustan.

Univen lies about 180 km to the northeast of Polokwane, the capital city of Limpopo Province. In addition, Univen is located approximately 160 km southeast of the Beitbridge border with Zimbabwe and 60 km to the west of the Punda Maria gate that connects Thulamela Municipality to the Kruger National Park. The university lies at a latitude of -22.9° and longitude of 30.4°.

The university's core business of research, teaching and learning, and community engagement is run across eight schools: Agriculture, Education, Environmental Sciences, Health Sciences, Human and Social Sciences, Law, Management Sciences, and Mathematical and Natural Sciences. All the schools are further disaggregated into 66 departments, centres and institutes. Currently, about 14 000 students are registered for various undergraduate and postgraduate qualifications within the university. The university's vision, which was adopted in 2007, is 'To be at the centre of tertiary education for rural and regional development in southern Africa'. However, as Francis et al.[14] contend, most schools within Univen continue to grapple with how best to make themselves relevant to the mainly rural grassroots constituency that the university is expected to serve.

Research methodology

Research design, population and sampling

This study was carried out using an exploratory case study research design. A purposive sample of 42 students was selected; the students were aged 19–24 years and were registered for undergraduate degrees in various academic disciplines in the eight schools at Univen. The students were drawn from the 70 who actively participated in a rural community-based programme called 'Amplifying Community Voices in Makhado Municipality'.[14] An open invitation was sent to the 70 students to participate in this study if they were interested and willing. In the invitation, the purpose, benefits of their participation and what the results of the study would be used for were explained. A total of 20 male and 22 female students voluntarily participated in the research. More information on the study was provided on the day of data collection. The students were accorded the opportunity to ask any questions on any contentious issues they wanted clarified. Each participant signed a consent letter to confirm their willingness to participate in the study. They were informed that they were free to withdraw from the study at any point. Ethical clearance was secured from the University of Venda's Research Ethics Committee (certificate number SHS/11/PH/002). The study was conducted as part of the community–university partnership programme (grant 71 231), which was one of the four national initiatives the Department of Science and Technology piloted to enable historically black universities to contribute more to socio-economic development.

Reflection circles (Figure 1) were constituted for data collection purposes. Table 1 shows the breakdown of the number of students in each reflection circle. When constituting the reflection circles, the sex of the students was used as a stratification variable. It was assumed that the views of male and female students might differ. Six reflection circles were constituted: men (2 groups), women (2 groups) and mixed men and women (2 groups). Each reflection circle deliberated on the same question, namely: 'What criteria would you use to assess the relevance of a rural-based South African university in community development?'
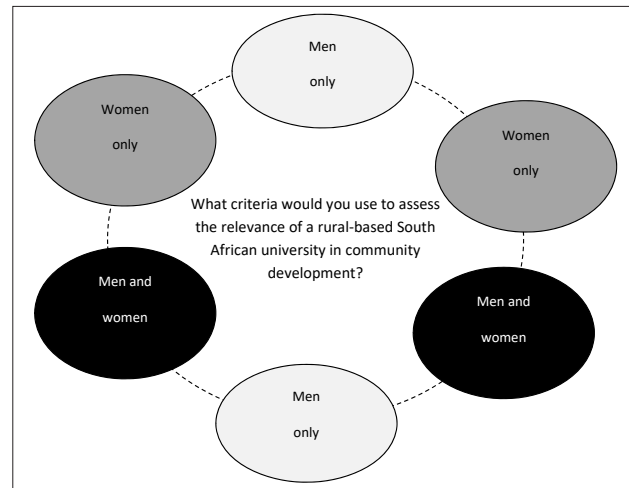
Figure 1: Schematic presentation of the reflection circles constituted in the exploratory study on the relevance of a South African rural area based university in community development.

Table 1: Composition of reflection circles of students formed to deliberate on the criteria to use when assessing the relevance of a South African rural-based university in community development

Respondent group	Number of students		
	Men	Women	Total
1	7	0	7
2	6	0	6
3	0	8	8
4	0	7	7
5	3	3	6
6	4	4	8
Total	20	22	42

Constituting two reflection circles per category was desirable because of the need for replication[40-42] and triangulation of perspectives[43]. This was also crucial in order to minimise the usually unavoidable bias that researchers often display.[42,44,45] Olsen[43(p.3)] defines triangulation as 'the mixing of data or methods so that diverse viewpoints or standpoints cast light upon a topic'. In this study, the meaning of triangulation was extended through replicating the number of cohorts of respondents.

Members of each reflection circle selected an individual from within the circle to facilitate their deliberations, focusing on the research question stated above. Also, each group selected a rapporteur or scribe. During the discussions, flipcharts and markers were used to record the results whenever consensus on a particular view was reached. Almost all the students who participated in this study were well-trained facilitators who had extensive experience of leading such deliberations, even at a village level at which literacy rates are generally low.[44] The main role of the facilitators was to ensure that the students who displayed domineering tendencies and also those who rarely expressed their views were appropriately managed in order to ensure that the collective voice of all members of the reflection circle was captured. An experienced principal researcher (J.F.) supervised the data collection. From time to time, the researcher clarified any issues that arose as data collection unfolded. It was important to supervise the data collection to ensure that the study was conducted as planned.

After about 1¼ hours of intense intra-reflection circle debates, the respective cohorts formed one reflection circle each for the purpose of consolidating their findings. For example, the two reflection circles

comprising only female students converged and shared their findings, resulting in the production of a consolidated set of results. The consolidated results were subsequently presented in plenary. The only male students and mixed students reflection circles followed suit. The principal researcher facilitated the discussions in the plenary session. As each group presented its findings, one of the participants wrote the various perceptions on a flipchart. In this way, the results of the reflection circles were consolidated and subsequently adopted as the composite set of the perceived criteria that should be used to assess the relevance of a rural area based university in community development. Post-plenary presentations, further discussion of the results ensued, leading to the adoption of the day's research output. The principal researcher thanked all the students for participating in the study, which took about 3 hours to complete.

Data analysis

The thematic content analysis technique of Cresswell[45] was used to consolidate the results of the study into sub-themes of criteria. This involved placing similar student perceptions into aggregate sub-themes. The number of reflection circles that identified each criterion was used as a measure of its popularity or importance. Therefore, this measure was adopted as the priority rank.

Description of results

The perceived criteria that the students identified for assessing the relevance of a university were classifiable into five sub-themes: (1) existence of active community-based programmes; (2) responsiveness to societal challenges; (3) strengthening social cohesion or capital; (4) community awareness of the university's roles; and (5) quality of programmes and graduates of the university. In Table 2, a more detailed picture of the various student perceptions per sub-theme is presented.

Across the sub-thematic areas, the most common criteria were:

- There are active long-term university-run programmes or projects in communities.

- University programmes addressing the needs of the community.

- Involvement of community members in decision-making regarding implemented projects.

- Improved social cohesion in the community mainly because of university development initiatives.

- Communities are aware of their developmental needs and the nature of the relevant support required from the university.

While arguing why his view on the latter criterion should be accepted, one male student remarked that:

> ....apart from having students renting rooms in most of the villages around this university, the communities do not feel the presence of Univen. So, can we surely say this university is relevant? I do not think so.

Another student said,

> It is unfair and unethical for any university to enjoy taxpayers' money without ploughing back. Anyway, even though there are so few of us trying to make a difference in the communities we are working in, one should see this as a positive sign of the direction this institution is taking. Maybe soon, we might find people having greater confidence in our university.

Table 2: Criteria for use in assessing the relevance of a South African rural-based university in community development

Criteria	Focus group of students			Total
	Women only	Men only	Mixed men and women	
Existence of active community-based programmes				
There are active long-term university-run programmes or projects in communities	••	•	•	4
Programmes or projects run by the university creating jobs for its graduates or other members of the community		••	•	3
Responsiveness to societal challenges				
University programmes addressing the needs of the community	••	•	•	4
Regular requests for university assistance from the community		•	•	2
Provision of bursaries or scholarships, specifically targeting students from surrounding communities		••		2
Strengthening social capital or cohesion				
Involvement of community members in decision-making relating to implemented projects or programmes	••	•	•	4
Improved social cohesion in the community mainly because of the university's development initiatives	•	••	•	4
Acceptance by society because the university understands and respects the culture of the community			•	1
Community awareness of the university's roles				
Communities are aware of their developmental needs and the nature of relevant support required from the university	•	•	••	4
Community members giving positive testimonies regarding the university work	•		••	3
Quality of programmes and graduates				
Increased interest by local school-leavers in being admitted into university programmes		•	•	2
Employers increasingly interested in and recruiting graduates of the university	•		•	2

•, affirmative for identification of that indicator

Discussion

Some interesting criteria to use when assessing the relevance of a rural area based university in community development were unearthed in this study. However, before explaining the results, it is important to restate that a university is expected to play a role in societal transformation and development through generating and disseminating knowledge to potential users.[13,30,46-48] In addition to this understanding, utilisation of the knowledge is of paramount importance because it helps the users to expand their work beyond traditional boundaries.[47] Universities provide both general and specific skills required to enhance societal transformation and development.[49] Moreover, they create an enabling environment for collective action via engaging in civic conversations with the concerned communities or beneficiaries.[16] This clarification is crucial because it helps to put into perspective the student-suggested criteria that should be used to assess the relevance of a university in rural community development.

The students who participated in this study pointed out that establishing long-term community-based programmes was an important criterion of relevance of a rural-based university. Presumably, the fact that the existence of active community-based programmes helps to build long-term relationships[49-51] informed this view, apart from improving the image of the university and related support. Students involved in such programmes tend to become more creative and energetic[52] – which are positive qualities that enhance the chances of recruiting and retaining these students. However, Butterfield and Soska[5], Torres and Reyes[13] and Bridger and Alter[15] caution against universities striving to serve their own interests and in the process use grassroots communities as subjects of their research. Smart partnerships with communities of place would ascertain enhanced relationships and justify implementation of long-term local programmes. Such programmes must benefit members of the community of place and the university and enhance the latter's relevance to its constituency.

Ideal and successful partnerships in community-based programmes should focus on mutual benefits[53], co-learning[54] and a strong sharing of knowledge[11,52]. The students' line of thinking was consistent with these views. Furthermore, other considerations seemed to have influenced the students' views on the criteria of relevance of a rural-based university: increased interest by local school leavers in being admitted into the university's programmes; university programmes or projects creating jobs for its graduates or other community members; and communities being aware of their developmental needs and the nature of relevant support required from the university.

Manzuri and Rao[53] rely on the work of Alderman[54] to base their argument that:

> The cornerstone of community-based development initiatives is the active involvement of members of a defined constituency in at least some aspects of project design and implementation. While participation can occur at many levels, a key objective is the incorporation of 'local knowledge' into the project's decision making processes.

However, some aspects of this assertion are contestable. For example, 'participation of members of a defined community in at least some aspects of project design and implementation' does not nurture genuine and fruitful partnerships. Such limited involvement is unlikely to yield sustainable community development. It is also not clear how the decision on the 'at least some aspects of project design and implementation' is made. Ruhiiga's[49] view of promoting the involvement of community members in all the programme's processes and activities should be embraced. This paradigm promotes co-learning and sharing of various perspectives and also develops transformative leadership for sustainable development among students, grassroots communities and academia. Possibly, when such collegial relationships are developed, there is a strong likelihood of 'increased interest by local school leavers in being admitted into the university programmes'. This might explain why the students viewed 'employers increasingly interested in and recruiting graduates from the university' as important. The student-distilled criteria provide valuable evidence of what to consider when assessing the quality of programmes and graduates of a rural-based university. When this happens it would not be surprising to find, as revealed in this study, 'community members giving positive testimonies' on the roles that an engaged university plays.

Badat[29(p.5)] highlights the need for a university to be 'responsive to its political, economic and social contexts and community engagement'. In this study, the context of responsiveness that the students seemed to have in mind was the ability of a functionally relevant university to respond to a request from a community in need or to transform its operational focus and character in order to deal with an existing or emerging local, national, regional or international challenge. The three measures of responsiveness that the students identified seemed one-sided because all of them suggested that the grassroots community was always the beneficiary. Taking into account their experience in community development practice as academics, Francis et al.[15] point out that establishment of sustainable partnerships builds a clearer understanding of the roles and hence the relevance of a rural-based university. Moreover, Judd and Adams[50] contend that university–community partnerships provide the opportunity for practical learning where students gain realistic experience. This prepares students to assume political roles in their communities.[55] Hunter and Mileski[56] add capacity-building and collective learning as the major benefits in the partnership, which enhance the relevance of a rural-based university to its surrounding rural communities and the broader society.[28]

Social capital or cohesion-related issues that the students identified highlighted the need for creating strong bonds between a community and the university in question. Netshandama[55] contends that the existence of a social contract between universities and the society is a clear indication of the commitment to work together. She argues that because the society requires scientifically generated knowledge, there exists a social contract between it and the institutions that produce it. Therefore, it is important for the two parties to work closely.

One of the most popular criteria cited in the current study regarding social capital – improved social cohesion in community principally due to university development initiatives – deserves special recognition, because there is considerable evidence of the existence of tensions and at times conflicts among traditional leaders, municipal ward committees, community development workers and civic structures in most communities in South Africa.[15,57,58] In the year 2008, the South African government launched a national strategy that recognises social cohesion as one of the nine critical pillars upon which efforts made to combat poverty should rest. Furthermore, the country adopted *A National Strategy for Developing an Inclusive and a Cohesive South African Society*[59] in 2012 whose overall goal is to create a caring and proud society, anchored on 13 indicators. All these national government efforts confirm that lack of social cohesion remains a challenge that pervades the country. For this reason, it is not surprising that the students who participated in this study highlighted various social cohesion-related criteria that they believed should be used to assess the relevance of a rural area based university in South Africa.

Based on our own personal experiences and observations, it can be concluded that academics involved in community-based work tend to shy away from tackling inherent tensions and conflicts in favour of focusing on the primary objectives of the research or community engagement they implement. The students' suggestion of including contribution to strengthening social cohesion should be applauded because communities that are not 'healthy', in particular as a result of infighting among its leaders, will always find it difficult to marshal sustainable energy to achieve self-driven development. Thus, inclusion of activities that seek to achieve and/or strengthen inter-leadership institutional bonds within grassroots communities is worth considering.

The arguments presented above resonate with those of Hobbs[60] and Manzuri and Rao[53]. These scholars cite the importance of Putnam's[61(p.167)] views, namely 'features of organisation such as trust, norms and networks that can improve the efficiency of society by

facilitating coordinated actions'. In this regard, social capital or cohesion refers to the ability of individuals to build 'bonds' within their own group and 'bridges' that link them with others. All this is deeply tied to the belief that the quality and quantity of group activity are key sources of a community's strength and its ability to work for its own betterment.[61,62] Based on the results of this study, it is prudent for a rural area based university such as Univen to deploy resources for use in efforts that strengthen social cohesion or capital so as to enhance the chances of achieving sustainable community development.

Conclusions

Various criteria that might be used to assess the relevance of a rural area based university in South Africa were unearthed in this study. The criteria were categorised into the following sub-themes: existence of active community-based programmes; responsiveness of the university to societal challenges; social capital or cohesion; community awareness of the university's roles; and quality of programmes and graduates of the university. These results have laid the foundation for further research that should guide the development, in partnership with rural communities and other stakeholders, of a tool to use when assessing the relevance of a rural-based university in South Africa and even beyond.

Acknowledgements

Authors' contributions

J.F. conceived the study, led the data collection and analysis process as well as the preparation of the research article. B.K. participated in the data collection and co-authored the article. P.N. finalised the writing and submission of the article.

References

1. Sawyerr A. Challenges facing African universities: Selected issues. Afr Studies Rev. 2004;47(1):1–59. http://dx.doi.org/10.1017/S0002020600026986

2. Naidoo R. Higher education as a global commodity: The perils and promises for developing countries. London: The Observatory on Borderless Higher Education; 2007.

3. Ijaduola KO. Driving policy of higher education in Nigeria towards relevance. Acad Leadersh J. 2011;9(1):1–7.

4. Boyer E. The scholarship of engagement. J Public Service Out-reach.1996;1(1):11–20.

5. Butterfield AKJ, Soska TM. University–community partnerships: Universities in civic engagement. New York: The Haworth Social Work Practice Press; 2004.

6. Olowu D. University-community engagement in South Africa: Dilemmas in benchmarking. S Afr Rev Sociol. 2012;43(2):89–103. http://dx.doi.org/10.1080/21528586.2012.694252

7. MacGregor K, Makoni M. Universities must be citadels not silos. University World News. 2010 May 02 [cited 2013 Jul 30]. Available from: http://www.universityworldnews.com/article.php?story=20100502103801345

8. Vasilescu R, Barna C, Epure M, Baicu C. Developing university social responsibility: A model for the challenges of the new civil society. Procedia Soc Behav Sci. 2010;2:4177–4182. http://dx.doi.org/10.1016/j.sbspro.2010.03.660

9. Fullerton C. University-community partnerships as a pathway to rural development: benefits of an Ontario land use planning project. J Rural Comm Dev. 2015;10(1):56–71.

10. Anderson JA. Engaged learning, engaged scholarship: A struggle for the soul of higher education. Northwest J Comm. 2014;42 (1):143–166.

11. Boucher G, Conway C, Van Der Meer E. Tiers of engagement by universities in their region's development. Reg Stud. 2003;37(9):887–897. http://dx.doi.org/10.1080/0034340032000143896

12. Kuhnen F. The role of agricultural colleges in modern society: The university as an instrument in social and economic development. Q J Int Agric. 1978;17:77–88.

13. Torres MN, Reyes LV. Challenges in engaging communities in bottom-up literacies for democratic citizenship. Educ Citizen Soc Justice. 2010;5(3):191–205. http://dx.doi.org/10.1177/1746197910382258

14. Francis J, Dube B, Mokganyetji T, Chitapa T. University–rural community partnership for people-centred development: Experiences from Makhado Municipality, Limpopo Province. S Afr J High Educ. 2010;24(3):357–373.

15. Bridger JC, Alter TR. The engaged university, community development and public scholarship. J High Educ Outreach Engagem. 2006;11(1):163.

16. Bodorkós B, Pataki G. Linking academic and local knowledge: Community-based research and service learning for sustainable rural development in Hungary. J Cleaner Prod. 2009;17(12):1123–1131. http://dx.doi.org/10.1016/j.jclepro.2009.02.023

17. Department of Education (DoE). The restructuring of the higher education system in South Africa. Report of the National Working Group; 31 January 2002.

18. Hansen JA, Lehmann M. Agents of change – Universities as development hubs. J Cleaner Prod. 2006;14(9–11):820–829. http://dx.doi.org/10.1016/j.jclepro.2005.11.048

19. Chow A, Whitlock M, Lucy P. Systemic Educational Change and Society. Int J Sc Soc. 2010;1(4):127–134.

20. Charlton JI. Nothing about us without us. Berkeley, CA: University of California Press; 1998. http://dx.doi.org/10.1525/california/9780520207950.001.0001

21. Nkomo M, Sehoole C. Rural-based universities in South Africa: Albatrosses or potential nodes for sustainable development? Int J Sust High Educ. 2007;8(2):234–246. http://dx.doi.org/10.1108/14676370710726689

22. De Ridder-Symoens H. A history of the university in Europe: Volume 1: Universities in the Middle Ages. Cambridge, UK: Cambridge University Press; 1992.

23. Hunt J. The university in medieval life, 1179–1499. Oxford: McFarland; 2008.

24. Etzkowitz H, Webster A, Gebhardt C, Terra BRC. The future of the university and the university of the future: Evolution of ivory tower to entrepreneurial paradigm. Res Pol. 2000;29:313–330. http://dx.doi.org/10.1016/S0048-7333(99)00069-4

25. Sachs FD. The end of poverty: Economic possibilities of our time. Eur J Dental Educ. 2008;12(s1):17–21. http://dx.doi.org/10.1111/j.1600-0579.2007.00476.x

26. International Fund for Agricultural Development (IFAD). Rural Poverty Report: New realities, new challenges: New opportunities for tomorrow's generation. Rome: IFAD; 2011.

27. Kamando A, Doyle L. Universities and community-based research in developing countries: Community voice and educational provision in rural Tanzania. J Adult Cont Educ. 2013;19(1):17–38. http://dx.doi.org/10.7227/JACE.19.1.3

28. Bunting I. The higher education landscape under apartheid. In: Cloete N, Fehnel R, Maassen P, Moja T, Perold H, Gibbon T, editors. Transformation in higher education: Global pressures and local realities in South Africa, Cape Town: Juta; 2002. p. 58–86.

29. Badat S. The role of higher education in society: Valuing higher education. Paper presented at the HERS-SA Academy, University of Cape Town Graduate School of Business, Cape Town; 2009 September 14.

30. South African Department of Education (DoE). A programme for higher education transformation. Education White Paper 3. Pretoria: DoE; 1997.

31. Statistics South Africa. Statistical release (revised) P0301.4: Census 2011. Pretoria, South Africa. Pretoria: Statistics South Africa; 2012. Available from: http://www.statssa.gov.za/publications/P03014/P030142011.pdf

32. Frank F, Smith A. The community development hand book: A tool to build community capacity. Hull, Canada: Labour Market Learning and Development; 1999.

33. Zadeh BS, Ahmad N. Participation and community development. Curr Res J Soc Sci. 2010;2(1):13–14.

34. Alinsky SD. Citizen participation and community organization in planning and urban renewal. Chicago, IL: Industrial Areas Foundation; 1962.

35. Freire P. Pedagogy of the oppressed. New York: Herder and Herder; 1970.

36. Wallerstein NB, Duran B. Using community-based participatory research to address health disparities. Health Promot Pract 2006;7(3):312–323. http://dx.doi.org/10.1177/1524839906289376

37. Chávez V, Minkler M, Wallerstein N, Spencer MS. Community organizing for health and social justice. In: Cohen L, Chávez V, Chehimi S, editors. Prevention is primary: Strategies for community well-being. San Francisco, CA: John Wiley and Sons; 2007. p. 95–120.

38. Maddison E, Laing S. The CUPP model in context. In: Hart A, Maddison E, Wolf D, editors. Community-university partnerships in practice. Leicester: National Institute of Adult Continuing Education; 2007. p. 8–20.

39. Downward P, Mearman A. Retroduction as mixed-methods triangulation in economic research: Reorienting economics into social science. Camb J Econ. 2007;31(1):77–99. http://dx.doi.org/10.1093/cje/bel009

40. Bryman A. Social research methods. Oxford, UK: Oxford University Press; 2001.

41. Jain SP, Polman W. A handbook for trainers on participatory local development. 2nd ed. Bangkok: Food and Agriculture Organization of United Nations; 2003.

42. Babbie ER. The practice of social research. 10th ed. Belmont, CA: Thomson/Wardsworth; 2004. http://dx.doi.org/10.1108/01443330410790849

43. Olsen W. Triangulation in social research: Qualitative and quantitative methods can they really be mixed. In: Holborn M, editor. Developments in sociology. Ormskirk: Causeway Press; 2004. p. 103–121.

44. Francis J, Mamatsharaga P, Dube B, Chitapa T. Making youth voices count in own development programming in Makhado Municipality, Limpopo Province. J Geo Reg Planning. 2011;4(5):297–304.

45. Creswell J. Research design: Qualitative, quantitative, and mixed methods approaches. Thousand Oaks, CA: Sage Publications; 2009.

46. Geoghegan W, Pontikakis D. From ivory tower to factory floor? How universities are changing to meet the needs of industry. Sci Public Policy. 2008;35(7):462–474.

47. Bok D. Universities in the market place. Baltimore, MD: Johns Hopkins University Press; 2003.

48. Boulton G, Lucas C. What are universities for? Leuven: League of European Research Universities (LERU); 2008.

49. Ruhiiga TM. Reverse empowerment in South Africa's comprehensive rural development programme. J Hum Ecol. 2013;41(2):165–174.

50. Judd AH, Adams MH. Lessons learned from a decade in a university-community partnership: Keys to successful engagement and outreach. J High Educ Outreach Engagem. 2008;12(3):117–127.

51. Harnish RJ, Bridges KR. University-community partnership: Teaching applied social psychology to foster engagement in strategic planning. J High Educ Outreach Engagem. 2004;10(1):107–119.

52. Kaufman R. A university-community partnership to change public policy: Pre-conditions and processes. In: Soska TM, Butterfield AKJ. University-community partnerships: Universities in civic engagement. New York: The Haworth Social Work Practice Press; 2004.

53. Manzuri G, Rao V. Community-based (and driven) development: A critical review. Washington DC: Development Research Group, World Bank; 2003.

54. Alderman H. Do local officials know something we don't? Decentralization of targeted transfers in Albania. J Public Econ. 2002;83:375–404. http://dx.doi.org/10.1016/S0047-2727(00)00145-6

55. Netshandama VO. Community development as an approach to community engagement in rural-based higher education institutions in South Africa. S Afr J High Educ. 2010;24(3):342–356.

56. Hunter R, Mileski K. Emerging leaders project: Connecting university resources to community-based organisations supporting refugee settlement. Adv Soc Work. 2013;14(2):613–628..

57. Mathekga R, Buccus I. The challenge of local government structures in South Africa: Securing community participation. In: Critical dialogue–public participation in review. Durban: Centre for Public Participation; 2006.

58. Fuo ON. Public participation in decentralised governments in Africa: Making ambitious constitutional guarantees more responsive. Afr Hum Rights Law J. 2015;15:167–191. http://dx.doi.org/10.17159/1996-2096/2015/v15n1a8

59. South African Department of Arts and Culture (DAC). Creating a caring and proud society: A national strategy for developing an inclusive and a cohesive South African society. Pretoria: DAC; 2012. Available from: http://www.gov.za/sites/www.gov.za/files/Social-Cohesion-Strategy-1.pdf

60. Hobbs G. What is social capital? A brief literature overview. Dar es Salaam: Economic and Social Research Foundation; 2000.

61. Putnam R. Making democracy work: Civic tradition in modern Italy. Princeton, NJ: Princeton University Press; 1993.

62. Doe SS, Lowery D. The role of evaluation in developing community-based interventions: A COPC project P71-88. In: Soska T, Butterfield AKJ, editors. University-community partnerships: Universities in civic engagement. New York: The Haworth Social Work Practice Press; 2004.

A bibliometric assessment of energy research in South Africa

AUTHOR:
Anastassios Pouris[1] iD

AFFILIATION:
[1]Institute of Technological Innovation, University of Pretoria, Pretoria, South Africa

CORRESPONDENCE TO:
Anastassios Pouris

EMAIL:
apouris@icon.co.za

The results of an effort to identify the performance of energy and fuels research in South Africa during the most recent period (2003–2013) are reported. Bibliometric approaches have been employed in order to assess the field of energy research. Energy research was identified to be improving over time, albeit from a small basis. The field appears to equally emphasise fossil and renewable energy research. Similarly, universities were identified to be producing a subcritical number of energy articles in comparison with international organisations. The relatively small activity in the energy field appears to affect the international collaboration of the field, which is well below the national average. International comparisons in terms of articles per GWH of electricity produced and articles per million population show that South Africa should increase substantially its effort in the field in order to be comparable with other countries.

Significance:
- This article makes a unique contribution in scientometrics to the field of energy research in South Africa which, given its multidisciplinary nature, is a generally neglected field of study in South Africa.

Introduction

Monitoring and assessment are integral parts of science, technology and innovation policy. Decision-makers need to know the performance of the various research disciplines so they can make intelligent decisions in the allocation of scarce resources. Historically, monitoring and assessment relied on expert opinions. Nowadays such assessments are based on quantitative information.

One of the most efficient and objective methods of assessing research and innovation performance is through bibliometric indicators. Bibliometric analysis, the quantitative study of the research system, is based mainly on publication indicators. In bibliometrics, the number of publications in a field is considered to be an indicator of research activity and the number of citations as an indicator of impact.

Bibliometric assessments have a number of advantages. For example, they are repeatable and verifiable exercises. They are not dependent on the choice of experts/peer reviewers or on their opinions which may vary. Their most important advantage is probably that they allow comparisons among different scientific disciplines and different countries. Both types of comparisons are not possible through peer-review approaches as it is almost impossible to find peers with expertise in different scientific fields and knowledge of the research systems in different countries.

Within this context, a number of South African related assessments have appeared in the open and grey literature recently.[1-3] There are limited mapping and comparative assessments in the field of energy nationally and internationally. Relevant literature includes Vlachy[4], Pouris et al.[5], Uzun[6] and Kostoff et al.[7]. As we have argued previously:

> This lack of research activity may have a number of adverse consequences for the economy. For example, it can be argued that the lack of expertise and independent advice (e.g. in the country's universities) may be partially the cause of the recent failure of Eskom to meet electricity demand in the country.[8(p.2)]

Similarly, the current debate related to the country's nuclear energy needs may have been better informed from independent, academic investigations.

The objective of this investigation is to assess the field of energy research in the country. More specifically, the objective is to assess the performance of the field during the most recent period from 2003 to 2013 to: (1) identify the sub-disciplines (e.g. fossil fuel research vis-a-vis renewable energies) emphasised in the country; (2) compare the country's performance with those of other countries and (3) elaborate on relevant research policy issues. These issues can guide policy in a number of domains. For example, the number of research publications produced is related to the expertise and financial resources available in the field in the country. Hence, one question that can be answered is whether the field is adequately supported. Similarly, the determination of sub-disciplines in which research is published can identify areas of inadequate support or emphasis. Comparisons of research outputs at institutional level can provide insights on the efficacy of existing instruments, approaches, etc.

Methodology

Bibliometric assessments require the availability of a database that covers adequately the subject under investigation. The Thomson Reuters databases – Science Citation Index Expanded, Social Sciences Citation Index and Arts and Humanities Citation Index – are the most often used for these types of investigations.

These databases combined cover comprehensively the most prestigious journals in the world in all fields of research and constitute a unique information platform for this objective. The most important advantage of these journals is that they constitute the most important (in terms of impact) journals in the world. Furthermore, the Thomson Reuters databases provide the corporate addresses of all co-authors in an article, and hence comprehensive coverage is possible.

In South Africa, the Department of Higher Education has approved the journals indexed by the Thomson Reuters databases for subsidy purposes; therefore universities receive approximately ZAR120 000 for each article published in one of these journals and provide incentives to their researchers to publish in such journals. Consequently, it is expected that the databases will cover not only the most important South African energy-related research but also the majority.

Two approaches were considered for the extraction of the relevant research literature: phrase-based query and journal-title-based query.

Thomson Reuters assigns the indexed journals to scientific categories. The energy-related journals are grouped under the title 'Energy and Fuels'. This group includes 138 journals. These journals can be considered as consisting of the 'core' journals of the field of energy in the Bradfordian sense.[9] It is emphasised that there are articles related to energy that are not published in the core journals. However, the most important and highest impact energy literature will be that published in the core journals, and hence this analysis aims to identify South Africa's contribution in the core energy literature. Keyword-based searches were used for the identification of research related to different types of primary energy (e.g. fossil and alternative energy; energy converters).

The platform was interrogated for the identification of South African authors publishing in the field of 'Energy and Fuels' during the period 2003–2013. The end of the period was chosen as 2013 as this was the most recent year with complete data. This investigation was initiated during 2014 and was completed during 2015.

The extracted information was analysed in order to identify trends over time; relative performance in comparison with other scientific disciplines in the country; research emphases to various primary sources of energy and technologies; prolific publishing organisations; co-authorship patterns with other countries and institutions; and relative performance vis-à-vis a set of comparator countries (i.e. Australia, Canada, New Zealand, Brazil and Russia). The comparator countries were chosen among those used for benchmarking exercises by the South African Department of Science and Technology.

Energy research in South Africa

Analysis of the core energy literature identified that 752 *articles* with at least one South African address appeared in the database during the 2003–2013 period. This figure represents 66.4% of all documents captured in the database (proceeding papers; meeting abstracts; editorial material, etc). A focus on articles prevents double counting, as a conference paper may also be published as a journal article.

Figure 1 shows the annual number of South African publications in the core energy literature for the period 2003–2013. It becomes apparent that South Africa's contribution to core energy literature is showing an increasing trend. The number of South African articles has grown from 24 during 2003 to 145 during 2013. It should be emphasised that this nominal increase takes place in an environment of increasing national research publications in general and international energy articles in particular.
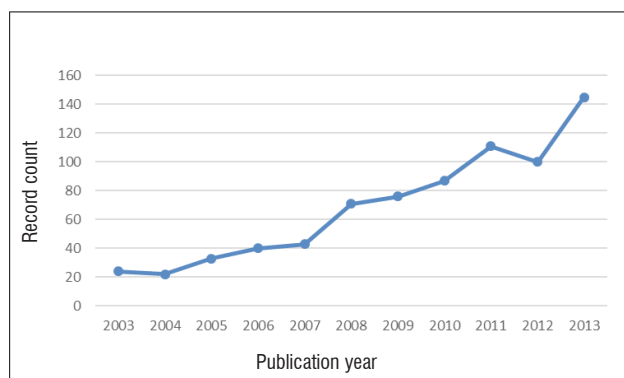
Figure 1:   The number of Energy and Fuels articles from South Africa in the period 2003–2013.

Figure 2 shows the percentage of energy articles of the country's total publications. The percentage of energy articles increased faster than those of the other scientific disciplines in the country during the last decade. Energy and Fuels articles increased from about 0.6% in the beginning of the 2000s to about 1% during the end of the period. This increase is partially the result of Thomson Reuters increasing the coverage of energy journals. Figure 3 makes this point. Even though the absolute number of energy articles increased, the country's share in the world energy publications has been static since 2008.
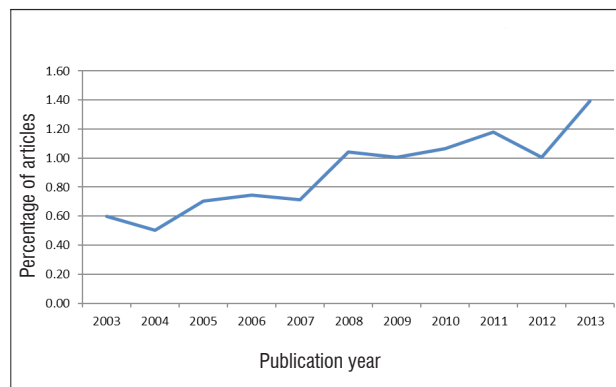
Figure 2:   Percentage of energy-related articles relative to the total number of articles from South Africa in the period 2003–2013.
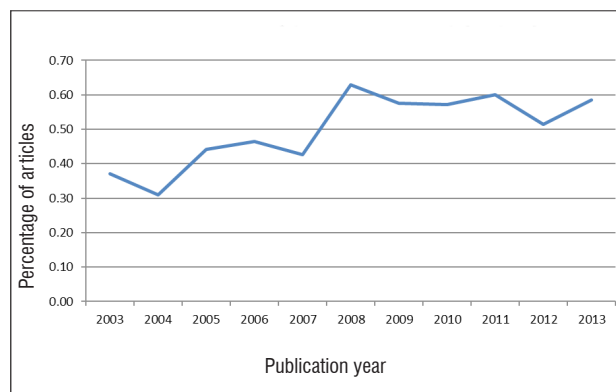
Figure 3:   Percentage of energy-related articles from South Africa relative to the total number of articles in the field during the period 2003–2013.

Figure 3 shows the percentage of South African Energy and Fuels publications within the total field. The relevant share increased from about 0.4% at the beginning of the period to about 0.6% at the end of the period. However, the trend has been static since 2008.

Table 1 shows the number of articles produced by South African researchers according to various research areas during the 2003–2013 period. The table reveals a priority ranking. Environmental Sciences and Ecology and Chemistry are on top of the list with more than 5000 articles over the period. Energy and Fuels appears 47th in the list with 752 articles. It is interesting to note that Water Resources – another multidisciplinary field – appears 22nd in the list. Jacobs et al.[2] argued that the existence of a funding body like the Water Research Commission is the underlying factor for the good performance of this research area. While there is an energy-related agency in the country (Sanedi), it suffers from limited financial support by the government. Similarly, other organisations like Eskom and the National Research Foundation provide limited support for energy research. Furthermore, changes in the various national incentives (e.g. THRIP) are expected to have an adverse effect on energy research publications.

Table 1: Number of research articles from South Africa in the period 2003–2013 according to research area

Rank	Research area	Record count
1	Environmental Sciences Ecology	5326
2	Chemistry	5099
3	Engineering	4644
4	Physics	4037
5	Plant Sciences	3046
6	Mathematics	2879
7	Science Technology Other Topics	2588
8	Agriculture	2566
9	Zoology	2541
10	Public Environmental Occupational Health	2431
11	Infectious Diseases	2407
12	Astronomy Astrophysics	2108
13	Psychology	2102
14	Geology	2072
15	Business Economics	2072
16	Immunology	1998
17	Biochemistry Molecular Biology	1979
18	Material Science	1894
19	Education Educational Research	1828
20	Pharmacology Pharmacy	1812
21	General Internal Medicine	1657
22	Water Resources	1618
23	Microbiology	1596
24	Biotechnology Applied Microbiology	1581
25	Veterinary Sciences	1579
26	Marine Freshwater Biology	1545
27	Religion	1268
28	Computer Science	1213
29	Entomology	1198
30	Food Science Technology	1167
31	Virology	1121
32	Crystallography	1070
33	Evolutionary Biology	1013
34	Mining Mineral Processing	948
35	Government Law	935
36	Surgery	924
37	Biodiversity Conservation	920
38	Metallurgy Metallurgical Engineering	908
39	Neurosciences Neurology	896
40	Genetics Heredity	889

Continued on next page

Continued from previous page

Rank	Research area	Record count
41	Social Sciences Other Topics	877
42	Paediatrics	864
43	Respiratory System	807
44	Meteorology Atmospheric Sciences	781
45	Linguistics	760
46	Health Care Sciences Services	757
47	Energy Fuels	752
48	Life Sciences Biomedicine Other Topics	744
49	Public Administration	743
50	Psychiatry	738
51	Cardiovascular System Cardiology	732
52	Geochemistry Geophysics	699

Table 2 shows the distribution of the South African Energy and Fuels publications to different scientific specialties. Articles are allocated to categories according to the journal in which they are published. Journals are categorised by Thomson Reuters' staff and may belong to one or more categories. The distribution of articles to disciplines shows the linkages of Energy and Fuels to other research areas. Approximately 40% of Energy and Fuels articles are based on Engineering and 12% on Chemistry. In comparison with worldwide emphases in the field, South Africa pays less attention to electrochemistry (9.7% versus 17% worldwide), biotechnology (5.8% versus 9.6%) and agriculture (5.1% versus 9%).

An important component of scientometrics is the capability to identify the subfields and/or technologies constituting a research field.

Table 2: Distribution to research area of Energy and Fuels research articles from South Africa in the period 2003–2013

Research area	Record count
Energy Fuels	752
Engineering	299
Chemistry	93
Electrochemistry	73
Environmental Sciences Ecology	69
Construction Building Technology	65
Thermodynamics	57
Biotechnology Applied Microbiology	44
Agriculture	39
Physics	34
Mechanics	25
Materials Science	23
Mining Mineral Processing	9
Geology	8
Optics	7
Nuclear Science Technology	5

Research fields are composed of many subfields and technologies which are constantly developing or becoming obsolete. The identification of these emphases is of importance for policymakers who need to guide the system according to their priorities and for researchers who need to know fledgling technologies which will shape the discipline in the future. It is also critical to emphasise that the composition of the research landscape is dynamic.

Researchers working in the field develop approaches in order to extract the relevant information from bibliometric and patent databases.[7,10]

Table 3 presents the classification of the South African articles in the research area Energy and Fuels. For the fossil fuels, the keywords used were: fossil; coal; oil; gas; gasification; liquefaction; alkylation; desulfurisation; electrocatalysis; liquid fuel; fluidised bed; emission; char; ash; combustion; pyrolysis; catalysis; incineration; engine; turbine. Among the Energy and Fuels articles, 237 articles had in their titles one or more of the keywords characterising fossil fuels and energy.

Articles related to renewable energy were identified by searching in their titles for at least one of the following keywords: solar; hydrogen; wind; geo*; bio*; hydro power; photovoltaic; tidal; waste fuel; fermentation; microbial desulfurisation; biosulfurisation; thermal decomposition; biodegradation; biomass gasification; water gasification; renewable. A total of 211 articles was identified. The number of renewable energy related articles increased from a handful in the beginning of the period to approximately 30 articles per annum during the end of the period.

The keywords: fuel cell; photovoltaic; solar collector; electrode; electrolyte; and membrane were used to search for direct electric converters and 66 articles related to such research were identified.

In addition, 85 energy-related articles were published in social science journals during the 2003–2013 period (Table 3).

Table 4 shows the frequency of appearance of certain terms in the list of topics of the South African energy and fuels research articles.

The topics related to 'efficiency' have the highest number of articles in the list (177). Figure 4 shows the increase in the number of energy efficiency articles during the period. 'Coal' follows with 135 entries; 'electricity' with 119 and 'oil and liquid fuels' with 113. Among the renewable energies, 'solar' and 'hydrogen' attract the most entries (98 and 73, respectively). It should be mentioned that 'solar', 'hydrogen' and 'biomass' have the common characteristic that they are not site specific. In contrast, 'wind' and 'geothermal' are constrained to geographical areas with favourable operating environments.

Table 3: Classification into broad categories of Energy and Fuels articles from South Africa in the period 2003–2013

Category	Number of South African articles	Share of South African total
Fossil	237	31.5%
Renewable	211	28.0%
Social Sciences	85	11.3%
Direct Energy Converters	66	8.7%
Total Energy and Fuels	752	100.0%

Table 4: Number of Energy and Fuel articles from South Africa in the period 2003–2013 according to specific topics

Topic	Number of articles
Coal	135
Electricity	119
Oil or liquid fuels	113
Natural gas	19
Solar	98
Hydrogen	73
Biomas*	53
Wind	34
Hydrop*	12
Geoth*	2
Thermal insulation; lighting; double glazing; water heating; space heating	59
Efficien*	177

*indicates all forms of the word
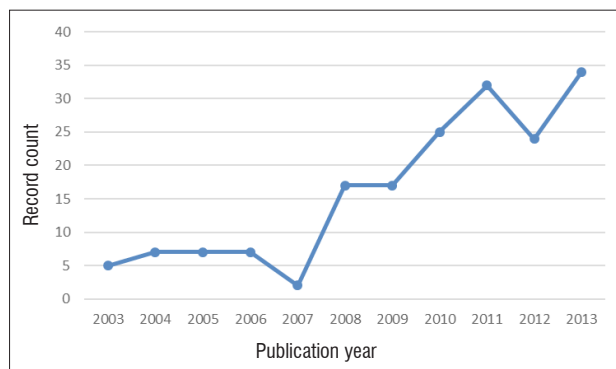
Figure 4: Growth in the number of Energy and Fuels articles related to efficiency from South Africa in the period 2003–2013.

Table 5 shows the South African organisations that are producing the majority of core energy and fuels literature. The Universities of Cape Town, Stellenbosch and Pretoria are on top of the list. Three divisions of Sasol also appear on the list. It should be noted that even the top institutions in the country produce less than 10 articles per year each on average during the period. The small number of publications makes the rankings volatile; and just a handful of publications could alter the rankings.

Table 5: South African organisations producing Energy and Fuels articles during the period 2003–2013

Organisation	Record count
University of Cape Town	107
Stellenbosch University	100
University of Pretoria	76
North West University	69
University of the Witwatersrand	66
University of KwaZulu-Natal	66
Council for Scientific and Industrial Research	59
University of the Western Cape	51
Tshwane University of Technology	46
SASOL Technology PTY LTD	27
University of Johannesburg	26
SASOL Technology RES DEV	26
Nelson Mandela Metropolitan University	24
Cape Peninsula University of Technology	19
SASOL Technology	16

The top five institutions produce more than 50% of the country's contribution to core energy and fuels literature. This result may be identified as a considerable dispersion, as a number of other scientific disciplines are concentrated in only one or two institutions in the country. For example, in the field of veterinary medicine/animal health, the University of Pretoria produces 61.68% of the country's research publications. The University of Pretoria also produces 49.15% of the country's publications in metallurgy and 46.96% in engineering mathematics.[11] As argued previously, it appears that political equity considerations in the country spill over in the research domain as well. To repeat it here, the issue is of particular developmental and research policy importance. Can a country leapfrog its science and innovation system to catch up with the rest of the world and compete internationally through a 'distributed' approach or should it concentrate its limited scientific expertise to a small number of focused research centres?
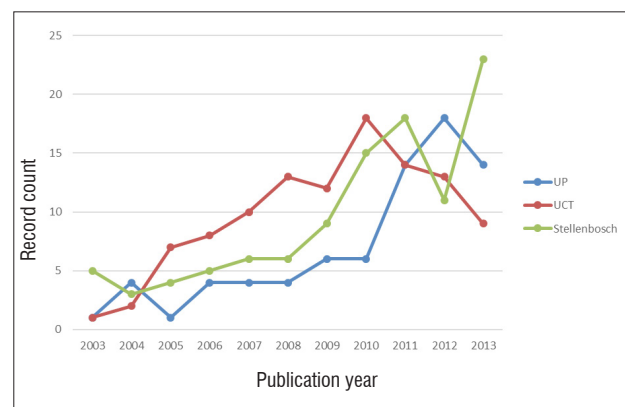
Figure 5: The number of Energy and Fuels articles produced by the Universities of Cape Town (UCT) and Pretoria (UP) and Stellenbosch University during the period 2003–2013.

Figure 5 shows the number of Energy and Fuels articles produced by the country's most prolific organisations. Again the limitation of a small number of publications should be emphasised.

Table 6 shows the number of Energy and Fuels articles produced by the most prolific organisations in India, Australia, Canada and South Africa. The South African organisations would need to increase their relevant research activity in the field fivefold in order to be comparable with those in the other countries.

Table 6: Prolific numbers of Energy and Fuels articles produced by organisations in India, Canada, Australia and South Africa during 2003–2013

Organisations	Country	Record count
India Institute of Technology	India	1560
Council of Scientific and Industrial Research India	India	1105
University of New South Wales	Australia	566
Commonwealth Scientific Industrial Research Organisation	Australia	440
University of Queensland	Australia	348
Western University of Western Ontario	Canada	575
University of Calgary	Canada	672
University of Alberta	Canada	589
University of Waterloo	Canada	435
University of Cape Town	South Africa	103
Stellenbosch University	South Africa	100
University of Pretoria	South Africa	76

Table 7: Countries collaborating with South Africa in Energy and Fuels articles

Country	Record count	% of 752
South Africa	752	
USA	59	7.846%
France	28	3.723%
People's Republic of China	25	3.324%
England	25	3.324%
Iran	20	2.660%
India	20	2.660%
Canada	18	2.397%
Nigeria	15	1.995%
Netherlands	13	1.729%
Germany	13	1.729%
Australia	13	1.729%
Malaysia	8	1.064%
Zimbabwe	6	0.798%

Table 8: Country ranking according to number of Energy and Fuels articles in 2003–2013

Rank	Country	Record count
1	People's Republic of China	33 971
2	USA	24 760
3	Japan	6280
4	Germany	6078
5	India	5721
6	Canada	5307
7	England	5120
8	South Korea	4722
9	Spain	4714
10	France	4326
11	Italy	4161
12	Australia	3579
13	Taiwan	3459
14	Iran	3127
15	Turkey	2928
16	Netherlands	2088
17	Brazil	2077
18	Sweden	1988
19	Russia	1913
20	Malaysia	1842
21	Denmark	1487
22	Norway	1341
23	Poland	1308
24	Greece	1267
25	Czech Republic	1241
26	Switzerland	1232
27	Thailand	1159
28	Mexico	1120
29	Singapore	1099
30	Portugal	1092
31	Belgium	982
32	Finland	928
33	Scotland	924
34	Saudi Arabia	793
35	**South Africa**	**752**
36	Austria	702
37	Egypt	664
38	Ireland	613
39	Algeria	547

Table 9: Comparison of the number of Energy and Fuels articles from South Africa and comparator countries in 2003–2013 according to electricity production (GWH) and population

Country	Energy articles 2003–2013	Articles/ GWH(000)	Articles/ population (million)
South Africa	752	2.9	13.9
Brazil	2078	3.7	10.2
Russia	1914	1.8	13.1
Australia	3603	14.5	152.6
Canada	5316	8.3	150
New Zealand	366	8.3	81

Table 7 shows the countries with which South African researchers collaborate for their research in the field of energy and fuels. USA, France, China and England are the top collaborating countries. It should be noted that Germany is lower down in the list (at 11th) even though it is third in collaborations with South Africa across all disciplines.[12] It should also be noted that 36.6% of the disciplinary articles were co-authored with international partners, which is lower than the country's average of 53%.[1] It can be argued that the low international collaborative activity in this field is the result of its low research activity in the country, although collaboration in the field has increased from 25% at the beginning of the period to 31% at the end of the period.

Table 8 shows the South African ranking according to the number of Energy and Fuels articles produced. South Africa is ranked 35, just above Austria and Egypt.

Table 9 shows the number of publications in the 'energy and fuels' core literature from South Africa and five comparator countries – Australia, Canada, New Zealand, Brazil and Russia – during the more recent period 2009–2013. The comparator countries are those with which South African authorities historically have compared the country.

Only New Zealand produces fewer Energy and Fuels articles than South Africa. In order to normalise the comparison, we estimated the number of energy and fuels publications per million population and per KWh of electricity produced. The two indicators provide evidence of the research that supports the population and the energy sector. Table 9 shows that South Africa compares unfavourably with the five comparator countries in terms of number of publications per KWh and it is half way in terms of articles per million population. It should be noted that South Africa is closer to Brazil and Russia and appears to be weaker than Australia and New Zealand. New Zealand appears to be stronger than South Africa in these comparisons even though it produces fewer research articles than South Africa.

Discussion and conclusion

A bibliometric assessment was undertaken of energy and fuels research in South Africa for the period 2003–2013. The major findings are summarised as follows:

- The South African national research system is producing a relatively small number of research publications in the international energy core literature. Energy research literature constitutes approximately 1% of the national output. Comparisons with the field of water research (another multidisciplinary field) support the argument that specified funding for particular research fields has the desirable effect of improving research performance.

- The number of South African energy research publications is following an increasing trend, albeit from a small basis. Furthermore, energy research is distributed across a number of universities with each organisation producing a small number of

research articles. The small number of research articles can be interpreted as a small number of relevant researchers. Hence, the move of one or two prolific researchers from one institution to another could drastically change the rankings of the institutions producing energy research. Comparisons with universities abroad indicate that the South African universities should aim to increase fivefold their energy-related publications (as well as their number of researchers in the field) if they wish to be comparable with similar institutions abroad.

- The top five most prolific institutions in the country produce more than 50% of the country's contribution to core energy and fuels literature. This spread is considered a high dispersion as a number of other scientific disciplines in the country are concentrated in only one or two institutions. It is suggested that political equity considerations in the country spill over in the research domain, which is of particular developmental and science policy importance. Can a country leapfrog its science and innovation system to catch up with the rest of the world and compete internationally through a 'distributed' approach or should it concentrate its limited scientific expertise to a few focused research centres? This issue has been identified as one of the major South African policy challenges by the Organisation for Economic Cooperation and Development in their review of the country's innovation policy.[13]

- Analysis of the specialisation patterns of energy research shows that fossil and renewable energy related research are equally emphasised in the country, albeit by a small number of articles. Topics related to energy efficiency appear to be following an ascending trend.

- Identification of the country's collaborative patterns indicates that energy researchers do not collaborate to the same extent as researchers in other research areas. It is suggested that this is a result of the limited energy research in South Africa.

- International comparisons in terms of articles per GWH of electricity produced and articles per million population show that South Africa should increase substantially its effort in the field in order to be comparable with other similar countries.

Probably the most important finding for policy is the small number of research articles in the field. The discipline's ranking is far below what was expected from a research field with a dedicated agency (Sanedi). Water research, another multidisciplinary field in the country with a dedicated agency, is ranked substantially higher than energy research. The discrepancy can be linked to Sanedi's limited budget. It is also interesting to identify whether the relevant departments are using publication outputs in order to decide priorities and funding.

The lack of independent academic energy research has critical impacts for the planning of energy in the country. In the past, it has been argued[8] that the inability of Eskom to meet electricity demand could have been prevented if there were independent researchers to argue in favour of additional power capacity. Currently, the lack of expertise in the field affects the research debate related to the need of nuclear energy in the Western Cape and the investigations related to future supply and demand for electricity in the country.

Acknowledgements

References

1. Pouris A, Ho Y-S. Research emphasis and collaboration in Africa. Scientometrics. 2014;98:2169–2184. http://dx.doi.org/10.1007/s11192-013-1156-8

2. Jacobs IM, Pouris A, Naidoo D. A scientometric examination of the performance of water research in South Africa. Water SA. 2014;40(4):631–637. http://dx.doi.org/10.4314/wsa.v40i4.8

3. Academy of Science of South Africa (ASSAf). The state of energy research in South Africa. Pretoria: ASSAf; 2014.

4. Vlachy J. World publication output in cross disciplinary physics: Materials science, physical chemistry, energy research, biophysics. Czech J Phys. 1983;33(2):247–250. http://dx.doi.org/10.1007/bf01605505

5. Pouris A, Pouris A. Structure of energy related literature: An application of citation analysis. S Afr J Sci. 1987;83:138–142.

6. Uzun A. National patterns of research output and priorities in renewable energy. Energ Policy. 2002;30(2):131–136. http://dx.doi.org/10.1016/S0301-4215(01)00084-2

7. Kostoff NR, Tshiteya R, Pfeil MK, Humenik AJ, Karypis G. Power source roadmaps using bibliometrics and database tomography. Energy. 2005;30(5):709–730. http://dx.doi.org/10.1016/j.energy.2004.04.058

8. Pouris A. Energy and fuels research in South African universities: A comparative assessment. Open Inform J. 2008;1:1–9.

9. Bradford SC. Documentation. Washington DC: Public Affairs Press; 1950.

10. Ziegler EB. Methods for bibliometric analysis of research: Renewable energy case study [ME thesis]. Boston, MA: Massachusetts Institute of Technology; 2008.

11. Pouris A. Assessing scientific strengths of academic institutions: The University of Pretoria. S Afr J Sci. 2006;102:23–26.

12. Pouris A. Swiss–SA Joint Research Programme and Germany–SA Science and Technology Cooperation Impact Study. Pretoria: Department of Science and Technology; 2014.

13. Organisation for Economic Cooperation and Development (OECD). Review of South Africa's Innovation Policy. Paris: OECD DSTI/STP; 2007.

Formulating tasks to develop HOTS for first-year calculus based on Brookhart abilities

AUTHORS:
Aneshkumar Maharaj[1] (iD)
Vivek Wagh[1]

AFFILIATION:
[1]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa

CORRESPONDENCE TO:
Aneshkumar Maharaj

EMAIL:
maharaja32@ukzn.ac.za

We describe an approach to develop higher-order thinking skills (HOTS) among first-year calculus students. The ideas formulated by Brookhart to develop HOTS were used to identify from the literature three core abilities that should be targeted. Then eight expected learning outcomes for the development of HOTS were documented, in the context of the study of first-year university calculus. Those expected outcomes were used to formulate sample tasks that were designed to target the development of the eight abilities. A pilot study was done to determine whether the tasks had the high mathematical demand envisaged. It was found that about 37% of the participants did not give any response to the tasks. Further it was found that about 31% of the participants were able to critically evaluate a given possible solution to a problem and make a value judgement. It is recommended that to promote HOTS among students, the formulation of tasks should focus on developing the following abilities: interpreting a general definition or statement in the context of a given model; translating a worded or graphically represented situation to relevant mathematical formalisms; identifying possible applications of mathematics in their surroundings; identifying linkages between groups of concepts and interpreting these linkages in the context of a model; working systematically through cases in an exhaustive way; critically evaluating one's and others' presented solutions to a problem; interpreting and extending solutions of problems; and using with reasonable skill available tools for mathematical exploration.

Significance:
- A large proportion of the student intake at university level is unable to answer mathematics questions that focus on HOTS. There is therefore a need to deliberately focus on and promote HOTS amongst the average students in the context of calculus.

Introduction and motivation

During the past few years we have had informal discussions with colleagues who lecture first-year university calculus. Those discussions gave the impression that much of their energy was and still is consumed in getting students to become comfortable with the prerequisites for calculus and the basics (skills and concepts) of calculus. The former is a result of an increasing number of students gaining entry to study university calculus. A large proportion of the student intake significantly lacks the required basic knowledge and skills to study university mathematics. The situation is so serious that some of the universities in South Africa require qualifying prospective first-year students to take a National Benchmarking Test for mathematics. However, there are some universities that refuse to implement this requirement and go strictly by the matric results. The University of KwaZulu-Natal is one such institution. This situation has resulted in a lack of focus on higher-order thinking skills (HOTS) in the context of the study of calculus. It is for these reasons that we decided to do a study on the formulation of tasks to develop among our students HOTS in the context of first-year calculus.

Research question

Our main research question was: What types of tasks could be formulated to target the development of HOTS among students enrolled for first-year university calculus? To answer this question we need to be clear on the answer to the following question: What are the possible outcomes with regard to HOTS in the context of first-year university calculus?

Review of literature

We review the literature on HOTS; expected learning outcomes; mathematical understanding; and calculus.

Higher-order thinking skills

There are a number of writings on HOTS.[1-6] Heong et al.[3] defined higher-order thinking as the expanded use of the mind to meet new challenges and noted that thinking skills are associated with the learning process. It should be noted here that the challenges should be new to the student. Karaali[4] argued that higher-order thinking behaviour is the creative formation of new knowledge based on old knowledge and the ability to apply owned knowledge to new situations. Polly and Ausband[5] noted that applying, analysing, generating, integrating and evaluating could be considered as HOTS. The study by Thompson[6] found that teachers who defined higher-order thinking as involving problem solving, discovering patterns, interpreting information, and conceptual understanding were much more likely to formulate higher-order thinking items than teachers who did not use those terms. We note that the Department of Education of Newfoundland and Labrador[2] gave a detailed document indicating the general outcomes, specific outcomes and achievement indicators for their calculus curriculum. Their examples of assessment items made use of the terms indicated above that were used to describe HOTS by the various authors mentioned. We unpack the ideas given by Brookhart[1] which we feel adequately summarises the abilities to be targeted in order to develop HOTS. HOTS involve the attainment of the following three abilities: transfer, critical thinking and problem solving. *Transfer* is conceptualised as a student's ability to relate their learning to other elements beyond those with which

they were taught to associate it. In the study of calculus, basic concepts (e.g. increasing, decreasing and concavity) related to the concept of derivatives are applied to various functions which model different situations. *Critical thinking* refers to a student being able to decide what to believe, reason, reflect and make sound decisions on their own and produce a reasoned argument. We believe that these attributes of critical thinking have subject-specific meanings and hence we will try to unpack these in the context of the expected outcomes for HOTS, in the context of calculus. *Problem solving* refers to the use of the above abilities to solve problems in different familiar and new contexts – which implies that the focus on questions in the context of calculus should include contexts that are familiar and new to the student.

Expected learning outcomes

Maharaj and Wagh[7] discussed the importance of documenting expected learning outcomes to guide the focus. We note for the reader the key issues here as relevant to this paper. Use of the term 'learning outcome' refers to a clear and detailed statement of what a student should be able to do if they have learnt the content of a particular topic. The Council of Regional Accrediting Commissions[8] (CRAC) in its discussion of student learning principles emphasised the importance of learning outcomes. With regard to learning outcomes towards which students are expected to aspire, CRAC pointed out that (1) these should be clear and easily available (made public) and (2) there should be reflection on such outcomes for a commitment to educational improvement. If one accepts these points then it follows that it is important for the learning outcomes relevant to the development of HOTS for calculus to be clearly documented, reflected on periodically and improved on if necessary, and also be made available at the outset to students. Those documented learning outcomes should guide the formulation of tasks which focus on the development of HOTS among students.

Next we looked at how some institutions addressed the issues of course outcomes and assessment for calculus. For example, a study of the University of New England[9] (UNE) course site revealed that they gave a general description of the unit and assessment tasks. Each of their assessment tasks related to a particular unit stated learning outcomes in the context of the content and graduate attributes. The latter is interesting as it gives an additional dimension to what would be required of university students. UNE listed the following five graduate attributes: (1) knowledge of a discipline, (2) communication skills, (3) information literacy, (4) problem solving and (5) team work. UNE also clearly indicated the expectation of a student for each of those attributes. For example, problem solving was unpacked as follows:

> The student will encounter in this unit a field of knowledge that is intensely problem based, and will acquire skills in connecting ideas within a network of logical relationships. A high emphasis will be placed on the development of analytical and deductive reasoning.[9(p.4)]

We note that for each of these five graduate attributes there was the intention for teaching, assessment and practise. In the context of the study of mathematics, we believe that a clear perception of the attributes of HOTS will contribute to the development of *problem solving* skills.

Mathematical understanding

A focus on mathematical understanding requires that one first answer the question: What is mathematics? In answering this question we found the view of Godino[10] to be useful. This view – which is also discussed by Maharaj[11] – is based on the following four assumptions which we summarise for the reader: (1) Mathematics is a human activity involving the *solution of problematic situations*. The problem situations could be external or internal. As responses or solutions to such problems are found and reflected on, mathematics emerges and evolves. (2) Mathematical problems and their solutions are shared in *specific institutions or collectives* involved in studying such problems. An example of such an institution is a university at which modules are offered that focus on the study of mathematics. (3) Mathematics is a *symbolic language* in which problem situations and the solutions found

are expressed. It is this symbolic language which represents coded information that allows for the communication of problems and possible solutions among those who study mathematics. This implies that the teaching and learning of mathematics should also focus on the study of this symbolic language. (4) Mathematics is a *logically organised conceptual system*. When a mathematical concept is accepted as a part of this system, it is considered as a textual reality and a component of the global structure. The implication here is that those who teach and want to study mathematics should adhere to the logically organised conceptual system. Suppose that one is required to analyse a function to provide a graphical representation. Then, to indicate the logic or thinking involved in deducing the key characteristics, explanations should be given and use made of connectives to link symbolic representations of mathematical concepts with mathematical symbols such as '∴' (therefore), '⟹' (implies) and '⟺' (implies and is implied by).

We now focus on what is meant by mathematical understanding. Skemp[12] identified two types of understanding: (1) *relational understanding* which he described as knowing what to do and why and (2) *instrumental understanding* which he described as rules without understanding. He noted that the process of learning relational mathematics leads to the building of a conceptual structure in mathematics. This is included in the logically organised conceptual system to which Godino[10] referred. Our opinion is that the focus in the teaching and learning of mathematics should be on relational understanding. The development of relational understanding requires that the focus should be on the three HOTS abilities identified by Brookhart[1]: transfer, critical thinking and problem solving. Suppose one is required to sketch the graph of the function $f(x) = \frac{x^2 - 16}{|x - 4|}$. A focus on the structure represented by the symbolic language implies that one should consider cases based on the structure of the denominator $|x-4|$, which requires an understanding of the absolute value concept. Because the denominator cannot be 0, the following cases need to be considered: $x - 4 > 0$ and $x - 4 < 0$. Each of these cases results in different implications. In the context of the Brookhart[1] abilities for HOTS one could view the sketching of the graph of f as targeting *problem solving* as this would include the abilities of: (1) *transfer* of learning on the absolute value concept to other elements beyond which students were taught to associate it with; and (2) *critical thinking* as the students would be required to reason, reflect, make sound decisions and produce a reasoned argument to arrive at the required graph of $f(x) = \frac{x^2 - 16}{|x - 4|}$.

Calculus

A prerequisite to study calculus is adequate knowledge and skills relating to algebra. This knowledge should include algebraic aspects of functions and their standard forms. The derivative of a function and integral of a function are both key concepts in the study of calculus. For details on some studies on students' understanding of the concept of a derivative of a function the reader could refer to Maharaj[13]. The review of relevant literature revealed that some of the main points that relate to teaching and learning of calculus, in particular the derivative of a function, could be summarised in five points. Firstly, understanding the concept of the derivative of a function is difficult for many students.[14,15] Secondly, one should be careful when distinguishing between a *description* of this concept (which specifies some properties of that concept) and the formal concept definition[16] of the derivative of a function. For example, a *description* of the derivative of a function, say $f'(x)$, could be:

the gradient of the function $f(x)$ at x_0 is the slope of the tangent line to the curve f at the point $(x_0, f(x_0))$.

The formal definition could be represented as:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}.$$

Thirdly, the understanding of students could be improved if one exposes them to several representations of the derivative.[17] Growth in understanding can be promoted by a variety of connections, both between and within representations, and also between a physical application and mathematical representations.[18] The teaching implication here is that

there should be a focus on representations and their relevant connections, as part of understanding derivatives. Fourthly, students seem to prefer graphical representation in tasks and explanations about derivatives.[19] This finding was supported by Tall[20] who argued that direct links between visualisation and symbolisation should be emphasised when teaching the concept of a derivative. Lastly, students' difficulties with the derivative increase in number and extent when the function they are confronted with is a composite function.[21] Difficulties in dealing with composition and decomposition of functions could be the reason the chain rule is one of the most difficult ideas in calculus to convey to students.[22,23]

A number of studies have focused on student understanding of the concept integration of functions and also on what could be done to improve student understanding.[24-29] It could be concluded from those studies that difficulties encountered by students in their understanding of the integration concept relate to two principle causes. Firstly, differentiation (finding the derivative of a function) could be viewed as a *forward* process in contrast to the reverse or *backward* process of integration. So any difficulties that students have with differentiation could compound their understanding of integration. Further the difficulties faced by students with regard to the concept of derivatives are not as complicated as those for the process of integration. Secondly, integration has a dual nature because it is both the inverse process of differentiation and a tool for calculation, for example, when required to determine the area or volume in the context of problem situations. The first teaching and learning implication from the above studies is that when introducing the antiderivative (improper integral) of a function this should be related to the concept of the derivative. This means that if $\int g(x)dx$ represents the general antiderivative of $g(x)$ then $\int g(x)dx = G(x) + C$ provided $G'(x) = g(x)$. Further, for the concept of the definite integral of a function, teaching should focus on the development of the spatial visualisation ability among students. The development of this ability could influence and strengthen the relationship between the graphical and the symbolic integral representation. Sevimli and Delice[27] noted that such an approach increases the performance of students when solving definite integral problems. So problems based on the visualisation of what a particular definite integral represents, could develop among students a useful strategy to conceptualise different definite integral contexts with which they are confronted. It should be noted that only when a student has a deeper understanding of the structure of the definite integral then the use of 'area under a curve' is helpful in problem solving.[30] This reinforces the assumption by Godino[10] that mathematics is the study of a symbolic language. These ideas influenced the formulation of the tasks that appear later in this paper.

Conceptual framework

The literature review[1-19] as well as our experience in teaching calculus to students and assessing their abilities in the context of calculus, guided the formulation of principles which give an overview of the conceptual framework for this study. In particular, the formulation of the following four principles were informed by the work of Brookhart[1], Maharaj and Wagh[7], the Council of Regional Accrediting Commissions[8], Godino[9] and Skemp[12]. (1) It is necessary to formulate and document expected HOTS outcomes on which teaching and learning should focus.[1,7] Those outcomes should be easily available to lecturers/teachers and students.[8] (2) The identified learning outcomes should inform the development of tasks[7] that help in the development of abilities related to HOTS.[1] (3) The key abilities relate to the transfer of knowledge and skills, critical thinking and problem solving[1] across sections in mathematics relevant to the context of studying calculus. The attainment of these should foster the development of relational understanding in students.[12] (4) HOTS are acquired and refined by practising them in the context of several different representations of the same calculus related concept.[17-20] So students need to be exposed to tasks that could help them develop the abilities related to HOTS. This implies that lecturers/tutors should plan and implement tasks that focus on those relevant abilities.

We now focus on the planning stage of sample tasks that could lead to the development of HOTS among students in the context of their study

of calculus. We undertook a brief pilot study to determine whether the sample tasks had the high mathematical demand that we envisaged.

Methodology

The general literature review led us to focus on abilities required for the development of HOTS identified by Brookhart[1]. We then unpacked those abilities in the context of HOTS required for the study of calculus, keeping in mind the main conclusions of past studies on the teaching and learning of calculus, especially those abilities that could foster relational understanding. A tabulation of those abilities and their relationship with the abilities of transfer, critical thinking and problem solving is given in Table 1. In our opinion, the abilities identified in column two of Table 1 are generic to mathematics. We then used those generic abilities to document the expected outcomes for HOTS in the context of calculus that we felt should be developed among our first-year students. Those expected outcomes then guided the framing of sample tasks that we felt could develop the identified HOTS among students.

Table 1: Unpacking of Brookhart's[1] abilities to develop higher-order thinking skills in the context of calculus

Brookhart abilities	Abilities in the context of calculus
Transfer	Interpret a general definition or statement in the context of a given model
	Translate a worded or graphically represented situation to relevant mathematical formalisms
	Identify possible applications of mathematics in their surroundings
Critical thinking	Identify linkages between groups of concepts and interpret these linkages in the context of a model
	Work systematically through cases in an exhaustive way
	Critically evaluate their and others' presented solutions to a problem/question
	Interpret and extend solutions of problems
Problem solving	Use of above abilities in problem-solving contexts, both familiar and unfamiliar
	Use with reasonable skill available tools for mathematical exploration

We then did an empirical pilot study on student responses to the tasks developed. That pilot study was conducted (in the month of August 2015) at a representative college, where average students seek admission to an undergraduate mathematics course, in the city of Nagpur in India. The institution also runs a postgraduate programme in mathematics. Permission to conduct the pilot study was obtained from the Principal of the college. The pilot study was part of the work required for the project 'Online diagnostics for undergraduate mathematics'. Ethical clearance for that project was provided by the Humanities and Social Sciences Research Ethics Committee of the University of KwaZulu-Natal (reference number HSS/1058/014CA).

The participants were 48 students, all of whom had exposure to the concepts that were covered by the tasks in the pilot study. There were 26 female students and 22 male students. The participants volunteered and were selected on the basis of their availability. Every question was explained before the students attempted the tasks for which an hour was given. The seven tasks are indicated in the sub-section 'Sample tasks for HOTS' below. We indicate for the reader what was clarified with the students before they attempted these seven tasks:

- Tasks 1 and 2: The concepts input and output were illustrated using the function of $p(x) = x^2$

- Task 3: The meaning of second derivative was made clear

- Task 4: What is meant by relevant mathematical formalism was demonstrated by focusing on 'two distinct numbers are added'

- Task 5: It was pointed out to the students how knowledge of trigonometry is used for finding height of a building

- Task 6: The fractional part of a number was demonstrated by discussing the example {1.123}=0.123

- Task 7: It was explained that the students need to identify if the given solution is correct or not and they should provide a reason for their judgement.

All of the participants submitted their attempts within 50 min. Their written responses were then looked at and sorted into the following categories: no response for all seven tasks; some written response for a task; correct responses; partially correct responses. The latter category was used to denote responses in which it was clear that the student reasoned correctly but the response was either incomplete or had a mathematical error.

Findings and discussion

Possible outcomes for HOTS

Our formulation of expected learning outcomes is that students should be able to: interpret a general definition or statement in the context of a given model (given concrete situation); work systematically through cases in an exhaustive way; identify linkages between groups of concepts and interpret these linkages in the context of a model (given concrete situation); translate a worded situation to relevant mathematical formalisms; translate a graphically represented situation to relevant mathematical formalisms; identify possible applications of mathematics in their surroundings; critically evaluate one's and others' presented solutions to a problem/question by identifying errors, redundancies, alternative solutions and how a solution could be improved; interpret and extend solutions of problems; and use with reasonable skill available tools for mathematical exploration (as a general consensus expected from the student). These tools include algebraic and trigonometric identities; standard limits and laws of limits; standard derivatives and their laws; and standard anti-derivatives and their laws.

Sample tasks for HOTS

The questions that we formulated to target the development of HOTS abilities indicated in the above outcomes are documented in the following sample tasks. We indicate next to each of these sample tasks the identified HOTS expected outcome that the task was designed to develop.

Task 1: Ability to work systematically through cases in an exhaustive way
How many polynomial functions are there whose coefficients are natural numbers and in which the output of their derivative is less than 10 on the restricted domain [0,1]? Justify your answer.

Task 2: Ability to interpret and systematically extend solutions to problems
How many polynomial functions are there whose coefficients are natural numbers and in which the output of their derivative is less than 20 on the restricted domain [0,10]? Justify your answer.

Task 3: Ability to interpret a general definition or statement in the context of a given model
Statement: If $f''(x)>0 \ \forall \ x \in I$ then the graph of f is concave upwards on the interval I.

Is the graph of the function defined by $f(x)=1+x^2+x^3$ concave upwards on the interval [-1,1]? Justify your answer.

Task 4: Ability to translate a worded situation to relevant mathematical formalisms
Find a function with domain the entire set of real numbers which satisfies the condition: the output at the sum of any two inputs is the sum of the outputs at those respective inputs.

Task 5: Ability to identify possible applications of mathematics in their surroundings
Identify at least five applications of the derivative in the context of your body. Explain how each is an application of the derivative.

Task 6: Ability to identify linkages between groups of concepts and interpret these linkages in the context of a model
Let x be a non-negative real number and $\{x\}$ denote the fractional part of x. For natural number n find the integral $\int_0^n \{x\}dx$.

Task 7: Ability to critically evaluate a presented solution to a problem/question
Examine critically the following question and a solution presented by a student. Indicate clearly with justification your comments on the presented solution. Also present your alternative solution to the question.

Question: Evaluate the integral $\int x \sin x \ dx$

Solution: $\int x \sin x \ dx = \int x \ dx \int \sin x \ dx = \frac{x^2}{2}\cos x+c$

Most (if not all) of the above tasks should be a challenge to the average student studying first-year calculus. Additional sample tasks that we formulated to target the development of the identified HOTS expected outcomes are presented in Appendix 1 of the supplementary material. Note that the seven sample tasks and the additional sample tasks given in Appendix 1 illustrate how the eighth expected outcome could be achieved. Those sample tasks illustrate how available tools could be used to develop reasonable skill in the context of mathematical exploration.

Pilot study on sample HOTS tasks

In Table 2 a summary of the students' responses to the seven tasks is given. It was noted that 18 students gave a blank sheet; meaning there were no responses from those students for any of the seven tasks. An informal chat with those students indicated that they were unable to respond to a task. When probed further, students gave the following as the main reasons for their lack of responses: they were never given a method to solve such questions and they had forgotten some of the things that they had studied. The first reason suggests that many of those participants were comfortable only when they were exposed to some sort of a method to deal with problems. It seemed that these students were exposed to only routine problems on which they worked according to the method they were exposed to. Note that Task 4 was not a calculus task, but focused on mathematical symbolisms relating to the concept of a function – which is a prerequisite to study calculus.

Table 2 indicates that all 48 participants lacked five of the seven HOTS abilities on which the tasks focused. These five HOTS abilities were: working systematically through cases in an exhaustive way; interpreting and systematically extending solutions to problems; interpreting a general definition or statement in the context of a given model; translating a worded situation to relevant mathematical formalisms; and identifying linkages between groups of concepts and interpreting identified linkages in the context of a model. It is also evident from Table 2 that only about 31% of the participants was able to critically evaluate and make a value judgement on a possible solution to a problem. Furthermore, only about 8% of the respondents was able to identify possible applications of mathematics in their surroundings; in this case, applications of the derivative in the context of their body. The implication from the pilot study is that the formulated tasks had the high mathematical demand in calculus that we envisaged.

Table 2: Summary of student (*n*=48) responses to the seven tasks

Task	Higher-order thinking skills outcome	Number of respondents	Number of correct responses	Number of partially correct responses
1	Ability to work systematically through cases in an exhaustive way	2	0	0
2	Ability to interpret and systematically extend solutions to problems	3	0	0
3	Ability to interpret a general definition or statement in the context of a given model	9	0	0
4	Ability to translate a worded situation to relevant mathematical formalisms	3	0	0
5	Ability to identify possible applications of mathematics in their surroundings	4	0	4
6	Ability to identify linkages between groups of concepts and interpret these linkages in the context of a model	11	0	0
7	Ability to critically evaluate a presented solution to a problem/question	25	15	0

Conclusions

We identified from the literature review the following three abilities that are central to the development of HOTS among first-year calculus students: transfer of knowledge and skills across sections, critical thinking, and problem solving. Based on the general literature review on critical thinking and a review of literature on the teaching and learning of calculus, we were able to document expected learning outcomes for the development of HOTS in the context of the study of first-year university calculus. Those expected outcomes were used to formulate sample tasks. The sample tasks were designed to target the development of abilities identified in the documented expected outcomes for HOTS, in the context of calculus. The pilot study indicated that the HOTS abilities that we identified for calculus, which are also generic to mathematics, were seriously lacking among the participants who were average students who had been previously exposed to the concepts on which the tasks were based.

Our next investigation will focus on using the sample tasks formulated to further research the development of HOTS among our students. These students could be South African as all first-year university calculus courses cover the same content, namely the derivatives and integrals of standard functions and their applications. In particular, we want to explore the use of technology to promote the identified HOTS abilities that the study indicates should be developed among students. We encourage others who may be interested to feel free to use or adapt the sample tasks to conduct research. We recommend that the formulation of tasks should focus on developing the following HOTS abilities: interpreting a general definition or statement in the context of a given model; translating a worded or graphically represented situation to relevant mathematical formalisms; identifying possible applications of mathematics in their surroundings; identifying linkages between groups of concepts and interpreting these linkages in the context of a model; working systematically through cases in an exhaustive way; critically evaluating one's and others' presented solutions to a problem; interpreting and extending solutions of problems; using these abilities in both familiar and unfamiliar problem-solving contexts; and using with reasonable skill the tools available for mathematical exploration.

Acknowledgements

Authors' contributions

This paper was conceptualised by both authors. All sections were formulated and drafted jointly. A.M. wrote the manuscript.

References

1. Brookhart SM. How to assess higher order thinking skills in your classroom. Alexandria, VA: ASCD; 2010.

2. Department of Education of Newfoundland and Labrador. Mathematics calculus 3208: Curriculum guide 2013 [document on the Internet]. c2013 [cited 2014 Jan 19]. Available from: http://www.ed.gov.nl.ca/edu/k12/curriculum/guides/mathematics/Calculus/Calculus_3208_curriculum_guide.pdf

3. Heong YM, Yunos JBM, Hassan RB, Othman WB, Kiong TT. The perception of the level of higher order thinking skills among technical education students. In: Proceedings of the 2011 International Conference on Social Science and Humanity; 2011 February 26–28; Singapore. Singapore: IACSIT Press; 2011. p. 281–285. Available from: http://www.ipedr.com/vol5/no2/62-H10167.pdf

4. Karaali G. An evaluative calculus project: Applying Bloom's taxonomy to the calculus classroom. PRIMUS. 2011;21(8):721–733. http://dx.doi.org/10.1080/10511971003663971

5. Polly D, Ausband L. Developing higher-order thinking skills through webQuests. J Comput Teach Educ. 2009;26(1):29–34. Available from: http://files.eric.ed.gov/fulltext/EJ856114.pdf

6. Thompson T. Mathematics teachers' interpretation of higher-order thinking in Bloom's taxonomy. Int Elect J Math Educ. 2008;3(2):96–109. Available from: http://www.iejme.com/022008/d2.pdf

7. Maharaj A, Wagh V. An outline of possible pre-course diagnostics for differential calculus. S Afr J Sci; 2014;110(7/8):27–33. http://dx.doi.org/10.1590/sajs.2014/20130244

8. Council of Regional Accrediting Commissions. Regional accreditation and student learning: A guide for institutions and evaluators. Atlanta, GA: Southern Association of Colleges & Schools; 2004. Available from: http://www.sacscoc.org/pdf/handbooks/GuideForInstitutions.PDF

9. University of New England. MATH101 Algebra and differential calculus [homepage on the Internet]. c2013 [cited 2013 Feb 02]. Available from: http://www.une.edu.au/courses/units/MATH101

10. Godino JD. Mathematical concepts, their meanings, and understanding [document on the Internet]. c1996 [cited 2013 Aug 02]. Available from: http://www.ugr.es/~jgodino/articulos_ingles/meaning_understanding.pdf

11. Maharaj A. A framework to gauge mathematical understanding: A case study on linear algebra concepts. Int J Educ Sci. 2015;11(2):144–153.

12. Skemp R. Relational understanding and instrumental understanding. Math Teach. 1976;77:20–26.

13. Maharaj A. An APOS analysis of natural science students' understanding of derivatives. S Afr J Educ. 2013;33(1):146–164. http://dx.doi.org/10.15700/saje.v33n1a458

14. Orton A. Students' understanding of differentiation. Educ Stud Math. 1983;14:235–250. http://dx.doi.org/10.1007/BF00410540

15. Uygur T, Özdaş A. Misconceptions and difficulties with the chain rule. In: Proceedings of the Mathematics Education into the 21st Century Project: Reform, Revolution and Paradigm Shifts in Mathematics Education; 2005 Nov 25– Dec 01; Johor Bahru, Malaysia. Johor Bahru: Malaysia University of Technology; 2005. Available from: http://math.unipa.it/~grim/21_project/21_malasya_Uygur209-213_05.pdf

16. Giraldo V, Carvalho LM, Tall DO. Descriptions and definitions in the teaching of elementary calculus. In: Pateman NA, Dougherty BJ, Zilliox J, editors. Proceedings of the 27th Conference of the International Group for the Psychology of Mathematics Education; 2003 July 13–18; Honolulu, HI, USA. Honolulu, HI: Center for Research and Development Group, University of Hawaii; 2003. p. 445–452. Available from: http://www.warwick.ac.uk/staff/David.Tall/pdfs/dot2003d-giraldo-carv-pme.pdf

17. Hähkiöniemi M. Perceptual and symbolic representations as a starting point of the acquisition of the derivative. In: Høines MJ, Fuglestad AB, editors. Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education; 2004 July14–18; Bergen, Norway; 2004. p. 73–80. Available from: https://www.emis.de/proceedings/PME28/

18. Roorda G, Vos P, Goedhart M. Derivatives and applications: Development of one student's understanding. In: Durand-Guerrier V, Soury-Lavergne S, Arzarello F, editors. Proceedings of CERME 6; 2009 Jan 28–Feb 01; Lyon, France. Lyon: Institut National de Recherche Pédagogique; 2010. p. 2296–2305. Available from: www.inrp.fr/editions/cerme6

19. Zandieh MJ. A theoretical framework for analyzing student understanding of the concept of derivative. CBMS Issues in Mathematics Education. 2000;8:103–122. http://dx.doi.org/10.1090/cbmath/008/06

20. Tall D. A sensible approach to the calculus. Plenary presented at: The National and International Meeting on the Teaching of Calculus; 2010 September 23–25; Puebla, Mexico. Available from: http://homepages.warwick.ac.uk/staff/David.Tall/downloads.html

21. Tall D. Students' difficulties in calculus. Plenary presented at: ICME-7: Proceedings of Working Group 3 on students' difficulties in calculus; 1992 August 17–23; Québec, Canada. Québec: ICME; 1993. p. 13–28. Available from: http://homepages.warwick.ac.uk/staff/David.Tall/downloads.html

22. Clark JM, Cordero F, Cottrill J, Czarnocha B, DeVries DJ, St. John D, et al. Constructing a schema: The case of the chain rule. J Math Behav. 1997;1:345–364. http://dx.doi.org/10.1016/S0732-3123(97)90012-2

23. Gordon SP. Discovering the chain rule graphically. Math Comput Educ. 2005;39:195–197.

24. Abdul-Rahman S. Learning with examples and students' understanding of Integration. In: Proceedings of the Mathematics Education into the 21st Century Project: Reform, Revolution and Paradigm Shifts in Mathematics Education; 2005 Nov 25–Dec 01; Johor Bahru, Malaysia. Johor Bahru: Malaysia University of Technology; 2005. Available from: http://math.unipa.it/~grim/21_project/21_malasya_AbdulRahman24-28_05.pdf

25. Ubuz B. Students' understanding of differentiation and integration [document on the Internet]. c1993 [cited 2011 Oct 25]. Available from: http://www.bsrlm.org.uk/IPs/ip13-3/BSRLM-IP-13-3-12.pdf

26. Orton A. Students' understanding of integration. Educ Stud Math. 1983;14:1–18. http://dx.doi.org/10.1007/BF00704699

27. Sevimli E, Delice A. The influence of teacher candidates' spatial visualization ability on the use of multiple representations in problem solving of definite integrals: A qualitative analysis [document on the Internet]. c2010 [cited 2011 Oct 25]. Available from: http://www.bsrlm.org.uk/IPs/ip30-2/BSRLM-IP-30-2-10.pdf

28. Haciomeroglu ES, Aspinwall L, Presmeg N. The role of reversibility in the learning of the calculus derivative and antiderivative graphs. In: Swars SL, Stinson DW, Lemons-Smith S, editors. Proceedings of the 31st annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education; 2009 Sep 23–26; Atlanta, GA, USA. Atlanta, GA: Georgia State University; 2009. p. 81–88.

29. Maharaj A. An APOS Analysis of natural science students' understanding of integration. REDIMAT. 2014;3(1):54–73.

30. Sealey V. Definite integrals, Riemann sums, and area under a curve: What is necessary and sufficient? In: Alatorre S, Cortina JL, Sáiz M, Méndez A, editors. Proceedings of the 28th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education; 2006 Nov 9–12; Mérida, Yucatán, México. Mérida: Universidad Pedagógica Nacional; 2006. p. 46–53.

A spatio-temporal analysis of fires in South Africa

AUTHORS:
Sheldon Strydom[1] iD
Michael J. Savage[1] iD

AFFILIATION:
[1]School of Agricultural, Earth and Environmental Sciences, University of KwaZulu-Natal, Pietermaritzburg, South Africa

CORRESPONDENCE TO:
Michael J. Savage

EMAIL:
savage@ukzn.ac.za

The prevalence and history of fires in Africa has led to the continent being named 'the fire continent'. Fires are common on the continent and lead to a high number of annual fire disasters which result in many human fatalities and considerable financial loss. Increased population growth and concentrated settlement planning increase the probability of fire disasters and the associated loss of human life and financial loss when disasters occur. In order to better understand the spatial and temporal variations and characteristics of fires in South Africa, an 11-year data set of MODIS-derived Active Fire Hotspots was analysed using an open source geographic information system. The study included the mapping of national fire frequency over the 11-year period. Results indicate that the highest fire frequency occurred in the northeastern regions of South Africa, in particular the mountainous regions of KwaZulu-Natal and Mpumalanga, and in the Western Cape. Increasing trends in provincial fire frequency were observed in eight of the nine provinces of South Africa, with Mpumalanga the only province for which a decrease in annual fire frequency was observed. Temporally, fires were observed in all months for all provinces, although distinct fire seasons were observed and were largely driven by rainfall seasons. The southwestern regions of South Africa (winter-rainfall regions) experienced higher fire frequencies during the summer months and the rest of the country (summer-rainfall regions) during the winter months. Certain regions – those which experienced bimodal rainfall seasons – did not display distinct fire seasons because of the complex wet and dry seasons. Investigation into the likely effects of climate change on South African fire frequency revealed that increased air temperatures and events such as La Niña have a marked effect on fire activity.

Significance:

- Fires have played a significant role in the morphology of the African continent.

- Fires provide a number of environmental services.

- Fires were observed in all months in all provinces in South Africa, although distinct fire seasons were observed and were largely driven by rainfall seasons.

- Global climate change will result in an increase in the frequency of fires.

Introduction

The African continent has for years been termed 'the fire continent' as a result of the high number of fires it experiences.[1] The development of advanced remote-sensing technologies has allowed the study of fires at a greater spatial resolution, in the context of vegetation communities and topography, and, more recently, in the context of climate change.[1] Remote sensing has also allowed for the study of fires over larger regions compared to the highly localised research areas such as the Kruger National Park and other national parks.

Fires in South Africa

Fire has always and will always be a natural and important phenomenon in environmental systems. However, because of an increase brought on through anthropogenic activities, fires are having a negative impact on the environment and more so on society and the economy. The likelihood of an 'accident' fulfilling the requirements to be classified as a disaster is increasing as a result of increased population densities and increased settlement in high-risk areas.[2] Fire disasters are of great concern in South Africa and one can conclude that these disasters are going to increase.[2] A large percentage of South Africa's population is located in rural areas, where they often are housed in close quarters, which allows fires to spread rapidly through housing structures. These rural areas are also generally situated in fire-prone regions of the country, making them vulnerable to fires.

A total of about 35 000 fires was reported during the 2008 fire season in South Africa, with a majority of fires originating from open flames during waste, grass or bush burnings; these fires led to close to 380 fatalities and more than ZAR2.3 billion in financial losses.[3] During the 2009 fire season, more than 40 000 fires were reported, and resulted in 376 fatalities and ZAR4 billion in financial loss – again with open flames during waste, grass or bush burnings identified as the source of most of the fires.[4] The 2010 fire season experienced similar statistics with a high number of fatalities caused by fires and a large financial loss.[5] These statistics clearly indicate the potential of fires to result in both loss of life and financial loss. The South African *National Veld and Forest Fire Act of 1998 (Act 101 of 1998)* specifies the prevention of wildfires through the implementation of a National Fire Danger Rating System under the responsibility of the Department of Water Affairs and Forestry. The National Fire Danger Rating System is currently operational and is being used by the South African Weather Service and other interested parties to mitigate wildfire outbreaks. Under the Act, fire prevention is considered the responsibility of the landowner and lack of regional coordination is visible.[6] While regional coordination is lacking, a number of regional fire prevention or protection agencies has been established. These agencies comprise mostly private landowners and agro-forestry managers working as 'umbrella' fire protection associations.[6] The *National Veld and Forest Fire Act of 1998* promotes the formation of these regional fire protection associations and requires all landowners to be members of local fire protection associations but coordination between different fire protection associations is minimal.[7]

Adverse effects of fires

Fire is associated with atmospheric pollution largely because of the release of carbon monoxide and ozone during biomass burning as well as emission of carbon dioxide.[8] Smoke haze has been linked to human respiratory diseases and poses a concern for agricultural crop productivity. The burning of biomass attributed to wildfires has a direct impact on climate change. Burning of biomass releases high concentrations of oxocarbons which act as a greenhouse gas by absorbing and re-emitting infrared radiation.[9] Black carbon has also been linked to positive feedback mechanisms when deposited on surfaces such as snow where it alters the surface radiation balance and hence the energy balance. Altering the surface radiation balance leads to a warmer temperature on the snow surface which leads to snowmelt which further exposes darker surfaces, leading to further warming.[9] It is acknowledged that fire is a necessity when used as a land surface management tool. However, by identifying areas prone to wildfires, one may, at the same time, identify areas that contribute significantly to atmospheric pollution and focus mitigation resources on these identified areas. Of concern is when a highly populated area experiences high fire frequencies, as the byproducts of the fires may adversely affect those populations.

Fire also has the ability to enhance or increase soil erosion. Particularly severe fires have been shown to remove surface vegetation and influence soil organic material and peat. This loss of vegetation effectively decreases the surface's ability to infiltrate surface runoff after a precipitation event.[10,11] Increased surface runoff then results in increased soil removal, but may also result in flash floods or increased sedimentation of water systems. Soil erosion may be exacerbated by fire activity when the area has already been overgrazed. Fires on mountain slopes have been shown to decrease infiltration by as much as 50%, thereby increasing the catchment's risk of flash flooding.[11] Much of South Africa is vulnerable to severe land degradation because of improper agricultural practices such as overgrazing.[10,12] Degradation in already vulnerable areas may be exacerbated through increased fire activity. For a province like the Eastern Cape, which is notorious for its severe soil erosion, further degradation could be mitigated or limited by identifying areas prone to fires and assessing the role of fires in removing land cover in those areas. Mountainous regions may also be vulnerable because of the steepness of slopes and the high fire frequencies associated with Berg winds. By identifying these areas, managers may be able to combat further degradation by limiting fire activity and ensuring the prevention of uncontrollable fires.

Need for a detailed spatio-temporal analysis of South African fires

Despite progress in fire mitigation and management, the country still experiences many fire disasters annually. Previous research has focused on the spatio-temporal characteristics of fires in protected areas which cannot be extrapolated to the rest of the country with reasonable accuracy.[1] A study of the spatio-temporal characteristics of South African fires is lacking, with only a few local studies having been published.[1,13,14] The mapping of areas prone to hazards is a common practice in disaster management and aids in identifying 'at risk' areas as well as providing insight to the spatial and temporal changes of natural hazards. A detailed spatio-temporal study can be used to identify areas within the country that are prone to fire activity and to identify times when fire activity is at its highest, with temporal scales ranging from months to years.

Motivation for the study

Fire has played a significant role in shaping the African landscape. Fires do indeed provide a number of environmental services, but increased anthropogenic fire activity may result in adverse effects such as loss of human life and property, enhanced environmental degradation and increased greenhouse gas concentrations. In order to mitigate the loss of human life and other adverse effects of wildfires, a spatio-temporal analysis of South African fires was undertaken to provide landowners and critical decision-makers with a better insight into the temporal and spatial variations of fires in South Africa.

Methodology

Data acquisition

One of the challenges faced when initiating the study was the lack of a national database of fires listing times and locations of fires. As no data set was available from government or independent research groups, it was decided to acquire raw data through the US National Aeronautics and Space Administration's (NASA) Earth Observing System Data and Information System (EOSDIS). The EOSDIS allows a registered user to download data in a specified format for a specified area during a user selected time frame.

Data characteristics

Archived fire data were accessed through NASA's Fire Information for Resource Management System (FIRMS). Based at the University of Maryland, FIRMS distributes fire hotspot information to a number of countries.[15] FIRMS data are obtained from the Moderate Resolution Imagining Spectrometer (MODIS) instruments aboard NASA Rapid Response satellites.[15] FIRMS incorporates remote sensing and geographic information system (GIS) technologies to produce MODIS fire data. MODIS fire data are provided in three unique ways: through an online mapping interface, through customised email alerts and through text messaging.[16] For this study, only Active Fire Data were used. The Active Fire product contains contextual algorithms which use data obtained from the mid-infrared (3.929–3.989 μm) and thermal infrared (10.780–11.280 μm) wavelengths as well as a 'fire radiative power parameter'. Products are generated daily at full resolution and plotted to 0.5° grids.[16]

Because the time span of the data set needed exceeded 8 days, archived data were needed. Using the EOSDIS Archive Data Tool, a polygon was inserted surrounding the area of interest – in this case South Africa. In order to select the entire South African land surface, a polygon larger than South Africa was drawn. Once an area of interest was selected, the calendar tool was used to select a time frame of 11 years, spaning from 1 January 2003 to 31 December 2013.

Limitations

As stated previously, one of the main challenges was acquiring a database containing national fire point locations with a temporal analysis. Reaching out to government departments and private research institutes proved fruitless; a database was therefore acquired through NASA's EOSDIS. Validation of archived remote-sensing data is a known challenge. Ground truthing or validation of fires detected by the MODIS satellites as far back as 2003 is impossible – which raises questions on the accuracy of the data used. A number of methods exist to validate satellite-derived fire products.[15] These methods require extensive field work shortly after the detection of a fire – essentially pointless when a data set is older than a few weeks. While ground truth validation was not possible in this study, the Active Fire product was validated using sample locations representative of the general area's vegetation, i.e. hotspots detected in South Africa were validated using a tropical savanna/grassland biome to mask out hotspots created by water bodies, mines and power stations.[16]

Data processing

Data were downloaded using the EOSDIS Archive Download Tool. For convenience, a shapefile format was selected to display the total number of fires detected within the area of interest between 2003 and 2013. The shapefile included an attribute table with information regarding time of detection, date of detection, location of hotspot, brightness of hotspot as well as a confidence level. The shapefile containing vector points of hotspots was projected in QGIS® using the Hartebeeshoek94 coordinate reference system. All shapefiles added thereafter to QGIS were projected using the Hartebeeshoek94 coordinate reference system to ensure spatial accuracy. The total hotspot shapefile was clipped to the borders of South Africa, eliminating any detected hotspots that were not within South Africa. To ensure the study focused on hotspots created by fires, only points with a confidence level of 100% were selected for analysis.

Analysing national fire distribution required the attribute table of 100% confidence level fires to be exported to Microsoft Excel, in which both univariate and bivariate descriptive statistics were applied. The QGIS was used to create a map illustrating the total distribution of 100% confidence level fires in the country between 2003 and 2013 which was overlaid with a shapefile of South African provincial boundaries. A shapefile of vegetation zones was added to the QGIS project to analyse fire distribution as a variable of vegetation. In order to calculate the total number of fires in each vegetation class, the QGIS Spatial Query tool was used. In order to better understand the role of vegetation in fires, the 11-year fire point data set was combined with a 1:50 000 vegetation shapefile which included 67 different vegetation classes. Fire points in each vegetation class were counted. This was done for all vegetation classes and then exported to Microsoft Excel for further investigation. Analysing provincial fire distribution used a similar method. Fire totals per province were calculated using the spatial query tool and then exported to Excel. As the attribute table included time of detection for each fire spot, it was possible to study temporal variations.

It must be noted that although only 100% confidence level fire points were considered in the study, the study may not have included all actual fires that occurred between 2003 and 2013 as a result of operational errors of the MODIS in which an actual fire may be detected but assigned a confidence level of less than 100% depending on its brightness and track. Cloud cover during the MODIS overpass may also reduce the number of active fires detected.

Results

National fire statistics

Using the 11-year data set of fire points, a map was created to indicate the total distribution of fires in the country per province. Figure 1 gives a clear indication that the northeastern and eastern regions of the country experienced the most frequent fires over the last 10 years. The southwest regions of the country also appear to have a high fire frequency compared to surrounding regions. From Figure 1 it is clear that topography and climate play a role in fire totals as evidenced by the high density of fire points along significant mountain ranges. For example, the main spread of fires in the southwest regions can be associated with the Cape Fold Mountains, while the western border of KwaZulu-Natal experiences a high concentration of fires along the Drakensberg mountain range. The Free State's highest fire concentration is situated along the Lesotho border, which is also part of the northern Drakensberg range. It should be noted that all fires, including those that occurred in close proximity to urban areas, were used in the analysis. Results produced from data excluding fires within a 20-km radius of an urban area can be found in the online supplementary material (Supplementary figure 1).

Vegetation statistics

As stated earlier, South Africa experiences two different rainfall seasons with some locations experiencing a bimodal rainfall distribution. These different rainfall seasons, associated with summer and winter months, have led to a wide variety of vegetation types in the country. The country's varied topography has also led to different vegetation patterns in mountainous areas, where air temperature and rainfall gradients vary over small areas resulting in pockets of different vegetation. Table 1 provides a list of the 'top ten' vegetation classes. These 10 vegetation types have the highest number of fires over the 11-year period. Table 1 also includes a percentage which indicates the ratio of fires in each vegetation class to the total number of fires.

Mountain grasslands in the northeastern regions of the country experienced the most fires over the 11-year period, followed by mountainous fynbos. The top ten vegetation types can easily be categorised into three main biomes: grassland, bushveld and fynbos. Of note is that while grasslands feature in the top ten, they are mostly associated with upland or mountainous areas such as western and southern KwaZulu-Natal or northwestern KwaZulu-Natal rather than the open plains type of grassland associated with the central Free State. The top ten can also be linked to the three provinces which experience high fire frequency. The vegetation types listed in Table 1 contributed to about 60% of all fires between 2003 and 2013, leaving the remaining 57 vegetation types contributing only 40%, again indicating the high concentration of fires in certain regions of the country.
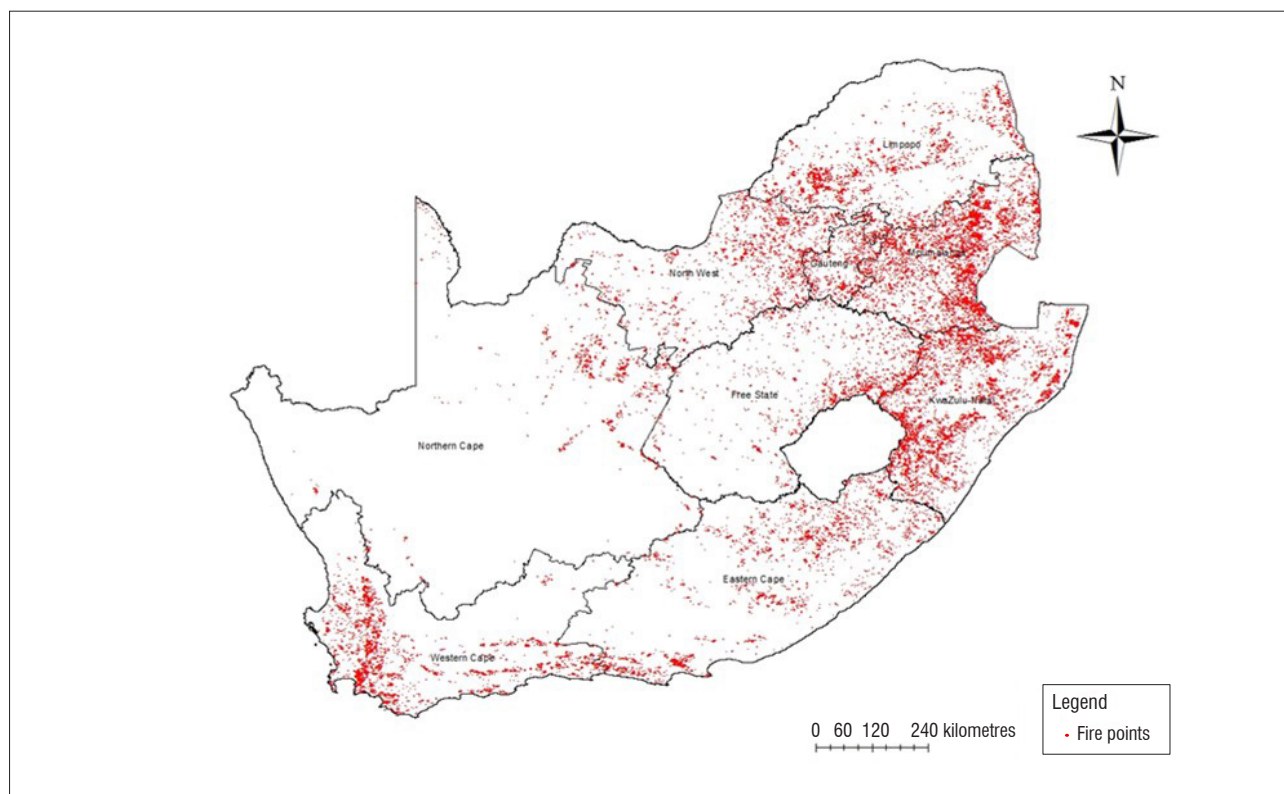
**Figure 1:** Fire distribution in South Africa between 2003 and 2013.

Vegetation	Number of fires	Percentage
Northeastern Mountain Grassland	4513	12.71
Mountain Fynbos	3291	9.26
Moist Upland Grassland	2711	7.63
Mixed Bushveld	2516	7.08
Sour Lowveld Bushveld	1633	4.60
Moist Sandy Highveld Grassland	1598	4.50
Rocky Highveld Grassland	1596	4.49
Moist Cool Highveld Grassland	1228	3.46
Natal Central Bushveld	1225	3.45
Wet Cold Highveld Grassland	1113	3.13

Annual fire trends

Acquiring an 11-year data set made it possible to analyse interannual trends in fire frequency. The effect of climate change on fire frequency is not known despite a number of existing scenarios. Figure 2 illustrates the total number of fires experienced in South Africa between January 2003 and December 2013. While no clear trend is evident in Figure 2 (slope = about 48 fires per annum, t=0.61, significant at 44% confidence level), there appears to be a stabilisation in fire totals over the last 3 years, beginning in 2011 and continuing through 2013. However, 3 years of stabilisation does not mean that fires are being better managed or that mitigation efforts are successful, and further investigation in the upcoming years will be needed to fully understand the trend, assuming the trend, or lack thereof, continues.

Also of note is the higher than average number of fires in the years 2005, 2007 and 2010. Years 2005 and 2010 have been classified as two of the warmest years in the past decade and the effect of these above-average air temperatures seems to have a strong correlation with fire frequency. The year 2006 has also been classified as the warmest La Niña year in the past decade, which may explain the high fire frequency in 2007.


$$y = 48.327x - 93810$$
$$R^2 = 0.0391$$
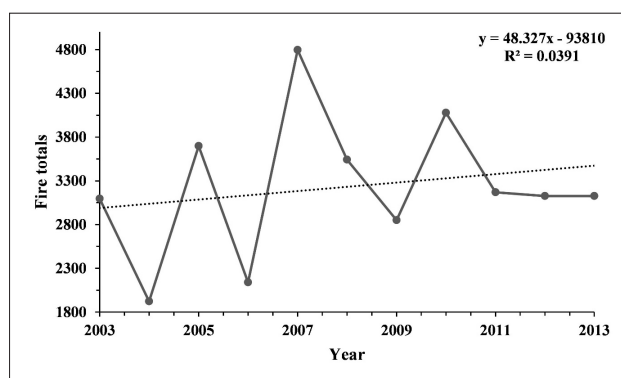
Figure 2: Annual fire totals for South Africa.

Monthly averages

South Africa is known to experience a number of fire seasons as a result of the variability of the country's rainfall seasons. As the western and southwestern parts of the country receive mainly winter rainfall, their fire season generally occurs in the summer months while the central and eastern portions of the country experience a winter fire season because of their summer rainfall pattern. Totalling the number of fires per month for each year in the data set and producing monthly averages shows

which months are prone to fires. Figure 3 illustrates the months which experience the highest number of fires. It is clear that the majority of fires occur in August and September; this finding therefore indicates that the central and eastern regions of the country experience higher fire numbers than the western portions, as August and September fall into the dry winter months in the central and eastern regions and the wet winter months in the west. While the highest fire numbers are seen in August and September, months associated with the western region fire season (December to March) also have relatively high averages.

It is important to note that while the winter fire season may appear to have higher fire frequencies, the size of the area is much larger than that which receives winter rainfall. Areas which receive winter rainfall are also largely semi-arid, thus limiting the fuel available to burn during a fire, which may limit fire activity. August and September can clearly be identified as the most severe months for fire outbreaks but Figure 3 also shows that fires are possible at all times during the year, which serves as a challenge to landowners and critical decision-makers.
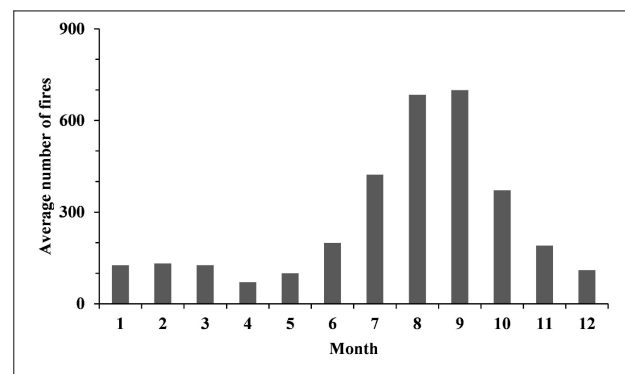
1, January…12, December

Figure 3: Monthly fire averages between 2003 and 2013 for South Africa.

Provincial fire statistics

The *National Veld and Forest Fire Act of 1998* promotes the establishment of regional fire protection association and requires all landowners to be members of a fire protection association. By understanding provincial trends in wildfires, these fire protection associations and landowners may be better equipped to mitigate and manage fire outbreaks in their respective provinces. As each province in South Africa has different topographic and climatic factors, one cannot expect the management plans of each province to be the same and each province needs to develop its own plan according to its own fire season and fire pattern.

Annual trends in provincial fire totals

Using GIS processing, the total number of fire spots per province was calculated and plotted over time. The calculations included yearly totals as well as monthly average fire numbers (Figure 4). Assessing the trends in yearly fire numbers may provide for better management and better allocation of funds and resources. All provinces display an increasing or stable trend in interannual fire totals, with the exception of the Western Cape which has experienced a small decline in fire totals since 2003. While the trend may be indicating a decrease, there has nevertheless been years in which the Western Cape experienced a significantly high number of fires. It should be noted that only the Northern Cape's annual fire frequency yielded a slope value that was greater than 0. The Northern Cape Province displayed the most noteworthy increasing trend in fire totals from 2003 (slope = about 16 fires per annum, t=1.87, significant at a 90% confidence level). While some years have fewer fires than the previous years, the overall trend indicates that fires may be becoming a common problem in the province. Other provinces do not display significant increasing trends and the Eastern Cape Province is the only province that appears to have a stable trend in fire totals since 2003, with an increasing slope of about two fires per annum. Table 2 provides an overview of the slope, t-value and confidence level of each province and the South African region as a whole.

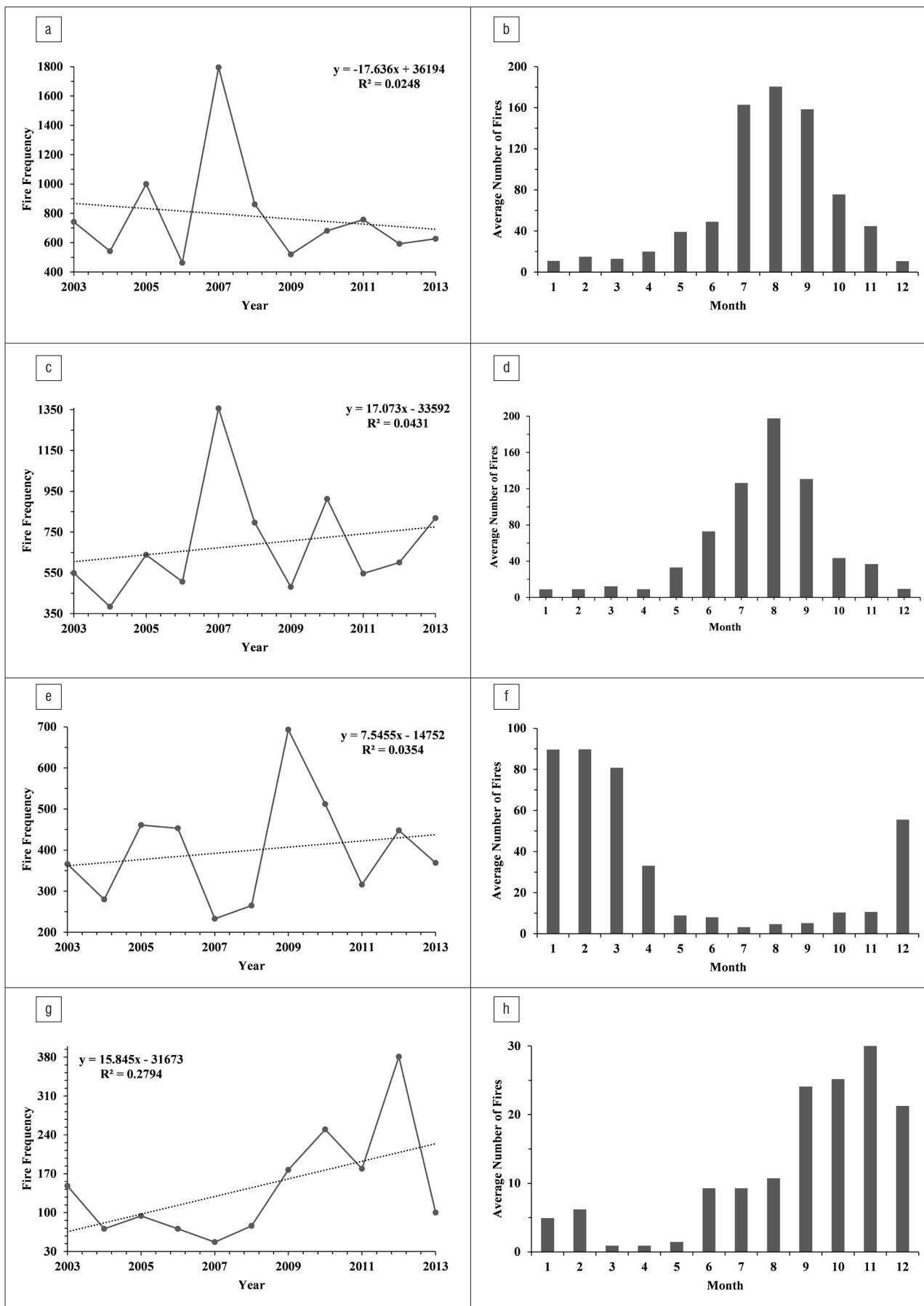**Figure 4:** Annual fire totals for selected South African provinces (a, c, e, g represent annual totals and b, d, f, h represent monthly averages): (a and b) Mpumalanga, (c and d) KwaZulu-Natal, (e and f) Western Cape and (g and h) Northern Cape.

Table 2: Statistics for the whole South African region and individual provinces

Area	Slope	*R*	*t*	Confidence
Northern Cape	15.85	0.53	1.87	90.54
Free State	7.70	0.26	0.80	55.84
KwaZulu-Natal	17.07	0.21	0.64	45.97
South Africa	**48.33**	**0.20**	**0.61**	**43.99**
Limpopo	8.02	0.19	0.58	42.70
Western Cape	7.55	0.19	0.57	42.04
North West Province	6.65	0.18	0.55	40.63
Mpumalanga	-17.64	-0.16	-0.48	35.62
Gauteng	1.16	0.14	0.43	32.58
Eastern Cape	2.05	0.04	0.12	9.30

Provincial variations in fire seasons

As stated previously, two seasonally distinct regions with characteristic fire seasons exist in the country: the cool, dry winter fire season experienced in the central and eastern regions of the country and the warm, dry summer season experienced in the west and southwest regions of the country. An analysis of provincial peaks in fire frequency, presented in Figure 4a–h, indicates slightly more discrete fire seasons. Figure 4 also illustrates different peaks in fire seasons around the country. The Western Cape experienced its peak in fire numbers in February, while the Northern Cape experienced a peak in November – two different peaks in fire season, although these two provinces both experience winter rainfall and thus a summer fire season. Mpumalanga and KwaZulu-Natal both experienced a peak in fire season in August while the North West Province, Gauteng, Limpopo, the Free State and the Eastern Cape experienced a peak in fire season in September (Supplementary figure 2).

While each province can be categorised into a specific fire season, it is important to note the severity of each province's fire season. Mpumalanga, KwaZulu-Natal and the Western Cape have been identified as the three provinces which experience the most severe fire seasons largely because of their native vegetation, topography and climate. KwaZulu-Natal and Mpumalanga have minimum monthly averages of about 10 fires and a maximum of close to 200 fires at their peak in the fire season. The rest of the provinces display similar minimum averages of close to one fire per month and have maximum averages ranging between 30 fires and 120 fires during the peak in the fire season (Supplementary figure 2).

Discussion

National distribution of fires

Figure 1 provides a stark indication of the severity of fire occurrences in South Africa over the last 11 years. The vegetation types which experience the highest frequency of fires occur in the northeastern and eastern regions of the country. Table 1 provides a list of the vegetation types most prone to fires. It is clear that grasslands and bushveld are common fire areas. A number of reasons can be cited for the link between these vegetation types and their high fire frequency. Vegetation such as grasslands and bushveld are able to grow rapidly and densely if provided with sufficient rainfall. When grasslands and bushveld grow, they produce a high fuel load as a result of the high biomass associated with the vegetation. During the dry season, the water content of the vegetation decreases, providing a fuel for fires to ignite and spread. The top ten vegetation types all have characteristics conducive for fires to occur. Many are classified as moist or wet, indicating an association with an above-average rainfall season (either summer or winter rainfall

season). Furthermore, many of these vegetation types are classified as mountainous or upland vegetation (the significance of which is discussed below).

Figure 1 also provides an insight into the role of topography in fire occurrences. The importance of topography in fire behaviour and fire likelihood is often overlooked by the South African fire community and is not taken into consideration when monitoring or calculating fire danger using the Lowveld fire danger index, evident by the lack of a fire behaviour model. Fire danger indices used by developed countries, such as the Australian McArthur Forest and Grassland Fire Danger Indices and the US National Fire Danger Rating System, consider topography when monitoring and predicting fire danger. The concentration of fire points along ranges such as the Drakensberg and Cape Fold Mountains indicates the need to account for topography as it has a marked influence on fires. Topography can also lead to dry mountain winds, known as berg winds in South Africa, which have been proven to exacerbate fire danger. Furthermore, it is believed that fires which originate on mountain slopes and spread upslope may be enhanced by the local topography. Under general meteorological conditions, one would assume that anabatic winds prevail during the day, essentially spreading any fire that starts upslope. As a fire moves upslope, the heat emitted causes the surrounding air mass to warm and rise, creating a ventilation system for the fire. The reasons stated above may explain the high concentration of fires along the major ranges of South Africa. Therefore, it is important for landowners, community managers and critical decision-makers to be aware of the topography of their area and assess its role in the local fire dynamics and develop plans accordingly.

Annual fire frequency trends

The annual fire frequency trends, displayed in Figure 2, are revealing in terms of the impact of predicted climate change and interannual climatic variations on fire. As seen in Figure 2, the years 2005, 2007 and 2010 experienced the greatest number of fires in the last decade. Years 2005 and 2010 have been cited as two of the warmest years on record in the southern hemisphere compared to a global average.[17] One can assume that South Africa's annual average air temperatures would be similar to that of the southern hemisphere as a whole.

Two possible scenarios have been established for fires under a warming climate. The first scenario states that under warmer air temperatures, heat waves and drought conditions may be more severe, which may result in vegetation desiccating at higher rates, leading to drier fuel loads resulting in increased fire numbers. The second scenario states that under a warming climate, rainfall may be significantly higher.[18] Higher rainfall totals may lead to increased rates of vegetation growth, leading to heavier fuel loads which results in more available fuel to burn and also increased rates of spread when fires occur. As 2005 and 2010 have been marked as years of above-average air temperature, they serve well to investigate the possible effect of global warming on South Africa's fire frequency. As both years have high fire frequencies, one can assume that the two scenarios are accurate.

While 2007 is not seen as a significantly above-average year in terms of air temperature, it still experienced a high fire count. However, 2006 has been cited as one of the warmest La Niña events. La Niña results in increased rainfall over the eastern half of South Africa. As 2006 has been cited as the warmest La Niña event, one can also assume that air temperatures were greater than average. Increased air temperatures and increased rainfall may have led to increased vegetation growth, which – during the 2007 winter (the eastern region's fire season) – could have resulted in higher fire frequencies which would be in agreement with the first possible scenario for fires under global warming. The South African Weather Service has made seasonal rainfall charts available (Figure 5). Analysis of these charts confirms that higher than normal rainfall, indicative of a La Niña event, occurred during the summer rainfall season between 2006 and 2007, which may explain the increased fire activity during the 2007 fire seasons. Despite other explanations as to why 2007 experienced the greatest fire totals, it is clear that global climate change will result in increased fires, which needs to be taken into consideration by government, landowners and fire protection associations.
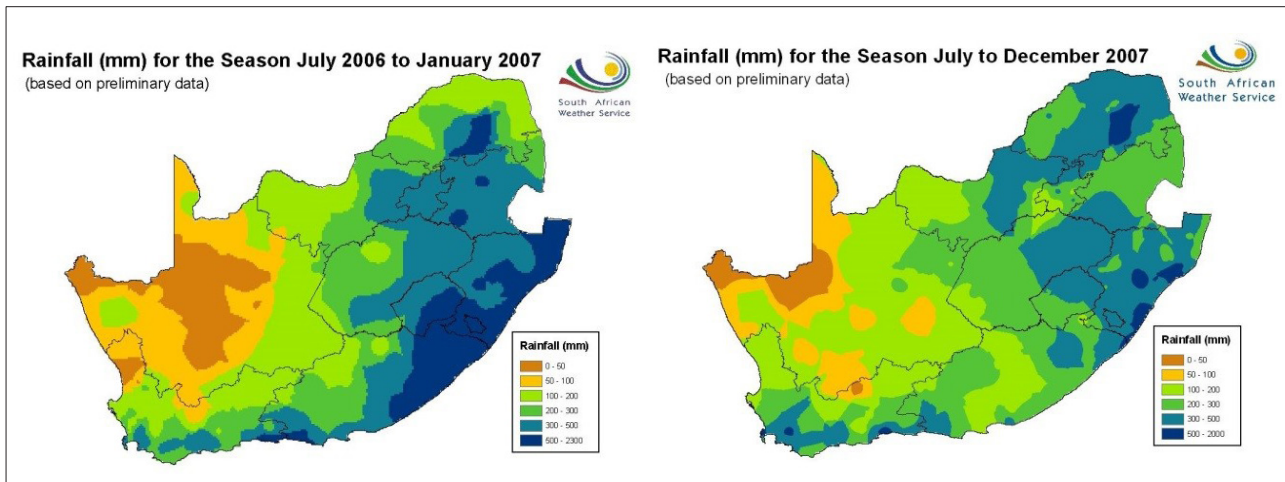
Charts courtesy of the South African Weather Service.

Figure 5: Seasonal rainfall prior to and during the 2007 fire seasons with (a) indicating the high rainfall totals experienced between July 2006 and January 2007 and (b) showing the total rainfall experienced between July and December 2007.

Provincial distribution of fires

Traditionally, two main fire seasons have been noted: the winter fire season of the eastern regions and the summer fire season of the western regions. The variation in fire seasons across the country is largely because of climatic conditions. As evident in the monthly distribution of fires for the various provinces, shown in Figure 4, there are four main peaks in fire seasons:

- Mpumalanga and KwaZulu-Natal, both situated in the far eastern regions of the country, experience a maximum in fire activity during August.

- The Western Cape on the southwestern seaboard experiences a maximum in February.

- The maxima of these three provinces have traditionally been identified as the peak of the two main fire seasons.

- Other provinces such as the North West Province, Free State, Gauteng and Eastern Cape experience maxima in September and the Northern Cape in November.

While these four distinct peaks occur in the conventional fire seasons, it is still important for landowners and decision-makers to take note of the months in which fire activity is at a maximum in their respective provinces so as to ensure the correct timing of resources, and to ensure the efficient provision of resources and infrastructure.

Annual trends in provincial fire activity

It is clear that while increasing and decreasing trends exist (as indicated by the slope), these trends are not statistically significant when using traditional significance levels. It is hypothesised, however, that the limited annual duration of the data sets ($n=11$) may have reduced the significance of the trends. The fact that fire numbers are influenced by anthropogenic activity may also have introduced fluctuations in the data set, essentially reducing the statistical significance of the slope relative to 0.

As previously stated, all provinces (with the exception of Mpumalanga) displayed a stable or increasing trend in fire activity between 2003 and 2013. As seen in Figure 1, Mpumalanga has displayed a high incidence of fire activity over the past 11 years, despite the decreasing trend and has been identified as the province most prone to fire activity. The decline in Mpumalanga's fire activity since 2008 may be, in part, a result of increased awareness of the dangers of fires and the better management and mitigation of fires by local fire protection associations under the Mpumalanga Umbrella Fire Protection Agency.

The Northern Cape Province has experienced the most notable increase in fire activity since 2008. An increase in fire activity in the Northern Cape came as a surprise as the province is largely classified as semi-arid with mainly succulent Karoo vegetation.[19] Provinces surrounding the Northern Cape have also experienced some kind of increase in fire activity over the last 11 years. As noted before, a data set with a time span of 11 years may not be sufficient to draw viable conclusions, but these increasing trends do provide relevant questions.

In the early 1950s, Acocks hypothesised the spread of desertification and land degradation from the southwest into the northeastern regions of the country.[12] Acocks believed that the grasslands of South Africa were under threat of incremental, progressive and irreversible degradation as a consequence of the spread of deserts coupled with the impact of human over-utilisation of grasslands.[12] The spread of semi-arid vegetation into grasslands has been theorised to occur along a front or boundary spreading northeast or to occur as a result of the coalescence of degraded patches.[19] If Acock's theory was true, then the succulent vegetation of the semi-arid Karoo should have invaded the productive grasslands of the northeastern provinces surrounding the Karoo region (namely the Free State and the North West Province). These two provinces, along with the Northern Cape, have however experienced increases in fire activity over the last 11 years. The succulent Karoo vegetation is believed to be less prone to fire than that of grasslands. The spread of semi-arid regions may have resulted in mixed vegetation communities in which both short grasses and the longer Nama Karoo grasses co-exist, which may result in higher fuel loads. A recent study, however, indicates that grasslands have encroached into the semi-arid Karoo landscapes,[20] which may explain the increased fire activity in the Karoo regions. As stated, increased human settlement in these regions may have led to increased agricultural activity, which is known to alter vegetation structures for livestock grazing.[2] Agriculture also brings with it the use of fire as a management tool, which may explain the increasing trend. The increases in fire activity may also be brought about by short-term variability in climate, such as La Niña or El Niño. The spread of semi-arid regions and the effects on fire activity will require further investigation in the future to assess the role of both climate change and human activity on fire activity.

Conclusions

Fire has always been a natural occurrence on the African continent and has led to the continent being nicknamed 'the fire continent'. Fires do provide an environmental service and are an integral part in the maintenance of the environment; but increased fire activity, brought about by anthropogenic activity, has led to adverse effects in the environment. Fires can be classified as South Africa's most common

hazard and have often led to disasters which result in both loss of human and animal life and loss of property. The study reported here provides an insight into the spatial (including topography) and temporal distribution of fires in South Africa over the past 11 years and the results indicate that the eastern and northeastern regions of the country are most prone to fires. Fire totals appear to be stabilising despite increased fire activity in certain provinces. The role of climate change and short-term climate variability has been analysed, and increased air temperatures and events such as La Niña have a marked effect on fire activity. In order for fire disasters to be mitigated, landowners and decision-makers need to ensure an efficient and effective allocation of resources, including human resources, to limit the spread of fires and to ensure that fire as a management tool does not result in adverse effects on the environment and society. The results provided here make it possible to identify high-risk areas. In order for the country to be resilient against fire disasters, we need to better understand the spatial and temporal characteristics of our fire seasons. This study has made way for further investigation into the spatio-temporal characteristics of South African fires.

Acknowledgements

Authors' contributions

S.S. was responsible for conceptualisation of the study, for writing the manuscript and undertaking the data analysis; M.J.S. was the supervisor, assisted with statistical analysis and made conceptual and editing contributions.

References

1. Archibald S, Nickless A, Govender N, Scholes RJ, Lehsten V. Climate and the inter-annual variability of fire in southern Africa: A meta-analysis using long-term field data and satellite-derived burnt area data. Global Ecol Biogeogr. 2010;19:794–809. http://dx.doi.org/10.1111/j.1466-8238.2010.00568.x

2. Arturson G. Fire disaster management. In: Masellis M, Gunn SWA, editors. The management of burns and fire disasters: Perspectives. Dordrecht: Springer; 2000. p. 53–56.

3. Fire Protection Association of South Africa. 2008 Statistics: Loss ratio poor against the national income [document on the Internet]. c2008 [updated 2010; cited 2013 Feb 25]. Available from: http://www.fpasa.co.za/images/FireStats/Fire_Stats_2008.pdf

4. Fire Protection Association of South Africa. South African national fire statistics 2009 [document on the Internet]. c2009 [updated 2011 Mar; cited 2013 Feb 25]. Available from: http://www.fpasa.co.za/images/FireStats/Fire_Stats_2009.pdf.

5. Fire Protection Association of South Africa. 2010 South Africa national fire losses [document on the Internet]. c2010 [updated 2012 Mar; cited 2013 Feb 25]. Available from: http://www.fpasa.co.za/images/FireStats/Fire_Stats_2010.pdf.

6. De Ronde C, Goldammer JG. Fire situation in South Africa: Fire environment, fire regimes, ecological role of fire. International Forest Fire News. 2001;25:105–111.

7. Working on Fire (WOF). Cape burns [media release on the Internet]. 2014 February 6 [cited 2015 Nov 25]. Available from: www.workingonfire.org/images/WoF_Media_Release_Cape_Burns.pdf.

8. Duncan B, Kangmann B, Textor C, Trentmann J, Van der Werf GR. Vegetation fire emissions and their impact on air pollution and climate. Atmos Environ. 2009;43:107–116. http://dx.doi.org/10.1016/j.atmosenv.2008.09.047

9. Keywood M, Kanakidou M, Stohl A, Dentener F, Grassi G, Meyer CP, et al. Fire in the air: Biomass burning impacts in a changing climate. Crit Rev Env Sci Tec. 2013;43:40–83. http://dx.doi.org/10.1080/10643389.2011.604248

10. Townsend SA, Douglas MM. The effect of three fire regimes on stream water quality, water yield and export coefficients in a tropical savanna (northern Australia). J Hydrol. 2000;229:118–137. http://dx.doi.org/10.1016/S0022-1694(00)00165-7

11. Merz SK. Impacts of bushfires on flood hydrology and erosion rates. Victoria, Australia: Department of Sustainability and Environment; 2004.

12. Meadows ME, Hoffman MT. The nature, extent and causes of land degradation in South Africa: Legacy of the past, lessons for the future? Area. 2002;34:428–437. http://dx.doi.org/10.1111/1475-4762.00100

13. Van Wilgen BW, Biggs HC, Potgieter ALF. Fire management and research in the Kruger National Park, with suggestions on the detection of thresholds of potential concern. Koedoe. 1998;41(1):69–87. http://dx.doi.org/10.4102/koedoe.v41i1.248

14. Van Wilgen BW. The evolution of fire and invasive alien plant management practices in fynbos. S Afr J Sci. 2009;105:335–342.

15. Tanpipat V, Honda K, Nuchaiya P. MODIS hotspot validation over Thailand. Remote Sens. 2009;1:1043–1054. http://dx.doi.org/10.3390/rs1041043

16. Justice CO, Giglio L, Roy D, Boschetti L, Csiszar I, Davies D, et al. Chapter 29 – MODIS-derived global fire products. In: Ramachandran B, Justice CO, Abrams MJ, editors. Land remote sensing and global environmental change. New York: Springer; 2011; p. 661–679.

17. National Oceanic and Atmospheric Administration (NOAA). Climate at a glance. National Climate Data Centre. No date [cited 2014 Sep 11]. Available from: http://www.ncdc.noaa.gov/cag/time-series/global

18. Intergovernmental Panel on Climate Change (IPCC). Summary for policy makers. In: Climate change 2014: Impacts, adaptions, and vulnerability. Part A: Global and sectoral aspects. Contribution of Working Group II to the Fifth Assessment Report of the IPCC. Cambridge, UK: Cambridge University Press; 2014. p. 1–32. http://dx.doi.org/10.1017/CBO9781107415416.005

19. Dean WRJ, Hoffman MT, Meadows ME, Milton SJ. Desertification in the semi-arid Karoo, South Africa: Review and reassessments. J Arid Environ. 1995;30:247–264. http://dx.doi.org/10.1016/S0140-1963(05)80001-1

20. Masubelele ML, Hoffman MT, Bond WJ, Gambiza J. A 50 year study shows grass cover has increased in shrublands of semi-arid South Africa. J Arid Environ. 2014;104:43–51. http://dx.doi.org/10.1016/j.jaridenv.2014.01.011

Deriving cues from human cognition for the modelling of shack boundaries in aerial imagery

AUTHORS:
Moreblessings Shoko[1]
Julian L. Smit[1]

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

CORRESPONDENCE TO:
Moreblessings Shoko

EMAIL:
moblesss@yahoo.com

Organic studies inspire cues for modelling logic in image processing and become a basis for the development of novel remote-sensing algorithms. Examples of applications of such paradigms include the growing application of techniques such as object-oriented analysis and neural networks in image analysis for which the logic was drawn from studying various components of organic systems in the human body. Here we document a key investigation based on a set of cognitive tests conducted using aerial imagery captured over Cape Town (South Africa). These tests were conducted to later draw parallels with a feature extraction algorithm for shack settlements. We found that the visual variables of 'pattern' and 'shape' display the most significant cognitive guide for shack boundary extraction. Although the focus here was on digital imagery, learning points can be selected for application in other scientific fields as well.

Significance:
- Provides an interest point for several image-processing and computer vision sciences.
- Contributes to geospatial studies and helps improve mapping and imaging initiatives.
- Directed towards solving the challenges that urban governors face in slum management.

Introduction

Organic or biological systems offer great models for imitating and learning from in the development of important technologies.[1] A classic example of this concept is flying – inspired by birds – on which aerial photogrammetry is based. Another example is graphic cable connectors whose titles 'male' or 'female' are inspired by human anatomy.[2] Similarly, in the earth observation based science of remote sensing, a range of ideas has been developed based on the behaviour of organic systems. Recently there has been a growing interest in relating the human ability to view and comprehend phenomena at different scales to feature extraction approaches on image-processing platforms.[3] Inspiration from the human visual system is significant in imaging efforts simply because the end user of most imagery is the human eye itself, therefore mimicking its optical capability creates an opportunity for optimised artificial intelligence.

In this paper, we focus on the spatial cognition of imagery with the aid of the human visual system. We do not deal in depth with the psychology of cognition but simply observe and summarise the image interpretation trends from three sets of tests conducted on focus groups. This observation is done in order to derive inspiration for use later in formulating logical processing steps, particularly for the extraction of informal settlements units from 2005 imagery flown in an aerial photogrammetry project over Cape Town, South Africa. The main objective of the study was therefore to develop a cognitive test kit that can be used to assess human visual interpretation of urban aerial imagery in order to derive cues for use in algorithm development. The algorithms to be developed are for a later study based on feature extraction. These algorithms will be used to extract and delineate detail such as the boundaries of informal settlements units and the living area extents from imagery using learning points from the cognitive tests discussed here. A particular focus is placed on extraction of informal settlements in urban imagery within the Cape Town data.

Informal settlements are of interest as they are a growing urban challenge. Many governors who seek to manage the proliferation of this urban typology require up-to-date information on their seeding and growth. This information can be spatial or attribute in nature. Spatial information comes in the form of map products that can be generated from image data as well as other in-situ data collection techniques. Imaging data such as aerial images provide significant product advantages over in-situ data collection techniques which are costly and time consuming. One of the main advantages of image data is applicability to image-processing functions that allow map generation to occur faster should optimum feature extraction algorithms be available. In this study, the target environments for end feature extraction were informal areas within 250-mm resolution urban imagery.

Informal settlements present a complex imaging scene with multiple land uses mixed closely in a small area of coverage. This scenario is a challenge to geo-information scientists who often are faced with the need to constantly update their feature extraction algorithms to match changing user requirements. In order to overcome these challenges, we propose a perspective that derives inspiration from the human visual system and its interpretation of urban aerial imagery. The learning points from the human interpretation of aerial imagery yield cues that are later imitated in an informal settlements land-cover classification algorithm. This end product classification algorithm is later implemented towards the creation of informal settlements maps or atlases that provide useful spatial information for stakeholders in the urban management matrix. The current study achieves its primary objective through documenting a detailed procedure of the development of a cognitive test kit in the form of a simple set of experimental exercises. These cognitive exercises draw inference from the human interpretation of 250-mm resolution optical imagery that contains scenes extracted from both formal and informal areas of the city by studying interpretation trends in the focus groups. Results from such studies can find much relevance in improving mapping and cartographic applications. There is also much application of this work in cue generation

for robotics and artificial intelligence. Vehicle or human navigation and way finding would also benefit greatly from the outcomes of this study. It is envisioned that the results derived through this research will be highly applicable to a broad spectrum of computer vision based applications.

Spatial cognition and the human visual system

Several researchers have documented the use of cognition in trying to decipher key image interpretation concepts using different testing approaches.[4-7] The limitation of these earlier studies is that most focused on the interpretation of general urban image scenes with very little attention paid to scenes containing the informal settlements component of urban fibre. Moreover, the target applications for these earlier studies have often been cartographic in nature with limited use of cognitive studies in linking to developing an organically inspired feature extraction algorithm. Here we specifically use observations from cognitive tests to inspire a novel algorithm for informal settlements extraction using aerial imagery from the City of Cape Town.

Informal settlements are an unstructured component of urban fibre. They are characterised by a somewhat random and irregular settlement environment that grows rapidly to dense nucleation. The human visual system on the other hand is much like other organic systems in biology. It is made up of millions of cells that are born, reproduce and grow. In addition to having metabolic characteristics, such organic systems are an intriguing combination of a scale-space environment in which there can be infinite levels of detail of sets of system components. By this description, an organic system of cells that grow close together to maturity is metaphorically similar to informal settlements. Informal settlements can be described as organic settlements as their nature mimics organic growth. In the same way in which organic tissue operates, these settlements are born (inception); require nutrition (mending of structures); move (unpredictable shift of shack locations); excrete (inhabitants require sanitation services); grow (rapidly change over time); reproduce (multiply to consolidation); sense their environment (this points to the fact that they represent an informal land rights or cadastral system with rules to avoid conflict) and die (some parts or entire informal settlements disappear with time). Several organic systems behave in the manner described here, and hence show morphological similarities with informal settlements. The human visual system was selected out of these numerous organic systems to draw inspirational parallels with the informal settlements environments because of its direct link to image interpretation. It is the human eye that consumes the imagery on which these unstructured settlements are often captured; therefore the human visual system is an ideal choice for inspiration and cue derivation.

The human visual system works in relation with the human cognitive system which coordinates the mental processes that take place as the mind digests visual detail captured by the eye. However, the scope of this study is not to study cognition in depth, but to capitalise on the growing trends that appreciate the strength of cognitive abilities as a foundation for robust image interpretation. Although cognitive research itself is not significantly documented in remote sensing and imaging sciences, it is well appreciated as being important in improving the understanding of the image interpretation process in remote sensing.[4-7] Cognitive abilities have also been used extensively in the related disciplines of cartography and environmental perception.[8] It is important to note that geospatial images and maps are among the most efficient media for communicating spatial knowledge, especially in unfamiliar environments.[9] They come with a diverse array of image parameters – such as dimensionality, resolution, orientation, accuracy, colour and theme – that become unique inputs in the human cognition computation when they are used to optimise feature identification. The image vantage point or perspective and visual realism are also key parameters worth mentioning that impact on knowledge extraction and cognition when using geospatial images. While vantage point focuses on the angle at which the navigator views the image, visual realism measures how closely an image itself represents the real world. Images themselves can be available in different formats – such as vertical, panoramic and oblique (low and high) – or ground-based representations often in two or three dimensions that also impact on

the realism parameter. All these image parameters mentioned above, together with other factors, collectively influence how information is extracted and cognitively relayed by the user in both analogue and digital environments. Research has shown that even additional parameters such as viewer or navigator sex or age can also impact on image interpretation for spatial cognition. Male subjects for example typically outperform female subjects on tasks dealing with mental rotation and spatial navigation, whereas female subjects tend to perform better on tasks dealing with object location, relational object location memory and spatial working memory.[10]

Various other studies have also documented how cognitive theories and paradigms give substance to the psycho-spatial process of how humans visualise or navigate through urban spatial environments in both two and three dimensions. To begin with, cognitive models for imaging applications are often formulated based on comparative experiments and exercises that test human visual and cognitive abilities on different types and forms of images or models. For example, Walther et al.[11] gave an interesting illustration of the growing capabilities of computer vision choices by simulating human vision concepts. The authors proposed that on entering a crowded classroom in order to localise and recognise someone, humans look around and scan everyone's faces without paying much attention to the interior design and room's furniture. However, when entering the same classroom with the intention of finding an available desk, humans look at the same scene, and yet their perception is skewed toward the arrangement of the furniture, and they mostly ignore the people present. This illustration shows that developing object-specific models of image interpretation can allow for adjustments by shifting the requirement parameters towards first extracting features of different interest. This shows the diversity and capabilities of human cognition for image interpretation and mapping.

The entire concept of human cognitive mapping is broad and cannot be fully captured in this study alone. Particularly, it refers to a series of psychological transformations by which an individual acquires, stores, recalls and decodes information about the relative locations and attributes of phenomena in their spatial environment.[12] Cognitive theories on facets such as way-finding and location identification using images or maps of the built environment summarise a repeated process that humans undertake daily within and between structures using both simple and complex mental encoding and decoding. When the image or map surrounding is familiar to the viewer, an internal cognitive map is recalled to facilitate what may ordinarily appear as a low-level task of interpretation. This internal map is simple in nature – it is not perfect; it is sketchy and somewhat distorted as it is derived from experiences and not precise measurements of the earth. This internal map is said to be characterised by four domains – recognition, prediction, evaluation and action – that assert its basic processing in the human mind.[13] However, when the viewer is in unfamiliar surrounds, a significant portion of the human cerebral system is employed into a more complex cognitive mental process. This shows a relationship between what the human visual system consumes and what the cognitive system then formulates. Making simple observations of how users behave as they relate the two systems can create interesting cues for algorithm development.

In contrast, there are also instances in which spatial cognitive abilities fail or are impaired, as medical experts have documented.[14] Such discussions show the importance of images of varying dimensionality as important tools in aiding spatial cognition in unfamiliar and familiar environments. The specifications of the images themselves contribute greatly to improved cognitive abilities. The mental process of how the human brain encodes and decodes data required to interpret location for the image test kits developed and tested in the study is described in the method below.

Materials and methods

Cities and built environments create interesting image outputs that capture some of the most complex relations between human constructed objects. As summarised in the introduction above, we sought to address the following three research questions:

1. To what extent can the human cognition inspired algorithms assist in differentiating between regular and irregular urban image elements as viewed from a scene extracted from a 250-mm resolution image?

2. Can there be a general flow or order in how humans use cognitive abilities to identify and visualise shack settlements?

3. Does the user accuracy and cognitive sequence vary with image-based skill and experience?

The first step of the study after identifying the research questions and scope was to develop the cognitive test kit and identify the focus group participants (using both systematic and random sampling techniques). We identified scenes of interest from a collection of vertical aerial imagery from 2005 captured in Cape Town, from which participants would categorise the targets of interest during the test based on the visible roof structures. Twenty sets of each image scene were printed on gloss print and typed descriptions to guide the interpreter on the interpretation task were attached per image set. In designing the test, participants would also be required to indicate a personal confidence rating for their categorisation of the structure marked in each image rated on a five-point Likert scale ranging from 1 ('not confident') to 5 ('completely confident'). Because similar studies have shown that response time is directly proportional to the difficulty of the task, a response time per question was recorded individually. This response time was judged by noting the time lapse between the initial viewing of the image and the participant's recording of a response. Prior to undertaking the task, some personal details of the participants were recorded: education level, age range, image interpretation experience and any practical contact with any geomatics work, air photo interpretation and remote-sensing exposure. Thereafter, participants were educated on the task objectives and they were also provided with a set of typed instructions explaining the task. A short training session was given to practise categorising other aspects of imagery such as vegetation and water bodies using visual interpretation.

Following the training, participants were informed of the importance of working quickly and accurately. Each participant was presented with the set of five unique image scene prints, handed out one at a time, and the participant was asked to assess the features highlighted using bright red marker in each image. A selection of four possible multiple-choice answers per question was provided for images B to E. All the participants received the same set of images although for each participant there was

effort to randomise the image order. The following is a detailing of the focus groups of the participants for the task:

1. Expert Group X comprised 13 men and 7 women randomly selected from third-year students who had taken a course in image processing.

2. Trained Group Y comprised 14 men and 6 women from the public who randomly volunteered. These individuals responded to a call for volunteers at a working table set up in an open shopping area and we had no a-priori knowledge of their image-processing skills.

3. Children's Group Z comprised 20 male and female teenage children between the ages of 16 and 18 selected by a patron from a group of children at a community youth centre.

The study was approved by the Midlands State University Research and Ethics body (clearance number SGEO 2016/02).

The three focus groups underwent the cognitive test using the five image sets as mentioned. Figure 1 illustrates the selected image scenes with bold red markings highlighting focus areas. Provided with Figure 1 Groups X and Y were asked the following questions (with multiple-choice responses included where applicable):

1. Using a maximum of two words per point, summarise in five points, what you see in the highlighted area in A.

2. How many units / structures are in the circled area of B?
   A. 0–5   B. 5–10   C. 10–15   D. 15–20   E. 20 or above

3. Identify the urban activity present in C.
   A. Residential area   B. Informal area   C. Business area
   D. Educational area   E. Industrial area

4. What type of activity is likely to take place in D?
   A. Residential area   B. Informal area   C. Business area
   D. Educational area   E. Industrial area

5. What is the first object you saw when you looked at image D?

6. What type of activity is likely to take place in E?
   A. Residential area   B. Informal area   C. Business area
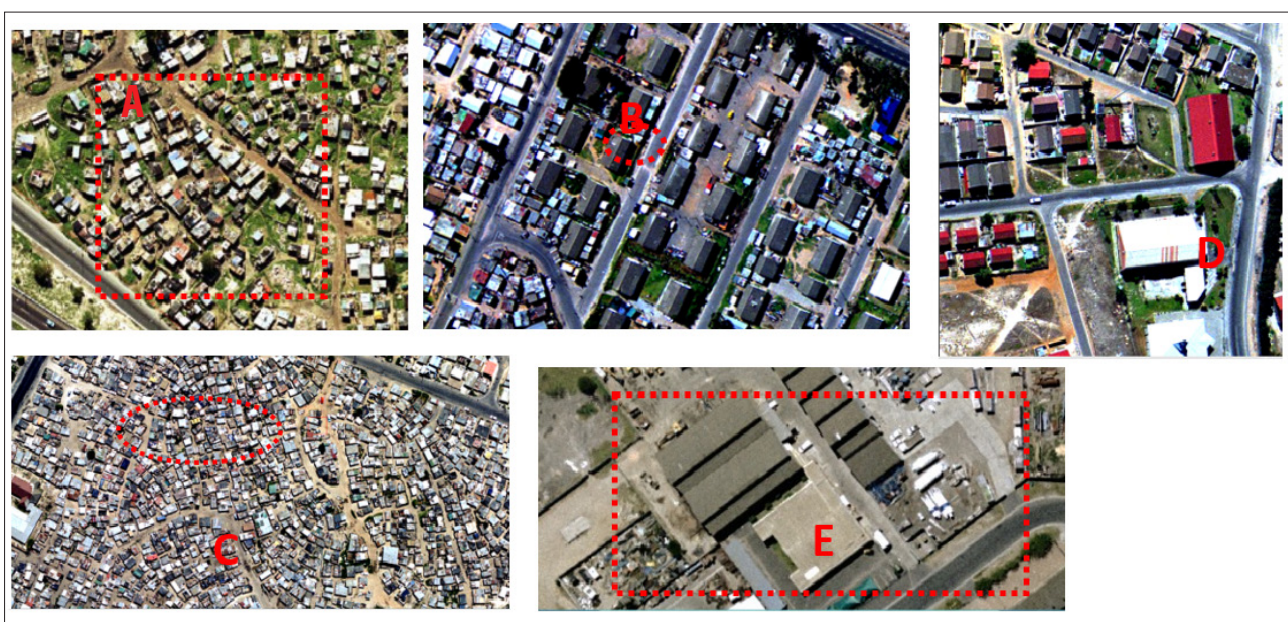   *D. Educational area   E. Industrial area*

Figure 1: The five vertical aerial photographs of 250-mm resolution used for the feature identification task.

Table 1: Percentage of correct responses in each group in the feature identification task

Image	Percentage of correct responses		
	Group X	**Group Y**	**Group Z**
B	85%	25%	10%
C	85%	30%	40%
D	**100%**	**90%**	**65%**
E	95%	75%	50%

High scores of interest are bolded.
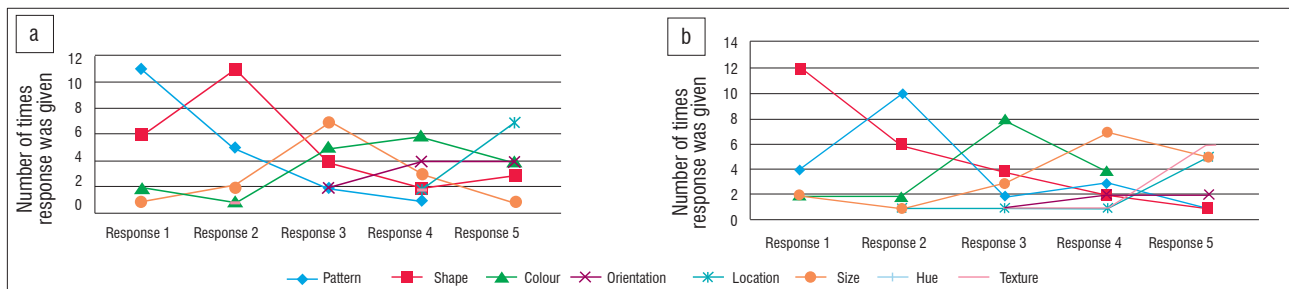
Figure 2: Responses from (a) Group X and (b) Group Y on image A, classified according to Bertin's descriptors.

Questions 2 to 6 were similar for Group Z but the language was simplified so that they could understand the questions more easily. Question 1 for Group Z was 'What do you see?', and did not include a request for a list. All answers were compiled, collated and assessed in the focus groups to make inferences on the image interpretation process. Participants were not from the area where the imagery had been captured. Responses to Question 1 in reference to image A were classified into Bertin's cartographic visual variables[15] to generalise how the responses simulate a mental cognitive process of an informal settlement. Responses were classified according to eight variables: size, shape, colour, orientation, pattern, texture, hue and location. For example, a response of 'Several squares' was taken as a shape descriptor.

Results

Groups X and Y can be seen to be representative of a neural network and supervised classifier, respectively, because Group X is an expert group which is the nature of neural network classifying systems which classify data through expertly learned steps, whereas Group Y is more like a supervised classifier as the participants were trained to classify other feature types prior to the task. Group Z is an experimental group included out of interest to assess the development of their cognitive skills in interpreting images of unfamiliar surroundings.

For the study, a response was rated as a correct response (CR) if the respondent correctly selected an accurate response for a multiple-choice question and as an incorrect response (IR) if the response was incorrect. Equations 1 and 2 were used to calculate the percentage of correct and incorrect answers, respectively:

$$CR = \frac{N}{P} \times 100 \qquad \text{Equation 1}$$

$$IR = \frac{N'}{P} \times 100 \qquad \text{Equation 2}$$

where N is the number of correct responses and P is the total group per focus population. Table 1 shows the confusion matrix of the classifications of images B, C, D and E. Figure 2 shows a grouped representation of the order of visual variables used to describe image A.

The children's group was simply an experiment to gauge the level of development of their cognitive skills in interpreting images. Their responses to Question 1 were collected and classified to Bertin's visual variables as best as possible, e.g. a response such as 'I see a jigsaw puzzle' was taken to refer to a fractal pattern and hence classified as 'Pattern'. Figure 3 represents the distribution of responses from the children with regard to the informal settlements environment in image A. Where a response contained more than one variable, e.g. 'Small squares', the first variable was taken as the response – so in this case, it would be assigned to size and not shape.
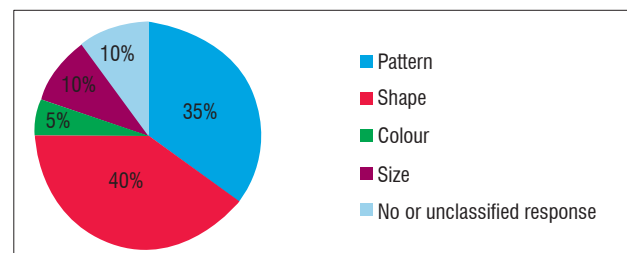
Figure 3: Group Z's responses to Question 1 on image A, classified according to Bertin's descriptors.

Based on the responses of Groups X and Y, and to some extent Group Z, two general cognitive variables for image interpretation were prominent: pattern and shape. Table 2 is a summary of the general flow of mental thoughts associated with viewing the images of an informal settlement. Pattern, shape and size are common to all three groups with shape scoring highest in Groups Y and Z. Location and colour appear to be of secondary significance to the urban image interpretation process. The unclassified responses – specific to the children's group – represent vague responses that either overlap with two variables or may not fall clearly into any variable category.

Table 2: Summarised flow of mental cognition counts of image A

Group X (experts)	Group Y (trained participants)	Group Z (children)
Pattern	Shape	Shape
Shape	Pattern	Pattern
Size	Colour	Size
Colour	Size	[†]Non-classified responses
Location	Texture	Colour

[†]*Non-classified responses are those which either overlapped between two categories or did not fall clearly into any category.*

Response times, as previously mentioned, can be a good indicator of task difficulty. The average response times per group per task are plotted in Figure 4. Group X, the experts, always took less time to respond than the trained group, Group Y, suggesting that they found the tasks less difficult than the latter group. It is notable that in most instances the children's response times closely matched the experts' response times. This could perhaps be because the children often did not give too much thought to the answer; hence their cognition of the interpretation was shorter but not necessarily accurate.
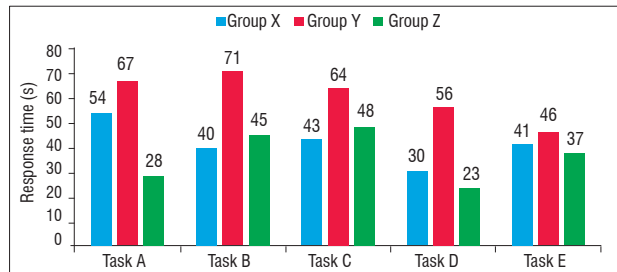

Figure 4: Average response time (in seconds) per task.

Discussion

In cartography and map making it is important to continuously update visualisation products so that they can match evolving consumer requirements. In this human interpretation of urban imagery, we studied the use of cartographic visual interpretation elements in cognitive interpretation of aerial data. Our findings show that pattern and shape are the strongest visual variables in cognition of an informal settlement within the built environment (Table 2). Pattern refers to the spatial arrangements of objects in a Euclidean space and it is a variable that the human eye captures easily, as shown by the results. Shape is the geometric characteristic of an object. This result is important in that it suggests that in the algorithm building phase, the use of methods that extract pattern- or shape-based information can be a significant first step towards informal unit extraction. These two variables can also be linked to texture and morphological operations as well as the application of object-based image segmentation which thrives on extracting objects in their complete geometry rather than as single pixels. In this study, the percentage of female participants in the expert, trained and children's groups was 40%, 50% and 60%, respectively. As there was no comparative testing of navigation skills, one cannot draw parallels with Levine et al.'s study[10] in which conclusions were made on the role of sex in cognitive tasks.

It is interesting to note the outcomes from the children's group. Children are exposed to elements of map reading in primary and early secondary education. Apart from formal education, technology and children's games are imparting spatial cognitive and map interpretation skills earlier in modern generations than in earlier generations. This has impacted much on the responses observed in the cognitive tests. The children gave encouraging interpretations showing their improved 'digital DNA' as compared to 'analogue times'.

Responses to Question 5 were not discussed much in the results section as they are likely to vary greatly based on user backgrounds, but it was noted that responses ranged from 'vegetation', 'white roof' and 'road' to 'order'. These responses suggest an element of unique personal individuality in cognition by which persons perceive an image differently. This finding creates the perception of a background subtraction model in the interpretation task by different individuals. A background subtraction model separates foreground – or what is attractive to an observer – from the background, which is the complement of the former. We conclude that texture and morphological operations could be a strong feature extraction approach to test in informal settlement unit extraction. Object-oriented analysis may also provide useful cues as it combines texture and morphologically based operations. In addition, the inclusion of background subtraction models in line with the outcomes of Question 5 could be a wise input into the informal settlement extraction problem.

The results can be improved by expanding the sample size and extending the study to a variety of locations that include different cultures and languages for comparison and to test the validity of perspectives such as the cultural and linguistic paradigm. It would also be interesting to make comparisons in cognitive abilities according to sex, age and profession. Further research could yield trends in the personalised aspects of spatial cognition by studying the backgrounds of the interpreters to draw inferences.

Authors' contributions

M.S. and J.L.S. conceptualised the study and designed the methodology; M.S. collected and analysed the data and wrote the first draft of the manuscript; and J.L.S. supervised the study, provided leadership and revised the manuscript.

References

1. Zheng Y. Bio-inspired color image enhancement model. Proc SPIE. 2009;7341, Art. #73410D. http://dx.doi.org/10.1117/12.819519

2. Meylan L, Susstrunk S. Electronic imaging. Bellingham, WA: International Society for Optics and Photonics; 2004. p. 46–56.

3. Bianchetti RA, MacEachren AM. Cognitive themes emerging from air photo interpretation texts published to 1960. ISPRS Int J Geo-Inf. 2015;4(2):551–571. http://dx.doi.org/10.3390/ijgi4020551

4. Estes J, Hajik E, Tinney L. Manual and digital analysis in the visible and infrared regions. In: Manual of remote sensing. 2nd ed. Falls Church, Virginia: American Society for Photogrammetry and Remote Sensing; 1983. p. 258–277.

5. Hoffman RR, Conway J. Psychological factors in remote sensing: A review of some recent research. Geocarto Int. 1989;4:3–21. http://dx.doi.org/10.1080/10106048909354230

6. Hodgson ME. What size window for image classification? A cognitive perspective. Photogramm Eng Rem S. 1998; 64:797–807.

7. Lloyd R, Hodgson ME. Visual search for land use objects in aerial photographs. Cartogr Geogr Inf Sci. 2002;29(1):3–15. http://dx.doi.org/10.1559/152304002782064592

8. Handy SL, Boarnet MG, Ewing R, Killingsworth RE. How the built environment affects physical activity: Views from urban planning. Am J Prev Med. 2002;23(2):64–73. http://dx.doi.org/10.1016/S0749-3797(02)00475-0

9. Kettunen P, Sarjakoski IT. A context-sensitive way finding ontology for hiking based on an empirical study. Poster presented at: Conference on Spatial Information Theory (COSIT11); 2011 September 12; Belfast, ME, USA.

10. Levine M, Marchon I, Hanley GL. The placement and misplacement of you-are-here maps. Environ Behav. 1984; 16:139–157. http://dx.doi.org/10.1177/0013916584162001

11. Walther DB, Cardigan E, Fei Fei L, Diane MB. Natural science categories revealed in distribution pattern of activity in human brain. J Neurosci. 2009;29(34):10573–10581. http://dx.doi.org/10.1523/JNEUROSCI.0559-09.2009

12. Downs RM, Stea D. Cognitive maps and spatial behaviour: Process and products. Chicago, IL: Aldine; 1973.

13. Kaplan MF, Miller CE. Judgments and group discussion: Effect of presentation and memory factors on polarization. Sociometry. 1977;40(4):337–343. http://dx.doi.org/10.2307/3033482

14. Edgin JO, Pennington BF. Spatial cognition in autism spectrum disorders: Superior, impaired, or just intact? J Autism Dev Disord. 2005;35(6):729–745. http://dx.doi.org/10.1007/s10803-005-0020-y

15. Koch WG. Jacques Bertin's theory of graphics and its development and influence on multimedia cartography. Inf Design J. 2001;10(1):37–43. https://doi.org/10.1075/idj.10.1.07koc

A morphometric analysis of hominin teeth attributed to *Australopithecus*, *Paranthropus* and *Homo*

AUTHOR:
Susan J. Dykes[1]

AFFILIATION:
[1]Evolutionary Studies Institute, School of Geosciences, University of the Witwatersrand, Johannesburg, South Africa

CORRESPONDENCE TO:
Susan Dykes

EMAIL:
Susan.Dykes@students.wits.ac.za

ARTICLE INCLUDES:
✓ Supplementary material
✗ Data set

Teeth are the most common element in the fossil record and play a critical role in taxonomic assessments. Variability in extant hominoid species is commonly used as a basis to gauge expected ranges of variability in fossil hominin species. In this study, variability in lower first molars is visualised in morphospace for four extant hominoid species and seven fossil hominin species. A size-versus-shape-based principle component analysis plot was used to recognise spatial patterns applicable to sexual dimorphism in extant species for comparison with fossil hominin species. In three African great ape species, variability occurs predominantly according to size (rather than shape), with the gorilla sample further separating into a male and a female group according to size. A different pattern is apparent for the modern human sample, in which shape variability is more evident. There is overlap between male and female modern humans and some evidence of grouping by linguistic/tribal populations. When fossil hominin species are analysed using equivalent axes of variance, the specimens group around species holotypes in quite similar patterns to those of the extant African great apes, but six individual fossil molars fall well outside of polygons circumscribing holotype clusters; at least three of these specimens are of interest for discussion in the context of sexual dimorphism, species variability and current species classifications. An implication of this study is that, especially in the case of modern humans, great caution needs to be exercised in using extant species as analogues for assessing variability considered to be a result of sexual dimorphism in fossil hominin species.

Significance:

- Caution should be exercised in using modern analogue species as proxies for fossil hominin species variability.

- Exceptionally wide ranges of molar variability between certain fossil hominin specimens currently allocated to the same species might indicate possible misclassification.

- Molar morphology in gorillas tends to reflect primarily size, rather than shape, variability between the sexes, which is a consideration in the context of assessing possible sexual dimorphism in fossil hominin species.

Introduction

Previous research has established that analyses of dental metrics and morphology on the post-canine dentition of extant hominoids are reasonably successful at differentiating between specimens at the species level and even at the subspecies/regional level.[1-7] Likewise, in the fossil hominin context, molar crown size, shape and cusp arrangements have traditionally been used as diagnostic tools to help to identify specimens attributed to different species of *Australopithecus, Paranthropus* and *Homo*.[8-14] However, taxonomic decisions cannot always be made with accuracy, particularly when the fossil record is incomplete, and boundaries between species are sometimes very indistinct.[15] It is common to use observed variability ranges in extant species as proxies for the quantification of expected variability in similar fossil species (for example, extant hominoids are often used as analogues for extinct hominins), but some caution needs to be observed in doing so.[16] Certain species, such as gorillas and orangutans, are known to be highly sexually dimorphic. In terms of their dental morphometrics, Uchida[1-2] has noted that although teeth vary greatly in size between the sexes, there are no significant differences between male and female gorillas and orangutans in terms of molar shape, in the context of their mean shape indices and cusp proportions. In the case of modern *Homo sapiens*, however, size differences are known to occur along regional or biogeographical lines, and although there may be regional variability and some sexual dimorphism in each region, certain groups globally have extremely large ('megadont') molars, while other groups have very small ('microdont') teeth by comparison.[17-19]

If molar morphology is linked to form and function, as researchers such as Kay[20] and Ungar[21] have postulated, then size reduction and shape changes are more likely to have occurred as a result of selective pressures over time in modern *H. sapiens* as diets and subsistence lifestyles have diverged between groups over millennia. Indeed, studies conducted on femora of modern human groups with differing lifestyles (e.g. hunter–gatherer; sedentary/farming; small-scale farmers) have noted that variability occurs as a function of subsistence lifestyle.[22,23] Other researchers have confirmed that this form–function variation along subsistence lifestyle lines is also found in molar metric variability as a result of long-standing divergences in diet in some groups after the Neolithic Revolution. As diets have become predominantly based on soft cereals and higher levels of cooking and food processing, tooth reduction has generally occurred in these groups[19,24-27], while other groups, such as Australian Aboriginal hunter–gatherers/terrestrial foragers, have retained large, robust molars[17,18,28]. Dietary and subsistence-lifestyle histories may not be the only factors at play in determining the wide variability in size and shape of modern human molars, but the fact remains that although there may be measures of sexual dimorphism within biogeographical groups individually,[29] if molar morphology were to be viewed in morphospace in the same way as that of other

hominoids, while gorilla teeth should separate (by species/sub-species) into a male group and a female group primarily by size,[1] modern human teeth would be expected to group by biogeographical population or by historical subsistence lifestyle divergences initially, and by sex thereafter.

In the case of *Pan* species, variation between the lower first molars of common chimpanzees and bonobos is linked to allometry, such that when the effects of allometry are factored out, 'chimpanzee and bonobo molars are not morphometrically distinguishable'[7]. It is not within the scope of this study to correct for allometry, but because mandibular molar morphometrics are strongly correlated with size in *Pan,* a visualisation in morphospace using size as the first principal component axis should achieve a good degree of discrimination between these two closely related species.

The aim of this study was to build upon previous studies by using geometric morphometric methods to provide a visual analysis of size-versus-shape variability patterns in the post-canine dentition of extant hominoid species and their implications for the analysis of molar variability within and between fossil hominin species. The goal was to obtain a general understanding, not only of patterns of size-versus-shape variability in the lower first molars of African ape species, but also of how these African ape variability patterns differ to those observed in *H. sapiens*. Understanding this difference is important because the typically high ranges of variability in modern human skeletal elements (including teeth) are often used as benchmarks for quantifying the expected range of variability in skeletal elements of fossil hominin species, and further, in species such as *Australopithecus afarensis,* modern human-like sexual dimorphism is cited as the primary factor to explain such high variability between specimens of this species.[30,31] In this context, Ferguson[32] strongly emphasises the need to take into account factors such as globalisation and differences in self-domestication between modern human populations before comparing dental variation between a fossil hominin species such as *Au. afarensis* and modern *H. sapiens.* His conclusion is that dental variation in modern *H. sapiens* is 'not evidence of normal dental variation in hominids'[32]. The aim of the present study was to provide a visualisation of both the range of variability and, more importantly, the pattern of general size/shape variability that would be expected of selected hominoid species, in the context of sexual dimorphism.

In particular, the following specific questions are addressed:

- Do specimens of selected extant hominoids group in morphospace in a way that confirms previous research (using a size-versus-shape principal components analysis to visualise the main axes of variability), particularly in the context of sexual dimorphism?

- Does the pattern of variability in morphospace (size versus shape) of modern *H. sapiens* differ from that of extant African great apes?

- Using equivalent principal components axes (size versus shape), do fossil hominin lower first molars group in morphospace in a similar way to those of gorillas, other African great apes, or modern *H. sapiens,* and if so, what conclusions should be drawn from these groupings?

- Are there certain instances in the fossil hominin record where the lower first molars of individual specimens attributed to a particular hominin species differ so significantly in size and/or shape from those of the other specimens in the group (including the type specimen or holotype of the species) that these specimens warrant further discussion in respect of species variability, sexual dimorphism or potential misclassification?

The aim of the present study was thus to test the predictions (1) that sexual dimorphism should be observable between lower first molars of male and female *Gorilla gorilla gorilla,* primarily according to size; (2) that variability between lower first molars of male and female modern humans may follow a different pattern in morphospace to that of African great ape species, possibly being observable primarily along biogeographical lines, and only secondarily according to sex; and (3) that certain specimens in the fossil hominin record may appear as outliers

from the typical individuals of their species, raising the possibility of misclassification.

Materials and methods

Digital two-dimensional images of 40 lower first molars (occlusal crown images) from 20 (10 male, 10 female, both antimeres) individuals each of *Gorilla gorilla gorilla, Pan troglodytes schweinfurthii, Pan paniscus* (from the Royal Museum for Central Africa in Tervuren, Belgium), and modern *Homo sapiens* (from the R. A. Dart Collection of the University of the Witwatersrand, Johannesburg, South Africa) were analysed to determine variability within and between species, and between sex within species, for comparison with 36 African Plio-Pleistocene lower first molars from 27 individuals [including five holotypes (*Australopithecus afarensis, Australopithecus africanus, Paranthropus robustus, Homo habilis* and *Homo erectus*) as well as Peninj 1, which is a mandibular proxy for the holotype of *Paranthropus boisei*]. Because all of the fossil specimens were from African Plio-Pleistocene hominin species, African great apes and modern humans from southern Africa were selected for comparative purposes.

Two-dimensional imagery was chosen for the study because holotypes of certain of the fossil hominin species were extremely worn (e.g. *Au. afarensis* and *P. boisei*), and would not have been able to be included in a three-dimensional analysis, but these specimens were still usable in a study relying on landmark analyses wherein homologous cusp intersections at the perimeter of the occlusal crown view were still discernible. Coupled with a geometric approach to landmarking the surfaces of the crowns, this enabled more fossil specimens (including holotypes and proxies thereof) to be included in the study, even if most of the topography and surface features on the crowns were worn or obliterated. Right antimeres were mirrored to appear as left molars.

Where possible, antimeres were included, because in many individuals there is odontometric asymmetry, which has been linked to tooth eruption patterns, masticatory loads and laterality (handedness) in modern humans.[33-35] As this asymmetry is generally manifested in the form of dimension differences between the two sides, the inclusion of antimeres enabled observations of the potential cause of spatial patterning differences to be controlled for, between sex as well as hemisphere, because for the fossil sample (where sex is unknown but laterality is known), in some cases only left lower molars or right lower molars were available, and to select only left or right specimens would cause a significant reduction in *n* for an already limited fossil hominin sample. Details for the specimens used are given in Tables 1–3.

Table 1: Lower first molar specimens included in the comparative study: Modern *Homo sapiens*

Catalogue number Dart Collection	Population/language group (as stated in the catalogue)	Sex	Age at death
A1263	Sotho (South Africa)	F	18
A1483	Tswana (South Africa or Botswana)	F	19
A3607	Mixed (European & African)	F	40
A84	Amafengu (South Africa)	F	38
A27	San ('Bushman' – South Africa)	F	N/A
A281	Sotho (South Africa)	M	18
A1264	Tswana (South Africa or Botswana)	M	23
A3421	Mixed (European & African)	M	60
A861	Amafengu (South Africa)	M	34
A173	San ('Bushman' – South Africa)	M	N/A

Table 2: Lower first molar specimens included in the comparative study: Extant African great apes

Species	RMCA (Tervuren) catalogue number	Sex
Gorilla gorilla gorilla	7732M15	F
Gorilla gorilla gorilla	7732M8	F
Gorilla gorilla gorilla	7732M5	M
Gorilla gorilla gorilla	7318M3	F
Gorilla gorilla gorilla	7732M7	M
Gorilla gorilla gorilla	7732M6	M
Gorilla gorilla gorilla	7732M3	F
Gorilla gorilla gorilla	7732M2	M
Gorilla gorilla gorilla	7732M1	M
Gorilla gorilla gorilla	7556M2	F
Pan troglodytes schweinfurthii	91060M422	F
Pan troglodytes schweinfurthii	91060M414	M
Pan troglodytes schweinfurthii	91060M410	F
Pan troglodytes schweinfurthii	91060M406	F
Pan troglodytes schweinfurthii	83006M37	F
Pan troglodytes schweinfurthii	83006M32	F
Pan troglodytes schweinfurthii	83006M22	M
Pan troglodytes schweinfurthii	83006M21	M
Pan troglodytes schweinfurthii	83006M17	M
Pan troglodytes schweinfurthii	83006M15	M
Pan paniscus	84036M11	F
Pan paniscus	84036M03	M
Pan paniscus	29055	F
Pan paniscus	29053	M
Pan paniscus	29050	M
Pan paniscus	29027	F
Pan paniscus	29028	M
Pan paniscus	29026	F
Pan paniscus	13021	F
Pan paniscus	11354	M

Table 3: Lower first molar specimens included in the comparative study: Fossil hominin specimens

| Specimen number	Side	Marker used	Current taxonomic designation	Location	Comments	Geological paper reference	Estimated geological age (Ma) (latest estimate)	Reference for estimated date
AL 145-35	L	♦	*Australopithecus afarensis*	Wits cast collection, Johannesburg	Cast	36	3.35	58
AL 128-23	R	♦	*Au. afarensis*	Wits cast collection, Johannesburg	Cast	36	3.25	58
AL 266-1	L	♦	*Au. afarensis*	Wits cast collection, Johannesburg	Cast	36	3.2	58
AL 266-1	R	♦	*Au. afarensis*	Wits cast collection, Johannesburg	Cast	36	3.2	58
AL 288-1	R	♦	*Au. afarensis*	Wits cast collection, Johannesburg	Cast	37	3.18	58
AL 333-W60	L	♦	*Au. afarensis*	Wits cast collection, Johannesburg	Cast	38	3.2	58
LH 2	R	♦	*Au. afarensis*	National Museums of Kenya, Nairobi	Cast	39,40	3.77	59
LH 4	L	♦	*Au. afarensis*	National Museum, Dar es Salaam	Holotype – original	39, 40	3.77	59
LH 4	R	♦	*Au. afarensis*	National Museum, Dar es Salaam	Holotype – original	39 40	3.77	59
MLD 2	L	▲	*Au. africanus/ (Au. prometheus)*	Wits fossil collection, Johannesburg	Original	41, 42	2.8	59
MLD 2	R	▲	*Au. africanus/ (Au. prometheus)*	Wits fossil collection, Johannesburg	Original	41, 42	2.8	59
Sts 52b	R	▲	*Au. africanus*	Ditsong National Museum, Pretoria	Original	43	2.3	60
Taung 1 (U.W. 1-1)	L	▲	*Au. africanus*	Wits fossil collection, Johannesburg	Holotype – original	44	2.7	59
Taung 1 (U.W. 1-1)	R	▲	*Au. africanus*	Wits fossil collection, Johannesburg	Holotype – original	44	2.7	59
OH 22	R	●	*Homo erectus*	National Museum, Dar es Salaam	Original	45	0.875	59
KNM-ER 806c	L	●	*H. erectus*	National Museums of Kenya, Nairobi	Original	46	1.49	9
KNM-ER 820	L	●	*H. erectus*	National Museums of Kenya, Nairobi	Original	46	1.6	9
KNM-ER 820	R	●	*H. erectus*	National Museums of Kenya, Nairobi	Original	46	1.6	9
KNM-ER 992	L	●	*H. erectus*	National Museums of Kenya, Nairobi	Holotype – *Homo ergaster* (African *Homo erectus*) original	47	1.49	59,61
KNM-ER 992	R	●	*H. erectus*	National Museums of Kenya, Nairobi	Holotype – *Homo ergaster* (African *Homo erectus*) original	47	1.49	9,61
OH 7	L	●	*H. habilis*	National Museum, Dar es Salaam	Holotype – original	48, 49	1.84	59
OH 7	R	●	*H. habilis*	National Museum, Dar es Salaam	Holotype – original	48, 49	1.84	59
OH 16	R	●	*H. habilis*	National Museum, Dar es Salaam	Original	50	1.74	59
KNM-ER 1802	L	●	*H. rudolfensis*	National Museums of Kenya, Nairobi	Original	51	1.89	9
KNM-ER 1802	R	●	*H. rudolfensis*	National Museums of Kenya, Nairobi	Original	51	1.89	9
KNM-ER 15930	L	■	*Paranthropus boisei*	National Museums of Kenya, Nairobi	Original	52	1.78	9
Peninj 1	L	■	*P. boisei*	National Museum, Dar es Salaam	Mandibular proxy for holotype (OH 5) – original	53	1.4	59
Peninj 1	R	■	*P. boisei*	National Museum, Dar es Salaam	Mandibular proxy for holotype (OH 5) – original	53	1.4	59
SK 6	L	■	*P. robustus*	Ditsong National Museum, Pretoria	Original	54	1.75	59
SK 6	R	■	*P. robustus*	Ditsong National Museum, Pretoria	Original	54	1.75	59
SK 23	L	■	*P. robustus*	Ditsong National Museum, Pretoria	Original	54	1.75	59
SK 23	R	■	*P. robustus*	Ditsong National Museum, Pretoria	Original	54	1.75	59
SK 63	L	■	*P. robustus*	Ditsong National Museum, Pretoria	Original	55	1.75	59
SK 63	R	■	*P. robustus*	Ditsong National Museum, Pretoria	Original	55	1.75	59
SKW 5	R	■	*P. robustus*	Ditsong National Museum, Pretoria	Original	56	1.75	59
TM 1517	R	■	*P. robustus*	Ditsong National Museum, Pretoria	Holotype – original	57	1.75	59

Photographic and image-processing methods

Digital photographs were taken using a Nikon D3200 digital SLR 24-megapixel camera, with an adjustable scale bar placed in each image at the height of the plane of the occlusal crown surface. The tooth being photographed was centred orthogonally below the lens of the camera, well in the centre of the frame of the image, and was aligned in the horizontal using the cervical plane as a horizontal guide; however, because it was not always possible to verify this plane, particularly along the buccolingual axis, and where immersion in levelled sand was not advisable because of the delicate nature of the specimens, visual alignment of the vertical lingual and buccal crown edges was used. The accuracy of the horizontal/vertical alignment of the tooth was tested by using three- to two-dimensional image superimpositions, wherein three two-dimensional digital images, taken for the same tooth at different times, were each inserted into the correct plane of a three-dimensional image of the same tooth using Amira® software, after which differences in the x-, y- and z-axes between the three alignments were measured. The resultant error (averaging 0.014° along the x-axis, 0.107° along the y-axis and 0.098° along the z-axis) was considered to be within acceptable limits, after correlations of landmark measurements at varying degrees of tilt had previously established that errors of tilt of up to 2° from the horizontal produced a correlation coefficient of 0.99 in relation to the same measurements at zero tilt. For purposes of the analyses and the landmark placements, the digital images were each aligned horizontally on screen along the longitudinal axis of the tooth, using the protocols of Wood[9] and Goose[62] as guidelines for this alignment (taking into account the 'normal' alignment[62] of the tooth in question as well as that of its immediate neighbours). A rectangle was superimposed over the image in the form of a bounding box (using Adobe Illustrator®) to act as a proxy for the mesiodistal (minimum) diameter and the buccolingual (maximum) diameter, respectively, according to definitions by Wood[9]. The centre of the bounding box was then calculated for purposes of a fan overlay placement, which would serve as a basis for the placement of landmarks at each 15° interval around the perimeter of the occlusal crown shape. The alignment and the position of this mathematical centre point were subjected to inter-observer tests for accuracy (average error: 0.295% in the x-axis and 0.316% in the y-axis).

Landmarking method

Five Type I (homologous or 'anatomical') landmarks[63] were sited at the intersection of each of the cusps at the perimeter outline of the occlusal crown surface in the image, and a further 44 Type III (mathematical or 'constructed') landmarks[63] were placed in such a way as to define (1) the buccolingual and mesiodistal diameters of each tooth (four landmarks placed on a bounding box enclosing the tooth plus a fifth at the mathematical centre of the bounding box); (2) the peripheral shape of the tooth (24 landmarks, every 15° around the tooth perimeter); and (3) the orientation of the cusps (five landmarks denoting the midpoint between cusp intersections and five landmarks measured equidistantly from the central landmark to the centre of the cusp arc at the perimeter, along the midline of the cusp, with a further five landmarks equidistant between these central cusp landmarks; in cases where a diagnostic sixth cusp was evident, the landmark normally sited midway between the landmarks at the centre points of the entoconid and the hypoconulid was sited instead at the midpoint between the landmark at the centre of the entoconid and the landmark at the intersection between the sixth cusp and the hypoconulid at the perimeter, so that the landmark would be sited over the sixth cusp). The position of the landmarks is shown in Figure 1.

The images were scaled and landmarks digitised using ImageJ® software and processed via Microsoft Excel® into IBM SPSS® and Morphologika® for purposes of performing Procrustes superimposition, principal components analyses and discriminant function analyses. Lastly, a custom-written macro for MS Excel®, named 'Professor Regressor'[64] was created to produce a high-speed throughput of pairwise regressions for purposes of conducting log se_m analyses to determine average conspecific variation.[65]
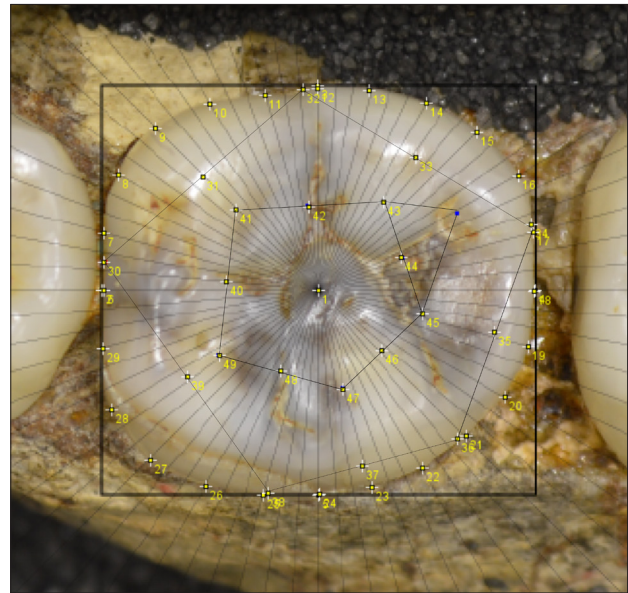
Figure 1: Landmark placements. The first five landmarks describe mesiodistal/buccolingual measurements. Landmarks 6–29 are sited to describe points every 15° along the outline of the perimeter of the tooth in the occlusal view. Landmarks 30–39 are sited at the points at which cusps intersect at the perimeter, together with the midpoints of the cords drawn across the arcs that are described by these cusps and describe an outer pentagon denoting peripheral widths and orientations of cusps. Landmarks 40–49 describe an inner pentagon (or hexagon in the case of a C6) denoting the geometric centre of each cusp as calculated from Landmark 1 to the perimeter of the tooth along the midline of each cusp arc. Thus general proportions (breadth/length), cusp size and cusp orientations are able to be landmarked, even if the crown surface is devoid of diagnostic features, provided that the five cusp intersections at the peripheral outline of the tooth are visible (Type 1 landmarks: 30, 32, 34, 36 and 38).

Methods: Analyses

A principal components analysis (PCA) was first performed on the sample from the four extant species' lower first molars in Morphologika after performing a generalised Procrustes superimposition wherein specimens are translated, rotated and scaled and then plotted on a graph showing the main axes of differentiation from a 'consensus' tooth; a second PCA was conducted using 'Procrustes form space', wherein size is factored back into the analysis by including the log of the centroid size for each shape as a variable in the analysis.[66] This second PCA thus provides a 'size–shape' analysis, aimed at visualising differentiation between similarly shaped, but differently sized, molars (e.g. male and female gorilla molars) on the graph. In a form space analysis, size becomes the predominant factor of variance along the first principal component (PC) axis (the x-axis). The second PC axis (the y-axis) summarises the main shape differences between specimens, statistically independently of size.[67] In the case of the extant species, PC2 summarised the primary shape variation to be a function mainly of relative breadth of the occlusal surface (high to low mesiodistal:buccolingual ratios), together with aspects of cusp orientation and perimeter shape differences. When fossil specimens were added to the analysis, examination of the thin plate spline warps for higher-order PCs showed that PC3 accounted for almost exactly the same variability factors as PC2 had done in the analysis for the extant species alone (relative breadth, cusp orientation and perimeter shape), and for all PC plots involving fossil specimens PC3 was selected as the y-axis component of the plot.

Thereafter, a discriminant function analysis (DFA) was carried out using the first eight PC scores on the four extant species and fossils together. The number of PC scores to include was decided on the basis

of a sensitivity test to determine the minimum number of PC scores needed before the four extant species in the analysis were classified 100% correctly. The data were previously verified as being normally distributed, using a Shapiro–Wilk test.

A further analysis of intra-species variability was carried out using the 'log se$_m$' methodology pioneered by Thackeray et al.[68] The basic premise behind this analysis is that within any one species, skeletal variability falls within a certain range, and although the range varies from species to species, there is an average range or central tendency of variability across species in general, that approaches what Thackeray et al.[15,65,69] call an approximation of a biological species constant (T=-1.61), and a comparison between any two specimens of unknown species group can be compared against this figure to establish a statistical probability of conspecificity[68,70]. To calculate the log se$_m$ values, pairwise regressions of measurements for specific skeletal elements are conducted, firstly with specimen A on the *x*-axis and specimen B on the *y*-axis, and then with specimen B on the *x*-axis with specimen A on the *y*-axis. The standard error of the slope *m* is calculated for the regression equation $y=mx+c$, and this is then log transformed to provide two paired log se$_m$ values, with the difference between the two values being designated as 'delta log se$_m$'.[64,65] Specimens of a similar size from the same species will be expected to have very low log se$_m$ values, because the standard error of the slope is a measure of 'scatter' around the regression slope between the two specimens, and there should be a high degree of correlation between points along a regression line for two specimens with similar shape, with more predictable expected *y*-values as a result; and because the specimens are similarly sized, both of the slopes would approach a gradient of 1, and the *x*-on-*y* and *y*-on-*x* values would be barely distinguishable from each other, thus the delta value should also be low. However, two specimens from different species with large shape and size differences between them will not only be poorly correlated in terms of shape (the standard error of the slope would be larger, as a result of the large amount of scatter around both lines), giving a high log se$_m$ value in at least one of the two slopes, but the delta value between the two log se$_m$ values would be high, as the two slopes for each pairwise comparison would have very different gradients as a result of the size differences.

A total of 760 pairwise comparisons of conspecific pairs of lower first molars of *G. gorilla*, *P. troglodytes*, *P. paniscus* and *H. sapiens* were analysed using measurements taken radially from the centre of the tooth to the landmark points as described above. A further 252 pairwise comparisons of conspecific pairs of lower first molars of the fossil hominin species were analysed, and the average species variability was compared against the average obtained for the extant species groups. After exclusion of any atypical (potentially misclassified) outlier specimens from the analysis, the remaining 176 pairwise comparisons of conspecific pairs were averaged and the results again compared with the results for the extant species groups.

Results

### Sexual dimorphism in gorillas identified in morphospace

Three main clusters in morphospace are detectable from the four extant species' samples using a shape-only PCA (Figure 2). Modern *H. sapiens* exhibit a high degree of shape variability, and have generally relatively wider lower first molars than those of the great apes (Figure 2). Gorilla molars exhibit a distinctive shape, while those of bonobos and common chimpanzees exhibit overlap in the PCA because their molars are less distinguishable between species in terms of shape when scaling is used (Figure 2).

As expected, sexual dimorphism is not evident even within the highly sexually dimorphic gorilla sample in the shape-only analysis, which suggests that lower first molars of male and female gorillas do not vary significantly in their shape. This finding confirms those of previous studies.[1,2] The second PCA, in which size was factored back into the analysis, is shown in Figure 3.

In Figure 3 (size versus shape), all four species are now relatively separate (as the inclusion of size now allows a better separation of bonobos from common chimpanzees). Additionally, there is a very clear separation between male gorillas and female gorillas as a result of differences in size of lower first molars (but as has been seen, not in overall shape).
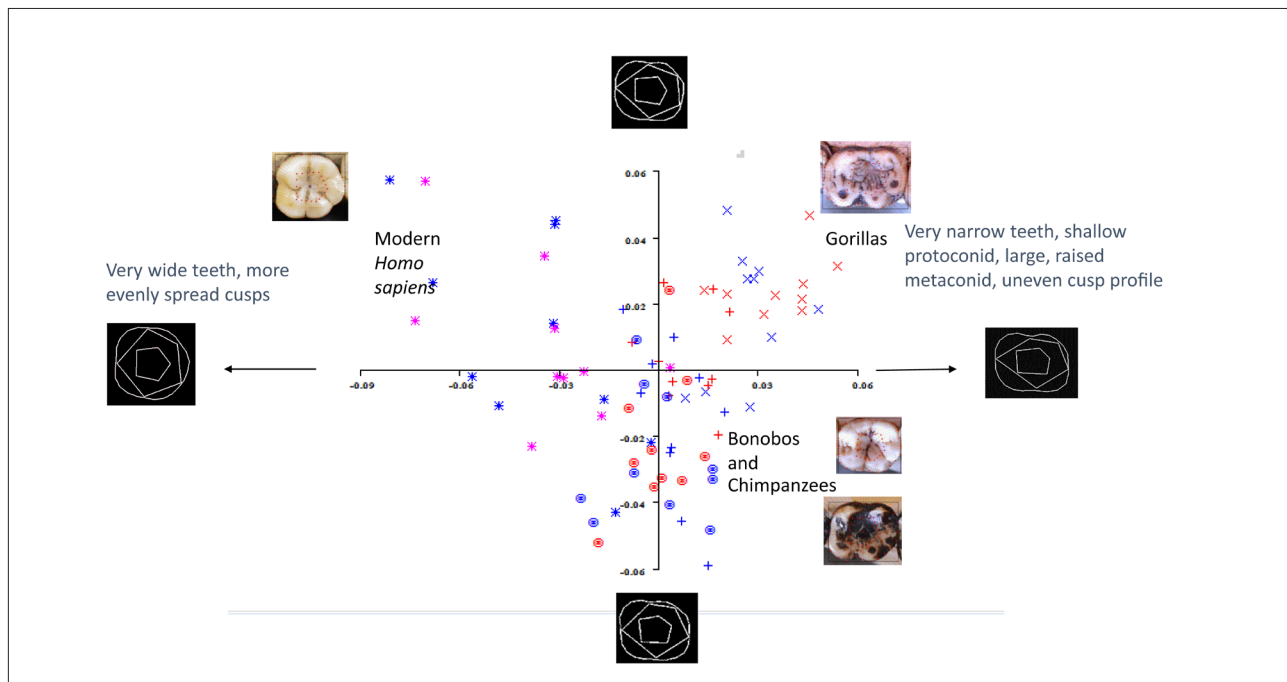
Figure 2: Principal components analysis of four extant species' lower first molars, based on Procrustes shape space (shape-only analysis). PC1 in this plot mainly accounts for the relative width of the tooth with some shape variability (28% of covariance). PC2 principally accounts for tooth shape variability and cusp orientations. *Homo sapiens* is represented by stars; *Gorilla gorilla gorilla* by diagonal crosses; *Pan paniscus* by upright crosses and *P. troglodytes* by circled target markers (pink/red=female and blue=male specimens). In this shape-only analysis, male and female gorilla specimens are not separated by sex, and there is considerable overlap between bonobo and common chimpanzee specimens.
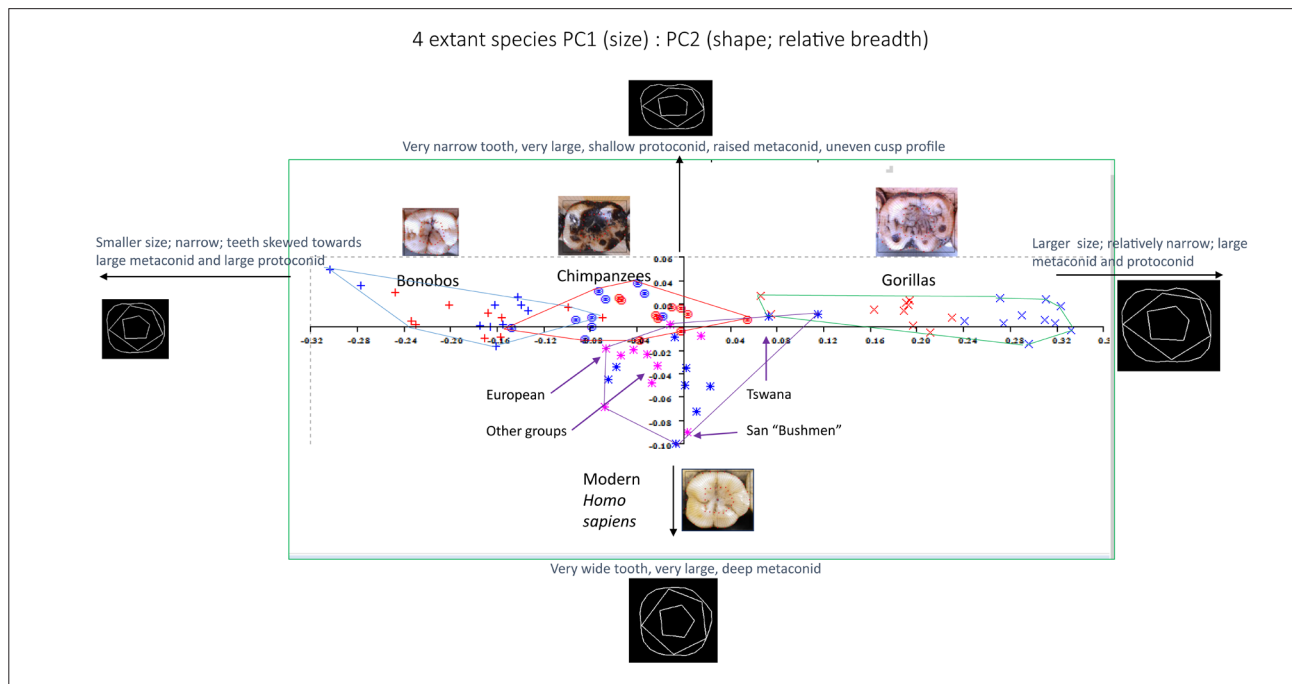
Figure 3: Principal components analysis (PCA) of four extant species' lower first molars, based on Procrustes form space (size-versus-shape analysis). PC1 in this plot mainly accounts for size differences; PC2 principally accounts for relative width of the tooth with some shape and cusp-orientation variability, formerly represented by PC1 in the shape-only PCA (Figure 2). *Homo sapiens* is represented by stars; *Gorilla gorilla gorilla* by diagonal crosses; *Pan paniscus* by upright crosses and *P. troglodytes* by circled target markers (pink/red=female and blue=male specimens). The three African great ape species vary predominantly along the *x*-axis direction (variability predominantly by size rather than by shape); male and female gorillas in the sample are separated into two distinct groups predominantly by size rather than shape; modern humans vary greatly by shape as well as size, and to a certain extent by population/linguistic groupings: the lower first molars of male and female San individuals are of medium overall size but very 'square' in occlusal crown shape; those of Tswana individuals are the largest and narrowest (the lower first molar of the male individual groups alongside that of female gorillas); lower first molars of specimens with European heritage are among the smallest and narrowest of molars in this sample.

Modern Homo sapiens*: Shape variability by population group*

All three African great ape species seem to have a limited range of shape variability within the species/subspecies selected for this study, which fall within narrow limits in terms of their relative length:breadth ratio (relative breadth of crown and cusp arrangements being the primary factors of variation shown along the *y*-axis). It is evident that the most variable within-species group is the modern *H. sapiens* sample, which varies mostly in overall proportion and shape (*y*-axis), rather than overall size (*x*-axis). On further analysis, the variability within this group appears to occur not primarily by sex (as in the gorilla sample, falling as it does into two distinct groups according to sex) but by biogeographical groupings, with teeth from the African San population being relatively square in shape (those from both male and female individuals), and those from the African Tswana group being very large and relatively very narrow in overall proportion – in fact, the teeth from two male Tswana individuals, chosen to represent megadont populations for the analysis, group with the female gorilla sample on the size-shape PC analysis because they are much larger and narrower.

When the same size-shape PC analysis as that applied to extant hominoid species is then applied to include the fossil hominin lower first molar sample (with the same shape parameters selected for the *y*-axis as were used for the extant species alone), the African great apes all group similarly as they had before – largely above the *x*-axis (narrow teeth), with size being the main source of variance between species and between male and female gorillas. The modern human sample again varies largely along the *y*-axis (wider range of shape variance – from wide to narrow across the breadth of the tooth) and most of the fossil specimens group in the bottom right quadrant (larger and wider teeth, as would be expected). *Paranthropus* species seem to follow a gorilla-like

grouping parallel to the *x*-axis (variability being predominantly according to size), with this species exhibiting tight cohesion as a well-defined group on the plot, while *Au. afarensis* specimens exhibit wider variance along both size and shape axes, with individuals overlapping with modern *H. sapiens*, *H. erectus* and *Au. africanus* (Figure 4).

After removing the extant species from the PCA plot, a much clearer picture of shape versus size variability in morphospace is obtained for the fossil hominin species. There is generally good discrimination between species, around their holotypes or holotype proxies (Figure 5).

The distribution of specimens within species groups in this plot shows groupings visually akin to those in the shape-size analysis of the PCA plot of the extant African ape species. The group of *Paranthropus* spp. (*P. robustus* in green squares and *P. boisei* in brown squares) cluster along the *x*-axis in a similar pattern to the gorilla sample (predominantly varying in size rather than shape). The fact that they seem additionally to group in two size clusters (smaller teeth on the left, larger on the right) might possibly hint at sexual dimorphism within this group, as with the gorillas. Specimens from *H. erectus*, the species with the smallest molars in the analysis, cluster visually in morphospace in a similar manner to bonobos.

In the case of *Au. afarensis*, the majority of specimens cluster horizontally around the holotype (varying mainly by size, parallel to the *x*-axis, in a similar manner to the gorilla grouping in morphospace), but two specimens fail to cluster with their group.

Anomalous 'outliers'

Six specimens among the 36 fossil hominin lower first molars failed to group in morphospace with the clusters located around the holotypes for each species.
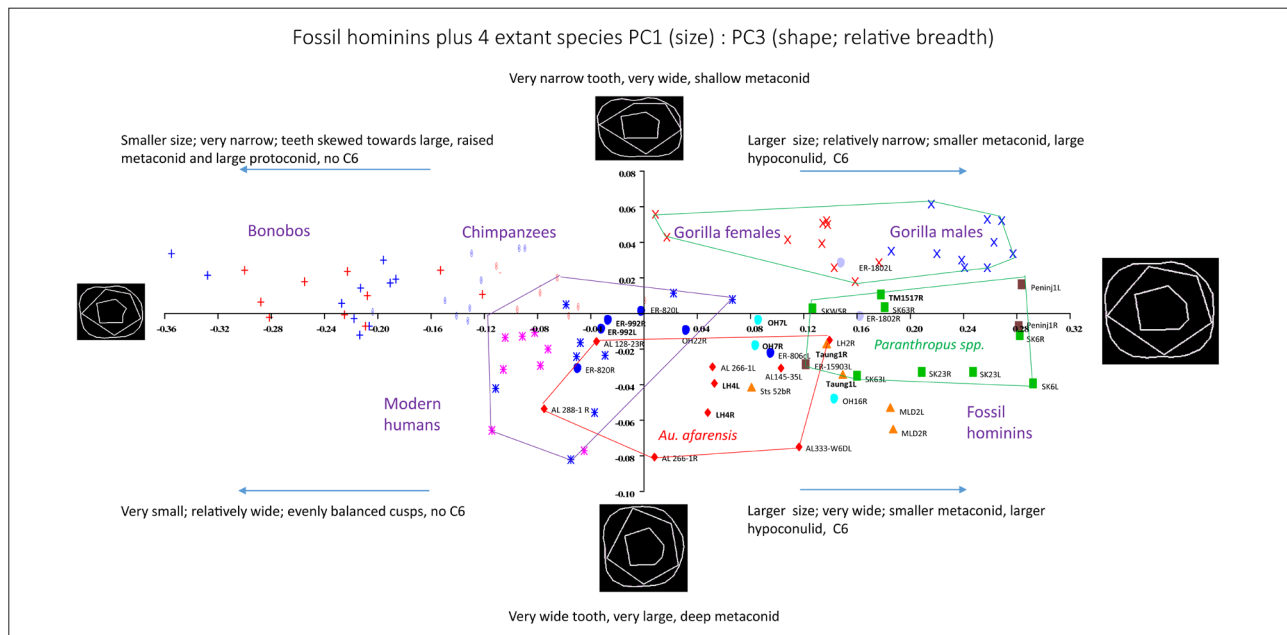
Figure 4: Principal components analysis of extant hominoid and fossil hominin species' lower first molars, based on Procrustes form space (size-versus-shape analysis). PC1 in this plot mainly accounts for size differences. PC3 principally accounts for relative width of the tooth with some shape and buccolingual cusp-orientation variability (the equivalent axis of variance that had been described along PC2 in the analysis for the extant species alone). Species markers are: red diamonds, *Australopithecus afarensis*; orange triangles, *Au. africanus*; lilac circles, *Homo rudolfensis*; turquoise circles, *H. habilis*; blue circles, *H. erectus*; green squares, *Paranthropus robustus*; brown squares, *P. boisei*. *Paranthropus* species plot in morphospace in a cohesive group with little overlap with other species, predominantly parallel to the *x*-axis (size differences as opposed to shape variability) and compare very well, spatially, to the patterning of the gorilla or other great ape samples. *Au. afarensis* specimens form a less cohesive group in morphospace, with individuals overlapping with numerous other species; there is a higher measure of shape variability alongside size variability in this species.
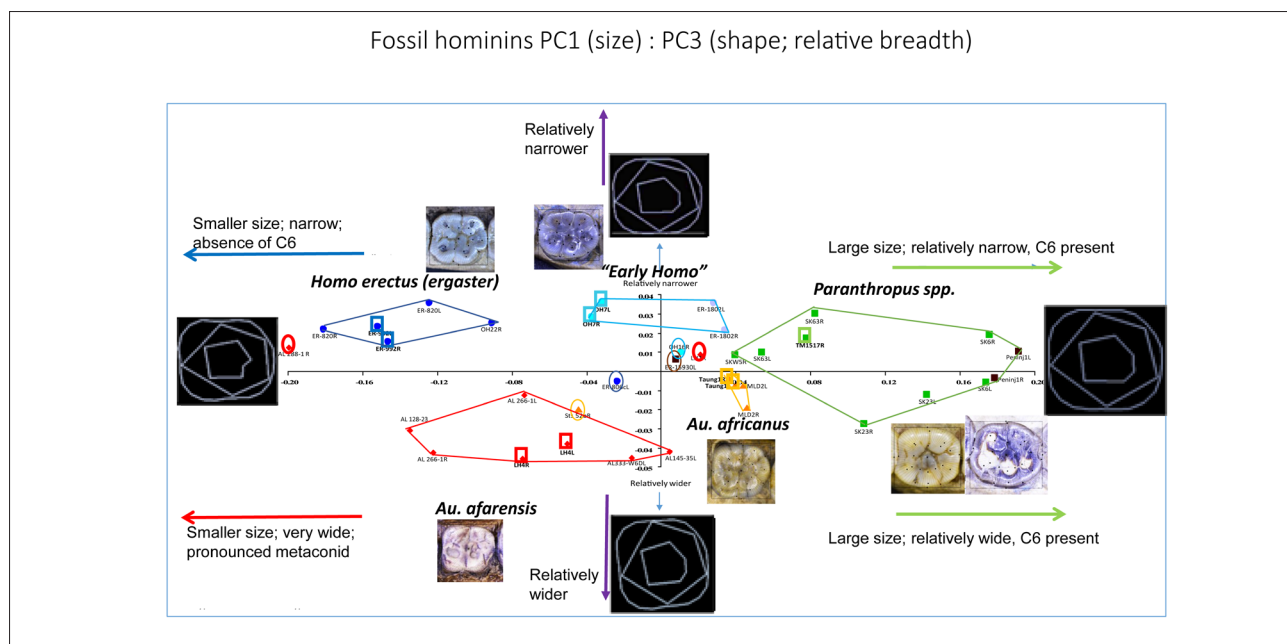
Figure 5: Principal components analysis of fossil species' lower first molars, based on Procrustes form space (size-versus-shape analysis). PC1 in this plot mainly accounts for size differences together with the presence or absence of a sixth cusp, with smallest specimens (lacking C6) at the negative extreme of the *x*-axis and largest specimens (C6 present) at the positive extreme of the *x*-axis; PC3 again principally accounts for relative width of the tooth with some shape and buccolingual cusp-orientation variability. Holotypes of species are marked by rectangular boxes. Species markers are: red diamonds, *Australopithecus afarensis*; orange triangles, *Au. africanus*; lilac circles, *Homo rudolfensis*; turquoise circles, *H. habilis*; blue circles, *H. erectus*; green squares, *Paranthropus robustus*; brown squares, *P. boisei*. Six specimens are marked with circles to illustrate that they do not group with the specimens of their currently allocated species that cluster around their holotype: AL 288-1 (currently allocated to *Au. afarensis* but groups more closely with *H. erectus*); Sts 52b (currently allocated to *Au. africanus* but groups more closely with *Au. afarensis*); KNM-ER 806c (currently allocated to *H. erectus* but groups more towards *Au. africanus* in general dimension, although smaller in size); OH 16 (currently allocated to *H. habilis*; wider buccolingually than the holotype); KNM-ER 15930 (currently allocated to *P. boisei*; extremely small in size for this group); LH 2 (currently allocated to *Au. afarensis* but groups closely with *Au. africanus*).

One outlier from the main cluster of *H. erectus* in morphospace is KNM-ER 806c, a specimen with wider buccolingual dimensions than the generally smaller and narrow molars typical of this species. This specimen also has a sixth and seventh cusp, a protostylid and a morphology extremely similar to that of the lower first molar of MLD 2, currently classified as *Au. africanus.*

Other specimens that do not seem to be typical of their species (which otherwise cluster well in morphospace) are Sts 52b, which groups with *Au. afarensis* rather than with *Au. africanus;* OH 16, which is wider, buccolingually, than the holotype of *H. habilis,* into which it has been classified; and KNM-ER 15930, which has been attributed to *P. boisei* but which is well outside of the normal size range for this species, being closer to the 'early *Homo*' group.

Notably, there are two specimens from *Au. afarensis* which group well away from the cluster around the holotype of the species in morphospace. At first glance, the *Au. afarensis* sample seems to mirror the spatial distribution of *Paranthropus* and gorillas, with several specimens clustering around the holotype (LH 4) in a polygon suggesting more size variability than shape variability. The first atypical outlier, LH 2, which is from a juvenile from the Laetoli area of Tanzania, groups closely with the *Au. africanus* cluster, being larger than the remaining specimens in the species group, and relatively more narrow in overall dimension. The second, AL 288-1, 'Lucy', groups into the quadrant in which *H. erectus* is located, but is even smaller in dimension than specimens clustering in this group.

A DFA conducted on all 116 specimens in the study confirmed that for the four extant species, the most variability was exhibited between specimens in the modern *H. sapiens* group as demonstrated by the mean squared Mahalanobis distance from the group centroid (modern *H. sapiens*: mean = 9.69±4.41 (n=20); *G. g. gorilla*: mean = 6.0±3.51 (n=20); *P. t. schweinfurthii*: mean = 4.51±1.94 (n=20); *P. paniscus*: mean = 5.36±3.20 (n=20)). For the fossil hominin species, the DFA confirmed that as a group, *Au. afarensis* was most variably distributed, with a mean squared Mahalanobis distance of specimens from their group centroid at 12.075±5.52 (n=9). *P. robustus*, with a mean of 6.49±3.11 (n=8) seems again to parallel the gorilla sample in terms of variability. *P. boisei*, with only three specimens in the sample, had a high mean Mahalanobis distance from the group centroid at 8.22±5.49 (n=3), as a result of the huge size difference between the smallest in the group (KNM-ER 15930) and the large Peninj molars that are more typical of this megadont species. The *H. erectus* group's mean squared Mahalanobis distance was also fairly high at 7.28±5.76 (n=6), but, if the one outlier (KNM-ER 806c) is excluded from the sample, this group, otherwise very homogeneous, would have a much lower mean distance from the centroid, as this single outlier's distance from the group centroid was 18.20. Another fossil hominin group with a high mean Mahalanobis distance from the group centroid was *Au. africanus* at 8.87±2.07 (n=5).

With respect to group classification predictions for individual specimens in the fossil hominin species, the DFA output confirms the anomalous status of AL 288-1 (grouped with *H. sapiens* because of its narrowness and relatively tiny size). KNM-ER 806c is predicted to classify, unsurprisingly, with the *Au. africanus* group, because of its extreme similarity with MLD 2. The third potential misclassification in this analysis is Sts 52b, which is more predictably classified as *Au. afarensis,* confirming the PCA plot results. One other instance of potential misclassification according to the DFA was that of Taung 1 (left and right antimeres), which classifies more readily with *Paranthropus.* However, this apparent anomaly is because this specimen has an obvious sixth cusp, and was landmarked with six cusps accordingly. The C6 or 'tuberculum sextum'[10,71-73] is diagnostic of *Paranthropus* spp., and all the specimens attributed to *P. robustus* and *P. boisei* were landmarked for this cusp, so it is not surprising that Taung 1 groups with the paranthropines.

Table 4 presents the summary of the results for the fossil hominin specimens. The full table, including the 80 extant species specimens, is provided in the supplementary material.

The results of the log se_m pairwise comparisons for the four extant hominoid species confirmed the results of the PCA plot: the widest ranges of values were shown by *G. g. gorilla* (low degree of shape, but high degree of size, and disparity between the smallest female and the largest male specimen, i.e. low average log se_m value coupled with high delta value) and by *H. sapiens* (most variability in shape, rather than size, of all four species, i.e. high average log se_m value with lower delta value). The average log se_m value for all conspecific comparisons was -1.6208, which is only a very slightly lower value than the average central tendency of average log se_m values for conspecific specimens of -1.61 as calculated by Thackeray[69]. The results are presented in Table 5.

Log se_m results similarly calculated for the lower first molars of fossil hominins included in the study also confirmed the PCA results and the DFA species-wide distributions of squared Mahalanobis distances. *Au. afarensis,* with its main 'holotype-like' group and two outliers/anomalies – one significantly tiny and narrow by comparison with the holotype, and the other more in the range of *Au. africanus* in dimension – had the highest average log se_m and delta values of all the species (indicating both shape and size disparity within the group as currently classified). Specifically, AL 288-1 ('Lucy') had a log se_m value of -1.278 in a pairwise comparison with LH 4 (the holotype of the species) and the equivalent value for LH 2 against LH 4 was -1.486. This result confirms the DFA results, in that the group mean log se_m showed extreme variability between specimens in the species sample, but that AL 288-1 was the specimen most likely not to be conspecific with LH 4. *H. erectus* and other early *Homo* species were the most cohesive of the groups, despite anomalous specimens in each group that both tended more towards *Au. africanus* in dimension. In particular, KNM-ER 806c was the main anomaly in the *H. erectus* group, with a log se_m value of -1.396, in comparison to KNM-ER 992 (the holotype of the species). The mean value for this group is -1.625±0.16 (n=30 pairwise comparisons), but if KNM-ER 806c had been excluded from the group, the mean value would have been -1.727.

The results for the species groups including and excluding the six anomalous specimens as identified by the PCA are presented in Table 6. Once the anomalous specimens are removed from the analysis, the mean log se_m value for conspecific comparisons is -1.607, with a standard deviation of 0.102 (n=176 pairwise comparisons), which is very much in line with Thackeray's mean log se_m value of -1.61 with a standard deviation of 0.230 for 70 species.[69]

Discussion

Sexual dimorphism evident in gorilla lower first molars

Lower first molars of male and female gorillas are undifferentiated in morphospace in a shape-only PCA (Figure 2), but are well separated when size is factored back into the analysis in a shape-and-size PCA (Figure 3). There was no overlap at all along the *x*-axis (PC1 accounting mainly for size) between male and female gorillas in the sample used for this analysis, with molars of all female gorillas being smaller than all molars belonging to male gorillas. The implication is that shape is not a determining factor in distinguishing between sexes within this sample from *G. g. gorilla*, but rather, the main difference is in the size of the molars. Bonobo lower first molars and common chimpanzee lower first molars similarly group together in morphospace in a shape-only analysis (Figure 2); but in a shape-and-size analysis (Figure 3), while separation is achieved between the species on the basis of size differences, there is no marked separation evident between male and female individuals, as there is with gorillas.

Variability between molars of Homo sapiens

The modern human molars included in the analysis showed the greatest within-species variability of the four species included in the study (Figure 2). When size was included in the analysis (Figure 3), these molars still failed to cluster closely, because of differences in relative mesiodistal:buccolingual ratios (relative width of the teeth) and in overall shape. Not only do modern human molars vary more in shape than the other modern African ape species according to the PC analyses, but molars belonging to male and female humans are not differentiated in the same way as gorilla molars are in morphospace within the species as a whole (distinct groups of male and female individuals).

Table 4: Results of discriminant function analysis for fossil hominin species

Specimen	Actual group	Predicted group	Probability	Squared Mahalanobis distance to centroid
AL 145-35 L	*Australopithecus afarensis*	*Au. afarensis*	1.000	18.235
AL 228-23 R	*Au. afarensis*	*Au. afarensis*	0.984	14.045
AL 266-1 L	*Au. afarensis*	*Au. afarensis*	0.939	11.886
AL 266-1 R	*Au. afarensis*	*Au. afarensis*	0.997	11.573
AL 288-1 R	*Au. afarensis*	*Homo sapiens*	0.547	11.957
AL 333-W6D L	*Au. afarensis*	*Au. afarensis*	0.989	6.529
LH 2 R	*Au. afarensis*	*Au. afarensis*	0.545	8.580
LH 4 L	*Au. afarensis*	*Au. afarensis*	0.958	21.809
LH 4 R	*Au. afarensis*	*Au. afarensis*	0.998	4.063
MLD 2 L	*Au. africanus*	*Au. africanus*	0.992	8.650
MLD 2 R	*Au. africanus*	*Au. africanus*	0.402	11.916
Sts 52b R	*Au. africanus*	*Au. afarensis*	0.498	6.568
Taung 1 L	*Au. africanus*	*Paranthropus robustus*	0.817	9.692
Taung 1 R	*Au. africanus*	*P. robustus*	0.958	7.498
OH 22 R	*H. erectus*	*H. erectus*	0.694	8.763
KNM-ER 806c L	*H. erectus*	*Au. africanus*	0.843	18.196
KNM-ER 820 L	*H. erectus*	*H. erectus*	0.960	5.324
KNM-ER 820 R	*H. erectus*	*H. erectus*	0.995	4.556
KNM-ER 992 L	*H. erectus*	*H. erectus*	0.986	2.097
KNM-ER 992 R	*H. erectus*	*H. erectus*	0.956	4.750
OH 7 L	*H. habilis*	*H. habilis*	0.970	3.534
OH 7 R	*H. habilis*	*H. habilis*	0.962	1.566
OH 16 R	*H. habilis*	*H. habilis*	0.992	8.395
KNM-ER 1802 L	*H. rudolfensis*	*H. rudolfensis*	0.995	1.601
KNM-ER 1802 R	*H. rudolfensis*	*H. rudolfensis*	0.964	1.601
KNM-ER 15930 L	*P. boisei*	*P. boisei*	0.729	14.276
Peninj 1 L	*P. boisei*	*P. boisei*	0.972	6.799
Peninj 1 R	*P. boisei*	*P. boisei*	0.991	3.584
SK 6 L	*P. robustus*	*P. robustus*	0.961	8.061
SK 6 R	*P. robustus*	*P. robustus*	1.000	5.432
SK 23 L	*P. robustus*	*P. robustus*	0.994	2.935
SK 23 R	*P. robustus*	*P. robustus*	0.881	7.439
SK 63 L	*P. robustus*	*P. robustus*	1.000	6.193
SK 63 R	*P. robustus*	*P. robustus*	1.000	3.367
SKW 5 R	*P. robustus*	*P. robustus*	1.000	12.829
TM 1517 R	*P. robustus*	*P. robustus*	0.983	5.639

Table 5: Log se$_m$ results for four extant hominoid species

	Gorilla gorilla	*Pan troglodytes*	*Homo sapiens*	*Pan paniscus*	Average of four species
Average log se$_m$	-1.6428	-1.6577	-1.5389	-1.6439	-1.6208
Standard deviation	0.13	0.09	0.13	0.10	0.12
Minimum value	-2.1135	-1.9291	-1.9035	-1.9142	
Maximum value	-1.2357	-1.4207	-1.2854	-1.3859	
Delta value	0.0707	0.0483	0.0469	0.0563	0.0555
n	190	190	190	190	760

Table 6: Log se$_m$ results for extinct hominin species groups

	Australopithecus afarensis	*Australopithecus africanus*	*Homo erectus*	*Homo habilis/ H. rudolfensis*	*Paranthropus boisei/ P. robustus*	Average conspecifics
	(all specimens)					
Average log se$_m$ (all specimens)	-1.437	-1.521	-1.625	-1.701	-1.566	-1.570
Standard deviation	0.125	1.134	0.160	0.096	0.110	0.125
Delta value	0.071	0.050	0.040	0.036	0.058	0.034
n	72	20	30	20	110	252
	Au. afarensis	*Au. africanus*	*H. erectus*	*H. habilis/ rudolfensis*	*P. boisei/ robustus*	Average conspecifics
	(corrected for anomalies)					
Average log se$_m$ (corrected for anomalies)	-1.584	-1.500	-1.727	-1.756	-1.602	-1.607

(*n* = 176; s.d. = 0.102)

On further inspection, shape and size variability of both male and female humans is noted at the level of individual population groups: the male and female San individuals in the sample both have molars with almost equal mesiodistal and buccolingual diameters (they appear almost square in the occlusal view), while the molars of the megadont male Tswana individual in this study were large and relatively narrow and overlapped with the smallest female gorilla teeth in the shape-and-size PCA. Molars of individuals with European heritage were among the smallest in the sample.

This finding would seem to fit with previous studies which have suggested that because modern humans have migrated globally and different populations have followed variable histories of subsistence lifestyles and diet, tooth size has evolved biogeographically.[17,26,28] Tooth size reduction, in particular, has occurred in specific regions (particularly Europe, North Africa, the Levant and the Anatolian area) as a result of changes from hunter–gatherer lifestyles to semi-sedentary herding, and particularly to large-scale farming lifestyles. This change in lifestyle involving sedentary or urban living has resulted in increased consumption of soft cereals and more efficient food-processing and cooking technologies, which set these groups on a different dietary trajectory from hunter–gatherer societies since the Neolithic Revolution.[24-27] Stark odontometric differences are also reflected in other skeletal elements such as femora (with European groups having long, gracile femora, with hunter–gatherers and certain Bantu groups having more robust, shorter femora.[22,23]) In the small sample chosen for this study, there are representatives from populations whose lifestyles, until recently, have revolved around hunter–gathering (the San), small-scale subsistence herder-–agriculturalists (Tswana and Sotho), and sedentary/ post-Neolithic large-scale farmers (individuals with European heritage).

Fossil 'outliers'

In all, from the PCA plot, there appears to be six lower first molar specimens that do not seem to cluster in morphospace with their own species groups: (1) AL 288-1 and (2) LH 2 (from *Au. afarensis*); (3) Sts 52b, which has been allocated to *Au. africanus* but whose classification has been questioned within this group by other researchers[74,75]; (4) OH 16, classified as *H. habilis* but which is notably larger and wider than the holotype and has been likened to molars of *Au. africanus* in size[50]; (5) KNM-ER 15930, which is classified into *P. boisei* but which is extremely tiny by comparison with the typically megadont examples of this species; and (6) KNM-ER 806c, classified as *H. erectus*,[9] but visually almost identical to (albeit slightly smaller than) MLD 2 – this particular specimen is currently classified as *Au. africanus* but has been identified as a 'larger-toothed' specimen,[42] and one of a group of specimens being considered for reclassification into a 'second species', *Au. prometheus*.[42]

Three of these specimens – Sts 52b, OH 16 and KNM-ER 15930 – are in species groups with low sample numbers in this study, and so it is difficult to draw firm conclusions. *H. habilis* as a species has been challenged since its first introduction into the literature; KNM-ER 15930 may represent extreme sexual dimorphism in *P. boisei,* as postulated by Leakey and Walker[52]. The status of the specimens representing *Au. africanus* in general warrants some further comment, as Taung 1 appeared to be anomalous in the context of the DFA, alongside Sts 52b. Unfortunately, the species in this particular study was represented by only five molars from three individuals, each of which has been the object of some discussion as to its inclusion within the species. Firstly, Taung 1, which is the holotype, has a sixth cusp more typical of *Paranthropus*, and its inclusion within the group of 'robust australopithecines' has

been previously discussed by researchers.[76] MLD 2 has characteristics associated with *Au. africanus*, including a protostylid, but its attribution to *Au. africanus* has recently been reconsidered by Clarke[42]. Sts 52b is a slightly damaged molar with a great deal of wear, but certain researchers have confirmed the anomalous status noted in this study of the specimen.[42,74,75] The high mean Mahalanobis distance for this species might be explained by the heterogeneity of this sample.

The other three anomalous specimens from the PCA are interesting for further discussion regarding potential misclassification and/or unusually high variability within one single species, as currently defined.

The striking morphological similarities between KNM-ER 806c (currently attributed to *H. erectus* but grouping closely with *Au. africanus* both on the PCA and the DFA) and MLD 2 (*Au. africanus* – or possibly *Au. prometheus*[42]) can be seen in Figure 6. In view of this similarity, and the dissimilarity of KNM-ER 806c with its current species holotype, KNM-ER 992, it would be interesting in future to look at comparisons of other molars of KNM-ER 806 to see with which species they group best.

Two of the atypical specimens identified from the shape-and-size PCA are currently classified into *Au. afarensis*. The first of these, LH 2, is from a juvenile mandible from Laetoli. In the PCA plots, LH 2 groups consistently towards *Au. africanus* in morphospace, being more robust in size and less square in relative dimensions in the occlusal view than the typical *Au. afarensis* molars, as shown in Figure 7. Morphologically, LH 2 shares lower first molar occlusal crown characteristics similar to those of Taung 1, the holotype of *Au. africanus*.

The second atypical specimen currently classified as belonging to the *Au. afarensis* species group is AL 288-1, or 'Lucy'. This molar clusters well away from the holotype of *Au. afarensis* (LH 4) in all of the PCAs, and this apparent misclassification is supported by the DFA and the log se_m analysis.

Donald Johanson, the team leader for the discovery of this specimen and whose PhD thesis was on primate molars, remarked that 'Lucy' was more chimpanzee-like[37,77] than the other specimens ultimately attributed to *Au. afarensis*, with an 'odd lower jaw', which he initially assigned to a different species than that assigned to the typically larger and squarer Hadar molars and those of the mandible from Laetoli, LH 4 (eventually designated as the holotype for *Au. afarensis*). Without wishing to suggest that the mandible of AL 288-1 is indeed from a chimpanzee, what is clear from the PCA plots of the occlusal crown morphometrics of the relatively tiny and narrow lower first molar of this specimen, is that the positions where this specimen consistently plots, away from the holotype group, would seem to add a point of initial agreement with Johanson that 'Lucy is different'[77]. Figure 8 shows occlusal views of the lower first molar of AL 288-1 between LH 4 (*Au. afarensis* holotype) on the one side, and of a chimpanzee and a modern human on the other side.

With these two outliers grouping in different directions in morphospace away from the holotype cluster, it might be interesting to revisit the question of how much variability is normal within any one species. Leonard and Hegmon[78] have suggested that, based on P3 morphology, vast differences between certain specimens can be explained if female individuals of the species were subject to different selective pressures than male individuals. This conclusion is rejected by Ferguson[32]. Perhaps, as Schmid argues[79], Lucy does not belong to the same species as the presumed 'males' of the species; or there may be more than one morphotype in this hypodigm.
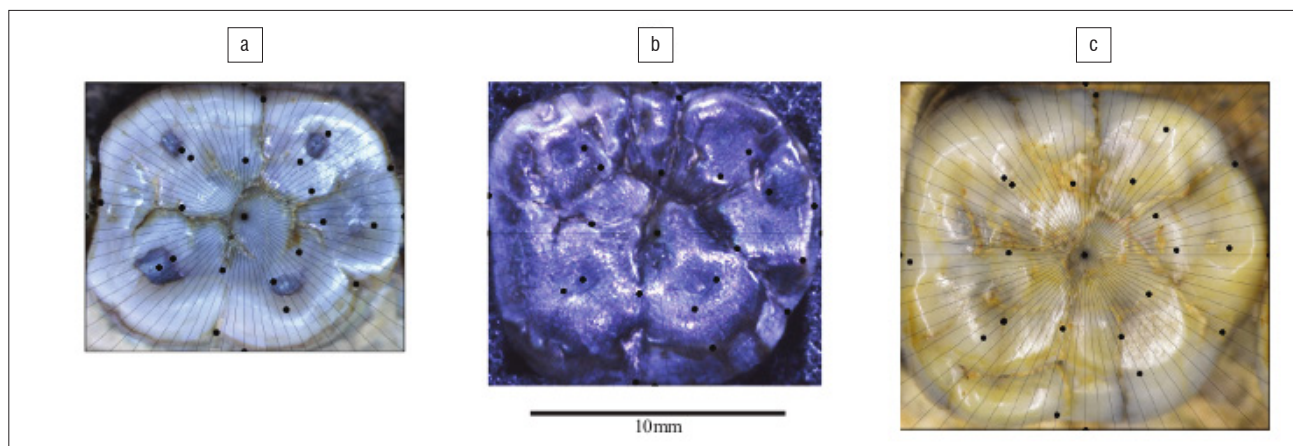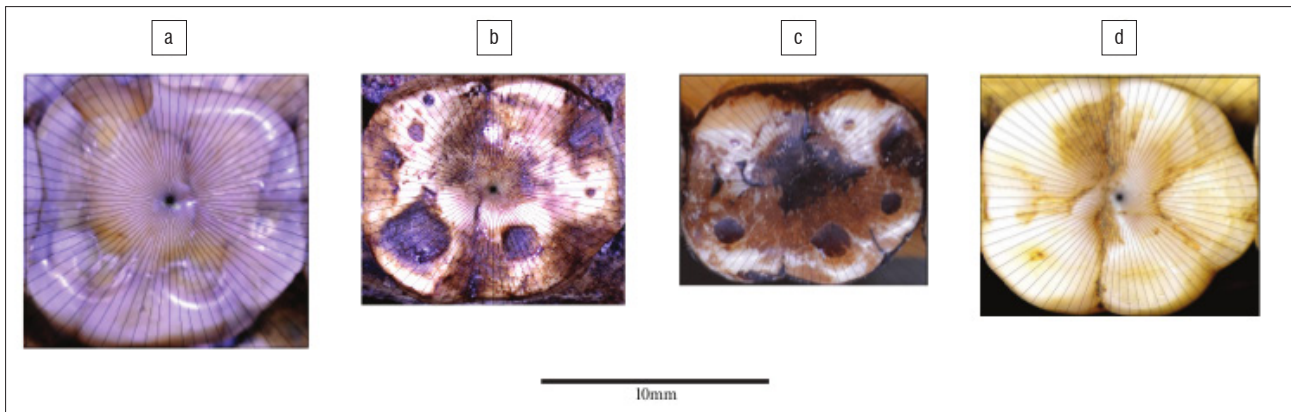

Figure 6: Comparison of lower first molar of KNM-ER 806c with other specimens. (a) KNM-ER 992, the *Homo erectus (ergaster)* holotype; (b) KNM-ER 806c (*Homo erectus*) and (c) MLD 2 (*Australopithecus africanus*). The occlusal views of KNM-ER 806c and MLD 2 show close affinities in morphology and in relative dimension.
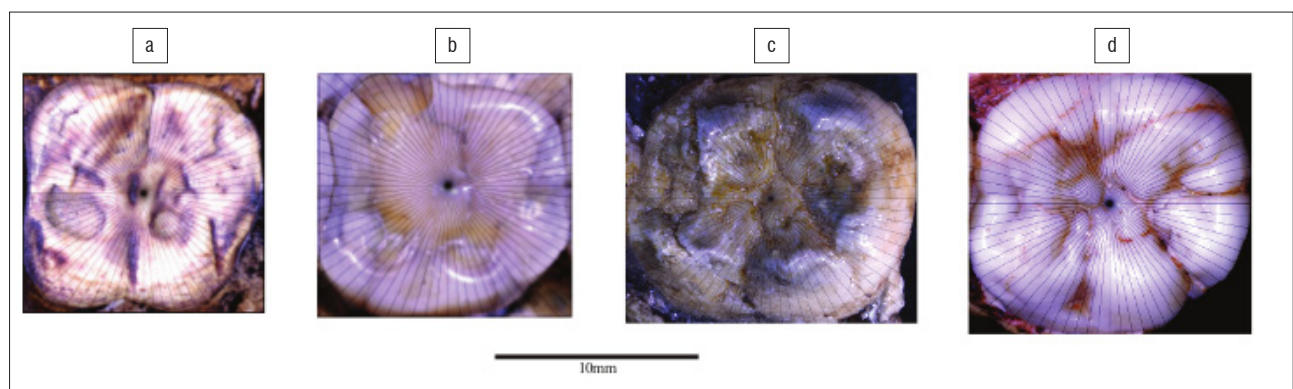
Figure 7: Comparison of lower first molar of LH 2 with other specimens. (a) AL 266-1 (*Australopithecus afarensis*), (b) *Au. afarensis* holotype LH 4, (c) LH 2 (*Au. afarensis*) and (d) Taung 1, the holotype of *Au. africanus*. In size and relative dimension, LH 2 has close affinities with *Au. africanus*, with a small metaconid and large entoconid, whereas the more 'typical' specimens of *Au. afarensis* are more 'square' in dimension (buccolingual diameters are almost equal to mesiodistal diameters), and have a very lingually oriented, large metaconid, coupled with a small entoconid.

Figure 8: Comparison of lower first molar of AL 288-1 with other specimens. (a) Holotype of *Australopithecus afarensis*, LH 4, (b) AL 288-1 (*Au. afarensis*), (c) lower first molar of a modern chimpanzee and (d) lower first molar of a modern human. In size, relative dimension and in wear pattern, the lower first molar of AL 288-1 appears to resemble the chimpanzee lower first molar more closely than it does the holotype of the species to which it is allocated.

Conclusions

Gorilla gorilla gorilla is well established as a highly sexually dimorphic species, and in the analyses the lower first molars demonstrate well-defined size differences between male and female individuals. However, as with the other two African great ape species in the study, the degree of shape variability is reasonably limited, particularly with respect to the mesiodistal:buccolingual proportions of the teeth in general. When extinct hominin species' lower first molars are landmarked and plotted in a similar analysis, species such as *H. erectus* and other early *Homo* specimens seem to follow the spatial variability patterning of extant ape species such as *P. paniscus* and *P. troglodytes. Paranthropus* species' lower first molars group together in morphospace within a limited range in respect of shape, but show some size disparity that is reminiscent of the way *G. g. gorilla* specimens clustered laterally on the shape-versus-size PCA plot. At first glance, *Au. afarensis* also shows signs of gorilla-like shape similarity, clustering around the very square-shaped holotype, the main difference being size variability (therefore possibly showing some sexual dimorphism observable by size differences rather than excessive shape variance between specimens), but on closer analysis, two very significant anomalies are plotted well away from the main cluster. The only other extant species group that displays such stark within-species shape and size differences in the analysis as a whole is the modern *H. sapiens* group, but unlike the African ape species and the fossil hominin species, modern *H. sapiens* has migrated globally, with individual groups exploiting extremely diverse environments and practising subsistence lifestyles that diverged from each other at least 12 000 years ago. In areas where farming groups have been exposed to soft cereals and have utilised more varied food-processing and cooking technologies than hunter–gatherers since the Neolithic Revolution, facial and tooth-size reduction have been reported (for example in Europe, the Middle East, North Africa and Anatolia). As this kind of dietary and subsistence lifestyle divergence within a single species cannot be applied by proxy as the cause of the range of variability seen between the molars of *Au. afarensis,* it might be argued that some measure of caution should be exercised before using modern *H. sapiens* as an analogue species for comparisons of ranges of variability in molar size and shape in fossil hominin species. If a more cautious approach is taken, a species with an arguably similar-sized range and dietary options available to it should ideally be chosen upon which to assess a likely range of variability of molar shape and the effects thereon of sexual dimorphism within a species. Based on the manner in which the molars of gorillas (the most sexually dimorphic species in the study) plot in morphospace, it could therefore be argued that anomalies or outliers from main species groupings that do not follow a similar clustering pattern (size variability, with limited shape variability) might indicate that some fossil hominin species as currently defined either consist of two (or three) distinct morphotypes within the same species, or that specimens currently attributed to single species belong, in fact, to several different species, or simply that certain specimens may have been wrongly classified. Future studies should include sample sizes that are enlarged sufficiently to encompass the full range of variability of extant species included in the study before confirming such conclusions. The sample representing the fossil species should also be expanded so that each species is adequately represented.

References

1. Uchida A. Variation in tooth morphology of *Gorilla gorilla*. J Hum Evol. 1998;34:55–70. http://dx.doi.org/10.1006/jhev.1997.0186

2. Uchida A. Variation in tooth morphology of *Pongo pygmaeus*. J Hum Evol. 1998;34:71–79. http://dx.doi.org/10.1006/jhev.1997.0187

3. Pilbrow VC. Dental variation in African apes with implications for understanding patterns of variation in species of fossil apes. New York: New York University; 2003.

4. Pilbrow VC. Patterns of molar variation in great apes and their implications for hominin taxonomy. In: Bailey SE, Hublin J-J, editors. Dental perspectives on human evolution. New York: Springer; 2007. p. 9–32. http://dx.doi.org/10.1007/978-1-4020-5845-5_2

5. Pilbrow VC. Population systematics of chimpanzees using molar morphometrics. J Hum Evol. 2006;51:646–662. http://dx.doi.org/10.1016/j.jhevol.2006.07.008

6. Pilbrow VC. Dental and phylogeographic patterns of variation in gorillas. J Hum Evol. 2010;59:16–34. http://dx.doi.org/10.1016/j.jhevol.2010.01.009

7. Singleton M, Rosenberger AL, Robinson C, O'Neill R. Allometric and metameric shape variation in *Pan* mandibular molars: A digital morphometric analysis. Anat Rec. 2011;294(2):322–334. http://dx.doi.org/10.1002/ar.21315

8. Bailey SE, Wood B. Trends in postcanine occlusal morphology within the hominin clade: The case of *Paranthropus*. In: Bailey SE, Hublin J-J, editors. Dental perspectives on human evolution: Dordrecht: Springer; 2007. p. 33–52. http://dx.doi.org/10.1007/978-1-4020-5845-5_3

9. Wood B. Koobi Fora Research Project. Volume 4: Hominid cranial remains. Oxford: Clarendon Press; 1991.

10. Wood B, Abbott SA. Analysis of the dental morphology of Plio-Pleistocene hominids. I: Mandibular molars: Crown area measurements and morphological traits. J Anat. 1983;136(1):197–219.

11. Wood B, Abbott SA, Graham SH. Analysis of the dental morphology of Plio-Pleistocene hominids. II: Mandibular molars – Study of cusp areas, fissure pattern and cross sectional shape of the crown. J Anat. 1983;137:287–314.

12. Wood B, Engleman CA. Analysis of the dental morphology of Plio-Pleistocene hominids. V: Maxillary postcanine tooth morphology. J Anat. 1988;161:1–35.

13. Gómez-Robles A, Bermúdez de Castro JM, Martinón-Torres M, Prado-Simón I, Arsuaga JL. A geometric morphometric analysis of hominin upper second and third molars, with particular emphasis on European Pleistocene populations. J Hum Evol. 2012;63(3):512–526. http://dx.doi.org/10.1016/j.jhevol.2012.06.002

14. Gómez-Robles A, Martinón-Torres M, Bermúdez de Castro JM, Margvelashvili A, Bastir M, Arsuaga JL, et al. A geometric morphometric analysis of hominin upper first molar shape. J Hum Evol. 2007;53(3):272–285. http://dx.doi.org/10.1016/j.jhevol.2007.02.002

15. Thackeray JF, Odes E. Morphometric analysis of early Pleistocene African hominin crania in the context of a statistical (probabilistic) definition of a species. Antiquity. 2013;87:1–2.

16. Ackermann RR. Using extant morphological variation to understand fossil relationships: A cautionary tale. S Afr J Sci. 2003;99:255–258.

17. Hanihara T, Ishida H. Metric dental variation of major human populations. Am J Phys Anthropol. 2005;128:287–298. http://dx.doi.org/10.1002/ajpa.20080

18. Kieser JA. Human adult odontometrics: The study of variation in adult tooth size. Cambridge: Cambridge University Press; 1990. http://dx.doi.org/10.1017/CBO9780511983610

19. Dempsey PJ, Townsend GC. Genetic and environmental contributions to variation in human tooth size. Heredity. 2001;86:685–693. http://dx.doi.org/10.1046/j.1365-2540.2001.00878.x

20. Kay RF, Hiiemae KM. Jaw movement and tooth use in recent and fossil primates. Am J Phys Anthropol. 1974;138:136–148. http://dx.doi.org/10.1002/ajpa.1330400210

21. Ungar PS. Dental topography and diets of *Australopithecus* and early *Homo*. J Hum Evol. 2004;46:605–622. http://dx.doi.org/10.1016/j.jhevol.2004.03.004

22. Püschel TA, Benítez HA. Femoral function adaptation: a comparison between hunter gatherers and farmers using geometric morphometrics. Int J Morphol. 2014;32(2):627–633. http://dx.doi.org/10.4067/S0717-95022014000200041

23. Siddiqi N. Comparison of osteometric femoral bone dimensions among the South Africans of different ethnic groups and South African whites. Egypt J For Sci. 2013;3:8–14. http://dx.doi.org/10.1016/j.ejfs.2012.11.001

24. Dahlberg AA. The dentition of the first agriculturalists (Jarmo, Iraq). Am J Phys Anthropol. 1960;18:243–256.

25. Emes Y, Aybar B, Yalcin S. On the evolution of human jaws and teeth: A review. Bull Int Assoc Paleodont. 2011;5(1):37–47.

26. Pinhasi R, Eshed V, Shaw P. Evolutionary changes in the masticatory complex following the transition to farming in the Southern Levant. Am J Phys Anthropol. 2008;135:136–148. http://dx.doi.org/10.1002/ajpa.20715

27. Pinhasi R, Eshed V, Von Cramon-Taubadel N. Incongruity between affinity patterns on mandibular lower dental dimensions following the transition to agriculture in the near East, Anatolia and Europe. PLoS One. 2015;10(2):e0117301. http://dx.doi.org/10.1371/journal.pone.0117301

28. Campbell TD. Dentition and palate of the Australian Aboriginal. Adelaide: Hassell Press; 1925.

29. Schwartz GT, Dean C. Sexual dimorphism in modern human permanent teeth. Am J Phys Anthropol. 2005;128:312–317. http://dx.doi.org/10.1002/ajpa.20211

30. Reno PL, Meindl RS, McCollum MA, Lovejoy CO. Sexual dimorphism in *Australopithecus afarensis* was similar to that of modern humans. Proc Natl Acad Sci USA. 2003;100(16):9404–9409. http://dx.doi.org/10.1073/pnas.1133180100

31. Reno PL, McCollum MA, Meindl RS, Lovejoy CO. An enlarged postcranial sample confirms *Australopithecus afarensis* dimorphism was similar to modern humans. Philos Trans R Soc Lond Ser B Biol Sci. 2010;365:3355–3363. http://dx.doi.org/10.1098/rstb.2010.0086

32. Ferguson WW. Critique of '*Australopithecus afarensis*' as a single species based on dental metrics and morphology. Primates. 1989;30(4):561–569. http://dx.doi.org/10.1007/BF02380881

33. Martinez-Gomis J, Lujan-Climent M, Palau S, Bizar J, Salsench J, Peraire M. Relationship between chewing side preference and handedness and lateral asymmetry of peripheral factors. Arch Oral Biol. 2009;54(2):101–107. http://dx.doi.org/10.1016/j.archoralbio.2008.09.006

34. Veloso-Durán A, Vazquez-Salceda MC, López-Jiménez J, Veloso-Durán M, Puigdollers A. A correlation between permanent tooth eruption pattern and the predominance of the motor function laterality. Med Oral Patol Oral Cir Bucal. 2014;19(5):473–477. http://dx.doi.org/10.4317/medoral.19567

35. Heikkinen T, Alvesalo L, Osborne RH, Tienari J. Tooth eruption symmetry in functional lateralities. Arch Oral Biol. 2001;46(7):609–617. http://dx.doi.org/10.1016/S0003-9969(01)00019-X

36. White TD, Johanson DC. Pliocene hominid mandibles from the Hadar formation, Ethiopia: 1974-1977 collections. Am J Phys Anthropol. 1982;57:501–544. http://dx.doi.org/10.1002/ajpa.1330570405

37. Johanson DC, Taieb M. Plio-Pleistocene hominid discoveries in Hadar, Ethiopia. Nature. 1976;260:293–297. http://dx.doi.org/10.1038/260293a0

38. Kimbel WH, Johanson, DC, Coppens Y. Pliocene hominid cranial remains from the Hadar formation, Ethiopia. Am J Phys Anthropol. 1982;57:453–499. http://dx.doi.org/10.1002/ajpa.1330570404

39. Leakey MD, Hay R, Curtis GH, Drake RE, Jackes MK, White TD. Fossil hominids from the Laetoli beds. Nature. 1976;262:460–466. http://dx.doi.org/10.1038/262460a0

40. White TD. New fossil hominids from Laetolil, Tanzania. Am J Phys Anthropol. 1977;46(2):197–229. http://dx.doi.org/10.1002/ajpa.1330460203

41. Dart RA. The Makapansgat proto-human *Australopithecus prometheus*. Am J Phys Anthropol. 1948;6(3):259–284. http://dx.doi.org/10.1002/ajpa.1330060304

42. Clarke RJ. Latest information on Sterkfontein's *Australopithecus* skeleton and a new look at *Australopithecus*. S Afr J Sci. 2008;104:443–449.

43. Broom R, Robinson JT, Schepers GWH. Sterkfontein ape-man *Plesianthropus*. Transvaal Mus Mem. 1950;4:502–505.

44. Dart RA. *Australopithecus africanus:* The man-ape from South Africa. Nature. 1925;115:195–199. http://dx.doi.org/10.1038/115195a0

45. Rightmire GP. Middle Pleistocene hominids from Olduvai Gorge, northern Tanzania. Am J Phys Anthropol. 1980;53(2):225–241. http://dx.doi.org/10.1002/ajpa.1330530207

46. Leakey REF. Further evidence of Lower Pleistocene hominids from East Rudolf, North Kenya, 1971. Nature. 1972;237:264–269. http://dx.doi.org/10.1038/237264a0

47. Leakey REF, Wood BA. New evidence of the genus *Homo* from East Rudolf, Kenya II. Am J Phys Anthropol. 1973;39(3):355–368. http://dx.doi.org/10.1002/ajpa.1330390304

48. Leakey LSB. Recent discoveries at Olduvai Gorge. Nature. 1960;188:1050–1052. http://dx.doi.org/10.1038/1881050a0

49. Leakey LSB. The juvenile mandible from Olduvai. Nature. 1961;191:417–418. http://dx.doi.org/10.1038/191417a0

50. Leakey LS, Tobias PV, Napier JR. A new species of genus *Homo* from Olduvai Gorge. Nature. 1964;202(4):424–427. http://dx.doi.org/10.1038/202007a0

51. Day MH, Leakey REF, Walker AC, Wood BA. New hominids from East Turkana, Kenya. Am J Phys Anthropol. 1976;45:369–435. http://dx.doi.org/10.1002/ajpa.1330450304

52. Leakey REF, Walker AC. New *Australopithecus boisei* specimens from East and West Lake Turkana, Kenya. Am J Phys Anthropol. 1988;76(1):1–24. http://dx.doi.org/10.1002/ajpa.1330760102

53. Tobias PV. The early *Australopithecus* and *Homo* from Tanzania. Anthropologie Paris. 1965;3:43–48.

54. Broom R, Robinson JT. Swartkrans ape man: *Paranthropus crassidens*. Transvaal Mus Mem. 1952;6:1–123.

55. Robinson JT. The dentition of the Australopithecinae. Transvaal Mus Mem. 1956;9:i–i79.

56. Grine FE, Daegling DJ. New mandible of *Paranthropus robustus* from Member 1, Swartkrans Formation, South Africa. J Hum Evol. 1993;24:319–333. http://dx.doi.org/10.1006/jhev.1993.1024

57. Broom R. The Pleistocene anthropoid apes of South Africa. Nature. 1938;142:377–379. http://dx.doi.org/10.1038/142377a0

58. Grine FE, Ungar PS, Teaford MF, El-Zaatari S. Molar microwear in *Praeanthropus afarensis*: Evidence for dietary stasis through time and under diverse paleoecological conditions. J Hum Evol. 2006;51(3):297–319. http://dx.doi.org/10.1016/j.jhevol.2006.04.004

59. Wood B, editor. Wiley-Blackwell encyclopedia of human evolution 2. New York: John Wiley & Sons; 2011.

60. Pickering R, Kramers JD. Re-appraisal of the stratigraphy and determination of new U-Pb dates for the Sterkfontein hominin site, South Africa. J Hum Evol. 2010;59(1):70–86. http://dx.doi.org/10.1016/j.jhevol.2010.03.014

61. Groves CP, Mazak V. An approach to the taxonomy of the hominidae: Gracile Villafranchian hominidae of Africa. Casopis pro Mineralogia Geologii. 1975;20:225–247.

62. Goose DH. Dental measurement: An assessment of its value in anthropological studies. In: Brothwell DR, editor. Dental anthropology. New York: Pergamom Press; 1963. http://dx.doi.org/10.1016/B978-0-08-009823-4.50012-4

63. Dryden IL, Mardia KV. Statistical shape analysis. Michigan: John Wiley & Sons; 1998.

64. Dykes SJ, Dykes R. 'Professor regressor': A computer programme for rapid processing of large sets of pairwise regression analyses in palaeolontological contexts. Palaeontol Afr. 2015;49:53–59.

65. Thackeray JF, Dykes SJ. Morphometric analyses of hominoid crania, probabilities of conspecificity and an approximation of a biological species constant. HOMO – J Comp Hum Biol. 2016;67(1):1–10. http://dx.doi.org/10.1016/j.jchb.2015.09.003

66. Mitteroecker P, Gunz P, Windhager S, Schaefer K. A brief review of shape, form and allometry in geometric morphometrics, with applications to human facial morphology. Hystrix: Ital J Mammal. 2013;24(1):59–66.

67. Mitteroecker P, Gunz P, Bernhard M, Schaefer K, Bookstein FL. Comparison of cranial ontogenetic trajectories among great apes and humans. J Hum Evol. 2004;46:679–698. http://dx.doi.org/10.1016/j.jhevol.2004.03.006

68. Thackeray JF, Bellamy CL, Bellars D, Bronner G, Bronner L, Chimimba C, et al. Probabilities of conspecificity: Application of a morphometric technique to modern taxa and fossil specimens attributed to *Australopithecus* and *Homo*. S Afr J Sci. 1997;93(4):195–196.

69. Thackeray JF. Approximation of a biological species constant? S Afr J Sci. 2007;103:489.

70. Thackeray JF. Probabilities of conspecificity: An application of a morphometric approach to extant and extinct hominoids. In: Strkalj G, Pather N, Kramer B, editors. Voyages in science essays by South African anatomists in honour of Phillip V Tobias' 80th birthday. Pretoria: Content Solutions; 2005.

71. Hillson S. Dental anthropology. Cambridge: Cambridge University Press; 1996. http://dx.doi.org/10.1017/CBO9781139170697

72. Aiello L, Dean C. An introduction to human evolutionary anatomy. London: Academic Press; 2002. http://dx.doi.org/10.1016/B978-0-08-057100-3.50017-1

73. Braga J, Thackeray JF. Early *Homo* at Kromdraai B: Probabilistic and morphological analysis of the lower dentition. Comptes Rendus Palevol. 2003;2:269–279. http://dx.doi.org/10.1016/S1631-0683(03)00044-7

74. Kimbel WH, White TD. Variation, sexual dimorphism and the taxonomy of *Australopithecus*. In: Grine FE, editor. Evolutionary history of the 'robust' australopithecines. New York: Aldine de Gruyter; 1988. p. 259–268.

75. Fornai C, Bookstein FL, Weber GH. Second maxillary molars confirm a dimorphism of *Australopithecus* at Sterkfontein Member 4. Paper presented at: European Society for the Study of Human Evolution Third Annual Meeting; 2013 September 19–21; Vienna, Austria.

76. Tobias PV. South African australopithecines in time and hominid phylogeny, with special 21reference to dating and affinities of the Taung skull. In: Jolly C, editor. Early hominids of Africa. London: Duckworth; 1978. p. 45–84.

77. Johanson DC, Edey MA. Lucy – the beginnings of mankind. London: Penguin Books; 1990.

78. Leonard WR, Hegmon M. Evolution of P3 morphology in *Australopithecus afarensis*. Am J Phys Anthropol. 1987;73(1):41–63. http://dx.doi.org/10.1002/ajpa.1330730105

79. Schmid P, editor How different is Lucy? In: Hominidae: Proceedings of the 2nd International Congress of Human Paleontology; 1987 September 28–October 03; Turin, Italy. Milan: Jaca Book; 1989.

AUTHORS:
Elisabeth K. Olsen[1,2]
Christopher K. de Cerf[1]
Godwin A. Dziwornu[1]
Eleonora Puccinelli[3,4]
Isabelle J. Ansorge[4,5]
Toufiek Samaai[5,6]
Laura M.K. Dingle[7]
Adrienne L. Edkins[7] (iD)
Suthananda N. Sunassee[1,5,8] (iD)

AFFILIATIONS:
[1]Department of Chemistry, University of Cape Town, Cape Town, South Africa
[2]MRC/NHLS Molecular Mycobacteriology Research Unit, University of Cape Town, Cape Town, South Africa
[3]Department of Zoology and Entomology, Rhodes University, Grahamstown, South Africa
[4]Department of Oceanography, University of Cape Town, Cape Town, South Africa
[5]Marine Research (Ma-Re) Institute, University of Cape Town, Cape Town, South Africa
[6]Oceans and Coasts Research, Department of Environmental Affairs, Cape Town, South Africa
[7]Biomedical Biotechnology Research Unit, Department of Biochemistry and Microbiology, Rhodes University, Grahamstown, South Africa
[8]MRC Drug Discovery and Development Research Unit, University of Cape Town, Cape Town, South Africa

CORRESPONDENCE TO:
Suthananda Sunassee

EMAIL:
sunny.sunassee@uct.ac.za

Cytotoxic activity of marine sponge extracts from the sub-Antarctic Islands and the Southern Ocean

Over the past 50 years, marine invertebrates, especially sponges, have proven to be a valuable source of new and/or bioactive natural products that have the potential to be further developed as lead compounds for pharmaceutical applications. Although marine benthic invertebrate communities occurring off the coast of South Africa have been explored for their biomedicinal potential, the natural product investigation of marine sponges from the sub-Antarctic Islands in the Southern Ocean for the presence of bioactive secondary metabolites has been relatively unexplored thus far. We report here the results for the biological screening of both aqueous and organic extracts prepared from nine specimens of eight species of marine sponges, collected from around Marion Island and the Prince Edward Islands in the Southern Ocean, for their cytotoxic activity against three cancer cell lines. The results obtained through this multidisciplinary collaborative research effort by exclusively South African institutions has provided an exciting opportunity to discover cytotoxic compounds from sub-Antarctic sponges, whilst contributing to our understanding of the biodiversity and geographic distributions of these cold-water invertebrates. Therefore, we acknowledge here the various contributions of the diverse scientific disciplines that played a pivotal role in providing the necessary platform for the future natural products chemistry investigation of these marine sponges from the sub-Antarctic Islands and the Southern Ocean.

Significance:

- This study will contribute to understanding the biodiversity and geographic distributions of sponges in the Southern Ocean.

- This multidisciplinary project has enabled the investigation of marine sponges for the presence of cytotoxic compounds.

- Further investigation will lead to the isolation and identification of cytotoxic compounds present in the active sponge extracts.

Introduction

Marine biodiscovery is the search for marine natural products with potential economic and societal benefits for use as agrochemicals, cosmetics, anti-fouling agents, nutraceuticals and pharmaceuticals.[1,2] Generally, marine invertebrates utilise their secondary metabolites or natural products as a form of chemical defence against predators and as a competitive advantage in their perpetual battle for limited resources, such as nutrients and space.[2,3] Over the past 50 years, marine invertebrates and their associated microorganisms have proved to be an important source of bioactive natural products for the development of new pharmaceuticals,[4,5] especially anti-cancer agents, e.g. Adcetris®, Halaven® and Yondelis®, that were discovered by initially screening for cytotoxic compounds against cancer cell lines.[6,7] Marine sponges (Phylum: *Porifera*) have arguably been one of the dominant sources of marine natural products that are utilised commercially for their beneficial pharmacological properties in biomedical research.[6] There currently are seven FDA-approved marine natural products derived drugs, of which four are anti-cancer chemotherapeutic agents, and a further 13 marine-derived agents that are either in human clinical trials or in advanced pre-clinical status, including a synthetic derivative of the South African marine sponge natural product Hemiasterlin.[6,7] The latest review by Newman and Cragg[8] highlights the significant contribution of natural products research to the field of drug discovery, with an estimated 77% of all anti-cancer drugs approved between 1940 and 2014 being small molecule natural products and natural product-derived compounds.

Southern Africa's marine biota is home to a wealth of biodiversity along the approximately 3000-km coastline of South Africa and around the archipelagos of the Southern Ocean, including the Prince Edward Islands (PEIs) which consist of Marion Island and Prince Edward Island. The biomedicinal potential of southern African marine invertebrates was initially revealed through the marine biodiscovery efforts of Robert Pettit (Arizona State University) and Yoel Kashman (Tel Aviv University).[9,10] The initial biological screening for cytotoxic compounds and subsequent natural products chemistry work by Pettit and his co-workers on their South African marine collections resulted in the isolation of two very important classes of bioactive compounds known as the cephalostatins and spongiostatins.[9,10] The cephalostatins and spongiostatins were isolated through cytotoxic activity-directed purification of crude extracts and are well known to be some of the most cytotoxic secondary metabolites (about 1 nM activity) ever to be screened by the United States National Cancer Institute (NCI) against their 60 cancer cell line panel (NCI-60).[9,10] Cephalostatin 1 and its closely related naturally occurring analogues

were isolated from aqueous and organic extracts of the South African marine tube worm *Cephalodiscus gilchristi*, while Spongiostatins 4 and 5 were isolated from the organic extract of the bright orange 'wall-sponge' *Trachycladus spinispirulifer* collected off the southern coasts of South Africa.[9,10] The collaborative biodiscovery programme in Sodwana Bay between Kashman, the Oceanography Research Institute and the pharmaceutical company PharmaMar involved the primary biological screening of crude extracts for cytotoxic compounds against the P-388, A549, HT-29 and MEL-28 cancer cell lines. This screening resulted in the discovery of 33 new bioactive natural product compounds, mostly isolated from marine sponges, of which 9 compounds were patented for their anti-cancer properties.[9,10] These compounds include two cytotoxic peptides – Hemiasterlin and Geodiamolide TA – isolated from the marine sponge *Hemiasterella minor* (Kirkpatrick); a synthetic analogue of Hemiasterlin was subsequently developed and has successfully entered phase one of human clinical trials as an anti-cancer drug.[6,9]

Between 1992 and 2012, the marine natural products research groups at Rhodes University, in collaboration with various other scientific research groups and institutes, led the search for, and provided important contributions to, the discovery of novel bioactive natural product compounds from southern African marine organisms.[9-12] Davies-Coleman and colleagues[13,14] have previously reported the screening of several extracts of southern African marine invertebrates for anti-oesophageal cancer activity, including five deep-water sponges collected from the PEIs that were tested against the WHCO1 cell line. The sponges were collected during the 2004 annual relief voyage by the SA *Agulhas I*.[13] To the best of our knowledge, this report was the first and only report of a multidisciplinary research consortium from exclusively South African institutions (Rhodes University, University of Cape Town and Council for Scientific and Industrial Research), searching for cytotoxic chemical compounds from marine sponges with potential anti-tumour properties from the Southern Ocean.[13] The natural product investigation by Davies-Coleman et al.[13] of the most promising extract was unfortunately hampered by the paucity of the sponge material collected (ca. 200 g), although the amount of crude organic extract derived was not given. As far as we know, a subsequent collection in 2005 has to date not yielded any new and/or bioactive marine natural product compounds.

The relatively unexplored natural product diversity and pharmaceutical potential of marine invertebrates from the sub-Antarctic Islands and Southern Ocean, therefore presents us with an exciting opportunity and we have thus initiated a new multidisciplinary effort by a concerted group of South African institutions to explore the cytotoxic potential of marine sponges from the sub-Antarctic Islands and the Southern Ocean. We report here the cytotoxic activity, against three different cancer cell lines, of both aqueous and organic crude extracts prepared from nine sponge specimens, collected from various depths around the PEIs in the Southern Ocean.

Materials and methods

Researchers from the University of Cape Town, Rhodes University, the Council for Scientific and Industrial Research, the South African Environmental Observation Network and the Department of Environmental Affairs participated in the annual relief voyage to the PEIs on board the research vessel SA *Agulhas II* from 6 April to 16 May 2015. As part of the inter-island biodiversity survey between Prince Edward and Marion Islands (46.77°S, 37.85°E), benthic trawls were performed using a steel dredge (1.0 m × 0.3 m ×1.0 m) at various depths (60–254 m) along four pre-determined stations (Figure 1). At each station, the dredge was towed behind the vessel at 0.5 knots for approximately 15–20 min.

From the various benthic dredges, nine sponge specimens were collected, labelled and photographed and the relevant collection information, e.g. GPS coordinates and depths, was recorded. The sponge specimens were placed in sealed bags, immediately frozen on board the research vessel until the end of the journey and transferred after the expedition to a freezer in the Department of Chemistry at the University of Cape Town. Histological and genetic samples (~5–10 cm³) were taken from each specimen for identification and stored in 96% (v/v) ethanol. Vouchers of the sponges are housed in the collection of Dr Toufiek Samaai (Department of Environmental Affairs).
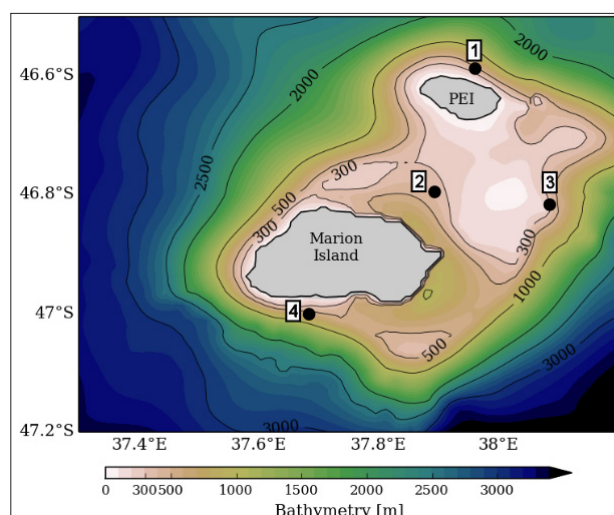
Figure 1: Survey area around the Prince Edward Islands, indicating the position of the four benthic dredge stations: (1) Prince Edward Island (PEI); (2 and 3) between the islands; and (4) south of Marion Island. Station depths are: 130 m (1), 65 m (2), 230 m (3) and 254 m (4).

All sponge specimens were processed individually in the same manner. A frozen portion (up to 90% of the wet mass) of each sponge was cut into small pieces, frozen in liquid nitrogen and lyophilised until the material was dry. The mass of the dried sponges was recorded and the sponge materials were then subjected to solvent extraction. Both aqueous (100% deionised water) and organic (50:50 mixture of methanol and dichloromethane) crude extracts of the sponges were prepared by following, where possible, the protocol for marine specimens established by scientists at the Developmental Therapeutics Program of the NCI.[15]

The resulting 18 crude extracts, one organic (suffix '-1') and one aqueous (suffix '-2') for each sponge, were screened for their cytotoxic activity against the cervical (HeLa ATCC CCL-2), lung (A549 ATCC CCL-185) and breast (MCF7 ATCC HTB-22) cancer cell lines. Cells were cultured in Dulbecco's Modified Eagle's Medium supplemented with 10% (v/v) foetal bovine serum, 2 mM Glutamax and 100 U/mL penicillin/streptomycin/amphotericin at 37 °C with 9% CO_2. Sub-culturing was carried out by removing the media, washing with phosphate-buffered saline (10 mM Na_2HPO_4, 1.8 mM KH_2PO_4, 2.7 mM KCl, 140 mM NaCl, pH 7.4) and lifting the cells using a 1% (v/v) trypsin in 0.3% (w/v) ethylenediaminetetraacetic acid (EDTA) solution. Cells were confirmed to be mycoplasma free by microscopy after staining with Hoechst 33342. The chemosensitivity of the cell lines to the extracts was tested using the WST-1 assay. For each extract, 5 x 10³ cells/well were seeded in a 96-well plate and allowed to sit overnight before being treated in triplicate with the extracts at a range of concentrations (0.1–100 μg/mL). The plates were incubated at 37 °C and 9% CO_2 for 72 h. All the media were removed and 100 μL WST-1 reagent (0.5 mg/mL) was added to each well after the 72-h incubation. The plates were incubated at 37 °C in 9% CO_2 and the absorbance at 450 nm read after 2 h, 4 h and 6 h. For all treatments the half maximum effective concentration (EC_{50}) for data from the reading at 6 h was calculated by non-linear regression using GraphPad Prism version 4.0. Spectroscopic analyses of portions of the crude extracts were acquired using standard pulse sequences on a Bruker 600-MHz nuclear magnetic resonance (NMR) spectrometer, equipped with a 5-mm Prodigy cryoprobe.

Results

Taxonomic identification of the nine sponge specimens collected revealed eight sponge species (Table 1) as two sponge specimens (MAR15-005 and MAR15-012), collected from two distant geographical locations (Dredge Stations 3 and 1, respectively; Figure 1), were identified as the same *Halichondria (Halichondria)* cf. *panicea* sponge species. The preliminary biological screening results obtained for the crude extracts are summarised in Table 1 and indicate that only three (MAR15-001, MAR15-005, MAR15-012) of the nine sponge specimens exhibited cytotoxicity against the cancer cell lines tested.

Table 1: Taxonomy, dredge details and EC_{50} ($\pm$ sem; μg/mL) values against HeLa, MCF7 and A549 cancer cell lines for nine species of sponges collected from the Marion Island region

| Sponge	Dredge station	Depth (m)	EC_{50} (μg/mL)		
			HeLa Cervical	MCF7 Breast*	A549 Lung*
Myxilla (Ectyomyxilla) kerguelensis					
MAR15-001-1	2	65	–	P	P
MAR15-001-2			11.7 ($\pm$1.1)	P	–
Halichondria (H.) cf. *panicea*					
MAR15-005-1	3	230	–	P	33.4 ($\pm$1.3)
MAR15-005-2			–	–	–
Acanthella erecta					
MAR15-006-1	3	230	–	P	–
MAR15-006-2			–	–	–
Geodia sp.					
MAR15-007-1-1	3	230	–	P	P
MAR15-007-1-2			–	–	–
Tetilla leptoderma					
MAR15-007-2-1	3	230	–	P	P
MAR15-007-2-2			–	–	–
Lissodendoryx sp.					
MAR15-008-1	3	230	–	–	P
MAR15-008-2			–	P	–
Hymeniacidon sp.					
MAR15-011-1	1	130	–	P	P
MAR15-011-2			–	P	–
Halichondria (H.) cf. *panicea*					
MAR15-012-1	1	130	–	P	35.2 ($\pm$4.4)
MAR15-012-2			16.1 ($\pm$1.1)	16.5 ($\pm$1.1)	–
Rosella antarctica					
MAR15-013-1	4	254	–	P	P
MAR15-013-2			–	P	P

Note: For each sponge specimen, suffix '-1' = organic extract and suffix '-2' = aqueous extract.

*P, proliferation

The aqueous extract of the sponge *Myxilla (Ectyomyxilla) kerguelensis* (MAR15-001-2) displayed cytotoxic activity (EC_{50} 11.7 μg/mL) against the HeLa cervical cancer cell line (Table 1). The organic extracts of the two sponges *H. (H.)* cf. *panicea* were similarly cytotoxic (MAR15-005-1; EC_{50} 33.4 μg/mL and MAR15-012-1; EC_{50} 35.2 μg/mL) against the A549 lung cancer cell line (Table 1). Interestingly, the aqueous extract of *H. (H.)* cf. *panicea* (MAR15-012-2) was similarly cytotoxic (EC_{50} ~16 μg/mL) to both the cervical (HeLa) and breast cancer (MCF7) cell lines, whilst the aqueous extract of the other *H. (H.)* cf. *panicea* specimen (MAR15-005-2) did not show any cytotoxic activity (Table 1) against any of the cancer cell lines. It is also interesting to note that a number of the samples, particularly the organic extracts, induced dose-dependent proliferation in the MCF7 and A549 cell lines (shown by P in Table 1).

Proton (^{1}H) NMR spectra (Figure 2) were acquired for both organic and aqueous extracts of the two different sponge specimens *H. (H.)* cf. *panicea* to examine their secondary metabolite profile and assist us in the interpretation of the bioactivity results obtained. A comparative analysis of the ^{1}H NMR spectra of the organic extracts for *H. (H.)* cf. *panicea* (MAR15-005-1 and MAR15-012-1) indicated a high degree of consistency in their spectroscopic profile (Figure 2a). A cursory examination of the ^{1}H NMR spectra recorded for the MAR15-005-2 and MAR15-012-2 aqueous extracts revealed some degree of variation in their secondary metabolite spectroscopic profile (Figure 2b).
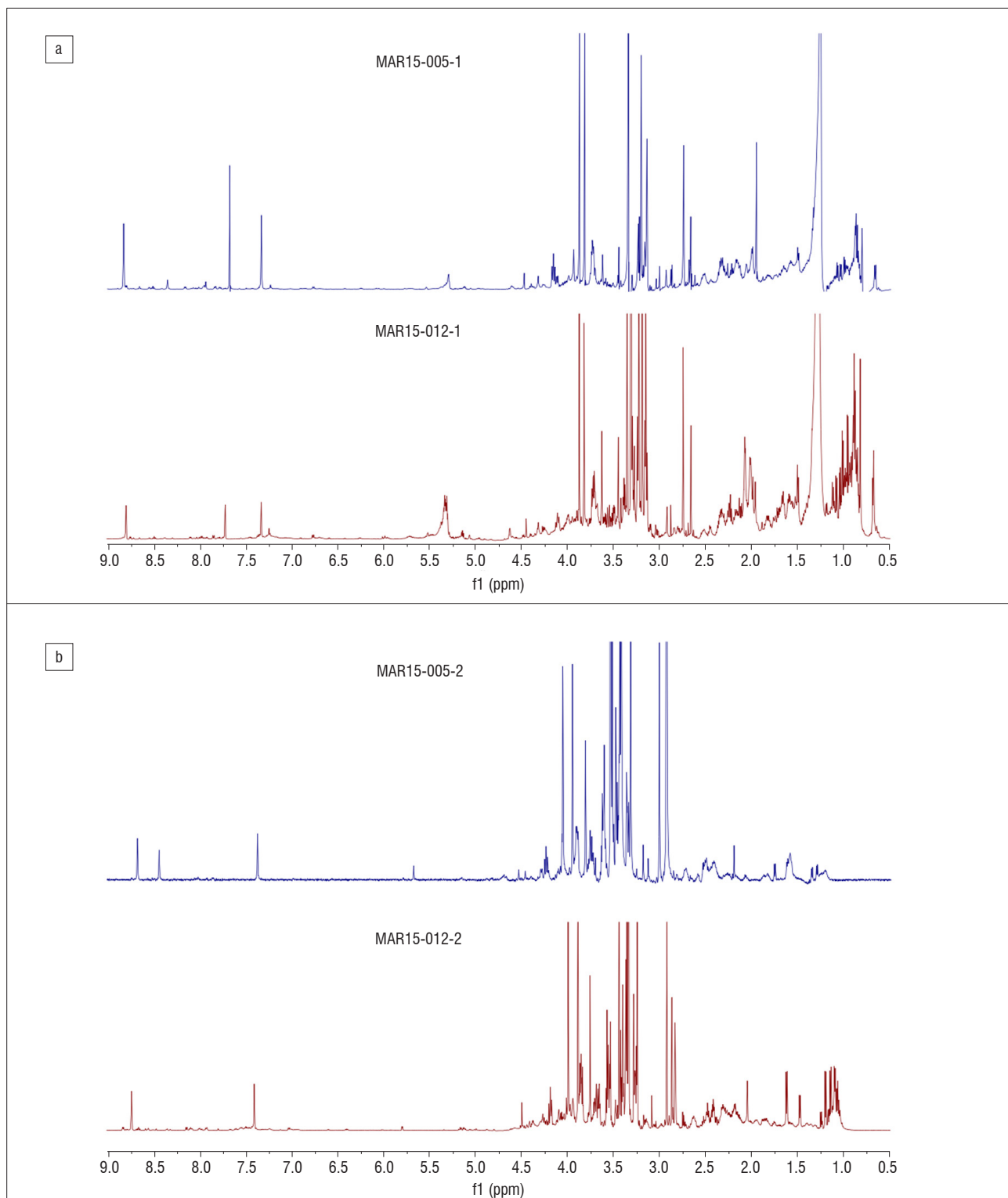
Figure 2: Comparison of the secondary metabolite profile of the two different specimens of *Halichondria (Halichondria)* cf. *panicea*. Stacked proton nuclear magnetic resonance spectra (δ_H 0.5–9.0 ppm) of (a) organic extracts (MAR15-005-1 and MAR-012-1) and (b) aqueous extracts (MAR15-005-2 and MAR-012-2).

Conclusion

The results obtained from the biological screening of the nine sponge specimens (eight species) collected off the PEIs are exceptionally encouraging as the potential presence of cytotoxic secondary metabolites of three sponge specimens has been revealed. The similar cytotoxicity ($EC_{50} \sim 34 \, \mu g/mL$; Table 1) of the organic sponge extracts MAR15-005-1 and MAR15-012-1 (collected at two different locations; Figure 1) against the A549 lung cancer cell line, coupled with their significantly congruent ^{1}H NMR profile (Figure 2a), is very interesting as it suggests the possible presence of the same active molecule(s) from allopatric specimens of *H. (H.)* cf. *panicea*, that is, different specimens of the same species from different locations. However, the aqueous extracts prepared from the same two *H. (H.)* cf. *panicea* sponge specimens exhibit different cytotoxic (Table 1) and spectroscopic profiles (Figure 2b), suggesting possible

different polar secondary metabolite constituents in these extracts. This variation in the biological activity and secondary metabolite profile of different specimens of the same sponge species was also observed in three different *H. (H.)* cf. *panicea* specimens by Davies-Coleman and co-workers, which can be attributed to the production of different secondary metabolites as a result of different environmental pressures such as predation and competition with neighbouring species.[13]

Marine sponges belonging to the genus *Halichondria* are known as producers of a wide range of bioactive secondary metabolites, including the well-known polyether macrolide class of highly cytotoxic compounds, such as the marine natural product Halichondrin B that showed excellent initial cancer cell toxicity and subsequently led to the development of the approved anti-cancer drug Halaven®.[6,7] Therefore, in our future work, we will aim to isolate and identify the cytotoxic compounds present in the two specimens of *H. (H.)* cf. *panicea*, as they could potentially be useful for further chemical and biological studies. The aqueous extract of the sponge *Myxilla (Ectyomyxilla) kerguelensis* (MAR15-001-2) displayed cytotoxic activity (EC_{50} 11.7 μg/mL; Table 1) against the HeLa cervical cancer cell line. A literature search revealed that the Marion Island sponge *Myxilla (Ectyomyxilla) kerguelensis* has not been extensively studied for its natural products chemistry composition and therefore presents us with a unique opportunity to address this knowledge gap. We have also recently extended our primary biological screening of these sponge extracts to include antiplasmodial and anti-tubercular bioassays, in which none of the extracts exhibited any activity. The results obtained in this study will now guide the purification, isolation and characterisation of the bioactive molecules responsible for the observed cytotoxicity against our panel of three cancer cell lines. Additionally, to supplement the paucity in biomass for the sponges collected during the 2015 expedition, more samples were collected from the same location on the recent 2016 survey to Marion Island and will be investigated in the same manner.

We highlight and acknowledge the various contributions from diverse scientific disciplines that have played a pivotal role in providing the necessary platform to enable this newly established marine biodiscovery effort to explore the natural products chemistry and biomedicinal value of marine sponges from the PEIs and Southern Ocean. We hope that our future collections of sponges and other marine invertebrates from the Southern Ocean will lead to the isolation and identification of highly potent cytotoxic marine natural product compounds such as Halichondrin B[6,7], Cephalostatin 1 and Spongiostatin 4[9,10], which have the potential to be further developed into anti-cancer agents. In our endeavour, we also hope to contribute to the understanding of the biodiversity and geographic distributions of these cold-water invertebrates.

Acknowledgements

Authors' contributions

E.K.O. was responsible for acquiring spectroscopic data on the sponge extracts; C.K.D.C. and G.A.D. were responsible for the preparation of the crude aqueous and organic sponge extracts; E.P. collected the sponge specimens; I.A. facilitated the annual relief voyage to Marion Island on board SA *Agulhas II*; T.S. identified the sponges collected from the 2015 survey; L.M.K.D. was responsible for designing and conducting the anti-tumour activity assays; A.L.E. was responsible for designing the anti-tumour activity assays and for analysis and interpretation of the data; S.N.S. was the overall project leader, oversaw the multidisciplinary process as a whole, especially the chemistry component of this project, and was responsible for conducting some of the experiments. S.N.S. wrote and edited the majority of the text and all authors provided input into the writing of the manuscript.

References

1. Fusetani N. Biotechnological potential of marine natural products. Pure Appl Chem. 2010;82(1):17–26. http://dx.doi.org/10.1351/pac-con-09-01-11

2. Bolton JJ, Davies-Coleman MT, Coyne VE. Innovative processes and products involving marine organisms in South Africa. Afr J Mar Sci. 2013;35(3):449–464. http://dx.doi.org/10.2989/1814232X.2013.830990

3. Simmons TL, Andrianasolo E, McPhail K, Flatt P, Gerwick WH. Marine natural products as anticancer drugs. Mol Cancer Ther. 2005;4(2):333–342.

4. Blunt JW, Copp BR, Keyzers RA, Munro MHG, Prinsep MR. Marine natural products. Nat Prod Rep. 2016;33:382–431. http://dx.doi.org/10.1039/c5np00156k

5. Leal MC, Puga J, Serôdio J, Gomes NCM, Calado R. Trends in the discovery of new marine natural products from invertebrates over the last two decades – where and what are we bioprospecting? PLoS One. 2012;7(1):e30580. http://dx.doi.org/10.1371/journal.pone.0030580

6. Gerwick WH, Moore BS. Lessons from the past and charting the future of marine natural products drug discovery and chemical biology. Chem Biol. 2012;19(1):85–98. http://dx.doi.org/10.1016/j.chembiol.2011.12.014

7. Newman DJ, Cragg GM. Marine-sourced anti-cancer and cancer pain control agents in clinical and late preclinical development. Mar Drugs. 2014;12(1):255–278. http://dx.doi.org/10.3390/md12010255

8. Newman DJ, Cragg GM. Natural products as sources of new drugs from 1981 to 2014. J Nat Prod. 2016;79(3):629–661. http://dx.doi.org/10.1021/acs.jnatprod.5b01055

9. Davies-Coleman MT, Sunassee SN. Marine bioprospecting in southern Africa. In: Chibale K, Davies-Coleman MT, Masimirembwa C, editors. Drug discovery in Africa. Berlin, Heidelberg: SpringerVerlag; 2012. p. 193–209. http://dx.doi.org/10.1007/978-3-642-28175-4_8

10. Davies-Coleman MT. Bioactive natural products from southern African marine invertebrates. In: Atta-ur-Rahman, editor. Studies in natural products chemistry. Amsterdam: Elsevier B.V.; 2005. p. 61–107. http://dx.doi.org/10.1016/S1572-5995(05)80054-7

11. Davies-Coleman MT, Beukes DR. Ten years of marine natural products research at Rhodes University. S Afr J Sci. 2004;100:539–544. http://dx.doi.org/10.1002/chin.200537245

12. Davies-Coleman MT. Natural products research in South Africa: End of an era on land or the beginning of an endless opportunity in the sea? S Afr J Chem. 2010;63:105–113.

13. Davies-Coleman MT, Froneman W, Keyzers RA, Whibley CE, Hendricks D, Samaai T, et al. Anti-oesophageal cancer activity in extracts of deep-water Marion Island sponges. S Afr J Sci. 2005;101(11–12):489–490.

14. Whibley CE, Keyzers RA, Soper AG, Davies-Coleman MT, Samaai T, Hendricks DT. Antiesophageal cancer activity from southern African marine organisms. Ann N Y Acad Sci. 2005;1056:405–412. http://dx.doi.org/10.1196/annals.1352.031

15. McCloud TG. High throughput extraction of plant, marine and fungal specimens for preservation of biologically active molecules. Molecules. 2010;15(7):4526–4563. http://dx.doi.org/10.3390/molecules15074526