An Exploration of Media Repertoires in South Africa: 2002-2014

Submitted as a partial fulfillment toward an MSc in Data Science

Department of Statistical Science

University of Cape Town

by

Hans-Peter Bakker

2018

Supervisor: Associate Professor Ian Durbach

Abstract

This dissertation explores trends in media engagement in South Africa over a period from 2002 until 2014. It utilises data from the South African Audience Research Foundation's All Media and Products Surveys. Using factor analysis, six media repertoires are identified and, utilising structural equation modelling, marginal means for various demographic categories by year are estimated. Measurement error is determined with the aid of bootstrapping. These estimates are plotted to provide visual aids in interpreting model parameters. The findings show general declines in engagement with traditional media and growth in internet engagement, but these trends can vary markedly for different demographic groups. The findings also show that for many South Africans traditional media such as television remain dominant.

Acknowledgments

Thank you to my supervisor Ian Durbach for his patience and wise guidance throughout the convoluted development of this dissertation. Thank you also to my former Rhodes University colleague and Statistics lecturer Jeremy Baxter for my being able to pop into his office at any time to ask advice. Thank you dear Athina for struggling along with me on what has been a very long and arduous path that started more than five years ago. And thanks to my mother who continues to support my intrepid adventures.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter provides initial broad descriptions of the project, statements of aims and objectives and the identification of a few specific examples of pertinent research questions. It also provides a project layout by briefly describing the content of each chapter.

This research is situated in the media environment in South Africa which, along with the global media industry, has undergone dramatic changes over the past decade - driven largely by the disruptive effects of the internet on traditional forms of media. The data in this project are drawn from national surveys of media and product usage, the All Media and Products Survey (AMPS), that have been conducted under the auspices of the South African Audience Research Foundation (SAARF) at least once a year from 1974 until 2014.

This project focusses on two main perspectives of the media environment in South Africa over the period under review:

- The first is aimed at identifying unobserved latent factors (called *repertoires* in this project) that underly the national multi-media environment. Respondents' aggregate degrees of engagement by demographic categories and by year are estimated for these *repertoires*, plotted and interpreted.

- The second is placing these *repertoires* in context by considering clustering of respondents as well as changes over time of respondents' aggregate degrees of engagement with broad media classes (referred to as *types* in this project). In particular, the *types* are: radio, television, newspapers, magazines, and the internet.

Given the nature of the data and the context of the research, various limitations needed to be considered. These included the construction of ordinal scales of engagement values that differed between broadcasting, internet and print media; the assumption that the sampling proportions were sufficiently representative to forgo the use of population weights; and further limitations that are described more fully in section 7.5 on page 90.

1.1 Aim

The aim of this work is to explore changes in media usage and media consumption behaviour in South Africa between 2002 and 2014. Although the intention was to consider media as *hard news and information*, as described by Anderson et al. (2012, p. 3) in section 2.1, separating *hard news and information* from entertainment was not always practically possible, confirming the ongoing blurring of news and entertainment as described by (Edgerly, 2015, p. 2) and Schröder (2015, p. 61, citing Bjur et al., 2013) in section 2.2. For example, while it was possible to leave out some internet activities such as the playing of games and accessing emails, it was not possible to identify television or radio programmes that only provided news and information or isolate the use of social media as a source of news and information. This sometimes hybrid nature of the data must be considered in reading the results.

1.2 Project Objectives

The objectives include:

- Identify and describe the main latent factors of cross media engagement (or media *repertoires*) through a dimension reduction of all national media vehicles in the dataset from 2002 until at least 2014;
- Model and describe changing aggregate levels of engagement on media *repertoires* for selected demographic categorical variables over time;
- Identify and describe clusters of relative engagement with five media *types* (newspapers, magazines, television, radio and the internet);
- Model and describe changing aggregate levels of engagement by media *type* for selected demographic categorical variables over time;
- Consolidate the perspectives of media *types* and media *repertoires* to gain an overview of shifts in media behaviour over the period under review.

1.3 Project Layout

Chapter 2 describes the contextual overview, which reviews the importance of media as a sector, the current media environment, and research in media; chapter 4 considers the methods used in this analysis, which, given the nature of the data and the aim of the study, are aimed mainly at developing a sound understanding of Structural Equation modelling (SEM); chapter 3 describes the extensive data preparation required for this project; chapter

5 explores latent structure, or what are referred to as media *repertoires* in this project, and changes of engagement for different demographic categories by factor or *repertoire* over the period under review; chapter 6 focusses on exploring changes of engagement by media type through clustering and linear regressions; chapter 7 offers a conclusion that contains a synthesis of the main results and considers some limitations of this study as well as suggestions for future research.

Chapter 2

Contextual Overview

This section serves to provide a broad contextual overview in which to situate this project. In section 2.1 the importance of media, in particular its role in the developing world, is considered; section 2.2 reviews the dramatic shifts impacting on the current global media environment, with specific reference to changed in the South African media industry, and section 2.3 considers some approaches to research in the contemporary media context.

2.1 The Importance of Media

A sound, functioning and strong media sector is important for societies and countries to build an open society. Nobel laureate and economist Amartya Sen observes that famine has never occurred in a democracy with a free press: "intimations of mass starvation are impossible to hide where journalists freely give voice to public criticism and warn of impending crises" (*Toward Economic Sustainability of the Media in Developing Countries* (Anon, 2007)).

Susman-Pena (2012), citing research over several years that show the coexistence of a sound media sector on the one hand and good developmental outcomes on the other, posits that media *matters*. She defines a *healthy* media sector as one that is "free, independent, produces high quality information, reaches all or most of the population, offers diverse perspectives, and provides the information people need to be able to make decisions and to be able to hold their government to account" (Susman-Pena, 2012).

Peters (2010, citing Jakubowicz and Sukosd, 2008), argues that media are an especially important focus of attention, "given the role they are often assumed to have in creating national identity and contributing to an energized democratic society". According to Conrad (2014), the notion that an independent media is the foundation of a functional democracy has been argued by liberal theorists, including John Locke, John Madison, John Milton, and John Stuart Mill, for decades. Democratic government relies on the ability of its citizens to make informed decisions and this requires access to information that is accurate, more often than not implying a "free and independent media" (*Toward Economic Sustainability of the Media in Developing Countries* (Anon, 2007)).

An effective media sector is also important to hold public officials accountable and in so doing help in the exposing of corruption (*Toward Economic Sustainability of the Media in Developing Countries* (Anon, 2007)). Susman-Pena (2012) lists governance institutions as an important outcome of a healthy media sector. More specifically, media oversight that provides information about government activities, decision-making and budgeting can ensure government accountability and "expose vice or incompetence". (Peters, 2010, citing Paul Starr, 2009; Adsera, Boix and Payne, 2003), argues that "corruption is more likely to flourish when those in power have less reason to fear exposure" and that a strong negative correlation exists between corruption and free circulation of newspapers in a country.

A healthy media environment also assists in promoting economic growth by disseminating information and ensuring transparency (*Toward Economic Sustainability of the Media in Developing Countries* (Anon, 2007)). Susman-Pena (2012) argues that the support and promotion of a healthy media sector is critical to "grow economies, alleviate poverty, and improve lives". A healthy media sector creates an information flow that is the lifeblood of a working and efficient society. In fact an environment that fails to support a free press, such as a regulatory environment that does not support the freedom of journalists to do their work and one that discourages independence and plurality, will also fail in achieving these outcomes (*Toward Economic Sustainability of the Media in Developing Countries* (Anon, 2007)). And, as posited by Susman-Pena (2012), "by increasing people's knowledge about their own and other societies, the media may strengthen bonds and common understandings among people".

In what Susman-Pena (2012) describes as "a remarkable commitment to supporting independent media, together with access to information, and freedom of expression" President Obama, in his May 2011 speech on the Middle East, declared:

> "Real reform does not come at the ballot box alone. Through our efforts we must support those basic rights to speak your mind and access information. We will support open access to the Internet, and the right of journalists to be heard–whether it's a big news organization or a lone blogger. In the 21st century, information is power, the truth cannot be hidden, and the legitimacy of governments will ultimately depend on active and informed citizens."

These arguments about the importance of media are in fact arguments for a particular kind of media and by extension for a particular kind of journalism. Anderson et al. (2012, p. 3) describe this by holding that not media as much as *journalism* matters: "Journalism exposes corruption, draws attention to injustice, holds politicians and businesses accountable for their promises and duties. It informs citizens and consumers, helps organise public opinion, explains complex issues and clarifies essential disagreements. Journalism plays an irreplaceable role in both democratic politics and market economies."

Anderson et al. (2012, p. 3) qualify this argument by stating that not all journalism matters, since "much of what is produced today is simply entertainment or diversion" . What matters

has variously been referred to as *hard news, accountability journalism*, or *the iron core of news.* Rather than trying to list or define the elements that separate hard news from what they refer to as "fluff", they adopt Lord Northcliffe's litmus test:

> "News is something someone somewhere doesn't want printed. Everything else is advertising" (Anderson et al., 2012, p. 3).

And state that:

> "Hard news is what distinguishes journalism from just another commercial activity. There will always be a public appetite for reporting on baseball, movie stars, gardening and cooking, but it's of no great moment for the country if all of that work were taken over by amateurs or done by machine. What is of great moment is reporting on important and true stories that can change society. " (Anderson et al., 2012, p. 3).

2.2 The Current Media Environment

The digitisation of media content and the expansion of the Internet has led to a large increase in the media sources available to consumers. In this proliferation of media, one of the main research questions has been how audiences respond to this explosion of choice. For example, does it lead to people consuming a steady and consistent diet of their preferred news genre or do they expand their consumption to a wider, more diverse range of sources (Xu et al., 2014, citing Gentzkow and Shapiro, 2011; Ksiazek, Malthouse and Webster, 2010)?

According to Xu et al. (2014, p. 100, citing Hollander, 2008; Iyengar and Hahn, 2009; Kziazek, Malthouse and Webster, 2010) a key finding on this question by various researchers is that consumers have responded to the increase in media sources using a strategy of "selective exposure" in which they consume more of similar news from a small number of news providers rather than consuming a greater variety from a larger number of news providers.

According to Schröder (2015, pp. 60-61), previous patterns and routines of news consumption, including trust in the news, are being transformed by a "nexus of innovative technologies and the revolutionising journalistic processes they afford. This nexus results in news media content which sets societal agendas and frames cultural issues in new ways for news audiences, irrespective of whether they get their news directly from social media platforms or not." Schröder (2015, p. 61, citing Bjur et al., 2013) argues that audiences are inherently cross-media and in the digital age, "emerging patterns of cross-media use are far more seamless and blurred, hybrid and complex than they used to be".

Edgerly (2015, p. 1) states that in a world where we have moved beyond a limited number of television channels, radio stations and print news outlet to an environment in which we make selection choices "amid hundreds of television channels, smart phone technologies, and

virtually unlimited news options available online", the most defining characteristic of the current media environment is *media choice.*

While the "low-choice" media environment was characterised by a "consistent approach to news presentation and style", the "high-choice" media environment is characterised by *diversity,* reflected by ideologically driven news and the blurring of news and entertainment (Edgerly, 2015, p. 2). Furthermore, Edgerly (2015, pp. 2-3, citing Baym, 2010; Shoemaker and Reese, 1996) identifies the "interplay of ownership influence, pressure to fill a large news hold, and increased competition" as resulting in "news content that shies away from traditional notions of neutrality and objectivity".

Edgerly (2015, pp. 3-4) also identifies soft news, daytime and late-night talk shows, and news satire as examples of a rise of "hybrid media" that "blur the line between news and entertainment" and expresses concern that "individuals are turning away from news altogether, or the news they do select is too entertainment oriented". Edgerly (2015, p. 4, citing Postman, 1986) cautions that the popularity of hybrid media combined with declines in traditional forms of news may indicate that we are "amusing ourselves to death".

The drop in news exposure among younger people is particularly evident. In this regard the Pew Research Center for the People and the Press (2010) found that only 23% of 18 to 29 year-olds regularly read a newspaper, compared to 55% of people over 65 - a pattern also evident in audiences of network evening news. The declining levels of news exposure among younger people may not be an indication that they are "tuned out" or "fleeing" from news, but that they are consuming "a different set of news content", for example from various online and social media sources and from entertainment news-driven television shows, such as the *Daily Show.* The concern is that the younger generations are replacing traditional forms of news with lower quality ones and that the declines in political knowledge and participation among this cohort reflects a change in news diets. "The overarching concern is not that these new types of news are inherently bad, but that the exclusive use of *only* ideologically driven news, or *only* media that mix news and entertainment is the cause for worry" (Edgerly, 2015, p. 4, citing Mindich, 2005; Patterson, 2008).

The explosion of media choice has according to Turcotte et al. (2015, p. 520, citing Prior, 2007; Stroud, 2011) resulted in audiences drifting away from mainstream media , exacerbated by "a steady decline of public trust in the institution of news" . Turcotte et al. (2015, p. 521, citing Gronke and Cook, 2007; Ladd, 2011) hold that "scholars have observed aggregate level declines of public trust in the news over the last few decades, transforming a once revered news profession to a subject of disdain and dissatisfaction". However, in spite of this aggregate decline of trust, not all news outlets are similarly impacted. For example, according to Turcotte et al. (2015, p. 521, citing Arceneaux, Johnson and Murphy, 2012; Gronke and Cook, 2007), "the public is more trusting of their preferred outlets for news and more trusting of local news outlets" and a 2014 survey released by Public Policy Polling found that *Fox News* was considered both the most and least trusted news source, suggesting that

news outlet credibility varies according to partisan predispositions.

The current era of expanded media choice, has led to people choosing news sources that are in agreement with their ideology - a selective exposure that may have made people too trusting of their preferred outlets while distrusting news sources that don't agree with their ideological leanings (Turcotte et al., 2015, pp. 521, citing Arceneaux et al., 2012). While demographics and political knowledge can also play a role in determining which news outlets one perceives as credible, media trust has been conceptualised in several different ways, including trust in content, trust in journalists or those responsible for delivering the news, and trust in media ownership (Turcotte et al., 2015, pp. 522, citing Williams, 2012). Trust in news can also directly influence political behaviour. As media distrust grows, the voting public becomes more dependent on partisan cues to determine their vote; and it is more likely to abandon mainstream news. Furthermore the dropping levels of media trust "fosters a heightened perception that the current political climate is a polarizing one" (Turcotte et al., 2015, pp. 522, citing Ladd, 2011; Ladd, 2013).

Edgerly (2015, p. 1-3, citing Mindich, 2005; Prior, 2007; and others) argues that the expansion of media options and therefore of media choice has led to a concern that it makes it easy to avoid news content altogether and that the "fragmented-nature" of news exposure can lead to the gravitation of audiences toward "sources that reinforce their ideological viewpoints and are of lower quality". She identifies a concern that ideological news "makes it easier for people to consume only content that agrees with their political views" and in reference to the red and blue divide signifying the Republican and Democratic political poles in the United States, argues that "despite all the colourful options the news media landscape offers, some audiences only see red media or blue media". Xu et al. (2014, p. 100, citing Ksiazek, Malthouse, and Webster, 2010; Stroud, 2008) also argue that due to the "increasing proliferation of news outlets, consumers with a particular political preference will be more likely to consume from news outlets that match their own value beliefs. This behavior results in a penchant for ideological segregation."

In a study on cross-media usage, Schröder (2015, p. 71) asks if the established mixtures and levels of news media qualify a given audience as "well-informed, resourceful and competent citizens". Starting with a question of what the media repertoire of a *competent* citizen would look like, he holds that traditionally daily newspaper readership would have been considered a *"sine qua non* of informed citizenship" and that the decline in newspaper readership would therefore indicate "democratic deficit... that could pose a serious challenge to the democratic health of a country". However this view is challenged by the changing face and growing consumption of online news services, many of which emanate from former print offerings, and the growing role of mobile platforms and social media that may indicate high levels of *digital literacies*, which in turn are increasingly being considered a "prerequisite of democratic citizenship, as well as civic agency" (Schröder, 2015, p. 71, citing Lund et al., 2014; Curren et al., 2009).

In such a complex media environment, while most news mediums struggle for audiences some social media sites are showing potential for growth as carriers of news. Facebook for example has become one of the fastest growing tools for gathering news with more than half its users consuming news on the site and 78% of its users reporting exposure to news while using Facebook for social and other reasons (Turcotte et al., 2015, pp. 521, citing Pew Research, 2014). Also in a media environment with so much choice, "one extremely important way (individuals) decide what to pay attention to is through recommendations that reach them through their online social networks". Turcotte et al. (2015, pp. 523, citing Mutz and Young, 2011), suggest that opinion-leaders play increasingly important roles in facilitating exposure to news; and 'friends' can play the role, traditionally reserved for journalists and editors, as information gate keepers, vetting the significance and relevance of news content and therefore help shape public agendas (Turcotte et al., 2015, pp. 524).

Xu et al. (2014, p. 98, citing comScore, 2012) argue that it is critical for advertisers for example understand any changes in news consumption behaviour since, after email and texting, accessing news in the 2012 study was found to be the most popular mobile data activity in the United States (U.S). The digitisation of news has also fundamentally reshaped the news industry. For example in the U.S. newspaper advertising revenues fell 47% from 2005 to 2009 as online advertising spending climbed to more than $100-billion in 2012 (Xu et al., 2014, p. 97).

The proliferation of new media outlets, resulting in growing numbers of people meeting their news needs through multiple outlets, has raised the problem of how best to target and reach consumers in such a multichannel environment. While online tracking technologies have made it possible to track online consumers, marketers still need to consider the placement of advertising on multiple outlets in order to reach consumers effectively. The emergence of these "disruptive channels" has made it imperative for to monitor changes in consumers' media and news consumption behaviours (Xu et al., 2014, p. 97, citing Ahonen, 2011; Athey, Calvano and Gans, 2013).

> "If you wanted to sum up the past decade of the news ecosystem in a single phrase, it might be this: Everybody suddenly got a lot more freedom. The newsmakers, the advertisers, the startups, and, especially, the people formerly known as the audience have all been given new freedom to communicate, narrowly and broadly, outside the old strictures of the broadcast and publishing models." (Anderson et al., 2012, p. 1)

2.2.1 Changing Media Environment in South Africa

The history of the All Media and Products Survey (AMPS) used extensively in this project is also a reflection of the history of change in South Africa - in particular change in media. To remain representative of the population, the survey continually adjusted its methods

to changing demographics. For example, according to Fulton (2017), in 1995 only 61% of households in South Africa had electricity compared with 94% in 2015; average household income in South Africa showed substantial shifts for different population groups over the study period with the number of black households in the top-income tier increasing dramatically from 17% in 2006 to 41% by 2015 (Fulton, 2017).

In a map of key events in South Africa's media landscape, Fulton (2017) identifies the arrival of smart phones in 2009, the launch of the *Daily Sun* in 2011 which, according to Fulton (2017), "shattered readership records in print recession with over 5 million readers"; and the beginning of online streaming of movies by *Netflix* and *Showmax* in 2015/2016. Since 1995 the number of media vehicles have also shown dramatic increases: from 84 print titles in 1995 to 163 in 2015; from 35 radio stations in 1995 to 297 in 2015; and from 14 TV channels in 1995 to 330 in 2015 (Fulton, 2017).

Another media channel that has shown very dramatic shifts impacting on the media industry over the period under review is the growth in numbers of cellphone users. In 2004 10.2 million (representing 34% of adult South Africans) had cellphones and by 2015, 34 million (representing 84% of adult South Africans) had cellphones (Fulton, 2017).

Fulton (2017) presented a plot, shown in figure 2.1, in which trends of proportions of the population consuming various media are shown. The plot shows that most South Africans still consumed radio and television at a relatively stable 92%, while internet (and to a lesser extent DSTV), which by 2014 was only consumed by 46% of the adult population, showed strong compound annual growth of 27% over the period 2009-2015. The measures for inclusion of respondents in these proportions are based on particular standards describe in the plot itself.

While the plot by Fulton (2017) in figure 2.1 considers changes in proportions, this study considers changes in intensity of engagement over largely the same time period. For example figure 2.1 shows relative stable proportions of respondents in print, this study will show declines in degrees of engagement and also how these rates of decline differ for different demographic categories.

2.3 Research of the Media Environment

Schröder (2015, pp. 60-61) argues that it is important to monitor "on a continuous basis precisely what the landscape of news looks like: what technological platforms and formats are receding and emerging, and which are dominant, as well as how people are accessing, navigating in, and making sense of the landscape of news."

Various attempts to "develop and operationalise" new conceptual frameworks for "mapping and explaining the cross-media practices of audiences", suggests that cross-media consumption can be researched from three perspectives:

- firstly in terms of *functional differentiation* that considers the extent to which different media complement and co-exist with each other, for example when one medium

specialises in fulfilling certain types of needs in order to differentiate itself from its rivals;

- secondly, research that is concerned with the building of *media repertoires*, i.e. how audiences create "personal constellations of media", reflecting a variety of media technologies, media genres and media brands or products, which jointly fulfill their everyday needs for information, diversion and sociability; and,

- thirdly research that considers *location ensembles* that adopts a "situational perspective" to study how media belong to or transcend specific socio-spatial contexts (Schröder, 2015, p. 62, citing Bjur, 2013).

Another challenge in audience research is situating media research in the debate about "mediatisation". Schröder (2015, p.62, citing Hepp, 2013) holds that we have been "stepping into the era of mediatisation, in which the role of the media across the range of social institutions and everyday life has grown in quantitative as well and qualitative terms". He identifies a plurality of "cultures of mediatisation", arguing that "the processes and 'logics' of mediatisation should not just be explored at the level of social institutions, but also in the everyday processes through which people encounter, acquire and make sense of the media in their dual appearance as technologies and multimodal discourses". For example, one such a 'logic' is *audience logic*, which can be operationalised using a notion of *worthwhileness*, described using seven factors or dimensions that determine why some news media and not others are chosen to become parts of an individual's news media repertoire (Schröder, 2015, p. 63).

The challenges of studying where and how, "the places and spaces", news media are being consumed has also become much more difficult. In the 1980's it was possible to send observers into homes in order to the map the uses - including concomitant social interactions - of various traditional media types in ethnographic or grounded theory-based research. However, today's multi-platform media world requires alternative research methods. For example various *multi-method* approaches have been attempted in which media use is observed using techniques involving a triangulating mixture of questionnaire-based surveys, focus groups and ethnological observation of participants' online practices, including methods described as *virtual shadowing* of users activities (Turcotte et al., 2015, pp. 524, citing Vittadini and Pasquali, 2014; Jensen and Sorensen, 2014).

Identifying media *repertoires*, as first outlined in the second research perspective described by Schröder (2015, p. 62, citing Bjur, 2013) above, and also mapping trends and changes in such repertoires may serve as a valuable starting point to gain a deeper understanding of the media environment on which to build further explorations of media usage that could serve marketers and other media analysts. In fact, Edgerly (2015, p. 4) encourages an approach grounded in work on media repertoires, arguing that "we can learn a lot about audiences by examining what combinations of media they choose over others".

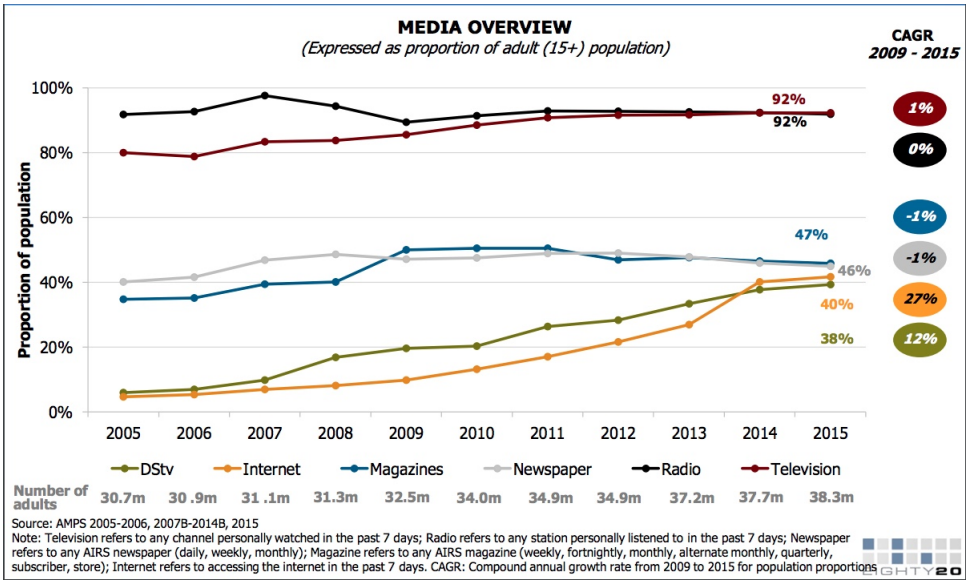
According to Edgerly (2015, p. 4), such a *repertoire* approach to media exposure was first

developed by Heeter in 1985 to describe channel-watching routines of television users. She identifies two main lines of research pertaining to repertoires: one focussing on repertoires within a single medium; another considering repertoires across media. Furthermore, to account for the ability to sample from many different types of news in repertoire research, news exposure is conceptualised as a "complex pattern of news use rather than a single media selection."

Edgerly (2015, pp. 1-2, citing Hasebrink and Popp, 2006, and others) describes a news repertoire approach to research in a cross-media environment as "identifying distinct ways that media users combine news across a wide array of media platforms and content". She holds that this approach is "less about exposure to a single news source, and more about the subset of news sources that people consume in tandem." And, "as such a repertoire approach provides a window into the decision-making strategies of audiences who are faced with increased media options". Using a national survey of the media usage of U.S. adults she identifies the existence of six distinct news repertoires, finding that while some are clearly ideologically based, spanning multiple media platforms, others have repertoires that function at a media level and others still show respondents who consume both politically conservative and liberal news. Furthermore the six repertoires show distinct audience groups in terms of media engagement and participation as well as socio-demographic profiles Edgerly (2015, p. 2). Schröder (2015, pp. 70-71), using both quantitative and qualitative methodologies on a relatively small sample in Denmark, identified seven types of news consumers, who all used a mixture of traditional and new sources of news.

This project applies a *repertoire* approach as described by Edgerly (2015, pp. 1-2, citing Hasebrink and Popp, 2006, and others), namely to considers *repertoires* in the sense of "identifying distinct ways that media users combine" media usage primarily in news and information (instead of *news* as in Edgerly (2015)*)* "across a wide array of media platforms and content"; and through this analysis also to offer a "window into the decision-making strategies of audiences who are faced with increased media options".

Figure 2.1: Media Trends



Source: Fulton (2017)

Chapter 3

Data and Data Preparation

3.1 Chapter Introduction

This chapter will describe the data and some of the transformations applied to the data.

This project made use exclusively of various All Media and Products Surveys (AMPS) that required substantial wrangling to prepare for analyses. The data used in this study were accessed through *DataFirst,* a research unit and data service based at the University of Cape Town. The service offered a dataset for 1995 and then - with the exceptions of 2006 and 2007 - for every year from 2002 until 2014 . Given the aim of the project of examining changes over time, it was important to find datasets that shared many of the same variables. The gap between 1995 and 2002 and the many changes that occurred in the survey over this period disqualified the use of the 1995 dataset. Although the intention was then to use datasets with two-year gaps starting from 2002, the absence of the more-detailed information on internet usage for the available years resulted in an unavoidably large gap between 2002 and 2008. Finally, the datasets used in this analysis were for the years: 2002; 2008; 2010; 2012; and 2014.

3.2 Overview of the All Media and Products Surveys (AMPS)

The AMPS is a single source survey that contains demographic information, media consumption, and product consumption for the same respondents (Corder, 2003). It has been conducted annually from 1975 until 2014 under the auspices of the South African Audience Research Foundation (SAARF). Corder (2003) described SAARF with its main product, AMPS, as the organisation that "provides the only common research currency for the advertising, marketing and media industries in South Africa."

In a presentation to stakeholders at the final 2014 AMPS release, Andrew Fulton of *Eighty20,* citing Barbara Cooke, described the vision of AMPS as being to provide "a single source

database (that) would be the only way to ensure a stable currency for the buying and selling of advertising in the media". He argued that this had been "far superior to each medium conducting their own research with no consistency in methodologies, sampling or definitions" Fulton (2017). Haupt 2012 argued that AMPS, the "currency survey" for marketing, needed to "contain extensive information on the characteristics of the users of mass media, their media consumption as well as information on their usage/purchasing of products, brands and services."

The SAARF website describes the AMPS sample as consisting of about 25 000 adults (15 years and older) per annum. Interviewing has been in-home and face-to-face and was generally conducted in two waves. Results were published every six months in the form of 12-month rolling data SAARF (2018). According to SAARF (1998), a "multi-stage, stratified, quasi-probability design" was employed. The universe from which the AMPS prior to 2009 was drawn, comprised adults in private households, resident in South Africa and aged 16-years or older. From 2009, the universe was expanded to include 15-year-olds, in that year this added 975 000 people to the adult population.

Throughout its history, various methods were applied to ensure proportional representation in the sample by using updated population estimates. Prior to 2012 the population estimates were based on the Bureau of Market Research (BMR) model, but since 2012 these population estimates were based on data provided by the International Health Services (IHS), which incorporated the 2011 Census results from StatsSA. (SAARF, 2014b).

Haupt (2012) described AMPS as a survey that was "constantly evolving" in response to "media proliferation and fragmentation" that has been updated every year since its inception in 1974 and that the opportunity to refine the survey over time has lead to it becoming a survey of "extremely high quality". According to Haupt (2012), in the eighties it became evident that the use of demographics alone was no longer adequate leading to the creation by SAARF of the Living Standards Measure (LSM), which Fulton (2017) describes as a "composite index comprising ownership of various appliances and access to services" and which was used as one of the six categorical variables in this project.

In 2002 SAARF launched the first Media Groups Measure (MGM) as "an examination of the reach of different media types" designed to "give insights into the build-up of media duplication not only in terms of media types, but also individual options within each medium" (Corder, 2003). Although this could be considered an attempt at a *repertoire* approach to media research in South Africa, the methods employed differed considerably with those applied in this project. Also, this researcher was not able to access the media groupings determined by SAARF and was therefore unable to consider comparisons with the results of this project.

3.3 Initial Data Preparation

In this project extensive use is made of the descriptions media *type* and media *vehicle. A* media *type* refers to a broad media category, in particular *newspapers, television, radio, magazines,* and *internet.* These media *type*s in turn consist of many different media products or, as they are described in this project, media *vehicles*, such as *The Star*, *Die Burger, Radio 5, SABC 3, internet for news,* or *internet for social.*

For every period of study (2002, 2008, 2010, 2012, 2014) datasets were created that combined available categorical demographic variables, variables reflecting engagement on different media *vehicles*, and composite variables reflecting total engagement by media *type*. Due to changes in surveys for different years, general descriptions of how these were determined are indicated in the sections below, but more details are available in various appendices identified in the summary table 3.2. As an illustration of the data, table 3.3 shows 10 cases with a small selection of the variables in the dataset.

3.3.1 Notes on Demographic Variables

The details are described in the various appendices pertaining to the datasets by year. Aggregated levels were created for *age*, *education*, *household income*, and *living standards measures* (*lsm*). Values for *attitudes* and *lifestyle* were not available until the 2008 survey and *lifestages* was not included in the 2010 survey. These variables were therefore excluded from the study. The income brackets used were standardised for those used between 2010 and 2014.

3.3.2 Engagement by Media *Vehicle*

Using available data, ordinal scales for engagement by media vehicle were created that would be used to identify media *repertoires* in chapter 5.

3.3.2.1 Print Media

For print media (*newspapers* and *magazines)*, one survey question asked how many issues of a particular newspaper or magazine the respondent had "personally read or paged through" in a given issue period. For example, for daily newspapers the issue period was five days of a week, while monthly magazines were considered over a six-week period. Respondents therefore had to select between zero and five issue for dailies and between zero and six issues for monthly magazines.

Another survey question asked respondents how *thoroughly*, on a six-point scale, a particular print product was read. Although this more detailed information would have been valuable for this project, it could unfortunately not be used since 2008 did not include this information; and the exclusion of 2008 was deemed to be unacceptable given the already large gap between 2002 and 2008.

The scales applied to newspaper and magazine were therefore based only on the number of issues in a given issue period. The more issues a respondent reads of a particular *vehicle* the more engaged they were.

Finally, a number of community newspapers were included only in the 2002 survey. Since these were not available in the other years of this study, they were excluded from this analysis.

3.3.2.2 Broadcast Media

Engagement measures for radio stations and television channels (*vehicles*) were based on recency. A survey question asked how recently (four weeks ago, seven days ago or yesterday) the respondent had *personally* listened to a particular radio station or watched a given television channel. For each of these electronic media *vehicles* a scale of 0-3 was developed to reflect the recency of engagement with that medium with 0 = *"not at all"* up to 3 = *"yesterday"*. Since many radio stations showed low coverage, the mean engagement value of all stations with less than a 5% sample response rate were relegated to an *other* variable. The assumption for this scale is that the more recently a respondent listened or watched a station or channel the more engaged they are with that media *vehicle.*

A number of community radio stations were included only in the 2008 survey and since these were not available for the other years of this study, they were excluded from this analysis.

3.3.2.3 Internet

Questions related to respondents' engagement with the internet differed considerably by period. However, all the periods under consideration offered a measure of recency, similar to the way in which electronic media was measured. A question asking how recently a respondent had accessed the internet either by mobile or computer resulted in either a four or a five-point scale. The five-point scale included an option for "in the past 12 months". To ensure consistency, only a four-point scale (from 0-3) was used throughout. Respondents in those surveys (2008, 2012, 2010, 2014) who identified that they has accessed the internet "in the past 12 months" were taken not to have accessed the internet at all.

A second level of engagement on internet was devised based on questions that aimed to gauge the purpose of a respondent accessing the internet. Given the focus of this project on media for the purposes of news, information and entertainment, the use of a computer or cellphone to access email, banking, dating or shopping was excluded. Binary data identifying whether or not a respondent accessed the internet for purposes of search, social networking, print media sites, listening to the radio, accessing news, or watching tv was extracted.

To create an engagement value for each of these purposes, the recency values (0-3) were multiplied by the binary values (0/1) of the purposes. This resulted in an engagement value based on recency for different purposes of accessing the internet.

3.3.3 Engagement by Media *Type*

Five composite engagement variables for each media *type* (newspapers, magazines, tv, radio, internet) were created by summing all the media *vehicle* engagement values. This was possible given the similarity of scales for the different media types. However, to create a composite *all* media variable for each year, these five different media *type* variables were summed after first being standardised to means = 0 and standard deviations = 1. These variables were used in the exploration of changing media *types* described in chapter 6.

3.3.4 Code for the Initial Data Preparation

The R code for this section can be accessed via the following urls:

- 2002: *https://raw.githubusercontent.com/hanspeter6/amps_2002/master/amps02dataIn.R*
- 2008: *https://raw.githubusercontent.com/hanspeter6/amps_2008/master/amps08dataIn.R*
- 2010: *https://raw.githubusercontent.com/hanspeter6/amps_2010/master/amps10dataIn.R*
- 2012: *https://raw.githubusercontent.com/hanspeter6/amps_2012/master/amps12dataIn.R*
- 2014: *https://raw.githubusercontent.com/hanspeter6/amps_2014/master/amps14dataIn.R*

3.4 Secondary Data Preparation

For the exploration of changing media *repertoires* in chapter 5, it was necessary to exclude most of the media *vehicle* variables. This was done to avoid the potential confusion of identifying factors driven by geographic or metropolitan areas; and to allow for modelling over time by ensuring the use of the same set of media *vehicles*. Therefore, only national *vehicles* that all the datasets had in common could be included. Accordingly, datasets were created that extracted only those media *vehicles*, shown by media *type* in table 3.1.

Table 3.2 illustrates summary information as well as page references for the appendices containing detailed information on the datasets for each of the periods considered in this study.

3.4.1 Code for Secondary Data Preparation

The R code used to generate the data for this section can be found at the following url:

- *https://raw.githubusercontent.com/hanspeter6/amps_nationals/master/nationals.R*

Table 3.1: National and Common Media *Vehicles* Used in *Repertoire* Analysis in Chapter 5

Type	Vehicle		Type	Vehicle
newspapers	Business Day		radio	Radio 5
	Mail and Guardian			Radio Metro
	Rapport		television	e TV
	The Sunday Independent			SABC 1
	Sunday Times			SABC 2
	Soccer Laduma			SABC 3
magazines	Drum			DSTV
	Huisgenoot		internet	int_social
	You			int_print
	Kickoff			int_search
	Bona			int_radio
	Car			int_news
	Cosmopolitan			
	Getaway			
	Sarie			
	Topcar			

Table 3.2: Summary Information on Datasets used in this Study. Details available in indicated Appendices

Year	Cases	Comments	Appendix Page
2002	29 791	Excludes *lifestyle* and *attitudes*. Full print and internet values. *FHM* excluded from time-based analyses	97
2008	21 083	Only print issues, but full internet values available. *Daily Sun*, *FHM*, *RSG*, *The Times* excluded from time-based analyses	101
2010	25 160	Excludes *lifestages.* Full print and internet values.	105
2012	25 108	Full print and internet values.	109
2014	25 584	Full print and internet values.	113

Table 3.3: Illustrating the Data: A selection of respondents, demographic categories, media *types*, (including the standardised values for *all),* and a small selection of media *vehicles*

ID	age	sex	edu	race	lsm	magazines	radio	tv	internet	all	Ilanga	Topcar	Kaya.FM	int_tv
16507	3	1	3	4	5	2	6	3	2	0	0	0	0	0
6352	4	1	1	2	4	0	3	13	0	-1	0	0	0	0
18933	1	2	1	2	5	15	2	6	6	2	0	0	0	0
3710	2	1	3	4	5	8	0	12	1	1	0	0	0	0
18596	4	1	1	2	3	0	1	12	0	-1	0	0	0	0
16862	2	1	1	4	4	0	9	11	9	5	0	0	0	0
8727	1	2	2	1	4	11	8	15	6	5	0	1	2	0
16071	1	1	1	2	4	10	3	12	9	4	0	0	0	0
24612	2	2	1	1	2	0	0	10	0	-3	0	0	0	0
3250	4	1	1	1	5	8	8	9	0	5	0	0	0	0

Chapter 4

Analytic Methods

4.1 Chapter Introduction

This chapter will provide details of the methods used in this project. In particular, it will focus on Principal Components Analysis (PCA), Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) - the main building blocks for describing Structural Equation modelling (SEM). It will also describe *k-means* clustering, bootstrapping and estimation of marginal means. Although the methods themselves are described in this section, the application of the methods can be found in sections 5.2.2 and 6.2.2.

4.1.1 Measurement

From the discussion on data preparation in chapter 3 it was indicated that the exploration of *repertoires* developed in chapter 5 as well as the composite variables used in chapter 6 would make extensive use of ordinal, scaled media *vehicle* values consisting of either 4- or 6-point scales. The use of such scales in statistical techniques such as SEM is contested.

Researchers are often interested in variables that cannot be directly observed, such as intelligence or attitudes and perceptions to a product or service - or for that matter media *repertoires* as in this study. These unobservable variables are described as *latent* variables, *factors* or *constructs* and researchers try to get information about them through observable variables, such as the response to a scaled question not unlike the media *vehicle* variables described for this project. Likert scales are examples of such measurement techniques that make use of descriptions for each scale point and can range from three to seven points (Berndt and Petzer, 2011, p. 189).

Factor analysis and SEM are techniques designed to reduce the number of observable variables into a smaller number of latent variables by considering the covariation of the observed variables (Schreiber, Nora, Stage, Barlow, and King, 2006, p. 323). The use of Likert-scale data for such analyses is controversial, with the debate hinging on two competing views: on the one hand, that Likert scales represent *ordinal* level data and hence they must be

analysed using non-parametric statistics; on the other hand, factors, comprising a number of items of Likert scales as opposed to individual Likert scales are *interval* in character and can therefore be analysed with the more powerful statistical techniques that apply to parametric data (Carifio and Perla, 2008, p. 1150).

Carifio and Perla (2008, p. 1150) refer to Monte Carlo studies of the F-test that were performed by Glass *et al.* (1972) that showed the F-test to be robust to violations of its assumptions as evidence that the F-test applied to ordinal data produces unbiased results. They also refer to various studies (Carifio, 1976; 1978) on the nature of Likert *scales*, particularly if they comprise at least eight items, that have shown the these can approximate ratio data.

In this study, the number of items (variables or *vehicles*) comprising factors are not as many as suggested by Carifio and Perla (2008, p. 1150) and must be considered a limitation of the study. On the other hand, the large sample size and the confidence expressed by Carifio and Perla (2008, p. 1150) in the use of Likert scales does provide some assurance.

4.2 Principal Components Analysis

4.2.1 Introduction

Kline (1994, p. 28) states that factor analysis without understanding is an "unmitigated evil" and argues that once the calculation of principal components has been understood, the nature of factor analysis becomes self-evident. For this reason Principal Components Analysis (PCA) will be considered in some detail.

Section 4.2.2 will begin with a description and the aims of PCA before outlining the derivation of identifying principal components (section 4.2.3). Section 4.2.4 describes some arguments for selecting and interpreting the principal components and their loadings.

4.2.2 Description and Purpose

PCA is used to identify underlying dimensions of multivariate data by describing a set of new variables, that are fewer than the original set, but yet explain most of the variance in the original sample (Grimm and Yarnold, 2004, pp. 99-100). If the original variables are nearly uncorrelated, then it would make no sense to consider PCA. According to Radloff (2015), PCA entails the finding of an orthogonal transformation of an original set of correlated variables to a new, reduced set of variables that is entirely uncorrelated. This new set of variables are called the *principal components*.

Each principal component is a linear combination of the original variables. A measure of the amount of information conveyed by each of the principal components is contained in its variance. The principal components are arranged, by construction, into descending order

such that the first component represents most of the explained variance and the last, the leas (Abdelmonem et al., 2011, p. 357).

Abdelmonem et al. (2011, pp. 357-358) offer a list of reasons for using PCA:

a) To reduce the dimensions of a problem without losing much of the information contained in the data. Here only the first few components would be selected for further analysis. This technique is attractive since the components are not inter-correlated and therefore allows for simpler analyses than working with complex interrelationships.

b) The principal components can also be used as an effective test for underlying multivariate normality. If the principal components demonstrate a normal distribution, the originating variables can be assumed to have a normal distribution.

c) Another use of PCA is to search for outliers: a histogram of each of the principal components can effectively identify specific subjects with measurements that are very large or very small.

d) As a means of overcoming multi-collinearity since it estimates orthogonal linear combinations of the orginal variables.

e) PCA can be used as an exploratory technique. Most of the examples of factor analyses in Kline (1994), use PCA as an initial exploration of the number of factors to consider in subsequent factor analyses.

In this project PCA will be used primarily to assess multi-variate normality and as an exploratory technique to determine the number of factors to use in subsequent EFA, which in turn will be used to define a measurement model for the SEM.

4.2.3 Deriving the Principal Components

In these sections the following conventions are used:

- Constants are indicated by small letters and their vector or matrix position by suscripts: a_1, b_{12}, p, k
- Vectors of random variables are indicated by capital letters with or whithout subscripts and in columns: C, X_1
- Matrices of observed variables are indicated with the use of bold capital letters: $\mathbf{X}$
- Vectors of estimated values are indicated by using greek lower-case letters: μ
- Matrices of estimated values are indicated by using greek-upper case letters: $\mathbf{\Lambda}, \Sigma$

4.2.3.1 Basic Terminology and Principles

This section draws on Abdelmonem et al. (2011, pp. 359-363).

Consider the case of n observations of just two variables, X_1, X_2.

PCA leads to the creation of two entirely new variables C_1 and C_2 , called the *principal components*, that can be written as linear combinations of X_1 and X_2, i.e.

$$\begin{aligned} C_1 &= b_{11}X_1 + b_{12}X_2 \\ C_2 &= b_{21}X_1 + b_{22}X_2 \end{aligned}$$

The coefficients, b_{ij}, are determined on the basis of three requirements or restrictions:

1. The variance of C_1 is as large as possible;
2. The values of C_1 and C_2 are uncorrelated;
3. For each component, the sum of the squares of the coefficients must equal unity in order to attain orthogonal eigenvectors, i.e. $b_{11}^2 + b_{12}^2 = b_{21}^2 + b_{22}^2 = 1$

Abdelmonem et al. (2011, p.361) demonstrate PCA graphically, showing that it amounts to rotating the original X_1 and X_2 axes to new C_1 and C_2 axes. The variances of C_1 and C_2 are the *eigenvalues* (alternatively, *characteristic roots*, *latent roots*, or *proper values*) and the sum of all the *eigenvalues* are equal to the sum of the variances from the original dataset. This relationship is retained in general. The total variance is therefore preserved when rotating the original dataset to its principal components.

The set of coefficients of the linear combinations, b_{ij}, are described as the ortho-normal eigenvectors of the variance-covariance matrix. Consider a matrix of independent random variables $\mathbf{X} = [X_1, X_2, \cdots, X_p]'$ of length p with an expected value $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, which is positive-semi definite. The aim of PCA is to identify a new matrix of variables, $\mathbf{C} = [C_1, C_2 \cdots C_p]'$ whose columns are uncorrelated, each of which is a linear combination of the original components of $\mathbf{X}$ and whose variances are arranged in decreasing order, thus

$$C_j = b_{1j}X_1 + b_{2j}X_2 + \ldots + b_{pj}X_p$$

or, in matrix notation, $C_j = \mathbf{b}_j'\mathbf{X}$, where $\mathbf{b}_j = [b_{1j}, ..., b_{pj}]'$ is a vector of "constraints", defined in such a way that $\mathbf{b}_j'\mathbf{b}_j = 1$ and $\mathbf{b}_j'\mathbf{b}_i = 0$ in order to achieve the desired orthogonal transformation.

To find the first component we want to choose $\mathbf{b}_1$ in such a way as to maximise the variance of C_1, but subject to the normalising constraint $\mathbf{b}_j'\mathbf{b}_j = 1$ and that the first $r \leq p$ eigenvalues of $Var(\mathbf{X}) = \boldsymbol{\Sigma}$ are distinct. That is, we want to maximise $Var(C_1) = Var\left(\mathbf{b}_1'\mathbf{X}\right) = \mathbf{b}_1'\boldsymbol{\Sigma}\mathbf{b}_1$.

The sum of the variances of the original variables is equal to the sum of the variables of the principal components, meaning that all the variance in the original dataset is retained

in the principal components. It also means that one can divide any of the variances of the elements of C , λ_i, by the sum of the variances of all the elements to determine the relative proportion of that variance (or the *eigenvalue)* and therefore the relative importance of a particular principal component. That is, for the i^{th} component:

$$\frac{\lambda_i}{\sum_{j=1}^{j=p}\lambda_j}$$

It is common to calculate the principal components from a set of variables that have first been standardised, i.e.,

$$Z_i = \frac{X_i - \bar{X}_i}{S_i}$$

where $\bar{X}_i$ is the mean and S_i the standard deviation of X_i. This would result in unit standard deviation and thus unit variance for Z_i (Radloff, 2015).

4.2.4 Selection of Components

A key question relates to how many components to select for subsequent analysis since one of the reasons for doing PCA is to reduce the sample space to fewer variables, making it possible for the dependent variable to be regressed using the principal components rather than the original variables (Abdelmonem et al., 2011, p. 363).

The question is which components should be used without losing too much of the information contained in the original dataset? According to Radloff (2015, p. 105), there are at least two alternative criteria to be considered, the choice of which depends on the purpose of the analysis:

- To delete those components that are relatively unimportant as predictors of the original independent variables. That is, delete the principal components with the smallest variance (or eigenvalues).
- To delete those components that are relatively unimportant as predictors of the dependent variable in the problem. That is, delete the principal components with the smallest absolute correlation coefficient with the dependent variable. In this case it's important to understand that the dependent variable need not be highly correlated with the principal components that have high eigenvalues in order for the explanatory power of the principal components regression to be high.

Abdelmonem et al. (2011, pp. 367-368), argue that one ideally wants to obtain a small number of principal components that explain a large part of the total variance, but that this ideal is rare in practice and that a compromise would be to choose as few components as possible to explain a reasonable portion of the total variance. They suggest that only those

principal components explaining at least $100/p$ percent of the total variance be selected, where p denotes the number of variables (Abdelmonem et al., 2011).

Furthermore, Abdelmonem et al. (2011, pp. 367-368) argue that it is important to realise that the eigenvalues represent *estimated* variances of the principal components and can show large sample variations and that small differences in cut-off values should not be taken too seriously. They conclude their discussion on the selection of components by stating that the selection of the number of principal components should ideally be based on underlying theory.

Grimm and Yarnold (2004, p. 103) describe different types of stopping rules to decide on the selection of components :

- The percentage of variance criterion: Specifying *a priori* that components will be included until some absolute percentage of the total variance has been explained;
- The *a priori* criterion: In cases where one is trying to replicate a previous study, one knows in advance how many eigenvectors to extract. This approach would also be appropriate when one has a theoretically motivated idea about the appropriate number of eigenvectors to extract.
- Kaiser's (1960) stopping rule: Extract only eigenvectors with eigenvalues of at least 1, that is the equivalent of the variance of a single standardised variable.
- The *scree test*: Catell (1966) proposed a graphic procedure, also described by Abdelmonem et al. (2011, p. 364) as the *elbow rule*, that consists of plotting the eigenvalues for successive components. These eigenvalues usually drop quickly after the first few before stabilising to a more gradual decline. The components in the steep decline are retained while those in the gradual decline are not (Grimm and Yarnold, 2004, pp. 103-104).

Abdelmonem et al. (2011, p. 364) describe rules that approach the subject from the opposite perspective, namely to discard principal components with the smallest variances. In this regard one such a rule is to discard all components that have a variance less than $70/p$ percent of the total variance, i.e., as opposed to the $100/p$ rule-of-thumb described above for selecting components. Another rule is to discard any components that explain only small proportions of the variance, for example less than 5%, since they may simply reflect random variations in the data.

Citing Stevens's (1986) summary of research on the accuracy of these stopping rules, Grimm and Yarnold (2004, p. 104) conclude that Kaiser's (1960) stopping rule should be used when there are fewer than 30 variables and where the communalities are at least 60%; alternatively they argue that the scree test should be used in applications for which there are at least 200 observations and where the communalities are reasonably large.

Abdelmonem et al. (2011, p. 363) argue that none of the stopping rules work well in all circumstances and that they should be used to provide guidance only and that since PCA

is often used as an initial exploratory technique, none of the so-called rules should be taken too seriously; and that one should retain as many components as one can either interpret or are useful in future analyses.

4.3 Exploratory Factor Analysis

4.3.1 Introduction

Once the number of components has been decided, how to interpret and use them become the next questions. Kline (1994, p. 39) states that since PCA produces an arbitrary general factor, the interpretation of results are difficult; and the positive and negative loadings make for difficult interpretation.. As a result, methods of simplifying PCA, such as the rotation of axes have been developed and will be described in this section.

The main purpose of Exploratory Factor Analysis (EFA) and the similarities of and differences between PCA and EFA will be discussed in section 4.3.2. This will be followed by a description of basic terminology and principals in section 4.3.3, which will be followed in turn by a brief discussions on rotation and the interpretation of factors in sections 4.3.4 and 4.3.6 respectively.

4.3.2 Description and Purpose

Johnson and Wichern (2002, p. 477) describe the essential purpose of Exploratory Factor Analysis as being to attempt to describe the covariance relationships among several variables in terms of a few underlying, but unobservable, random quantities, called *factors.* They argue that the factor model is motivated by the following argument:

> "Suppose variables can be grouped by their correlations. That is, suppose all variables within a particular group are highly correlated among themselves, but have relatively small correlations with variables in a different group. Then it is conceivable that each group of variables represents a single underlying construct, or factor, that is responsible for the observed correlations ."

Exploratory Factor Analysis (EFA) is closely related to Principal Components Analysis (PCA). As in PCA, EFA is a dimension-reduction technique that aims to identify a small set of readily interpreted eigenvectors that explain most of the variation in a given dataset. While these eigenvectors are called *components* in PCA, they are called *factors* in EFA (Grimm and Yarnold, 2004, p. 106).

Abdelmonem et al. (2011, pp. 379-380) describe PCA and EFA as being similar in that they are both techniques for examining and exploring the interrelationships among multiple variables and neither of these techniques make a distinction between dependent and independent

variables. They differ, however, in their main objectives: In PCA this is to select a number of *components* that explain as much of the variation in the data as possible, while in EFA the *factors* are chosen mainly to explain the interrelationship among the original variables.

According to Kline (1994, p. 36), the terms *components* and *factors* are often used interchangeably. Citing Harman (1976), he argues that although there is a real difference that should be understood, the difference between *components* and *factors* become trivial as sample sizes increase. Kline (1994, p. 36) defines a *factor* as a linear combination of variables and *factor loadings* as the correlations of the original variables with the *factors.*

Kline (1994, p. 36) argues that *components* are *real* factors in that they are derived directly from the data, while the common factors from EFA are *hypothetical* since they are estimated from the data and; and that while PCA explains all the variance in the data, factor analysis does not necessarily do so. According to Kline (1994), this can be an advantage since it is unlikely that *factors* can explain all the variance in a dataset and that the full account of principal components would most certainly be contaminated by error.

According to Grimm and Yarnold (2004, p. 107), the difference between EFA and PCA relate to their assumptions: While the key assumption in PCA is that the total variance in a given sample is described by the sum of the explained and the error variances; in EFA the total variance is made up of the sum of three different kinds of variances, namely the *common*, the *specific* and the *error* variances.

The *common* variance is a reference to that portion of the total variance that is shared with other variables in the analysis; the *specific* variance refers to that portion of the variance that does not correlate with other variables; and the *error* variance - as in PCA - describes random variation (Grimm and Yarnold, 2004, p. 107).

In this project, while PCA was used to determine underlying dimensions or *repertoires*, EFA was used to determine the meaning of each of these *repertoires* by considering the loadings on different media *vehicles* and how these interrelate with each other, as described by Grimm and Yarnold (2004, p. 99).

4.3.3 Basic Terminology and Principles of EFA

This section describes some of the basic terminology and principles of EFA .

Consider, as in section 4.2.3.1 above, a set of independent random variables, $\mathbf{X} = [X_1, X_2, ..., X_p]$. The aim of factor analysis is to represent each of these variables as a linear combination of a smaller set of *common* factors plus a specific factor that is unique to each variable. According

to Abdelmonem et al. (2011, pp. 381-383), this *factor model* is described as

$$\begin{aligned} X_1 &= \lambda_{11}F_1 + \lambda_{12}F_2 + ... + \lambda_{1m}F_m + e_1 \\ X_2 &= \lambda_{21}F_1 + \lambda_{22}F_2 + ... + \lambda_{2m}F_m + e_2 \\ &\vdots \\ X_p &= \lambda_{p1}F_1 + \lambda_{p2}F_2 + ... + \lambda_{pm}F_m + e_p \end{aligned}$$

where:

1. m represents the number of common factors, ideally smaller than the number of variables in the original dataset p.
2. $F_1, F_2, ..., F_m$ are the common factors. They are assumed to have zero means and unit variances.
3. λ_{ij} denote the coefficients of F_j and are called the *factor loadings* of the i^{th} variable on the j^{th} common factor.
4. $e_1, e_2, ...e_p$ are the unique or specific factors, each relating to one of the original set of variables.

The factor model in matrix notation can be described as

$$\mathbf{X} = \mathbf{\Lambda F} + \boldsymbol{\epsilon}$$

This factor model breaks each response variable X_i into two parts: one due to the common factors and another due to the unique factor. This also breaks the variance of X_i into two parts: the *communality*, or that part of the variance that can be attributed to the common factors and the *specificity*, that part of the variance that is due to the unique factor e_i . Since the data has been standardised, the sum of these two variance parts is equal to 1. The *communality* is denoted by h_i^2 and the *specificity* by u_i^2 and therefore $Var(X_i) = h_i^2 + u_i^2 = 1$.

Factor analysis is concerned with finding estimates of the factor loadings, λ_{ij}, and the communalities, h_i^2. According to Abdelmonem et al. (2011), there are many ways in which these estimates can be determined in processes that involve an initial extraction of factors, followed by rotation to generate new factors that assist in the interpretation.

In all the methods the number m of common factors is a required input, which is ideally known *apriori*. If it is not known, a default option in most computer programmes sets the number of common factors equal to the number of eigenvalues greater than 1 (Abdelmonem et al., 2011). Alternatively, as demonstrated by Kline (1994), the use of PCA as a preliminary investigative technique can be used to establish a sense of the underlying number of common factors to use in subsequent EFA. According to Abdelmonem et al. (2011), Gorsuch (1983) can be considered for a review of several other methods for choosing m numerically.

If one is confident that the factor model is valid and that the variables have a multivariate normal distribution, the maximum likelihood method (ML) should be considered. This method makes it possible to consider tests of hypotheses and determine confidence intervals ((Abdelmonem et al., 2011, p. 390), citing Gorsuch, 1983). Another advantage of ML is that the estimates of the factor loadings are invariant to changes in scale of the original variables.

The ML procedure is also what is used in Confirmatory Factor Analysis (CFA) (see section 4.4) in which the constraints on the outcome of a factor analysis are specified *a priori* (Abdelmonem et al. (2011, p. 390), citing Long, 1983; Bartholomew & Knott, 1999). According to Kline (1994, pp. 49-50), ML factor analysis differs from PCA and other methods of factor analysis in the following ways.

1. ML computes a set of factors, each of which in turn explains as much variance as possible of the population correlation matrix, as estimated by the sample correlation matrix. This, as opposed to PCA or previously described methods, which explain as much variation as possible in the sample matrix only. ML is considered a statistical method since inferences are made from a sample about a population. The consequence of this is that large samples are even more essential than in PCA or other factor analytic techniques.

2. According to Kline (1994, p. 50), when test reliabilities or *communalities* are high, the difference between ML factor analysis and PCA becomes "trivial".

3. Kline (1994, p. 50) holds that the strongest argument for using ML analysis is that it has statistical tests for the significance of each factor as it is extracted.

The ML method of deriving the loadings estimators allows for a formal test of the *goodness-of-fit* or adequacy of a given m-factor model. In most cases one does not know the number of common factors m. According to Morrison (2005, p. 328), one could begin with a low estimate and then try successively larger numbers until either the hypothesis is no longer rejected or the procedure fails to converge.

In this project the number of common factors m were established using PCA; and the ML method of estimating factors was applied to identify media *repertoires* after first considering the distribution of the data.

4.3.4 Rotation

In this section the issues of rotation in order to achieve *simple structure* will be considered since, as Grimm and Yarnold (2004, p. 105) claim, when *simple structure* has been achieved, interpretation of the factor can be relatively straightforward.

Kline (1994, pp. 56-64) provides simple algebraic examples that show that the rotation of factors change the factor loadings and therefore the meaning of factors, but that the different

factor solutions are mathematically equivalent in that they explain the same amount of variance in each variable and as a result in the matrix as a whole; and that rotated factors also reproduce the original correlations. He also demonstrates that there are infinitely many ways in which factor axes can be rotated.

The properties of *simple structure,* as outlined by Thurstone (1974), cited in Grimm and Yarnold (2004, p.105) are:

1. Each variable should have at least one loading that is near zero on at least one of the vectors; and for situations with four or more eigenvectors, most of the variables should have loadings that are near zero on most of the eigenvectors.

2. For each eigenvector there should be at least as many variables with loadings that are near zero as there are eigenvectors.

3. For every pair of eigenvectors, there should be several variables that load on only one eigenvector. In general, variables should have a large loading on one eigenvector.

Grimm and Yarnold (2004, p. 105) describe different types of rotations, broadly distinguished by whether they are *orthogonal* or *oblique*.

The most commonly used orthogonal rotations include *varimax*, which aims to make as many values in each column of the factor loading coefficient table as close to zero as possible and *quartimax*, which aims to make as many values in each row of the table as close to zero as possible. Both try to achieve as much simple structure as possible without losing the independence between the eigenvectors.

In *oblique* rotations, independence is forfeited, leading to factors that may be correlated. Once simple structure has been achieved, the rotated coefficients need to be considered carefully to find the *central dimension* identified by the eigenvector Grimm and Yarnold (2004, p. 105). In this project, given the inherent overlap of media consumption, *oblique* rotations were used.

For oblique rotations, the *promax* rotation is fast and conceptually simple. It tries to fit a target matrix which has a simple structure and requires two steps. The first, mostly using a varimax rotation, defines the target matrix by forcing the structure of the loadings to become bipolar. The second step involves a least-square fit from the varimax solution to the target matrix. The results of oblique rotations are generally interpreted not by graphical means, but by considering the correlations between the rotated axis and the original variables, which are interpreted as loadings (Abdi, 2003, p.6).

4.3.5 Factor Scores

Following the extraction and rotation of factors, it would be necessary to obtain aggregate scores for each case/respondent in a dataset with the factors to allow for further analyses of the data. In this project, these factor (or *repertoire*) scores were aggregated for different

demographic categories and presented visually for interpretation. The scores can be obtained in various ways. The simplest is to add the values of the variables loading heavily on a given factor, alternatively a *regression* procedure can be used to determine factor scores according to which the inter-correlations among the X_i variables are combined with the factor loadings to produce *factor score coefficients.* These factor scores can then be used as data for further analyses (Abdelmonem et al., 2011, p. 396).

4.3.6 Interpretation

One of questions raised in interpretation is what value of factor loading (coefficient) is required for a variable to be considered as a constituent of a given eigenvector? Grimm and Yarnold (2004, p. 106) hold that researchers typically consider variables with factor loadings of at least the absolute value of 0.3, that is the variable and the eigenvector share $(0.3)^2$ or 9% of their variance. However the practice of only considering factor loadings of greater than 0.3 ignores the sample size, which should be considered since the statistical significance of the correlation between a variable and an eigenvector depends on the sample size.

Variables with negative factor loading coefficients are negatively correlated with the eigenvector and eigenvectors that have positive and negative loadings are called *bipolar* eigenvectors Grimm and Yarnold (2004, p. 106).

When items are virtual paraphrases of each other, that is for example when a statement is simply rephrased, they will load as a single factor described as *bloated specifics* that look like factors but are really only specific variance. It is possible to discriminate between specific and common factors by the fact that the specific factors correlate with no other factors or external material (Kline, 1994, pp. 12, 112, citing Cattell (1978)).

Finally, when a component correlates with only one variable, this may indicate that the variable should be used as is and that it is not part of the latent variable structure of the data.

4.4 Structural Equation modelling and Confirmatory Factor Analysis

4.4.1 Introduction

Many statistical techniques, including multiple regression, multivariate analysis of variance and factor analysis only examine a single relationship between dependent variables and independent variables. Structural Equation modelling (SEM) on the other hand examines a series of dependent relationships simultaneously. It can be useful when one dependent variable becomes an independent variable in other relationships (Hair et al., 1998, p. 576-578).

Hair et al. (1998, p. 578) hold that SEM is an attractive technique in many fields of study since it offers methods for dealing with multiple relationships; and it provides a transition from exploratory to confirmatory analysis, while providing statistical efficiency at the same time. SEM encompasses a family of models that include covariance structure analysis, latent variable analysis and confirmatory factor analysis that have been developed as integral tools in many fields of applied and basic research. All SEM techniques are distinguished by two characteristics: a) the estimation of multiple interrelated dependence relationships; and b) the ability to represent unobserved, or latent, concepts in these relationships, while accounting for measurement error in the estimations.

Confirmatory Factor Analysis (CFA) forms the foundation from which the structural model is formulated (Hair et al., 2006, p. 18). According to Brown and Moore (2016, pp. 5-6), CFA should be used as a precursor to structural equation models. SEM consist of two major components: the CFA component or the *measurement model* specifying the number of factors, how the various indicators are related to the factors, and the relationships among the indicator errors; and a *structural model* which specifies how the various factors are related to one another. When poor model fit is encountered in SEM it is more likely that this will be due to mis-specification in the *measurement* portion of the model than from the *structural* component since there are usually more things that can go wrong in the measurement model than in the structural model, such as problems in the selection of observed measures, mis-specified factor loadings, and additional sources of covariation among observed measures that cannot be accounted for by the specified factors (Brown and Moore, 2016, pp. 5-6).

The details of how SEM is used in this project is described more fully in

Section 4.4.2 will first provide a description of CFA and its use before offering a brief introduction into SEM itself in section 4.4.3.

4.4.2 Confirmatory Factor Analysis

This section will begin with a description and purpose of CFA before considering aspects related to the conducting of a CFA in section 4.4.2.2, which is followed by notes on the interpretation of CFA in section 4.4.2.4.

4.4.2.1 Description and Purpose

CFA (Confirmatory Factor Analysis) is a type of structural equation that deals specifically with measurement models; that is, the relationships between observed measures and latent variables or *factors*. As a confirmatory technique it is driven by theory in which the planning of the analysis is driven by the theoretical relationships among the observed and unobserved variables (Schreiber et al., 2006, p. 323).

The difference between CFA and EFA is that while EFA aims to find the single underlying factor model that best fits the data, CFA can test more precise hypotheses. One can, for

example, specify which items belong to which factors and how the factors relate to each other and it can be used to identify the model that offers the best fit (Bryant and Yarnold, 2004, pp. 12, 109). According to Bryant and Yarnold (2004, p. 109, citing Bollen, 1989, Hayduk, 1987, Long, 1983), EFA is primarily concerned with theory building, whereas CFA is used more for theory testing.

Brown and Moore (2016, p. 2) argue that the two differ fundamentally by the number and nature of *a priori* specifications and restrictions made on the latent variable measurement model. EFA is used as an exploratory or descriptive technique to ascertain the number of common factors and to determine which measured variables are reasonable indicators of the various latent dimensions; while in CFA the number of factors, the pattern of indicator-factor loadings as well parameters such as those related to the independence or covariance of the factors and indicator unique variances are specified. This pre-specified factor solution is evaluated on the basis of how well it reproduces the sample covariance matrix of the measured variables (Brown and Moore, 2016).

In practice CFA and EFA are often used together. For example, according to Brown and Moore (2016, p. 3), EFA is often used early in the process of scale development and construct validation while CFA is used later when the underlying structure has been established on prior empirical and theoretical grounds. The two techniques can be viewed as opposite, yet complimentary, sides of a coin. Or, as described by Bryant and Yarnold (2004, p. 109): "whereas EFA involves hindsight, CFA requires foresight".

CFA can also be used effectively in multi-sample hypothesis testing, for example by testing whether the same factor structure holds across multiple groups, known as *simultaneous* CFA, which allows one to test hypotheses such as the invariance of factor loadings and unique error terms for a given model across independent samples (Bryant and Yarnold, 2004, p. 112, citing Alwin and Jackson, 1979, and others).

Unlike EFA, which extracts factors from the data to maximise the common variance, CFA uses whatever model is specified to generate a predicted set of item interrelationships (Bryant and Yarnold, 2004, pp. 110-111) or, as (Schreiber et al., 2006, p. 323) describe, an estimate of a population covariance matrix that is compared with the observed sample covariance matrix, with the researcher aiming to minimise the difference between the observed and estimated matrices.

According to Brown and Moore (2016, p. 6), all CFA models contain the parameters of factor loadings, unique variances, and factor variances; where factor loadings reflect the regression slopes for predicting the indicators from the factors, unique variance is the variance in the indicator that is not accounted for by the factors, and the factor variance which is the dispersion of the sample responses on a particular latent variable or factor Brown and Moore (2016, p. 6).

4.4.2.2 Conducting CFA

A CFA requires an input matrix of either correlations or covariances of the original data as well as information about factor loadings, factor interrelationships, and measurement errors Bryant and Yarnold (2004, p.115).

With regard to the input matrix, standardisation of the raw data to produce correlation matrices is important when the observations reflect different units or scales of measurement or when data from separate groups are combined. If however the groups have significant differences in their means, simply combining them can result in "spurious correlations". It is generally better not to standardise data within groups when one is interested in exploring structural differences between groups since group differences in variability contained in covariances can be obscured by using correlation matrices (Bryant and Yarnold, 2004, p. 115, citing Cunningham, 1978 and Joreskog and Sorbom, 1989).

The number of hypothesised latent factors and the pattern of item loadings that define each factor need to be stipulated. In deciding which factor loadings to include in a CFA model one should aim to develop parsimonious models in which the items load on as few factors as are necessary to provide a reasonable fit of the data (Bryant and Yarnold, 2004, p. 115).

The nature of the relationship among the latent factors also need to be specified. These can either be independent (orthogonal), inter-correlated (oblique), or a combination of the two in which some are orthogonal and others oblique. By considering the chi-square values from models with orthogonal vs oblique factors, one can test for the hypothesis that the factors are interrelated (Bryant and Yarnold, 2004, pp. 115-116). The latent variables can also be defined as either *endogenous* or *exogenous* variables.

Brown and Moore (2016, p. 6) argue that when the CFA solution has two or more factors, a factor covariance is generally specified to estimate the relationship between latent dimensions.

Unlike EFA, the nature of the relationships among the indicator unique variances can be modeled in CFA. When measurement error is specified to be random, (i.e. the unique variances are uncorrelated), the assumption is that the observed relationship between any two indicators loading on the same factor is due entirely to the shared influence of the latent variable (Brown and Moore, 2016, p. 4). According to Bryant and Yarnold (2004, p. 116), researchers often allow for correlated measurement error in their CFA models, especially when these improve the fit of models that are already grounded in theory.

In a CFA model parameters can be specified in three different ways: *free*, *fixed* or *constrained*. A *free* parameter is unknown and the analysis strives to find its optimal value that minimises the difference between the observed and predicted variance-covariance matrix, a *fixed* parameter is specified *a priori,* most often either 1 or 0, a *constrained* parameter is unknown, but while its not free to be any value, the values it can assume are restricted. The most common example of a constrained parameter is when it is restricted to be equal to another parameter (Brown and Moore, 2016, p. 9).

The output of CFA can give parameter estimates as *completely standardised*, when both

the latent variable and indicator are standardised and as a result the factor loading of an indicator that loads on only factor can be interpreted as the correlation between the indicator and the factor. Results can also be *unstandardised* reflecting the original input metrics or *partially standardised,* where either the indicator or latent variables are standardised (Brown and Moore, 2016, p. 10).

To be able to estimate a CFA solution, the measurement model must be *identified*, which can occur if on the basis of the sample variance-covariance matrix a unique set of estimates for each parameter in the model can be obtained. Two issues that can inhibit identification are the *scaling* of the latent variables and *statistical identification.* To understand scaling its important to realise that the units of measure of the latent variables need to be identified by the researcher. This is done in one of two ways: the most widely used method is the *marker indicator* approach according to which the unstandardised factor loading of one observed measure per factor is fixed to a value of 1; alternatively, the variance of the latent variable is fixed to a value of 1 (Brown and Moore, 2016, p. 10). Bryant and Yarnold (2004, pp.117-118, citing Alwin and Jackson, 1979, 1980), suggests fixing the measurement scale for each latent factor by constraining either the factor variances to one, providing a standardised solution, or by fixing one loading on each factor (usually that of the highest loading indicator) to one .

Statistical identification relates to the fact that a CFA solution can only be estimated if the number of freely estimated parameters (i.e., factor loadings, uniquenesses, factor correlations) does not exceed the information in the input matrix (e.g., the number of sample variances and covariances). In this regard, a model is *over-identified* when the number of known elements exceeds the number of unknowns and it is *under-identified* when the opposite is the case. The difference in the number of known and unknowns denote the model's *degrees of freedom* (df). An over-identified solution (i.e., when the df is positive), can result in an output and accompanying goodness-of-fit evaluation, but an under-identified solution (when the df is negative) cannot be estimated Brown and Moore (2016, p. 11).

To overcome a problem of under-identification, one can specify or fix the values of some parameters to reduce the number of unknowns. In particular, for a model to be identified a minimum of k^2 elements must be fixed in which k is the number of latent factors in the model. On the other hand if a model is over-identified, meaning multiple estimates can be derived for free parameters one can increase the number of unknowns by freeing more parameters Bryant and Yarnold (2004, pp.117-118, citing Joreskog and Sorbom, 1989).

A CFA model can be properly identified, but still be *mis-specified,* which relates to whether a model is different from the structures that the observed data would support. Model mis-specification can result in solutions in which unique errors are negative, factor inter-correlations are more than one, or parameter estimates are very large. While trivial mis-specification, for example erroneously including a factor loading when the correct model has none, has little effect on goodness-of-fit indexes, more substantive mis-specification, for

example leaving out an important factor loading, can dramatically lower the value of the goodness-of-fit indices (Bryant and Yarnold, 2004, p. 118, citing La Du and Tanaka, 1989 and Bagozzi and Yi, 1988).

Finally, an important difference between EFA and CFA is that in EFA all possible relationships (factor loadings) between the factors and indicators are freely estimated, allowing for *cross-loadings* where an indicator is predicted by more than one factor. In a CFA all cross-loadings are fixed to zero. As a result, while EFA models with two or more factors are subjected to rotations in order to achieve *simple structure* and therefore more accessible interpretations, in CFA, given the absence of cross-loadings, rotation does not apply Brown and Moore (2016, p. 12).

4.4.2.3 Estimation

The estimation in CFA (and, in general in SEM) involves a *fitting function* that operates to minimise the difference between the sample and the hypothesised variance-covariance matrices Brown and Moore (2016, p. 14). According to (Bryant and Yarnold, 2004, p. 116), CFA uses mainly three methods of estimation: Maximum Likelihood (ML), Generalised Least Squares (GLS) method, and the Unweighted Least Squares (ULS) method. The ULS method has the advantage of being invariate to the scale of the variables; while the ML and GLS methods allow for testing of model fit with an overall chi-square statistic (Bryant and Yarnold, 2004, p. 116).

The most commonly used estimation method, and also the default option in most factor analysis software, for both CFA and SEM is the maximum likelihood method (ML), which aims to maximise the likelihood of the parameters given the observed data. ML methods estimate the parameters in a CFA model by an iterative process that begins with starting or initial values, which are either specified or generated by the software, and then refined until stable convergence is attained between the sample variance-covariance matrix and that of the model-implied matrix (Brown and Moore, 2016, p. 15).

ML in these cases require large sample sizes, indicators that have been measured on continuous scales (i.e., approximate interval-level data), and multivariate normal distributions of the indicators as well as linear combinations of the indicators (Brown and Moore, 2016, p. 15). Violations of normal distribution can distort goodness-of-fit indexes by resulting biased standard errors that can invalidate the conclusions from the statistical tests. If non-normality is extreme (as can happen with the use of Likert items when a majority of respondents select the lowest choice such as 1 out of 5), the ML method will produce incorrect parameter estimates (Brown and Moore, 2016; Bryant and Yarnold, 2004).

Citing Bentler, 1995, Brown and Moore (2016, p. 15) advise that in the case of non-normal but continuous indicators, it is better to use ML that has been made robust to violations of normality. Such an estimation would provide the same parameter estimates as the standard ML, but the goodness-of-fit statistics such as the overall maximum likelihood χ^2 and the

standard errors would be corrected for non-normality. If one or more of the observed indicators are categorical or if non-normality is extreme, normal theory ML should be used where estimators such as a weighted and unweighted least squares methods should be considered (Bryant and Yarnold, 2004, p. 116, citing Muthen, 1993) and Brown and Moore (2016, p. 15).

4.4.2.4 Interpretation

Various output values are used in the interpretation of CFA, these include the standardised root mean square residuals, parameter estimates (factor loadings, error variances and covariances, and factor variances and covariances) as well as various measures and indexes of overall goodness of fit. For each estimated parameter in the model, CFA provides the likelihood that the given parameter is different from zero. This can be used to decide which observed indicators can be eliminated from the model while retaining model reliability (Bryant and Yarnold, 2004, p. 111).

According to Brown and Moore (2016, p. 16) three main aspects of the results should be examined in considering the CFA model: a) overall goodness-of-fit; b) specific points of poor fit; and c) the interpretability, size and statistical significance of the parameter estimates.

4.4.2.5 Overall Goodness-Of-Fit

Determining goodness-of-fit involves considering how well the model parameter estimates (factor loadings, factor correlations, error covariance) are able to reproduce the observed relationships.

> "For example if the standardised loadings of a factor on two variables X_1 and X_2 were $\lambda_1 = 0.760$ and $\lambda_2 = 0.688$ and the standardised factor variance was given by $\phi_1 = 1$, the model-implied correlation of these indicators would be the product of their factor loading estimates and the factor variance. That is $Cov(X_1, X_2) = \lambda_1\phi_1\lambda_2 = (0.760)(1)(0.688) = 0.523$. Assuming the sample correlation of X_1 and X_2 was 0.516, the model-implied estimate differs by only 0.007 standardised units" Brown and Moore (2016, p. 17).

CFA determines an overall maximum likelihood χ^2 and an associated p-value, which describes the probability that the matrix of fitted residuals generated by the model is different from zero (Bryant and Yarnold, 2004, pp. 111-112).

In general factor, CFA require large samples, since small samples can inflate the model's apparent goodness-of-fit Bryant and Yarnold (2004, p. 117). Citing Alwin & Jackson (1980), Bryant and Yarnold (2004, p. 113) warn that due the sensitivity of the χ^2 statistic to sample size, it may not be that useful to consider overall goodness of fit when large samples are used since "even reasonable models are likely to produce significant χ^2 values" and that the

difference in χ^2 may be more informative. Citing Jöreskog (1978, p448), Bryant and Yarnold (2004, p. 113) give the following advice: If the χ^2 statistic is large compared to the degrees of freedom, the fit can be explored by considering the residuals, which may suggest an option of introducing more parameters, which could lead to a smaller χ^2 value. If however the drop is large compared to the difference in the degrees of freedom, the change may indicate an improved model. Should the drop in the χ^2 value be close to the difference in the numbers of degrees of freedom, the apparent improvement could be considered as "capitalising on chance", implying the added parameters may not be significant. A useful heuristic for comparing relative fit is to use a ratio of χ^2 divided by degrees of freedom and decreases approaching zero imply improvement of the fit.

In searching for an appropriate CFA model, one typically considers a range of alternative models; from very restrictive or *null* models with no latent factors and therefore no factor loadings and no factor variances or covariances, to unrestricted or *fully saturated* models in which all factor loadings and factor interrelationships are free to be estimated (Bryant and Yarnold, 2004, p. 115). When competing models are *nested*, i.e., when the model that is more restrictive can be obtained by imposing constraints on a more general, less restricted model, their chi-square test statistics can be directly compared by simply subtracting the chi-squared statistic for the more general model from that of the more restricted one, giving $\Delta\chi^2$. Similarly the difference in the degrees of freedom, Δdf, can be determined. Using the difference in χ^2 and the difference in the df in an ordinary chi-squared test for significance the models can be compared for best fit, with the model with the smaller chi-squared statistic considered a better fit (Bryant and Yarnold, 2004, pp. 110-112, citing Bentler and Bonner, 1980, and others).

Although χ^2 is effectively used in nested model comparisons it is seldom used as the only measure of fit since, as already described, it is very sensitive to sample size. It is however utilised in the calculation of various other indexes, such as the standardised (or unstandardised) root mean square residual (SRMSR), root mean square-error of approximation (RMSEA), the Tucker-Lewis index (TLI) and the comparative fit index (CFI). Each of these should be considered since they provide different information about model fit. Considered together they provide a more conservative and reliable evaluation of goodness-of-fit (Brown and Moore, 2016, pp. 17-18).

Citing Hu and Bentler (1999), Brown and Moore (2016, p. 18) suggest the following guidelines to acceptable model fit, although, citing Marsh, Hau & Wen (2004), they warn that some researchers consider these guidelines to be too conservative:

- SRMR values close to or below 0.08;
- RMSEA values close to or less than 0.06 ;
- CFI and TLI values close to or greater than 0.95

Kenny (2015) utilising guidelines from Hu and Bentler (1999) suggests a value of 0.8 for the Incremental Fit Index (IFI) values as measure of "good" fit. This measure can be compared with an R^2 as a measure of proportion variance explained,

Citing MacCallum, Browne and Sugawara (1996), Kenny (2015) posits that Root Mean Square Error of Approximation (RMSEA) values of 0.01, 0.05 and 0.08 denote, respectively, *excellent*, *good*, and *mediocre* fits, with 0.1 described as a "cutoff for poor fitting models".

The Standardised Root Mean Square Residual (SRMSR) is an absolute measure of fit defined as the "standardised difference between the observed correlation and the predicted correlation". Kenny (2015), citing Hu and Bentler (1999), suggest that while a value of zero indicates perfect fit, values less than 0.08 are generally considered a good fit.

4.4.2.6 Specific Poor Fit

The goodness-of-fit statistics and indices can only provide overall indications of model fit. Sometimes, especially in more complex models, while overall indicators may show acceptable model fit, some relationships are less than acceptable; or, alternatively overall goodness-of-fit is rejected, but the reasons for this rejection are not clear. Two statistics that are used to identify specific areas of misfit are *standardised residuals* and *modification indices (Brown and Moore, 2016, p. 19).*

Standardising the residuals in order to consider them independent of their units of measurement and examining their distribution is one way of judging how well a CFA model fits the data. The difference between each of the predicted interrelationships and the actual observed interrelationships is referred to as a *fitted residual*. The goodness-of-fit of a particular model can be assessed by examining the overall size of the fitted residuals and the proximity of these residuals to zero (Bryant and Yarnold, 2004, pp. 110-111).

A standardised residual is analogous to standard scores in a sampling distribution and can be interpreted as z-scores. For example a standardised residual of an absolute value of 1.96 or more would indicate significant additional covariance between indicators that was not reproduced by the model's estimates.

The *modification index* (MI) represents the predicted decrease in the χ^2 value if a given parameter no longer constrained the model. A given parameter's MI can be evaluated by for example comparing it with 3.84, the critical value of a χ^2 distribution with one degree of freedom at $p < 0.05$. An MI value greater than this value would suggest overall fit can be significantly improved if the fixed or constrained parameter was freely estimated. However, since MI's are sensitive to sample size, such revisions should only be considered if they can be justified on empirical or conceptual grounds. Revising a model purely on the basis of large standardised residuals or MIs can result in further mis-specification and over-fitting (Bryant and Yarnold, 2004; Brown and Moore, 2016, pp. 111-114, citing Jöreskog and Sörbom (1989)).

4.4.2.7 Interpretability, Strength and Statistical Significance of Parameter Estimates

The parameter estimates such as factor loadings and factor correlations should only be interpreted in the context of a fitted model solution since the parameter estimates of a poorly fitting model are likely to be biased. Assuming a good fit, the parameter estimates should first be evaluated to confirm that they make statistical as well as substantive sense. They should for example not take on values such as negative indicator variances, which could be indicative of model mis-specification, and they should be of a size and direction that is in line with the conceptual or empirical aspects of the model. For example very small estimates may indicate unnecessary parameters, while very large estimates may question the existence of distinct constructs (Brown and Moore, 2016, p. 22).

4.4.3 Introducing Structural Equation modelling

Gefen et al. (2000, p. 3, citing Gerbing and Anderson, 1988) describe SEM as an example of a second generation data analysis technique that can model the relationships among multiple independent and dependent constructs simultaneously; in contrast to what they describe as first generation tools such as regression analysis. Although the focus of structural modelling is on estimating relationships among hypothesised latent constructs, it is also used to test experimental data. SEM allows for the testing of theoretical propositions about how constructs are theoretically linked as well as the directions and significance of these relationships (Schreiber et al., 2006, p. 326).

Schreiber et al. (2006, p. 325), cite Ullman (2001) in describing SEM as a combination of EFA and multiple regression, but they argue that SEM more accurately combines CFA and multiple regression since SEM is more confirmatory, although they concede that it can also be used for exploratory purposes.

SEM extends the possibility of relationships among latent variables and, as was noted already, can be considered as consisting of two components: a *measurement* model, and a *structural* model.

The measurement model, which reflects the CFA part of SEM, relates to the pattern of observed variables for the latent constructs in the hypothesised model. The measurement model (as with CFA) tests the reliability of the observed variables and is used to analyse the interrelationships and covariation among the latent constructs. As part of this process, factor loadings, unique variances and, should the model be changed, modification indexes are estimated in order to derive the best indicators of latent variables before testing for a structural model (Schreiber et al., 2006, p. 325).

The structural model, or the multiple regression part of SEM, considers the interrelationships among latent constructs and observable variables in the hypothesised model in the form of a succession of structural equations, similar to performing several regression analyses Schreiber

et al. (2006, p. 325). Structural equation models are "sets of linear equations used to specify phenomena in terms of their presumed cause-and-effect variables. In their most general form, the models allow for variables that cannot be measured directly" (Johnson and Wichern, 2002, p. 524).

Schreiber et al. (2006, p. 325) however warn against using terms such as *causal modelling* when referring to the relationships that are considered in SEM. They argue for using descriptors such as *direct*, *indirect,* and *total effects* among latent constructs, where a *direct* effect represents the effect of an independent (or exogenous) variable on a dependent (endogenous) variable; an *indirect* effect describes the effect of an independent variable on a dependent variable through another mediating variable; and a *total* effect on a dependent variable is the sum of the *direct* and *indirect* effects.

The use of SEM is predicated on a strong theoretical model according to which latent constructs are defined and as a result the CFA model forms the foundation from which a structural model is defined. Once the measurement model is well established, in order to move to a structural model two fundamental decisions need to be made: a) to distinguish between exogenous and endogenous constructs; and b) to specify the relationships between constructs (Hair et al., 2006).

4.4.3.1 The Structural Equation Model

Johnson and Wichern (2002, p. 525, citing Jöreskog and Sörbom (1996)) describe the LISREL (Linear Structural Relationships) that make up the structural equation model as:

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \tag{4.1}$$

$$\mathbf{Y} = \boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\epsilon} \tag{4.2}$$

$$\mathbf{X} = \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta} \tag{4.3}$$

with

$$\begin{aligned} E(\boldsymbol{\zeta}) &= 0;\ Cov(\boldsymbol{\zeta}) = \boldsymbol{\Psi} \\ E(\boldsymbol{\epsilon}) &= 0;\ Cov(\boldsymbol{\epsilon}) = \boldsymbol{\Theta}_{\boldsymbol{\epsilon}} \\ E(\boldsymbol{\delta}) &= 0;\ Cov(\boldsymbol{\delta}) = \boldsymbol{\Theta}_{\delta} \end{aligned}$$

The quantities $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ in equation 4.1 are respectively the cause-and-effect variables normally representing the unobserved latent variables. The quantities $\mathbf{Y}$ and $\mathbf{X}$ in equations 4.2 and 4.3 are measured variables that are linearly related to the latent variables represented by $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ through coefficient matrices $\boldsymbol{\Lambda}_y$ and $\boldsymbol{\Lambda}_x$. Their observed values constitute the data in a SEM and are sometimes referred to as the measurement equations (or model). The random or unique error terms $\boldsymbol{\zeta}$,$\boldsymbol{\epsilon}$,$\boldsymbol{\delta}$ in equations 4.1, 4.2, and 4.3 are mutually uncorrelated.

4.4.3.2 Path Diagrams

Johnson and Wichern (2002, p. 526) describe path diagrams as useful aids for formulating structural models since they indicate both the direction and the nature of causality and force a researcher to consider the relationships in a problem.

In SEM a distinction is made between *exogenous* variables that are not influenced by other variables in the system and *endogenous* variables that are affected by others. Each of the dependent *endogenous* variables is associated with a residual. The following conventions adapted from Johnson and Wichern (2002, p. 526) and Schreiber et al. (2006) determine the meanings of a path diagram.

- Observed variables are traditionally depicted as squares or rectangles
- Unobserved variables are depicted as circles or ovals
- Directed arrows represent a path
- A straight arrow identifies each dependent (endogenous) variable from each of its sources. For example a straight-line arrow from a unobserved latent variable to a measured item implies causality.
- A straight arrow is also drawn to each dependent variable from its residual
- A curved, double-headed arrow is drawn between each pair of independent (exogenous) variables thought to have nonzero correlation.

The path diagram in figure 4.1 below refers: Variables ξ_1, ξ_2, ξ_3 represent the independent (exogenous) variables with ϕ_1, ϕ_2, ϕ_3 denoting the correlations between them. Variables η_1 and η_2 denote the dependent (endogenous) variables with γ_1 for example representing the coefficient between the independent (exogenous) variable ξ_1 and the dependent (endogenous) variable η_1 which also depends on ξ_2 quantified by the coefficient γ_2. The random effect on or residual of η_1 is represented by ζ_1.

The path diagram relates to the structural (LISREL model) equations that are described in section 4.4.3 .

4.5 K-means clustering

Cluster analysis are techniques for grouping individuals into *unknown* groups and different methods can lead to different groupings (Abdelmonem et al., 2011). The clustering technique found to be most useful for this project was *k-means* clustering, which can be used to partition a dataset into k distinct and non-overlapping clusters in such a way that the within-cluster variation, summed up over all k clusters is as small as possible (James et al., 2013) .

James et al. (2013) describe the *k-means* algorithm that require the number of clusters (k) to be specified at the start:

1. Randomly assign a number, from $1 - k$ to each of the observations in the set. These serve as an initial cluster assignment for the observations.

2. Iterate until the cluster assignments stop changing:

   (a) For each of the k clusters, compute the cluster *centroid*. The k^{th} cluster *centroid* is the vector of the p features means for the observations in the k^{th} cluster.

   (b) Assign each observation to the cluster whose *centroid* is closest (where *closest* is defined using Euclidean distance)

According to James et al. (2013), it is important to run the algorithm several times with different starting values since the algorithm tends to find a local rather than a global optimum.

4.6 Estimated Marginal Means

Given the size of the model and the large number of estimated parameters for the categorical predictors, interpreting these coefficients directly would be difficult and, as Lenth (2018c) demonstrates, the calculation of marginal means using unbalanced research designs can be misleading.

Estimated Marginal Means (EMM) can be used to either ensure better estimates in unbalanced research designs; or to provide more useful aggregations to interpret model parameters (Ko, 2017). Since unbalanced data was intentional in the AMPS survey designs, EMM will primarily be used as a tool for interpretation in this project.

EMM is also referred to as *least squares means, predicted means*, or *expected means* and are not calculated directly from the data, but determined using a model, such as a linear regression model (Ko, 2017). An EMM is based on a reference grid that consists of all combinations of factor levels. Predictions of the outcome variable are aggregated over such a reference grid (Lenth, 2018c).

An important question relates to how the model parameters should be weighted in such procedures. Although there are other methods, for this project only *equally* or *proportionally* weighted options will be considered and compared (Ko, 2017). For *equally* weighted estimations the calculations assume an underlying balanced design. Therefore, for example for a five-level categorical variable, the model coefficients would be multiplied by the equal level proportions of 0.2, while for a two-level variable, they would be multiplied by 0.5. For *proportionally* weighted estimations the model coefficients would be multiplied by the proportion of cases in that categorical level (Ko, 2017).

In this project, the only interaction effect that will be considered in the modelling is the interaction of categorical variables on years and, since the EMM will be determined for particular levels *by year*, the interaction effects as described in Lenth (2018b) do not apply.

Finally, although where possible both *proportional* and *equally* weighted means are indicated and compared, where a choice needed to be made, the *proportional* values were used, since proportional representation is a central component of the sampling methodology (see section 3.2).

4.7 The Bootstrap

Santana (1916) define the *bootstrap* as:

> "a technique that can estimate population parameter and distributional properties of statistics by substituting the population mechanism used to obtain the parameter with an empirical equivalent. These estimates can be obtained analytically, but they are mostly obtained through the use of resampling and Monte-Carlo methods carried out on a computer."

The *bootstrap* is an "extremely powerful statistical tool" that can be used to quantify the uncertainty associated with a given estimator (James et al., 2013, p. 187). A simple example would be to use it to estimate the standard errors of the coefficients from a linear regression fit. Although R's *lavaan* package does provide estimates of the standard errors of the model coefficients, a problem arises when those coefficients are in turn used to estimate the marginal means (EMMs). This would be an example of utilising the "power of the *bootstrap*" to respond to the requirement of determining "measures of variability that are otherwise difficult to obtain or not automatically output by statistical software", as identified by (James et al., 2013, p. 187).

This project makes us of the *Basic* or *Backwards Percentile Method* of estimating parameter confidence intervals. Santana (1916, p. 102) cautions that this procedure can only approximate the interval for the statistic estimate and not for the population parameter and that the resulting Monte-Carlo approximate interval is in fact an *estimator* for the true interval of the population parameter.

The algorithm used to determine these *Basic Percentile Method* confidence intervals in this project was obtained from Santana (1916, p. 103) and started with drawing 1 000 random samples of the full dataset with replacement. These 1 000 samples were then used to estimate the same SEM model 1 000 times, generating 1 000 sets of model coefficients. These model coefficients were then used to calculate 1 000 sets of EMMs. These were then sorted and the 2.5th and 97.5th percentiles used to mark the upper and lower confidence intervals of the EMMs.

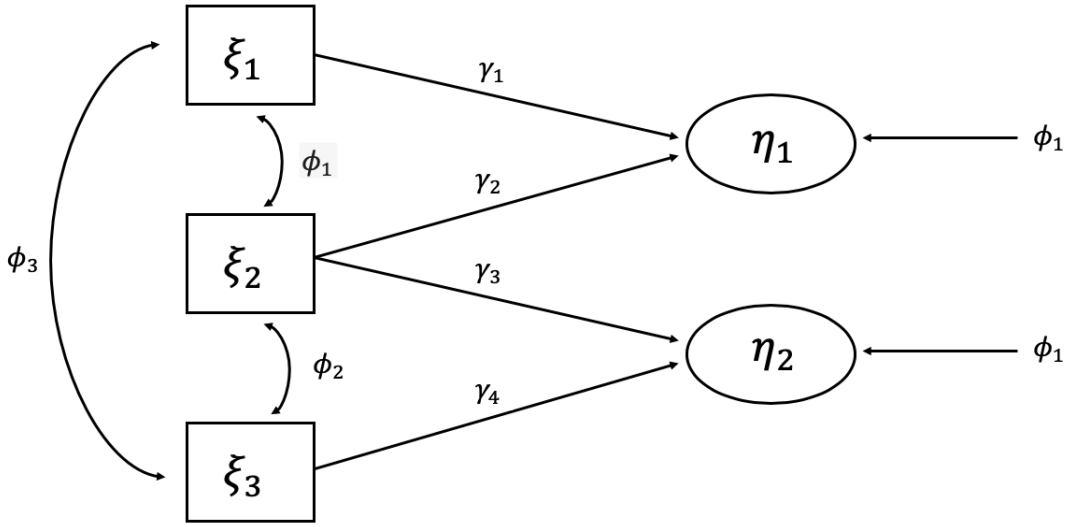


Figure 4.1: Path Diagram (Johnson and Wichern, 2002, p. 526)

Chapter 5

Exploring Media Repertoires

5.1 Chapter Introduction

The aims of this chapter are first to identify media *repertoires* before exploring changes by demographic categories and levels over the 2002-2014 period of this study. Here media *repertoires* refer to possible latent or unobservable *factors* that extend over the media *vehicles* feature space.

After first utilising PCA to identify the number of factors to extract, common EFA (or EFA utilising Maximum Likelihood estimation) was applied to identify and then interpret the selected number of latent factors (hereafter referred to as *repertoires*). The *repertoire* model was then considered for goodness-of-fit by considering various measures of fit from a CFA. Once the model structure was confirmed, it was used as the measurement model component in a SEM; with a structural component consisting of various demographic variables and years, regressed onto each of the *repertoires* as outcomes. These resulted in the estimation of SEM model parameters that were used to estimate marginal means, which were plotted to aid interpretation.

5.2 Data and Methods

5.2.1 Data

For this chapter the five annual datasets as described in 3.4 were combined into a single set containing 126 726 records. The variables that were included were six demographic variables (sex, age, race, education, household income, and LSM), aggregated as described in 3.3.1; and 28 media *vehicle* variables containing raw (ie., unstandardised) values. The list of media *vehicles* can be reviewed in table 3.1.

5.2.2 Methods

5.2.2.1 Principal Components Analysis

PCA was used to identify the number of *repertoires* to consider in a subsequent EFA. The decision on the number of components was based on several criteria that included the use of R's *nFactors* package to identify an "optimum" and to present an informative screeplot.

5.2.2.2 Exploratory Factor Analysis

Following the decision on the number of *repertoires* EFA was applied using Rs *psych* package to identify which *vehicles* loaded most strongly on which *repertoires* and to interpret each of the *repertoires.* However, as described in section 4.3, various methods of estimating factors in EFA exist. For this analysis, given the use of R's *lavaan* package to estimate the SEM and the variations in scales of the original data, the method of Maximum Likelihood (ML) was preferred, but a key assumption of this method is that the original variables required a multivariate normal distribution (MVN).

The *shapiro wilks* test of multivariate normality applied to several random samples of 5 000 cases in the data and utilising R's *mvnormtest* package, indicated rejection of the null hypothesis of the data having a multivariate normal distribution. Given the nature of the data and the sample size this result was not unexpected. Two alternative approaches to the *shapiro wilks* test were considered:

- Based on the argument presented in section 4.2.2: if the principal components demonstrate a multivariate normal distribution, the originating variables can be assumed to have a normal distribution, frequency histograms were plotted of the first six components separately and combined (see plots in figure 5.1). Although not ideal, the histograms do appear to show some bell-shaped behaviour.

- Correlations of loadings from the ML and the Principal Components (PC) methods of estimation showed a near 100% correlation.

Given the outcome of these additional approaches, an EFA using the *psych* package and ML method of estimation with an *oblimin* rotation to allow for correlations between factors - as would be expected given the nature of the data - generated the loadings shown in table 5.1. These loadings were used to interpret the six factors or *repertoires.* In this case, given the large sample size, all positive loadings larger than 0.2 were considered, although all loadings larger than 0.1 are shown in the table to assist in interpretation. Loadings larger than 0.3 were indicated in red to illustrate stronger loadings.

Figure 5.1: Frequency Histogram of Combined 1-6 Principal Components Scores

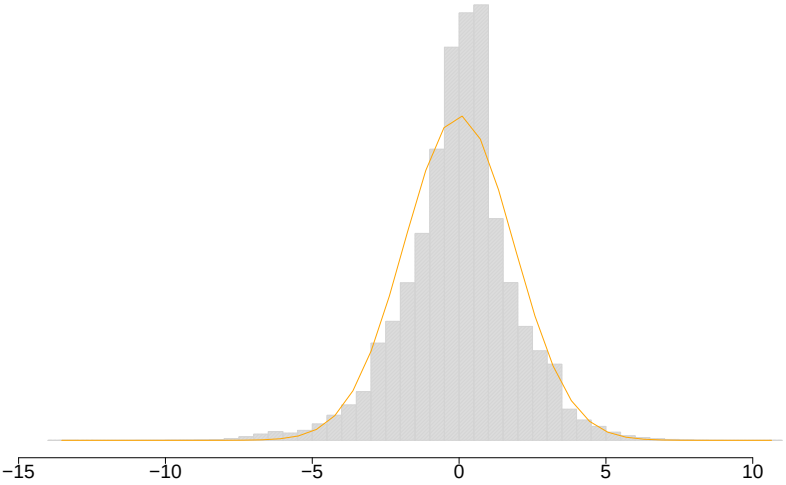
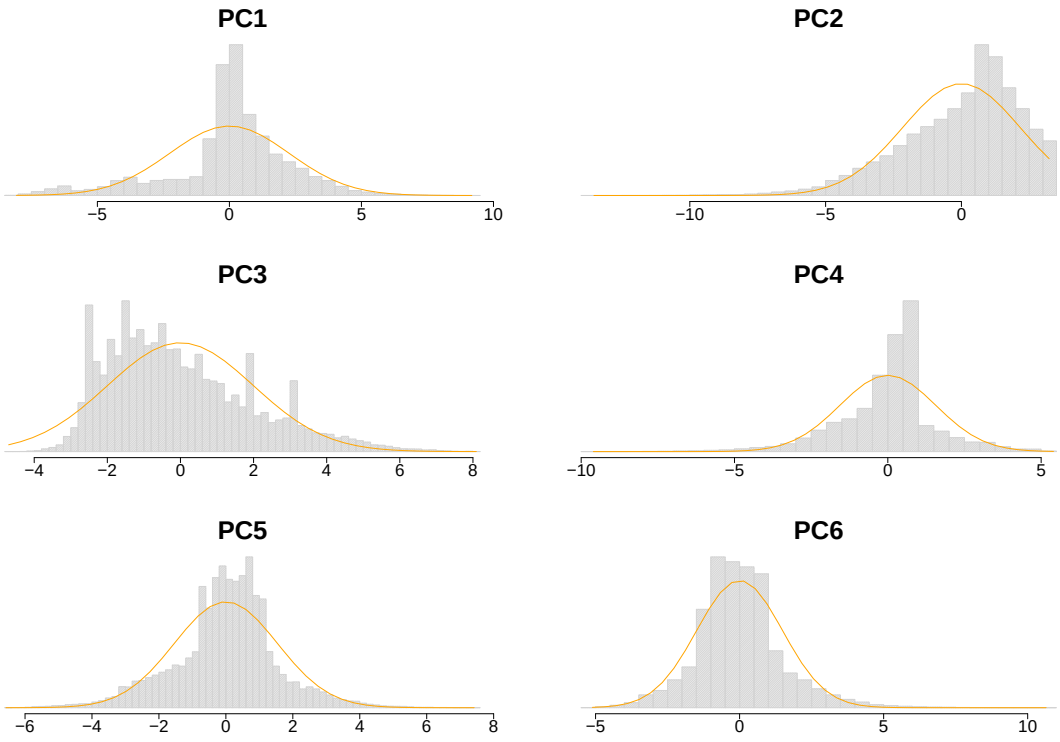


Figure 5.2: Separate Frequency Histograms of Principal Components Scores 1-6



5.2.2.3 Confirmatory Factor Analysis

In order to consider fit and whether the underlying latent structure identified in the EFA applied by year, separate confirmatory factor models, each with six latent factors and the same loadings structure, were estimated for each of the years in the dataset. This use of CFA is described in section 4.4 as *simultaneous* CFA.

The CFAs were run using the *lavaan* package with estimation arguments set for standardised latent variables and the inclusion of mean structures, to ensure latent variables with unit variances and zero means, but allowing for factor covariances to be estimated. Various fit measures, described in section 4.4.2.5, were applied to consider the model's fit.

5.2.2.4 Structural Equation modelling, Estimated Marginal Means and Bootstrapping

A SEM using R's *lavaan* package with a measurement model as confirmed by the CFA and a structural component consisting of linear regressions with the six estimated latent factors (*repertoires)* as outcome variables and demographic variables as independent predictors - including interaction effects for all demographics on *year* - was applied to 1 000 bootstrapped samples of the dataset (for an explanation on bootstrapping see section 4.7). The estimation of latent variables in the SEM used the same arguments as applied to the CFA in subsection 5.2.2.3, namely the latent variable estimations were limited to unit variance and zero means, while they were free to covary with one another.

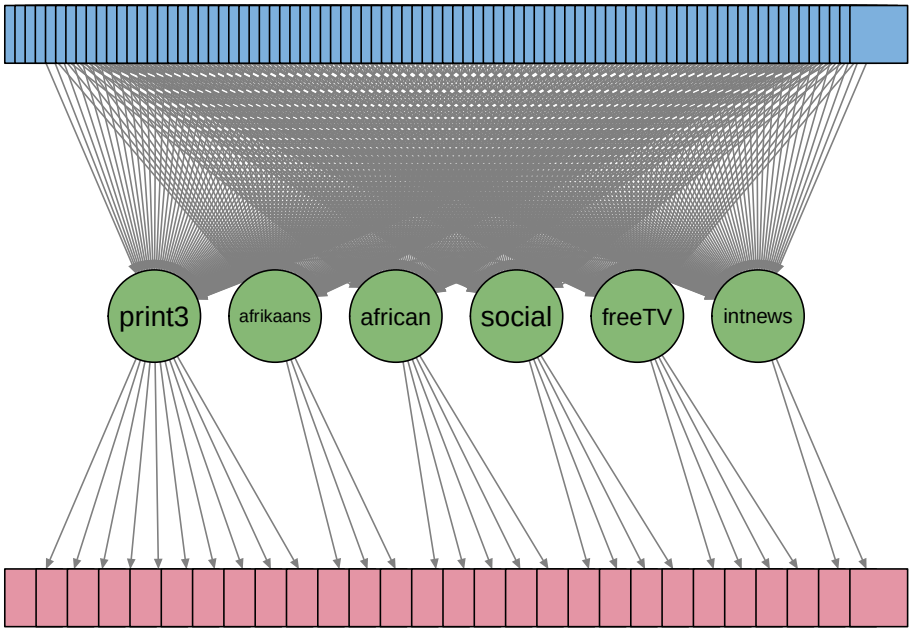
The coefficients estimated in the structural component of the 1 000 bootstrapped models were then used to manually determine 1 000 sets of estimated marginal means (EMM), using both *equal* and *proportional* weightings to identify mean engagement values by demographic levels by year. For an explanation on EMMs, see section 4.6. Measurement errors were accounted for with a 95% confidence intervals of the means, represented by the 2.5th and 97.5th percentiles of the 1 000 bootstrapped means. The EMMs with their confidence intervals by demographic category levels and by year were then plotted to aid in the interpretations.

The complexity and size of the model as well as the size of the dataset meant that the estimation of 1000 SEMs was not feasible on the researcher's own computer. Accordingly extensive use was made of Amazon Web Server's cloud computing services.

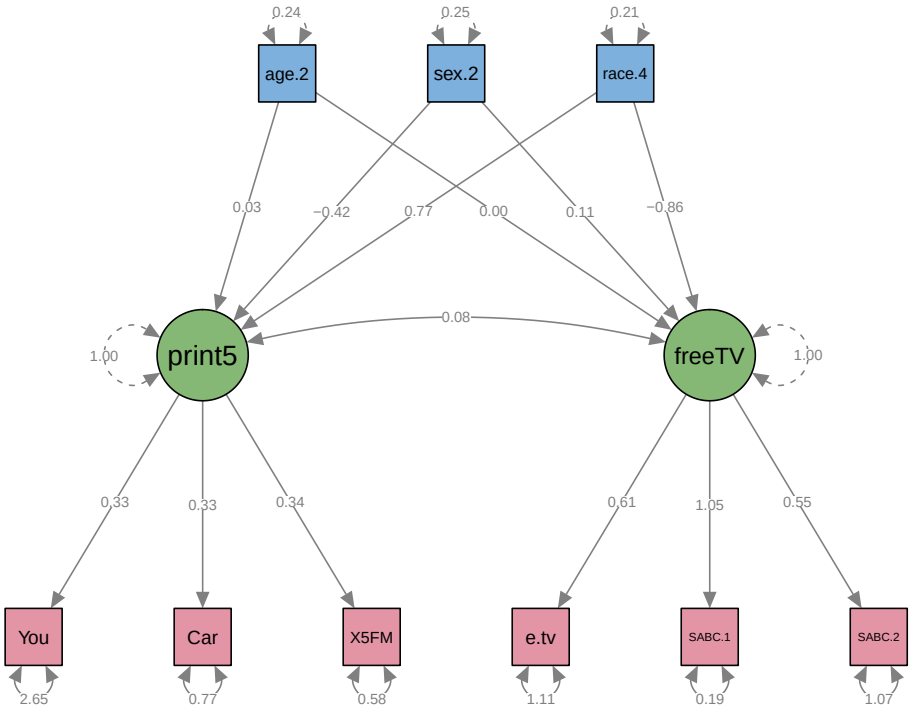
To illustrate, figure 5.3 shows the scale of the estimation, using the descriptions of latent variables identified in section 5.3.1, with the top plot showing all the relationships in the SEM; while the bottom plot shows a fictitious, partial model - purely as illustration of what lies within the mesh of relationships in the full model. In the fictitious, more detailed path diagram the unit variance and the covariance (0.08) of the latent variables can be seen. The variances of the input variables can also be seen and the estimated coefficients are indicated.

Figure 5.3: SEM Path Diagrams: The top plot showing all the relationships; the bottom plot (purely as illustration) demonstrates the detail.

SEM SHOWING ALL THE RELATIONSHIPS: DEMOGRAPHICS IN BLUE, LATENT VARIABLES IN GREEN, MEDIA VEHICLES IN PINK



PARTIAL DETAIL, PURELY AS ILLUSTRATION: DEMOGRAPHICS IN BLUE, LATENT VARIABLES IN GREEN, MEDIA VEHICLES IN PINK



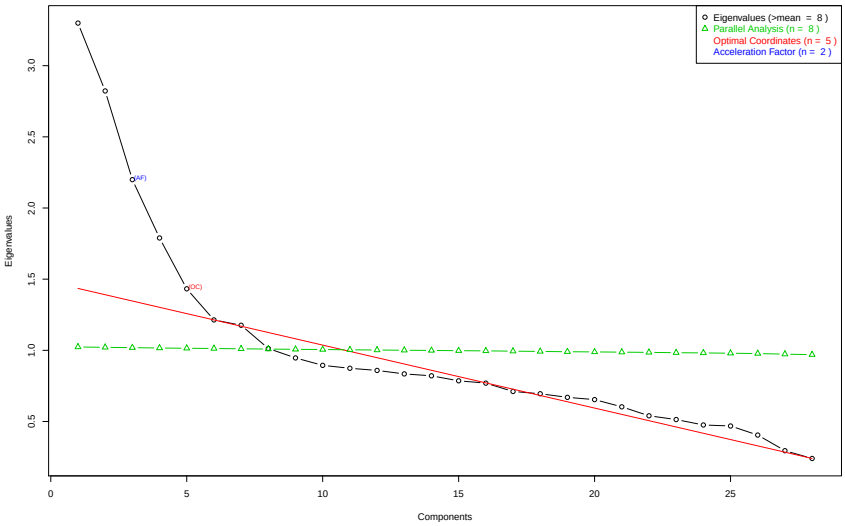
5.3 Results

5.3.1 Identifying and Interpreting Repertoires

5.3.1.1 PCA to Determine Number of Factors

An initial exploration using PCA, see figure on this page, suggests an optimal latent structure of five factors, while the application of Kaiser's stopping rule (see section 4.2.4) would suggest eight factors. Applying the "elbow" rule, also described in section 4.2.4, suggests anything between six and nine. After considering the EFA loadings for models with between five and eight factors, it was decided to continue with a model comprising of six factors. This number compares with the six used by Schröder (2015) and the seven by Edgerly (2015).

Figure 5.4: PCA Scree Plot: N = 126 726



5.3.1.2 EFA to Identify and Interpret Factors

The application of EFA generated the loadings shown in table 5.1. The loadings in red represent the loadings larger than $0,3$ as suggested in section 4.3.6. Given the large sample size, other loadings larger than $0,1$ were retained for purpose of comparison. These data were used to describe and name six media *repertoires:*

- A first *repertoire*, designated as *freeTV,* shows relatively strong positive loadings for SABC1, SABC2, SABC3, and eTV; weaker positive loadings for You, Drum and Metro FM, and slightly stronger negative loading for DSTV. The strong positive loadings consistently on free-to-air national television best describes this factor.

- A second *repertoire*, designated as *social,* shows strong positive loadings on internet social and internet search, which would suggest strong use of the internet for social

media purposes. The weaker, but still substantive loadings for DSTV and internet radio should be considered in further interpretations of this factor.

- A third *repertoire*, designated as *intnews,* shows only two strong loadings, namely on internet print and internet news, with weaker loadings on internet radio. This factor would be dominated by internet engagement, in particular by respondents with an interest in hard news.
- A fourth repertoire, designated as *afrikaans,* shows strong loading on Afrikaans-language media, in particular on Huisgenoot, Rapport and Sarie. When radio station RSG is included in the mix for exploratory factor analysis by year, it also loads onto this factor. Weaker loadings are apparent for SABC 2, Getaway and Car magazines.
- A fifth repertoire, described as *african,* shows strong loadings on Kickoff, Soccer Laduma, Drum and Bona, with weaker but still substantial loadings on Metro FM. What ties these together is the fact that their audience is predominantly 'African' black. Exploratory analysis with seven factors show a splitting of this factor into one loading strongly on Kickoff and Soccer Laduma and another dominated by Drum and Bona. When including the important Daily Sun newspaper on analyses by period, it loads strongly on this factor for six dimensions and splits most strongly along with Drum and Bona for seven factors.
- A sixth and final repertoire collects the remaining print vehicles, with weaker loadings on Radio 5 - hence the naming designation as *print5*

5.3.1.3 CFA to Confirm the Measurement Model

Table 5.2 indicates consistently strong correlations for the *simultaneous* CFA loadings by year and with the full set. Although not conclusive, these correlations do provide some measure of confidence in the use of this measurement model when applied to the full dataset, especially given the aim of determining estimations of marginal means by year. Various measure of fit were also considered, the results of which are displayed in table 5.3. The assessment of fit was based on the literature reviewed in section 4.4.2.5. These would serve to confirm the measurement model that was used in the SEM that followed.

5.3.2 Bootstrapped Structural Equation Model

In this section the results from the SEMs and the resultant EMMs from the 1 000 bootstrapped datasets are presented. The plots will be discussed by *repertoire* before considering summaries by demographic category. The range and scale of engagement of the plots are identical to allow for comparison of variation, but the standardised estimation procedures applied in the SEM imply that it would make no sense to compare aggregated coefficients between the different *repertoires* directly.

Table 5.1: Selected Loadings: Exploratory Factor Analysis

	ML2	ML1	ML4	ML3	ML5	ML6
Business.Day						**0.281**
Mail.n.Guardian			0.111			**0.320**
Rapport				**0.647**		
The.Sunday.Independent						**0.291**
Sunday.Times						**0.455**
Soccer.Laduma					**0.698**	
Drum	0.135			-0.115	**0.302**	0.119
Huisgenoot				**0.760**		
You	0.105					**0.387**
Kickoff					**0.725**	
Bona					**0.281**	
Car				0.112		**0.355**
Cosmopolitan						**0.350**
Getaway				0.118		**0.319**
Sarie				**0.490**		
Topcar						**0.350**
X5FM		0.198				**0.230**
Metro.FM	0.142			-0.185	**0.228**	
e.tv	**0.678**					
SABC.1	**0.647**			-0.184	0.125	
SABC.2	**0.688**			0.183		
SABC.3	**0.725**					0.118
DSTV	-0.273	**0.323**				0.112
int.print			**0.858**			
int.radio		**0.204**	0.188			
int.news			**0.815**			
int.social		**0.872**				
int.search		**0.851**				

Table 5.2: Correlation Matrix of Loading Coefficients

	2002	**2008**	**2010**	**2012**	**2014**	Full
2002	1	0.96	0.94	0.86	0.79	0.93
2008	0.96	1	0.99	0.95	0.89	0.98
2010	0.94	0.99	1	0.97	0.92	0.99
2012	0.86	0.95	0.97	1	0.98	0.98
2014	0.79	0.90	0.92	0.98	1	0.95
Full	0.93	0.98	0.99	0.98	0.95	1

Table 5.3: CFA Fit Measures by Year and for the Complete Dataset

SET	**IFI**	**RMSEA**	**SRMR**	**NNFI**	**TLI**	**ASSESSMENT OF FIT**	**OVERALL**
2002	0.70	0.07	0.07	0.66	0.66	ifi below the *good* cut off; RMSEA suggests *mediocre;* SRMR suggest *good*	MEDIUM
2008	0.81	0.05	0.05	0.78	0.78	ifi suggests *good* ; RMSEA suggests *good;* SRMR suggest *good*	GOOD
2010	0.82	0.05	0.05	0.80	0.80	ifi suggests *good* ; RMSEA suggests *good;* SRMR suggest *good*	GOOD
2012	0.87	0.05	0.05	0.85	0.85	ifi suggests *good* ; RMSEA suggests *good;* SRMR suggest *good*	GOOD
2014	0.86	0.05	0.05	0.84	0.84	ifi suggests *good* ; RMSEA suggests *good;* SRMR suggest *good*	GOOD
Full	0.85	0.05	0.05	0.83	0.83	ifi suggests *good* ; RMSEA suggests *good;* SRMR suggest *good*	GOOD

5.3.2.1 Profiles of Engagement on *freeTV*

This section references plots in figure 5.5. The media *vehicles* comprising this factor were: eTV, SABC 1, SABC 2, & SABC 3.

The LSM coefficients in linear models for television as *type* are generally higher than others suggesting they are stronger determinants of television viewing; also, the *proportional* weightings are slightly higher than *equal* weightings for LSM. It would follow therefore that the non-LSM demographic categories would show a consistent pattern of higher values for *proportional* than *equal* weightings, while the LSM values show very little difference between these.

For some demographic levels profiles for *equally* and *proportionally* weighted marginal means are quite different. For example the *equally* weighted marginal means for 15-24 year olds shows consistent decline, while the *proportionally* weighted means for the same age group shows some increase before a decline. Given the argument in section 4.6, this interpretation would be based on the *proportional* weightings.

Comparing the plots with the EMM of TV as media *type* in figure 6.9, would suggest that while engagement of TV as a media *type* may have grown among most demographic levels over the period under review, the engagement on *freeTV* appears to show a more static profile with some suggestion of declines in later years. The vertical bars in the plot indicate

the 95% bootstrapped confidence intervals from the bootstrap procedure described in section 5.2.2.4.

In particular, engagement on *freeTV* for gender, age, education and income shows slight increases until 2010, after which engagement either appears to stabilise (for ages 45-54; and income <R11000) or show some declines. Among black respondents engagement on this factor shows growth until 2010, with stability after this period. Although engagement among coloured respondents are generally higher than for any of the other population groups, a steep initial growth in 2008 appears to have settled and even declined toward the end of the study period. White respondents show lower levels than other population groups and while these held until 2010, they declined subsequently. Finally, while LSM 5-8 show the highest levels of engagement, the quite substantial growth shown for TV as medium *type* in contrast shows a decline for *freeTV,* especially for LSM 7-10 since 2010. It should be borne in mind that ownership of appliances, including television sets for LSM1-2 should not be assumed.

5.3.2.2 Profiles of Engagement on *intnews*

This section references plots in figure 5.6. The media *vehicles* loading onto the *intnews* factor (or *repertoire*) are internet to access print media or other news sites. There appear to be little difference between *proportional* and *equal* weightings in these cases. The vertical bars in the plot indicate the 95% bootstrapped confidence intervals from the bootstrap procedure described in section 5.2.2.4.

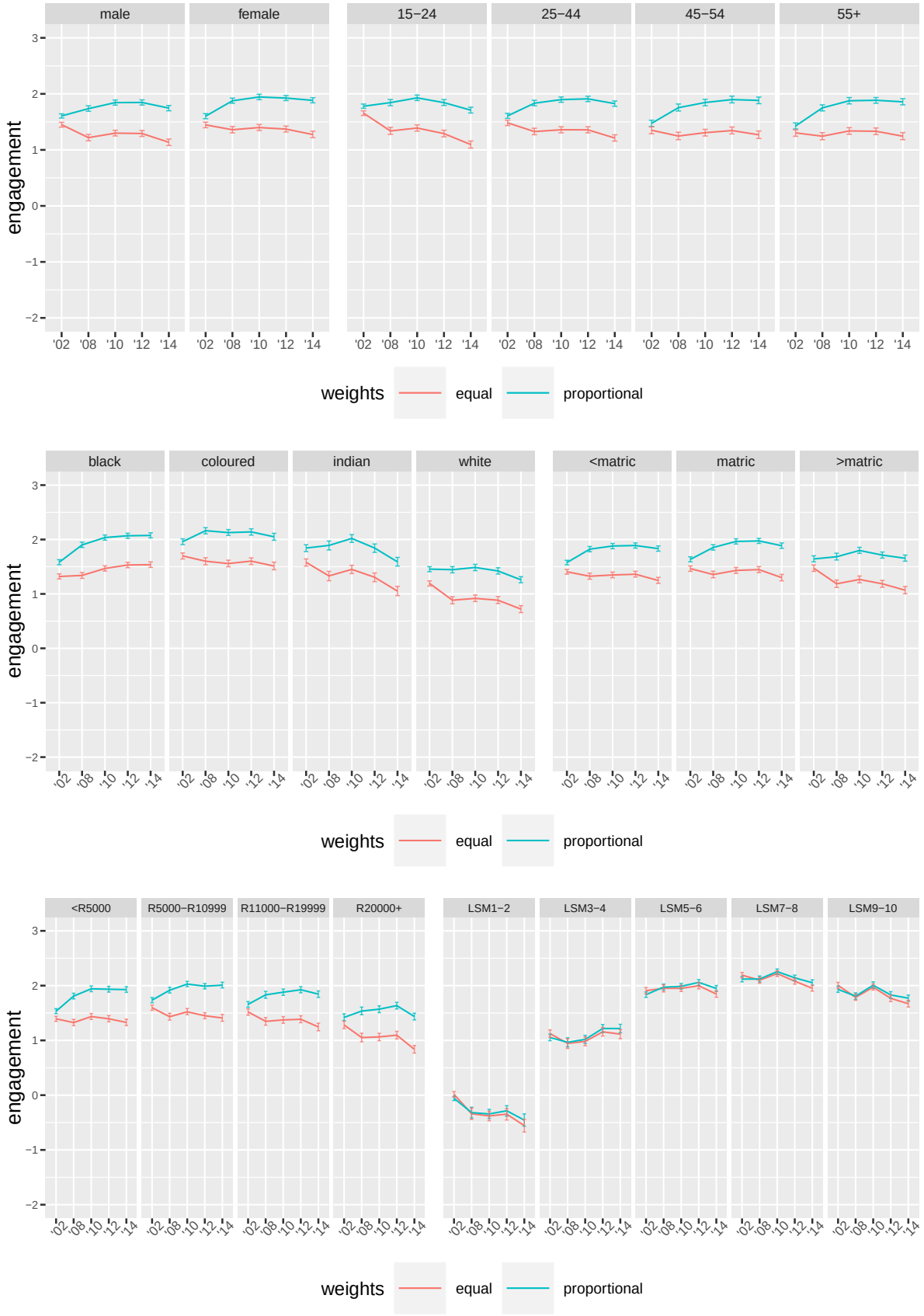
In most cases it would appear that from stable or slight growth in engagement on this *repertoire* from 2002 until 2010, considerably steeper growth is apparent after 2010. Although small differences are apparent for gender and race, with male and white respondents showing marginally higher levels of engagement, these two categories do no appear to be particularly significant differentiators. Age, education, income and LSM show some profile differences:

All age groups start at relatively low levels, but with 45+ slightly lower and 15-44 somewhat higher. All age groups appear relatively stable until 2010, followed by growth, but with marked differences in the steepness of these increases. The younger the age group, the steeper the increases.

Respondents with post matric showed higher levels of engagement at the start of the study period and also relatively steeper growth until 2010, followed by steep increases until 2014. In contrast, respondents with only matric or lower than matric showed little increase in engagement until 2010 when both levels grew their engagement, although respondents with matric increased their engagement at a faster rate than those with less than matric.

The profiles for respondents below R20 000 income levels showed very similar profiles: starting from low bases, followed by slow growth until 2010, followed in turn by sharper increases. Respondents with household incomes higher than R20 000 started off at higher levels, but only really showed steep growth after 2010.

Figure 5.5: Estimated Marginal Means with error bars showing bootstrapped 95% confidence intervals: *free TV*



The profiles for the three aggregated LSM levels 1-6 show quite similarly stable profiles until 2010, followed by similar rates of increase. LSM 7-8 show very similar levels to the lower brackets until 2010, after which growth among these respondents increases more steeply. LSM 9-10 respondents show a higher starting level but also steeper growth than other LSM brackets.

5.3.2.3 Profiles of Engagement on *african*

This section references plots in figure 5.7. The media *vehicles* loading onto the *african* factor are Drum, Bona, Metro FM, Soccer Laduma, and Kickoff. The vertical bars in the plot indicate the 95% bootstrapped confidence intervals from the bootstrap procedure described in section 5.2.2.4.

Given that the generally higher values for *proportional* relative to *equal* weighting are a reflection of the fact that race (in particular black) is a main driver of this factor in terms of model coefficients, *proportional* values were used in this interpretation.

Although the vehicles defining this factor are either print or radio, it is interesting to note that the general declines obvious in *print5* and also in the print and radio media *types* are not reflected here. Although there are some declines and some increases, for most category levels it would appear engagement on this *repertoire* has been reasonably stable over this period.

Males are somewhat more engaged than females and both levels appear to be stable over the period under review. 15-44 year-olds are engaged at a slightly higher level than the other age groups. All engagement on this *repertoire* appears relatively stable over the period under review.

For the lowest levels of education and income and for LSM 1-4, the levels of engagement appear to be somewhat lower than for the higher levels of the respective categories, which appear to be at quite similar levels of engagement. Apart from declines for LSM 3-6, all other levels appear to be quite stable over the period.

5.3.2.4 Profiles of Engagement on *afrikaans*

This section references plots in figure 5.8. The media *vehicles* loading onto the *afrikaans* factor are Rapport, Huisgenoot and Sarie. The *proportional* and *equal* weightings do not differ substantively in this case and where they do differ somewhat, the margins of error appear large enough to make it difficult to make any substantive conclusions about the differences. The vertical bars in the plot indicate the 95% bootstrapped confidence intervals from the bootstrap procedure described in section 5.2.2.4.

Generally, engagement on this factor shows some declines until 2008, when they appear to stabilise or at least continue at a slower decline. As would be expected white and coloured

Figure 5.6: Estimated Marginal Means with error bars showing bootstrapped 95% confidence intervals: *intnews*

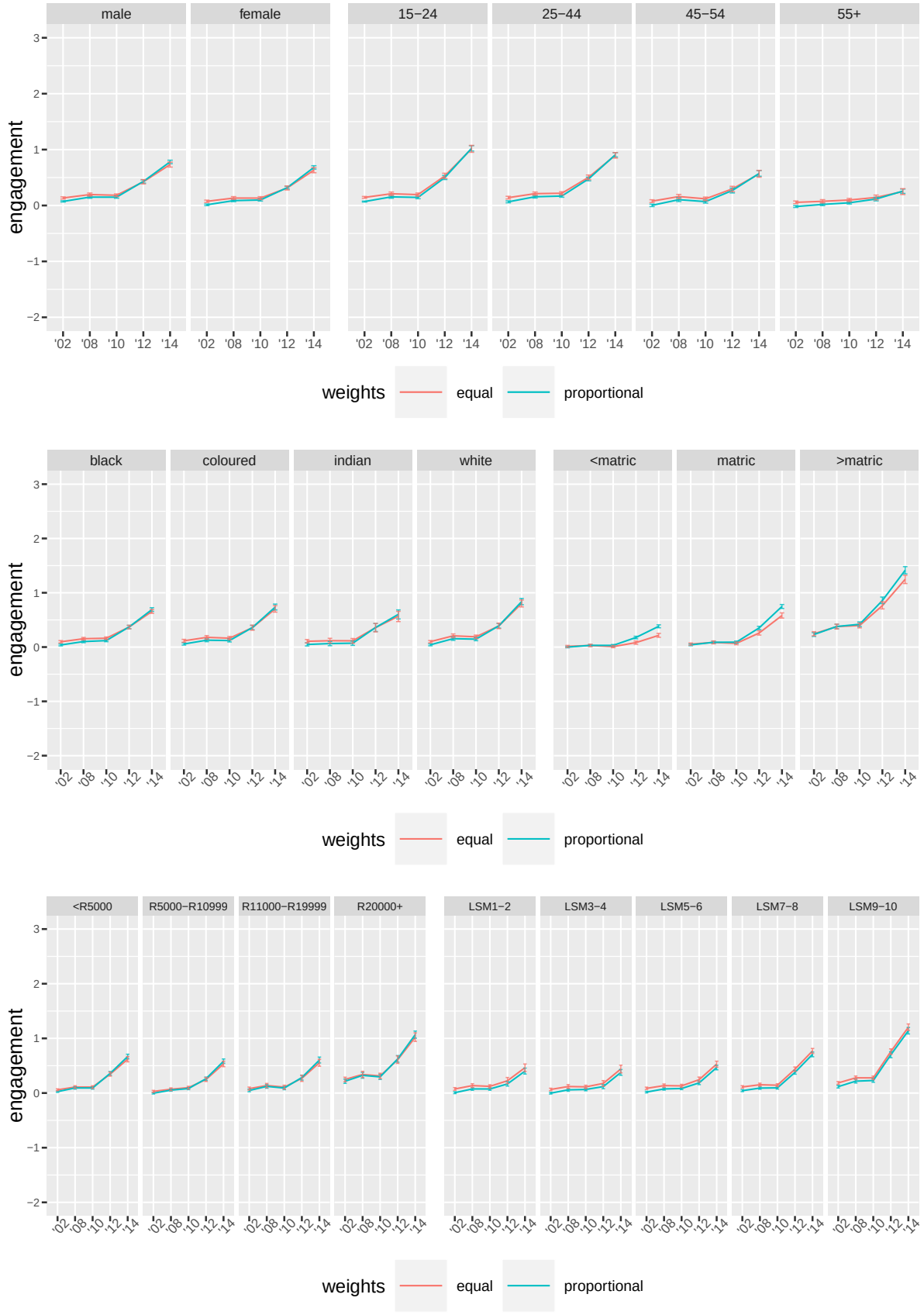
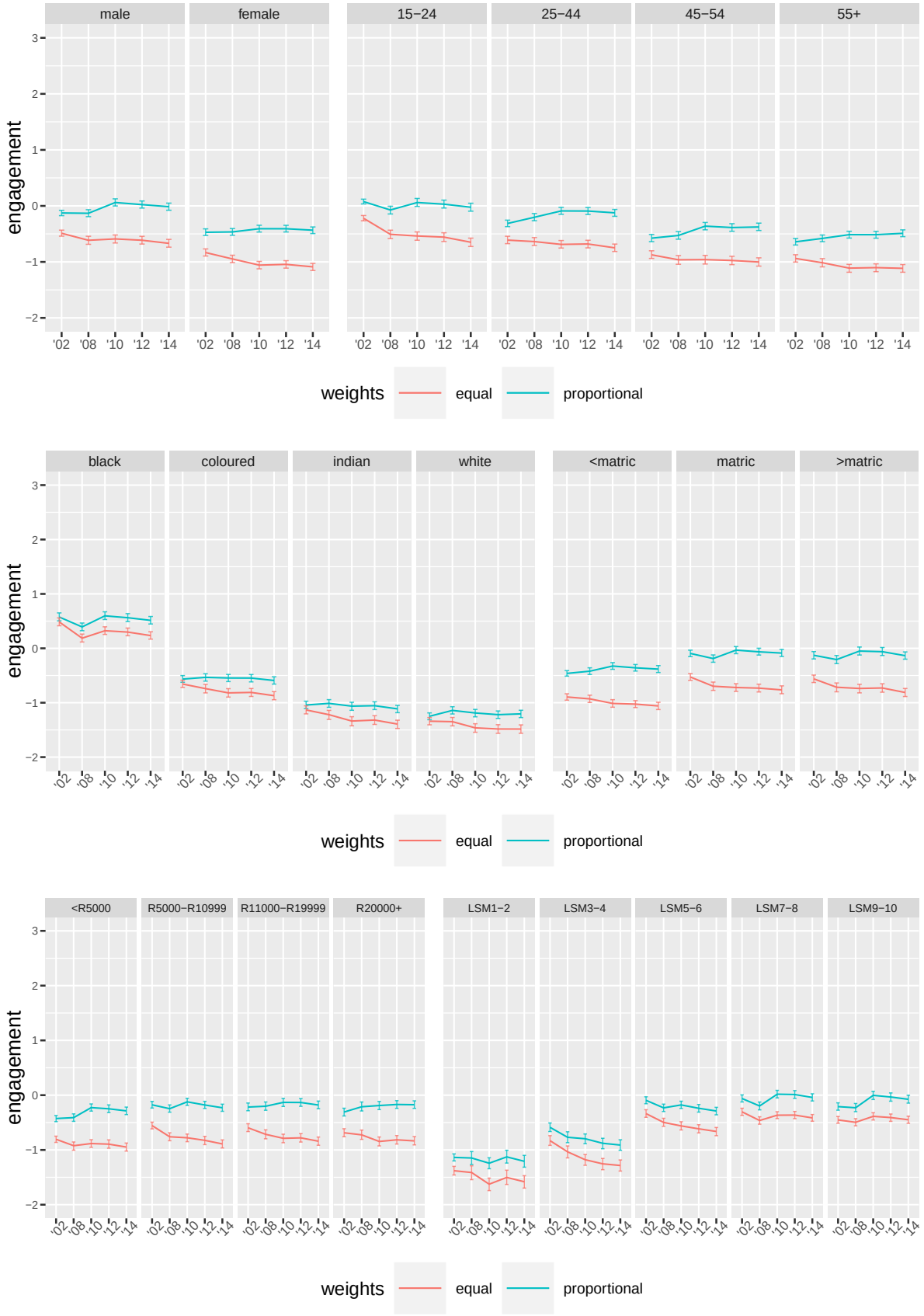


Figure 5.7: Estimated Marginal Means with error bars showing bootstrapped 95% confidence intervals:*african*



respondents showed considerably higher engagement that indian or black respondents. Respondents with post-matric and higher income levels showed slightly lower levels of engagement on this factor than the other levels. In terms of LSM, the plots suggest the highest levels of engagement occurred among LSM 9-10, which also saw the steepest initial declines.

5.3.2.5 Profiles of Engagement on *social*

This section references plots in figure 5.9. The media *vehicles* loading onto the *social* factor are DSTV and internet use for the purposes of social media, radio streaming and search. The vertical bars in the plot indicate the 95% bootstrapped confidence intervals from the bootstrap procedure described in section 5.2.2.4.

In contrast to the slight differences between gender and race shown in *intnews*, for this factor the differences seem to disappear. The profiles for males and females and for the different population groups are virtually identical and seem to mirror the levels and profiles for internet as media *type* (see section 6.3.3.5).

The engagement levels and profiles for all the other demographic categories closely mirror those of the internet as media *type*, described in section 6.3.3.5, namely that engagement shows generally consistent and steep increases over the study period, most markedly after 2010. The only exception appears to be considerably lower increases for the 55+ age group. As engagement on internet generally, an interesting feature is the apparent slowing down between 2008 and 2010, that may the result, as suggested in 6.3.3.5 of a recession, followed by the wider application of wifi. Also as for the internet as media *type*, while most levels start out at relatively similar low engagement, some category levels attain considerably higher engagement by the end of the study period. In particular - and as was expected - for younger, better educated and higher income and higher LSM respondents.

The systematic growth may also be partly explained by the inclusion of DSTV in this factor. Growth in DSTV engagement would also partially explain the disparity between TV as media *type* and the *repertoire freeTV* (compare sections 6.3.3.3 and 5.3.2.1).

5.3.2.6 Profiles of Engagement on *print5*

This section references plots in figure 5.10. The media vehicles loading onto the *print5* factor are Business Day, Mail and Guardian, The Sunday Independent, Sunday Times, You, Car, Cosmopolitan, Getaway, Topcar and Radio 5 FM. The vertical bars in the plot indicate the 95% bootstrapped confidence intervals from the bootstrap procedure described in section 5.2.2.4.

In most cases the interpretation between *proportional* and *equal* weighting do not materially affect the interpretation. Where this may be the case, *proportional* values will be used on the basis that the samples were taken based on proportional representation of the population;

Figure 5.8: Estimated Marginal Means with error bars showing bootstrapped 95% confidence intervals: *afrikaans*

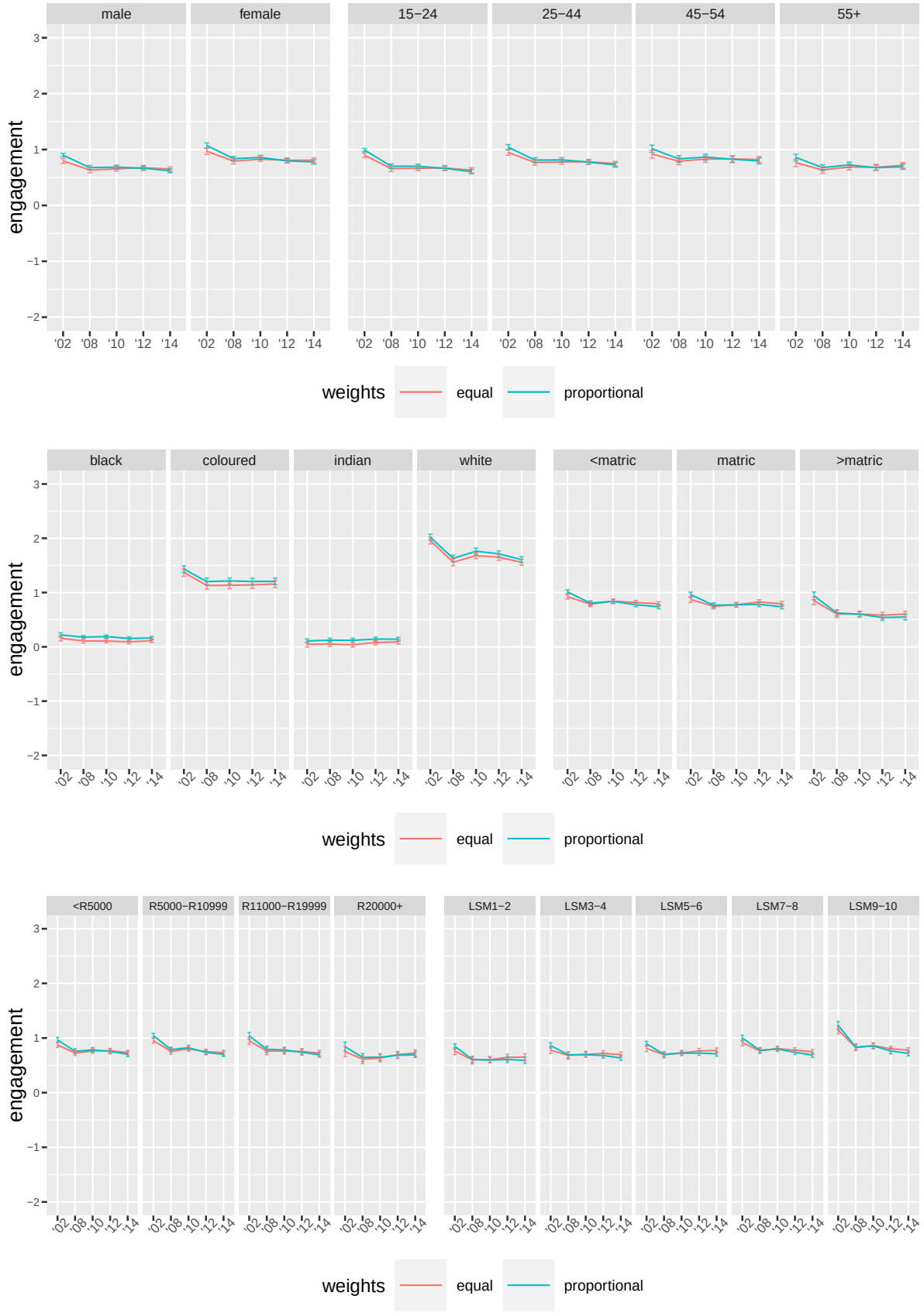
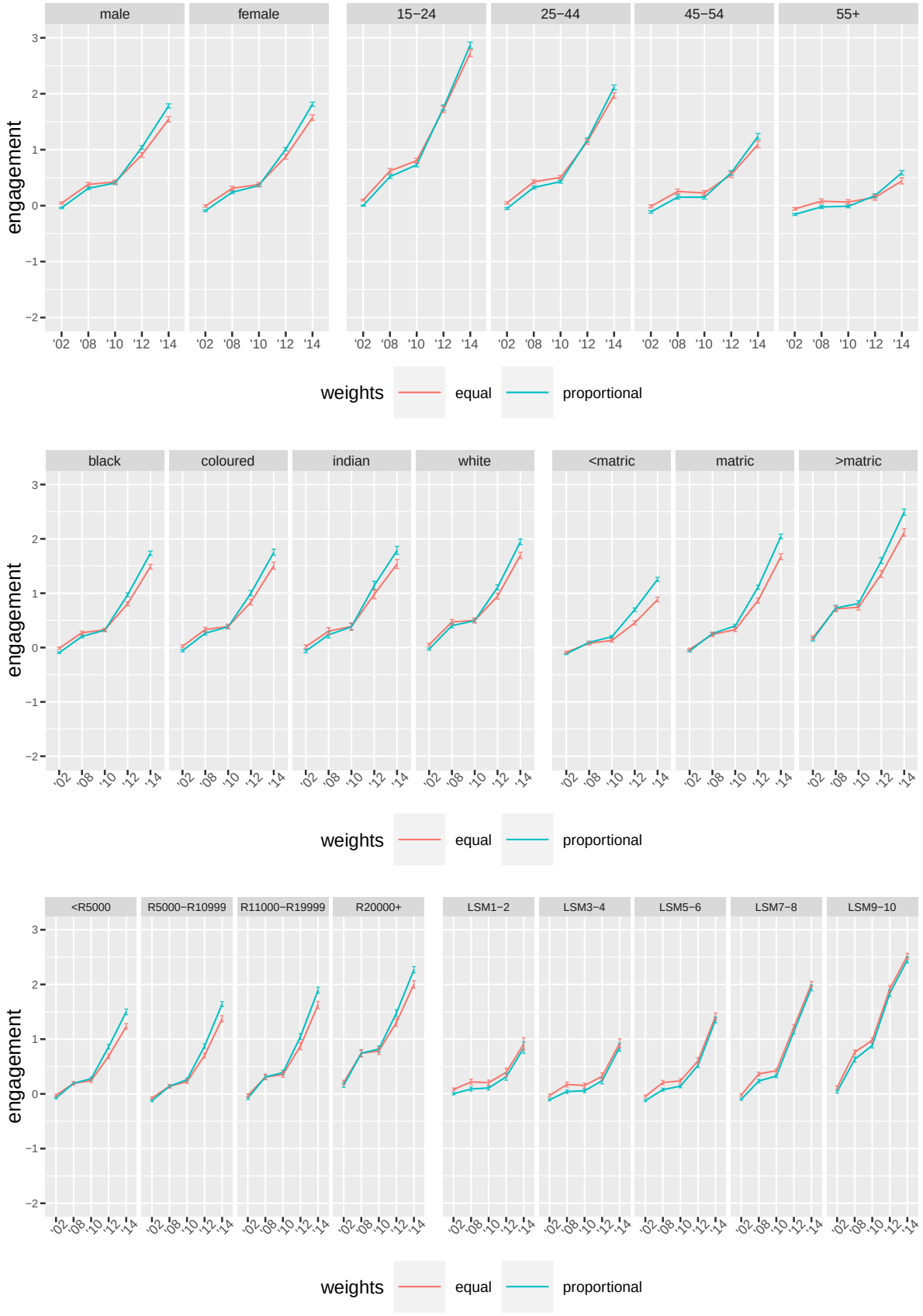


Figure 5.9: Estimated Marginal Means with error bars showing bootstrapped 95% confidence intervals: *social*



and, since for LSM where higher category levels are generally more engaged with print, *equal* estimates are consistently higher than for *proportional*.

As expected, given that this factor comprises mainly print media, engagement shows general declines of engagement for all levels since 2008. Prior to 2008 there do appear to be some increases, mainly among black, female, older and lower income respondents.

Also as could be expected, the levels of engagement for different levels of education, income and LSM increase for higher levels of the categories. In all cases the declines also appear to be somewhat steeper for higher category levels.

5.3.3 Summary of Repertoire Engagement by Demographic Categories

5.3.3.1 Gender

There appears to be little difference between genders on all but *african*, which shows a male bias, but with stable profiles. Both genders showed slight increases followed by some declines on *freeTV*. On *intnews* and *social* both genders showed steady increases until 2010, followed by steep growth until 2014. Both genders also showed similar declines on *print5.*

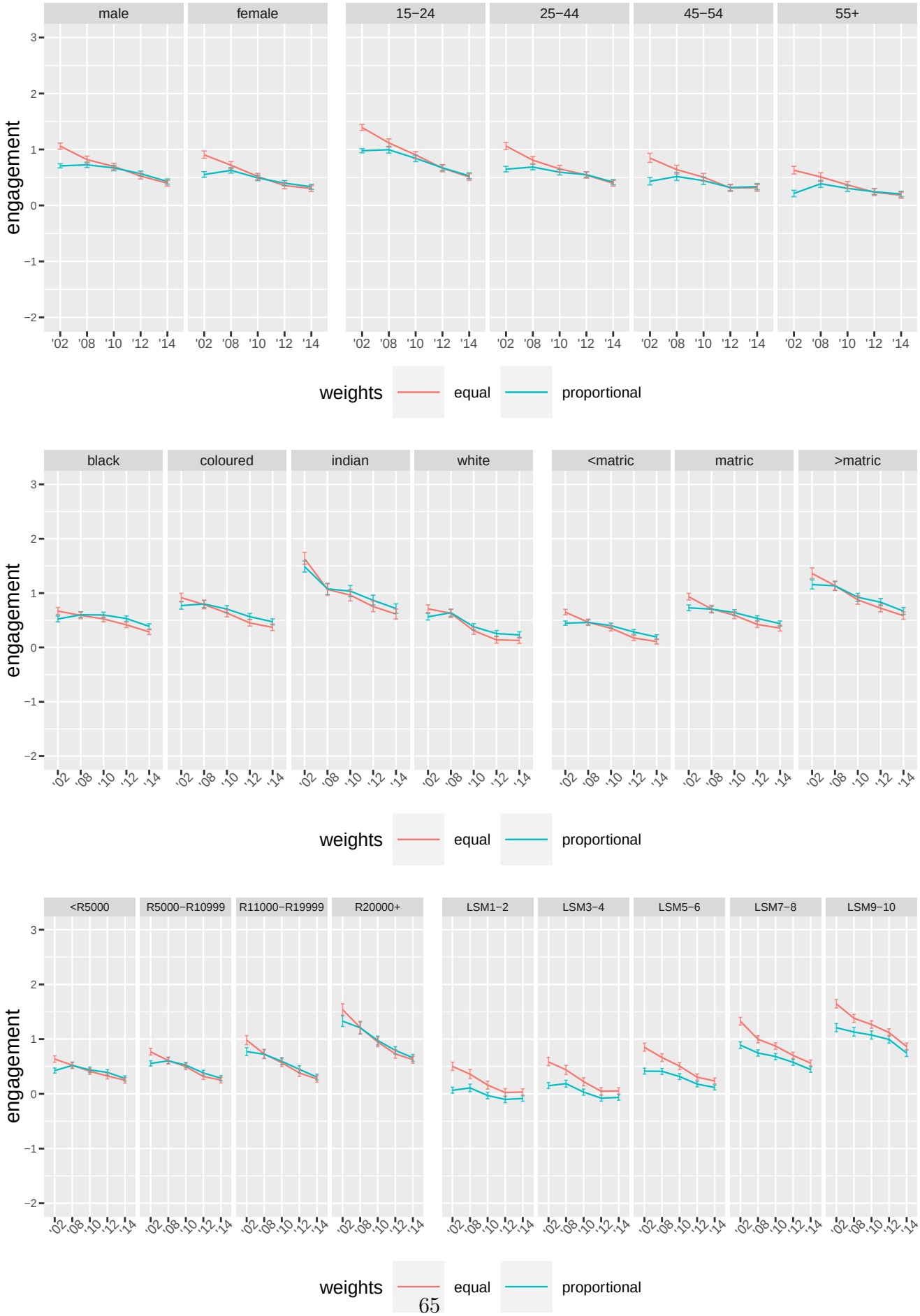
5.3.3.2 Age

All age groups showed slight increases on *freeTV* until 2010. After this all age groups, with the exception of relative stability for 45-54, showed declines. On *intnews* all age groups in 2002 appear to be at similarly low level, although those over 45 slightly lower and under 45 slightly higher. The profiles and appear quite stable until 2010 when all groups show growth, but with quite different profiles of growth: the younger the age group, the steeper the growth profile. On *african* respondents under 44 showed higher levels of engagement, but engagement on this *repertoire* as well as on *afrikaans* appears quite stable across different age groups. In contrast to *intnews,* engagement levels on *social* appeared to grow more between 2002 and 2010, but as with the former, they showed growth, differentiated for different age groups after 2010. As for *intnews*: the younger the age group, the steeper the growth profile. *print5* showed similar rates of decline across age groups, but with younger respondents starting out at slightly higher levels of engagement.

5.3.3.3 Education

For all education groups engagement on *freeTV* appears to show slight increases until 2010, followed by declines. On *intnews* respondents with post-matric qualifications show slightly higher engagement, with slightly steeper growth until 2010, followed by even steeper growth post 2010. Respondents with matric also showed slightly steeper rates of growth on *intnews* than those with less than matric. On *african* and *afrikaans* respondents with lower levels

Figure 5.10: Estimated Marginal Means with error bars showing bootstrapped 95% confidence intervals: *print5*



of education showed slightly lower engagement, but levels of engagement on these *repertoire* remained quite stable over the period. On *social* respondents with higher levels of education showed generally higher rates of growth; while on *print5,* all education groups appeared to show declines, those with higher levels of education show somewhat steeper declines.

5.3.3.4 Population Group

Black respondents showed growth on *freeTV* until 2010 followed by stability, while coloured respondents showed high engagement levels on *freeTV,* with steep initial growth that settled and even declined toward the end of the period. White respondents appear less engaged than the other population groups and appeared to remain stable until 2010, followed by declines. Different population groups showed very little difference in terms of both levels and profiles on both *intnews* and *social.* The profiles indicate slow growth on these *repertoires* until 2010, followed by steep growth. On *african,* as would be expected, black respondents showed higher levels of engagement that other groups, but the profiles over the period of study remained stable. For *print5* all population groups showed similar rates of declines. The only exception is that indian respondents declined from higher levels than other groups.

5.3.3.5 Household Income

Respondents at all levels of household income showed slight increases on *freeTV* until 2010. After this, those earning less than R11 000 appear to be stable, while others declined. Respondents earning less than R20 000 started from similarly low bases in 2002, showing slight growth on *intnews* until 2010, followed by steeper increases. Those earning more than R20 000 started at higher levels in 2002, showing some growth until 2010, but with steeper growth until 2014. Respondents of all income groups showed generally stable engagement on *african* and on *afrikaans,* with higher levels showing slightly steeper growth profiles on *african* and lower levels on *afrikaans.* Higher household income levels showed generally steeper profiles of growth on *social,* while this is reversed on *print5*, where higher income levels decline more than lower levels.

5.3.3.6 Living Standards Measure

Respondents in LSM 5-8 showed the highest level of engagement compared to other LSM brackets on *freeTV.* Since 2010 all levels showed declines, although respondents from levels 1-7 showed steeper declines over this period. Respondents from levels 1-6 showed stability on *intnews* until 2010, followed by similar rates of growth. Respondents in levels 7-8 showed similar starting levels, but steeper growth than the lower levels, while respondents from LSM 9-10 showed higher starting levels and steeper growth profiles than other levels. Respondents at lower levels of LSM showed lower engagement on *african,* but all levels remained stable over the period. Respondents at levels 9-10 showed highest engagement on *afrikaans,* but

also steeper declines, while respondents at higher LSM levels showed higher growth rates on *social*. All LSM levels showed similar rates of declines on *print5,* with higher LSM levels starting at higher levels.

5.4 R Code for Chapter 5

The R code used to identify the *repertoires* can be found at the following url:

- *https://raw.githubusercontent.com/hanspeter6/amps_level2/master/explore_sem_pooled.R*

R code used to run the 1000 bootstrapped SEMs and manually determine EMMs can be found at the following url:

- *https://raw.githubusercontent.com/hanspeter6/amps_level2/master/boot_Para_17Sept.R*

R code used to visualise the EMMs can be found at the following url:

- *https://raw.githubusercontent.com/hanspeter6/amps_level2/master/boot_outs.R*

Chapter 6

Exploring Changes in Media Type

6.1 Chapter Introduction

The aim of this chapter is to explore changes of media as *type*, in particular how engagement by demographic categories have changed over the period of study on *newspapers, magazines, television, radio,* and the *internet*. This results from this chapter are seen as complementary to the *repertoire* analysis undertaken in chapter 5. The chapter starts in section 6.2.1 with a description of the data that was used and then briefly outlines the main methods and techniques that were applied to the data in section 6.2.2. These included some exploratory techniques as well as cluster analysis and EMM.

6.2 Data and Methods

6.2.1 Data

The dataset used in this chapter consists of 126 726 records comprising the combined 2002-2014 AMPS survey results. The variables are the same six demographic variables as was used in chapter 5 (sex, age, race, education, household income, and LSM), also aggregated as described in 3.3.1; and the five media *type* variables (*newspapers, magazines, tv, radio, internet)* with the combined *all* variable as described in section 3.3.3. It is important to note that while the media *vehicle* variables used in chapter 5 consisted only of those *national* media *vehicles* that all the years in this study had in common, these composite *type* variables were determined using all the media *vehicles* that were included in the sets, as described in section 3.3.2.

6.2.2 Methods

To gain a better understanding of the data, some exploratory and descriptive techniques were first applied. The proportions of the levels of different categorical variables were examined to

consider differences over the time period of this analysis. This was followed by determining correlations between the media *types* to gain an understanding of the strength and direction of linear relationships and how these may have changed over time.

Following the basic exploratory descriptive analyses, the data was considered for clusters, using a *k-means* clustering algorithm (see section 4.5), which were interpreted with the aid of graphical displays.

To examine changes of media *type* by different demographic categories and levels over the period under review, estimated marginal means (EMM), derived from linear regression models in which demographic variables and years - with years interacting with demographic variables - were regressed as predictors on the five media *type* variables as outcomes. These results, with estimation errors, were then graphically displayed to aid interpretation (details on the EMM can be reviewed in section 4.6). The R package *emmeans* was used to estimate the EMMs and to estimate their 95% confidence intervals.

6.3 Results

6.3.1 Descriptive Statistics

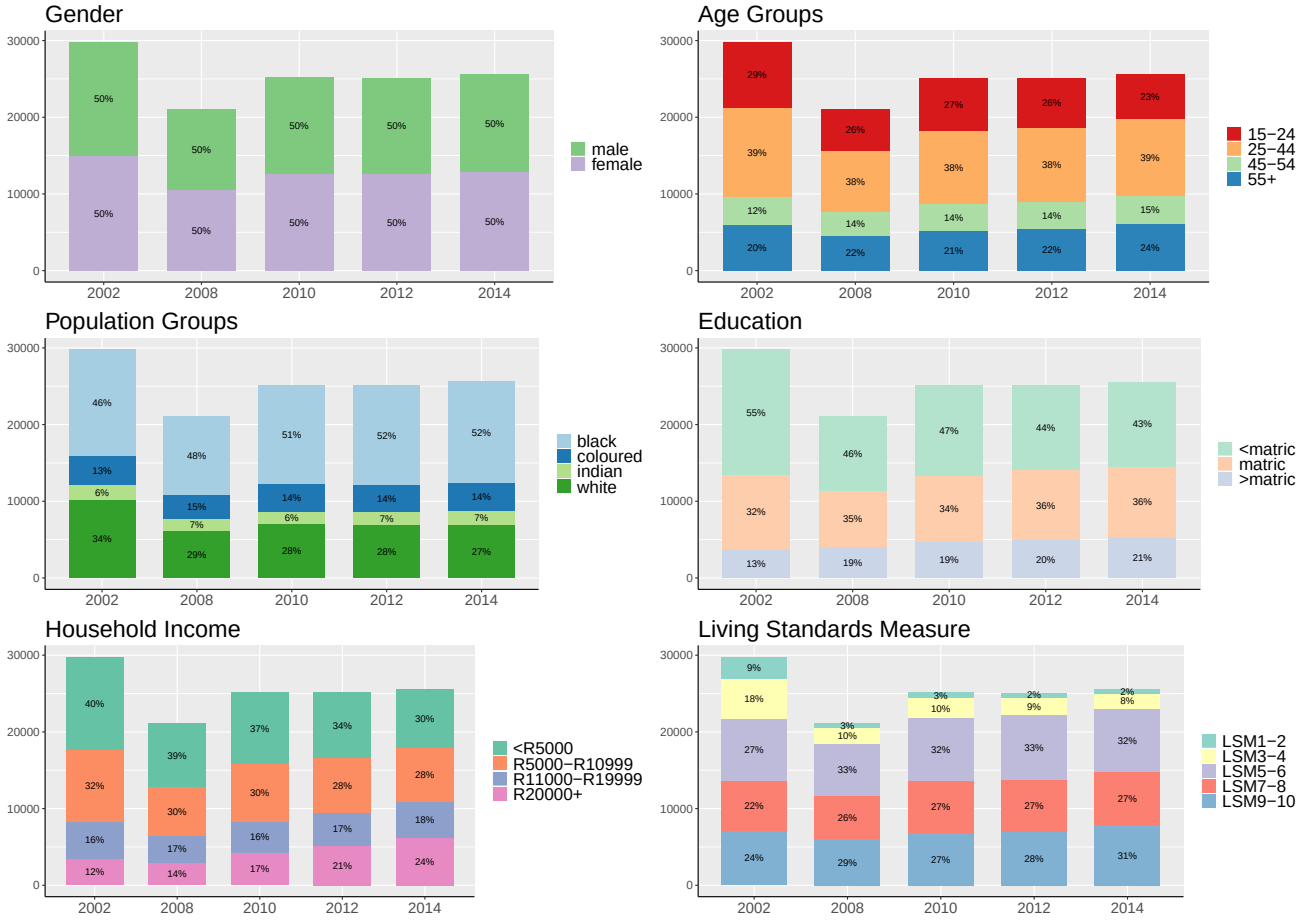
6.3.1.1 Proportions of Demographic Levels

From figure 6.1 it would appear that although the total number of respondents varied somewhat - from a low of 21 083 in 2008 to a high of 29 791 in 2002 - the relative proportion of levels remained reasonably consistent over time. This is particularly true with regard to gender and age. There are however also some marked differences: in population groups whites accounted for 34% of the sample in 2002 compared with 27% in 2014; in education those with post-matric qualifications increased from 13% to 21% over the study period, while those with less than matric decreased from 55% to 43%; household income also showed differences, with those earning less than R10 000 a month decreasing from 72% to 58% and those earning more increasing from 28% to 42%; for LSM, the lower groupings LSM 1-4 decreased from 27% to 10%, while LSM 7-10 increased from 46% to 58%. The descriptions of AMPS sampling methodology in section 3.2 suggests that where these differences occur, they are the result of estimated changes in the sample universe and therefore meaningful.

6.3.1.2 Correlations

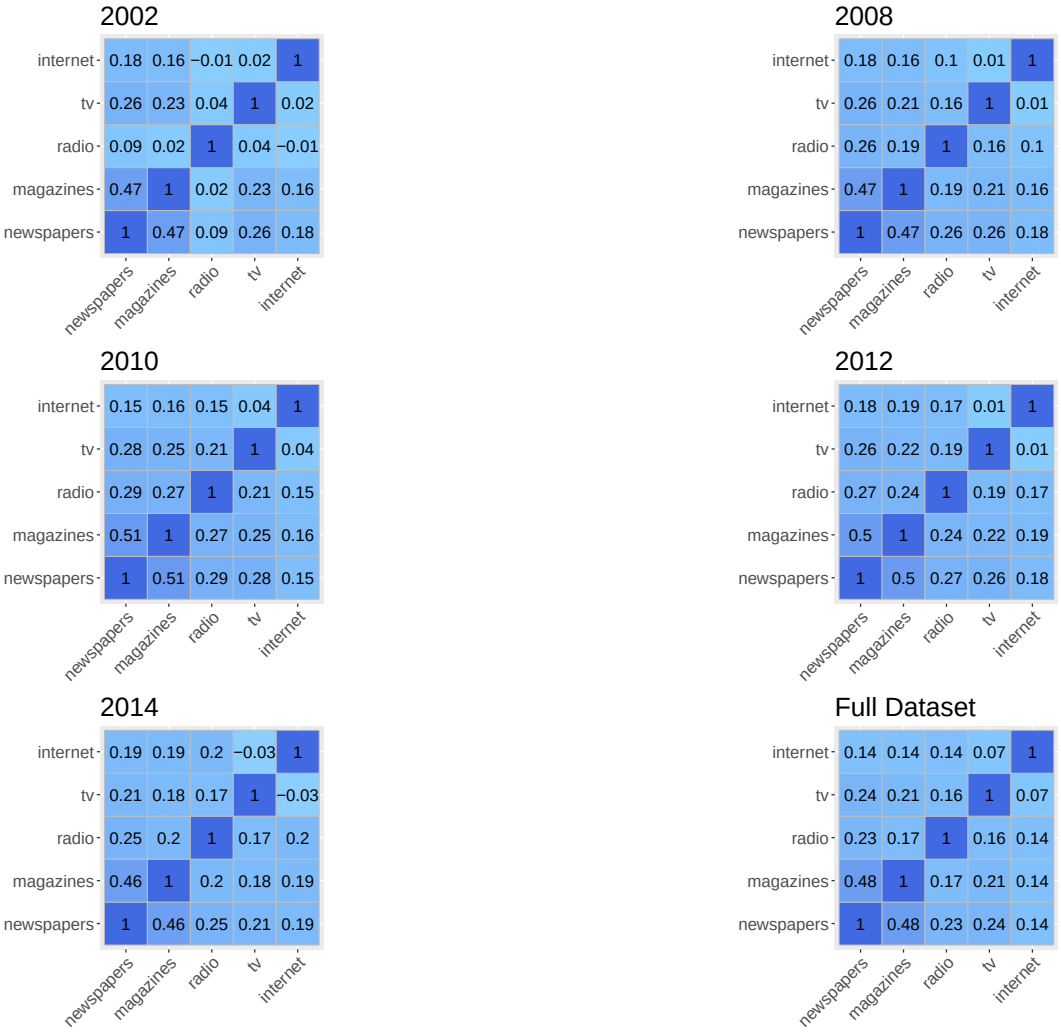
To gain a sense of the relationship between media *types* and to consider changes of these relationships over time, within-year correlations between engagement on each of the media *types* was considered. The resulting correlations matrices are displayed in figure 6.2. The correlation matrices generally show strong correlations between newspapers and magazines. Newspapers are also generally more correlated with other media, with the marked exception

Figure 6.1: Sample Proportions by Demographic and by Year: Counts and Percentages



of radio. Also, TV engagement appears to be uncorrelated with internet. Radio engagement in 2002 also correlated poorly with internet, but contrary to TV, this relationship appears to strengthen slightly over the period under review. Engagement with newspapers also appears to correlate most strongly with all other media types, while internet engagement correlates poorly with all other media types.

Figure 6.2: Correlations of Engagement Between Different Types of Media by Year and for the Full Dataset: Darker shades of blue signify higher positive correlations



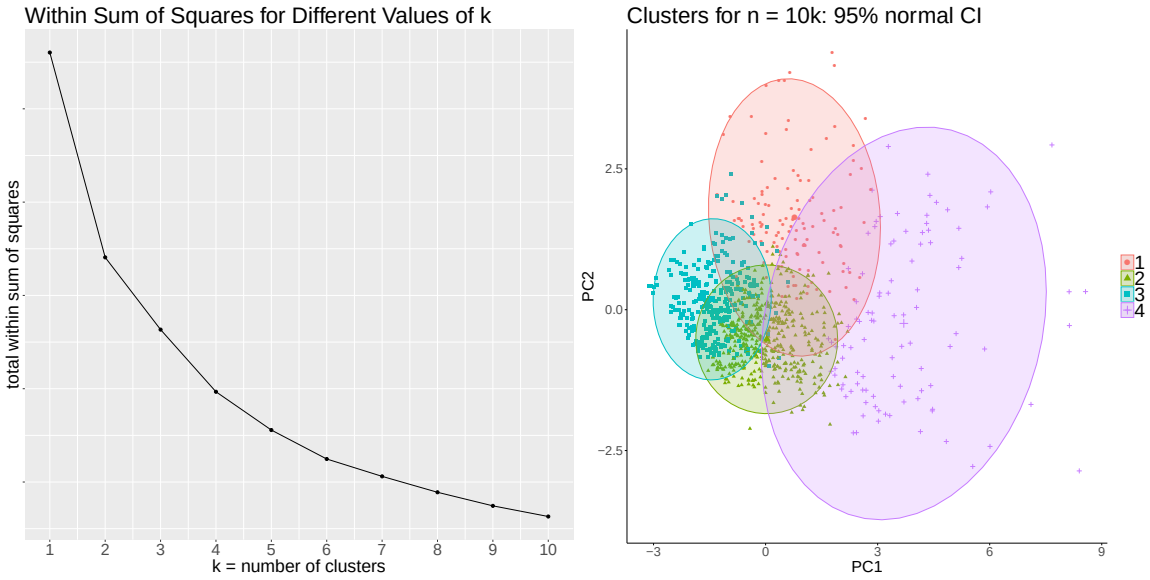
6.3.2 Clustering

The left-hand plot in figure 6.3 is the result of applying *k-means* clustering for different values of k and shows declining values of total within cluster sum of squares (wss) for increasing values of k, the number of cluster centers. The screeplot does not suggest a very clear candidate for the number of clusters. After considering both four and six clusters, four was selected due to a more balanced distribution of cases on the clusters.

After labelling each respondent in the dataset according to their membership of the four

clusters, this separation was visualised by randomly selecting 10 000 respondents and plotting the scores of the first and second principal components, including 95% normal confidence intervals. This visualisation is shown in the right-hand plot of figure 6.3 and confirms the considerable degree of overlap while at the same time illustrating four quite separate medoids.

Figure 6.3: K-means Clustering Scree Plot and Sample Visualisation



6.3.2.1 Interpreting Clusters

To interpret the meaning of cluster membership, as shown in table 6.1, box plots for different media *types* were plotted by cluster as shown in figure 6.4. Demographic profiles of clusters were determined with the aid of plots in figure 6.5.

The box-plots in 6.4 were used to interpret the clusters and to give each cluster a descriptive name:

1. Respondents who are members of the first (red) cluster are differentiated from the other cluster members by showing the highest median engagement on all media *types* with the exception of *internet* where it showed the second highest - hence the description of this cluster as ***"heavy"***.

2. Respondents who are allocated as members of the second (green) cluster are differentiated from those in other clusters primary due to their considerably higher median engagement on *internet*. When considered by year (not shown) members of this cluster also showed early adoption of *internet* as media type - hence a descriptor of ***"internet"***.

3. Respondents allocated as members on the third (blue) cluster are somewhere in the middle between high and median engagement for all media *types* - hence the descriptor ***"medium"***.

4. Respondents allocated as members on the fourth (lilac) cluster are differentiated from other clusters primarily by the fact that they show the lowest median engagement for all media *types* - hence a descriptor of ***"light"*** in direct contrast to ***"heavy"***.

Table 6.1: Media *Type* Clusters Summary Demographic Profiles

Cluster	Descriptor	Sample %	Summary Demographic Profile: based on the plots in figure 6.5
1 (Red)	"Heavy"	11.89%	More male, younger (15-44), mainly black, matric, all incomes, LSM 5-10.
2 (Green)	"Internet"	12.08%	More male, younger(15-44), more white, matric and post matric, higher income, higher LSM, especially LSM 9-10.
3 (Blue)	"Medium"	43.73%	More female, both 15-44 and 55+, mainly black then white, lower than matric, low income, mainly LSM 5-6
4 (Lilac)	"Light"	32.29%	Slightly more female, 15-44, more black then white, lower than matric, low income, across LSM groups, most strongly represented in LSM1-2.

Considering changes over time shown in figure 6.6 it would appear that the relative proportion of media *type* in the samples reflect the growing importance of the internet as a medium for information and news. This increase in the cluster size for *internet* (from 1% of the sample in 2002 to 31% in 2014) contrasts with the decrease in all three the remaining clusters: the cluster *heavy* decreasing from a high of 14% to a low of 9%, *medium* from a high of 48% to a low of 36%, and *light* showing a drop from 2002 (42%) to 2008 (33%), with a further drop to 24% in 2014. From these figures it would appear that the growth in internet engagement has disrupted all other media categories in South Africa over the period under review.

6.3.3 Estimated Marginal Means on Models of Engagement in Different Media Types from 2002-2014

The EMM values based on both *equal* and *proportional* weighting for the predicted values by demographic level and conditioned on *year* were plotted by media *type* to provide visual means of interpreting the differences between demographic levels and media *type* engagements over the period under consideration. 95% confidence interval error bars were added to the plots.

6.3.3.1 Radio

The following interpretations are based on figure 6.7 below. The vertical bars in the plot indicate the 95% bootstrapped confidence intervals from the EMM procedures described in section 4.6:

Figure 6.4: Box-plots per Media Type and by Cluster

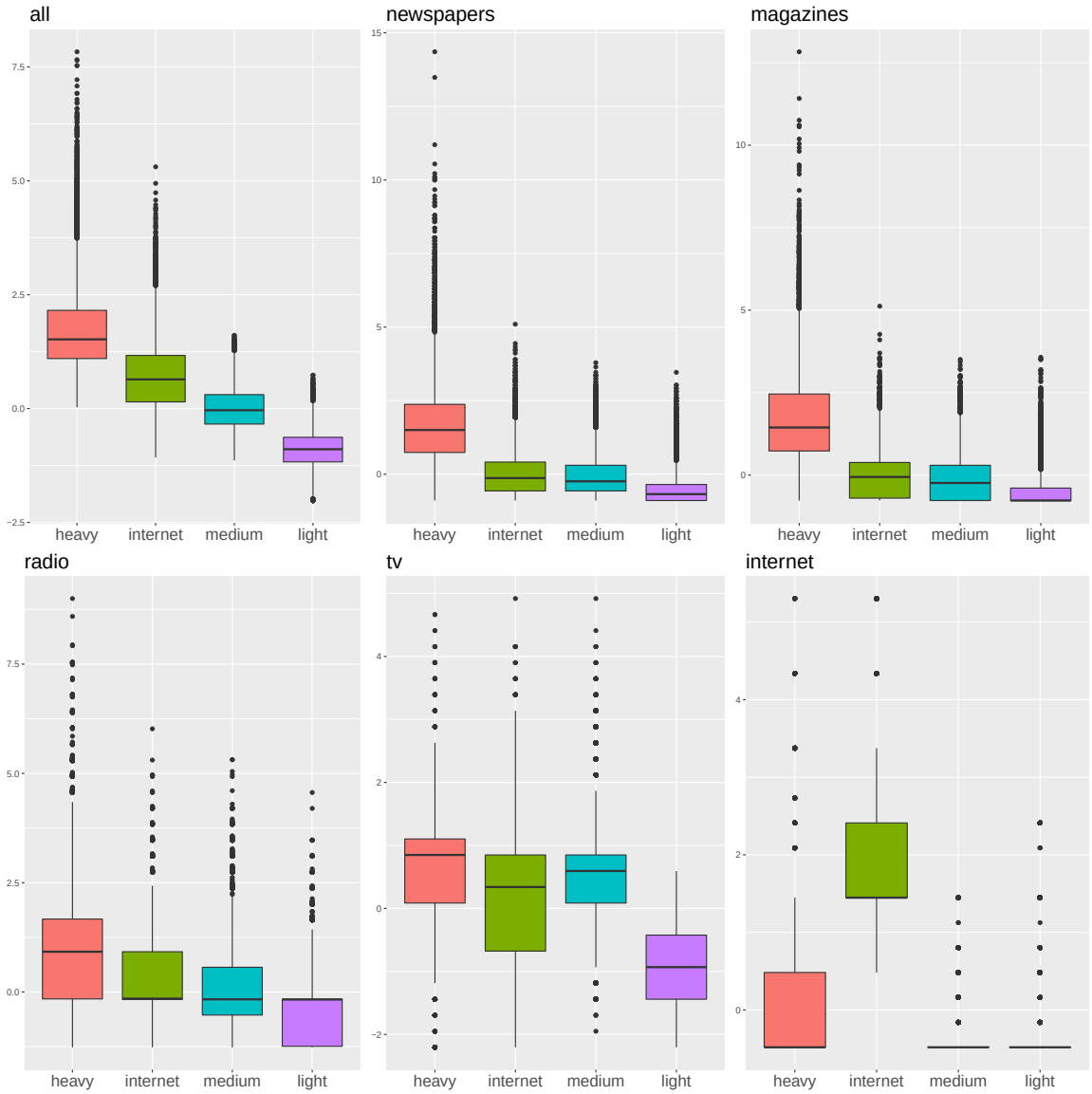


Figure 6.5: Demographics Categories by Count and by Cluster

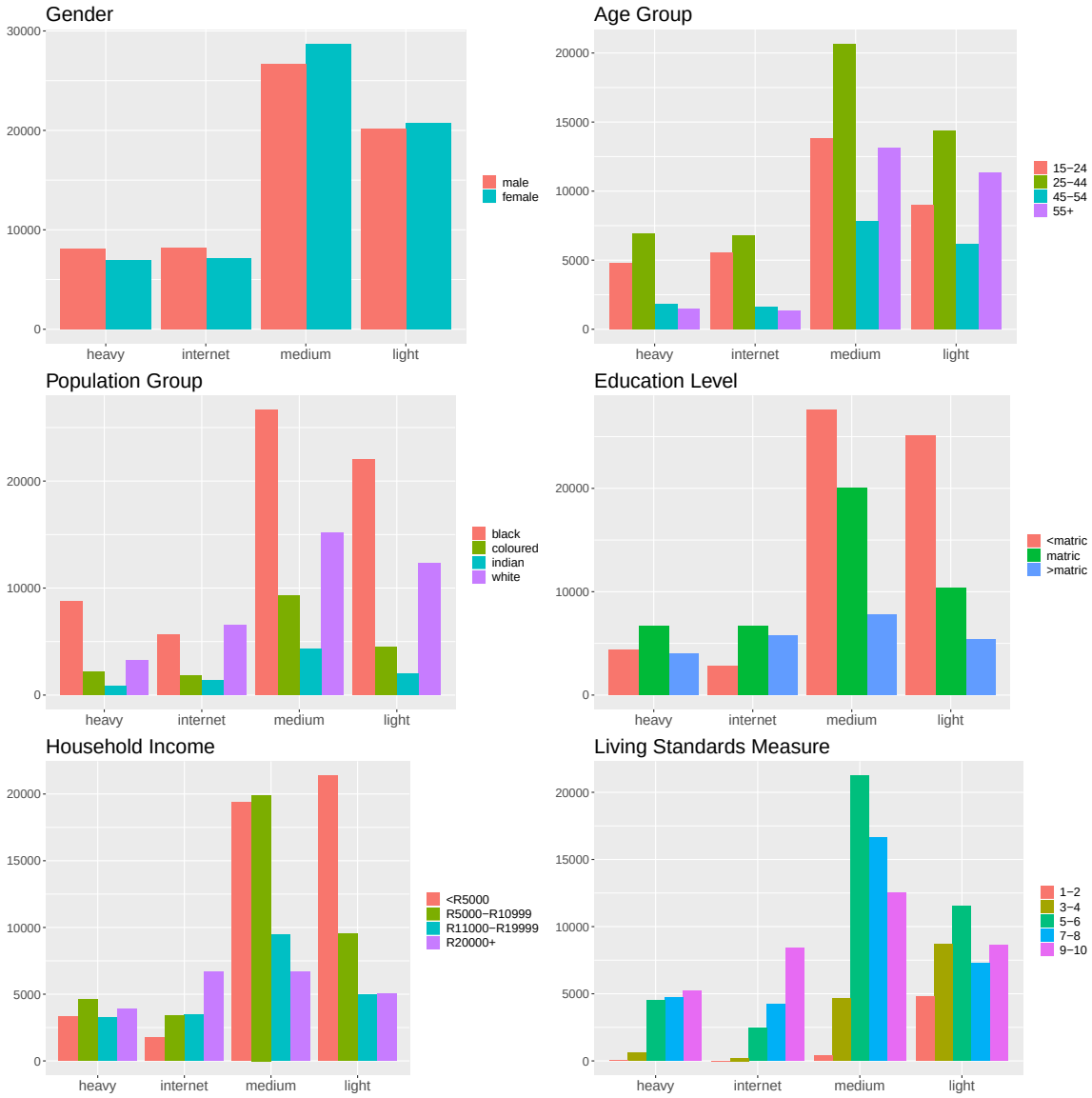
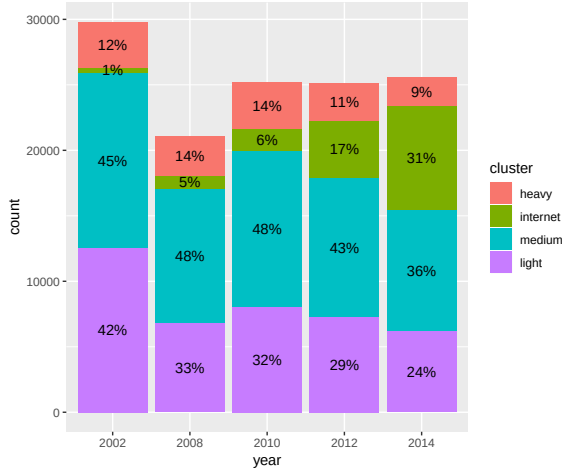


Figure 6.6: Cluster Proportions by Year



In all cases the profiles for *proportional* and *equally* weighted estimated marginal means closely resemble one another, although their levels differ. The difference in levels can be understood by the largest coefficients in a linear model being based on race (especially for the reference category "black") and secondly on age (especially under 25). Furthermore, the proportion of "black" used in the *proportional* weighting is nearly double what it would be for *equal*. This would also explain why the differences between *proportional* and *equal* for race are less than for the other levels.

In most cases it would appear that engagement in radio increased somewhat between 2002 and 2008, but showed a steady decline since then. Given the dearth of appropriate surveys between 2002 and 2008, it is difficult to assume that this increase is valid. The increase in proportional engagement on radio over this period is however supported by figure 2.1. Although males show a slightly higher level of engagement than females, the profiles of initial growth with subsequent decline pose 2008 are very similar. The profiles for different age groups show markedly different levels with engagement in radio, generally declining with age. The decline in radio engagement since 2008 appears to be less marked among older respondents than among the 15-24 year olds, with respondents older than 45 showing some degree of stabilisation, albeit at a lower level that younger age groups. It would appear that among different population groups the very steep growths from 2002 to 2008 are not reflected among black and indian respondents. Levels of radio engagement among black respondents are considerably higher than among any other population groups. Although black, coloured and indian respondents all showed similar declines in radio engagement since 2008, white respondents appear to show some degree of stabilising and even some growth from 2010 to 2014. Education, income and LSM levels show quite similar patterns of engagement. The higher the category level, the higher the level of engagement; and all showed a similar rate of decline since 2008, with post matric, incomes lower than R2 500 and LSM 9-10 indicating a somewhat slower decline than the other levels.

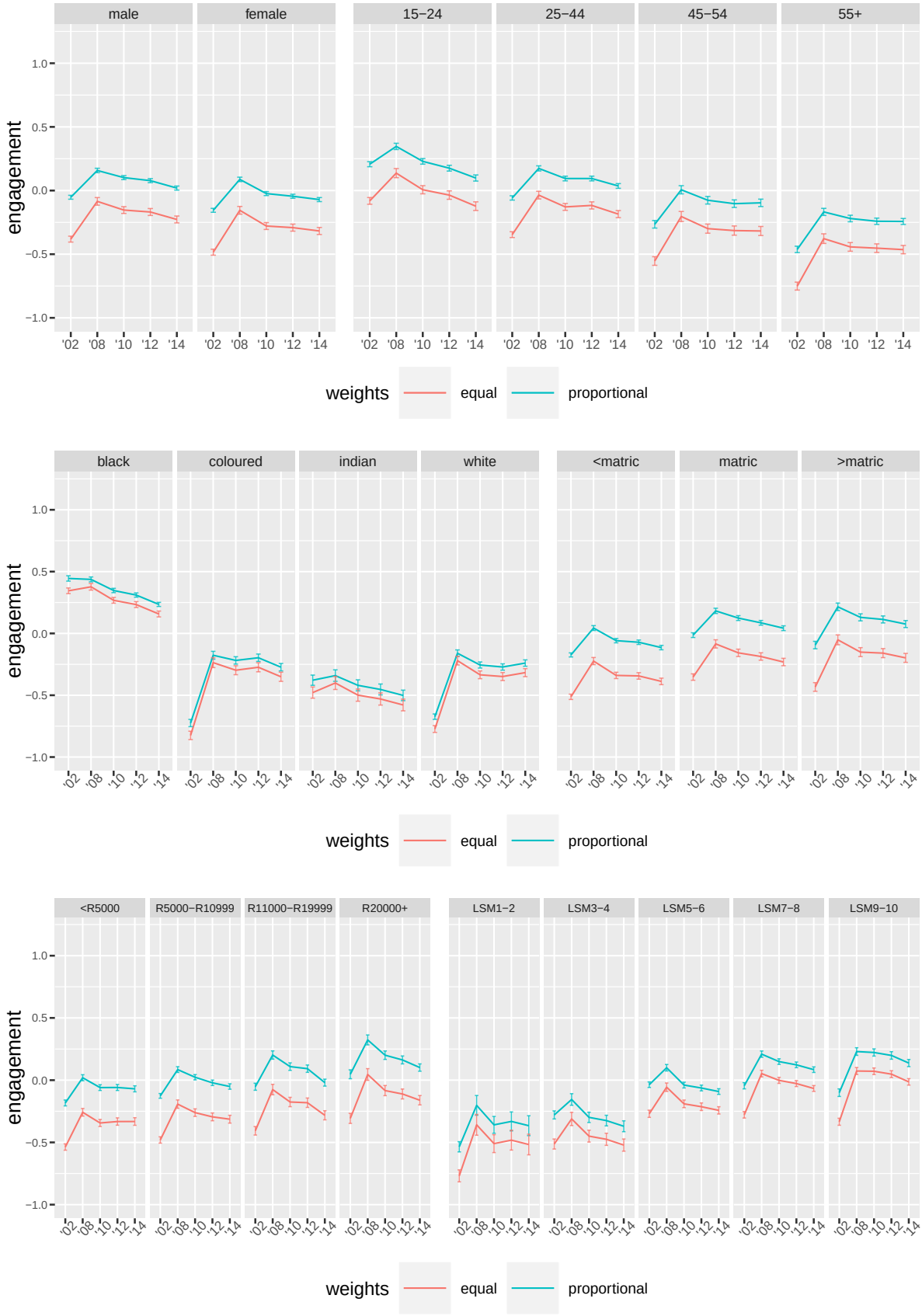
6.3.3.2 Newspapers

The following interpretations are based on figure 6.8. The vertical bars in the plot indicate the 95% bootstrapped confidence intervals from the EMM procedures described in section 4.6:

As for radio, the difference in weighting does not materially affect the interpretation:

Although some levels would indicate increases between 2002 and 2008, most markedly among black respondents and possibly primarily driven by the launch of tabloids, the Daily Sun and The Times, the marginal means generally show quite dramatic declines in engagement levels for newspapers - especially since 2008. The profiles between males and females show quite similar rates of decline, although the levels for males are higher than for females. A similar interpretation would hold for the different age groups, although it would appear that the 15-24 year-old respondents show generally lower levels and steeper declines of engagement

Figure 6.7: Estimated Marginal Means with error bars showing 95% confidence intervals: Radio



than the other age groups. Among population groups, the profiles suggest considerably lower levels of engagement among white respondents, but the same declines across all groups. As in radio, the profiles for newspaper engagement appear quite similar for education, income and LSM groupings. In all cases, the levels of engagement increase quite sharply with higher levels of education, income and LSM, but the relatively steep declines are apparent for all levels with the possible exception of lower levels of income and LSM, which appear to decline somewhat less steeply than for other levels in the category.

6.3.3.3 Television

The following interpretations are based on figure 6.9 below. The vertical bars in the plot indicate the 95% bootstrapped confidence intervals from the EMM procedures described in section 4.6:

Based on the arguments presented in section 5.3.2.1 in which the differences between *proportional* and *equal* weightings are described as being driven primarily by LSM, the *proportional* values were used in this interpretation.

Generally, engagement in TV appears to have increased over the period under review with the exception of indian and white respondents as well as respondents classified as LSM 1-2. The profiles for LSM 1-2 are not visible in the plots since they fall below the engagement axis limits. The limits were set also in recognition of the fact that the definition of membership to these levels implies that respondents do not have television sets in their homes. Female respondents appear to become more engaged over the period than males. And younger respondents do not engage as much as older respondents, although increasing levels of engagement over all age groups is apparent. Population groups differ quite markedly with regard to their engagement on TV: Black respondents show the highest increase and highest level of engagement, while coloured respondents, although also exhibiting some growth have not changed their levels of engagement as much as black respondents. indian and white respondents appear to have remained relatively stagnant or even shown some declines in levels of engagement. It would appear that respondents with post-matric qualifications did not grow their engagement as much as those with matric or less than matric. This also appears to be the pattern for levels of household income, although all levels exhibit increases, the higher income levels have grown more slowly over the period of the study. LSM appears to be a good indicator of TV engagement. LSM 3-4 show initial declines but then increases over the period under review; LSM 5-6 showing growth from relatively high levels, but not as much as those for LSM 7-8. Respondents categorised as LSM 9-10 appear to show lower levels than those from LSM 7-8 as well as slower growth, even leveling off from around 2010.

6.3.3.4 Magazines

The following interpretations are based on figure 6.10 below. The vertical bars in the plot indicate the 95% bootstrapped confidence intervals from the EMM procedures described in

Figure 6.8: Estimated Marginal Means with error bars showing 95% confidence intervals: Newspapers

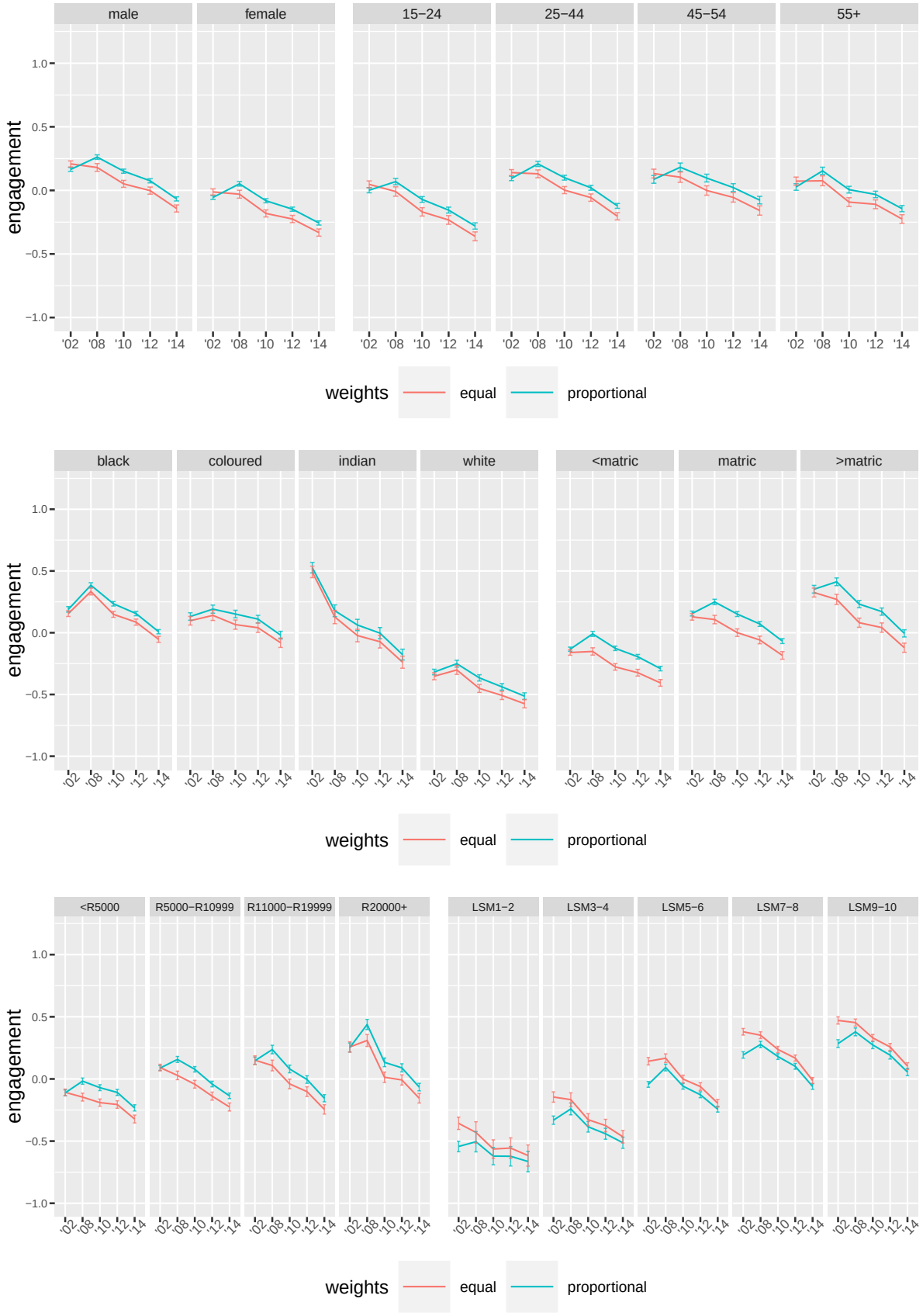
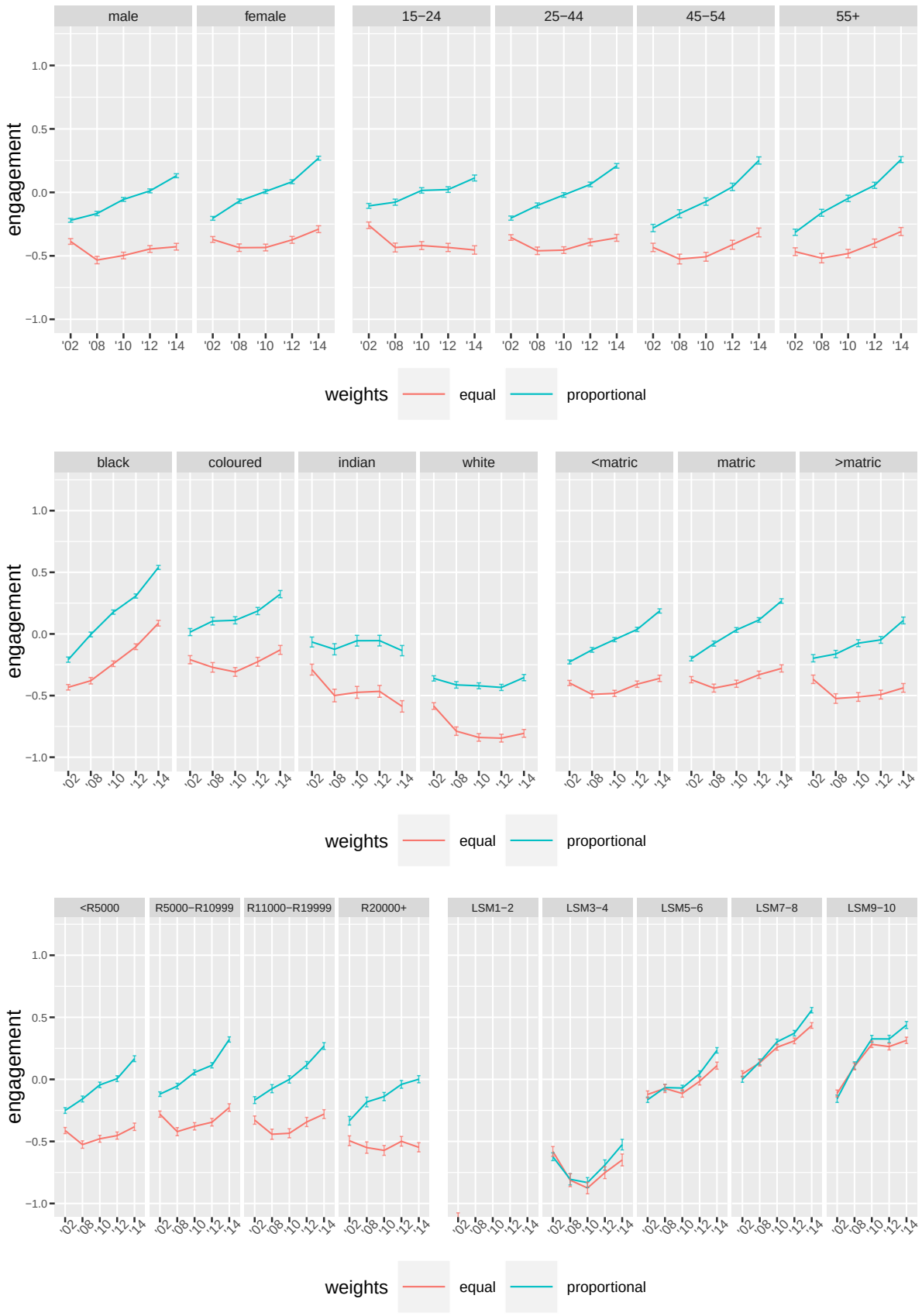


Figure 6.9: Estimated Marginal Means with error bars showing 95% confidence intervals: Television



section 4.6:

For engagement in magazines, the differences between *proportional* and *equal* weightings do not substantively change the interpretations. Furthermore, the close similarity between these weightings for LSM compared with other categories and levels would suggest (as for TV above) that LSM offers stronger predictors of magazine engagement than others.

In most cases the decline in engagement seen for newspapers is also apparent in magazines. The plots suggest that females have a generally higher engagement on magazines than males and that the decline appears to be less steep. From high levels for 15-24 year olds, this age group also shows the steepest decline in engagement levels. Although the levels of engagement for age groups older than 45 appear lower than for the younger brackets, they show a degree of stability compared to declines from higher engagement levels for the younger brackets. While black and coloured respondents appear to show quite stable profiles, those of indian and white respondents suggest steep declines, although white respondents started off at higher levels of engagement. For education, income and LSM groupings, the pattern of higher levels of engagement in magazines for higher levels of the categories appears to be consistent with previous media *types*. The declines in engagement also appear to be quite consistent, with the exception of respondents at the lowest levels of income and LSM.

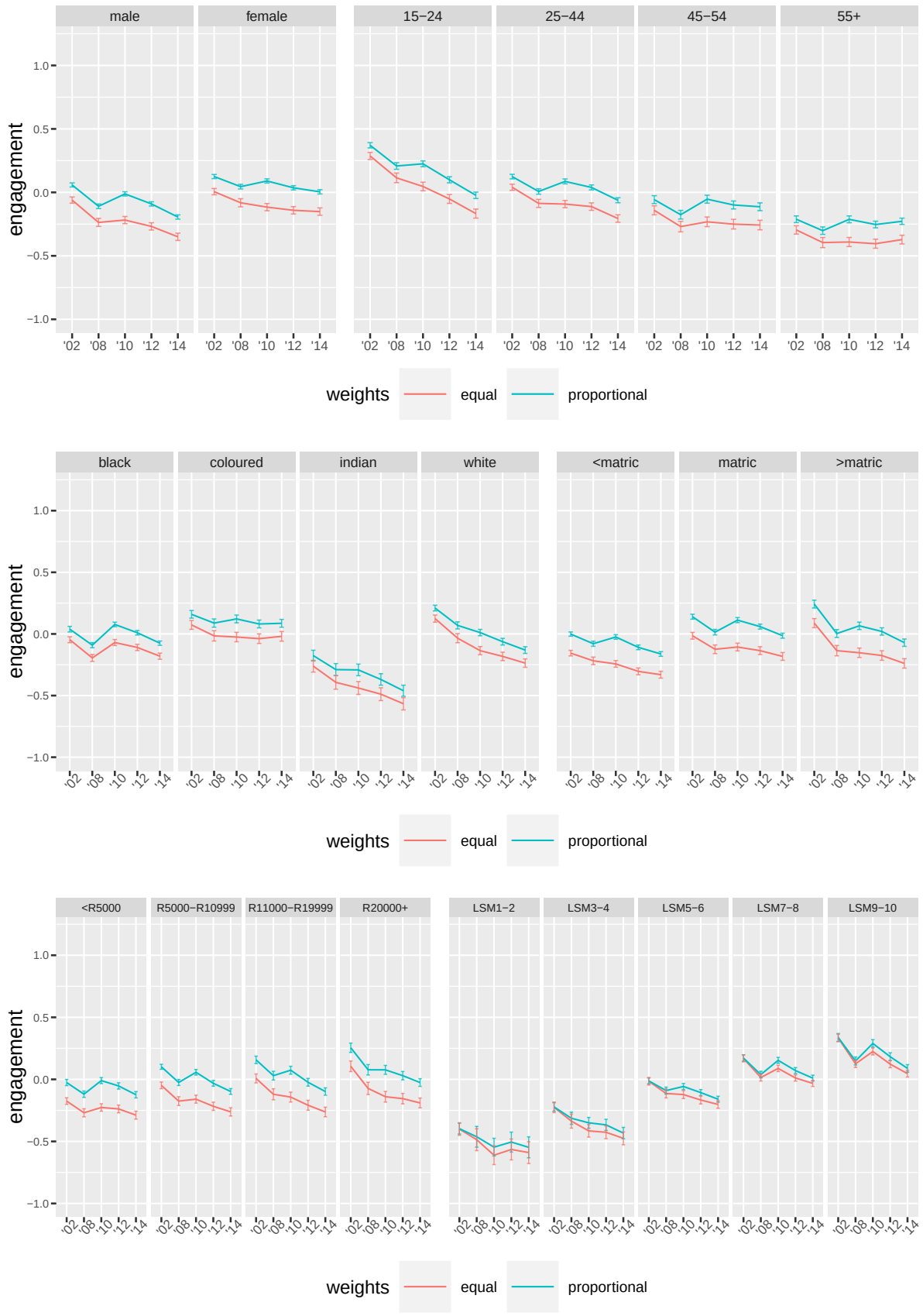
6.3.3.5 Internet

The following interpretations are based on figure 6.11 below. The vertical bars in the plot indicate the 95% bootstrapped confidence intervals from the EMM procedures described in section 4.6:

From an inspection of coefficients on linear models with internet engagement as outcome variable, it would appear that the largest influences on internet engagement are age and years. This perspective is supported by the plots below with what appear to be steep changes over time and with the most obvious changes in relative steepness of the profiles being those for different age groups. And the profiles shown for *proportional* and *equal* weightings in the case of engagement on internet are similar in terms of both level and profile.

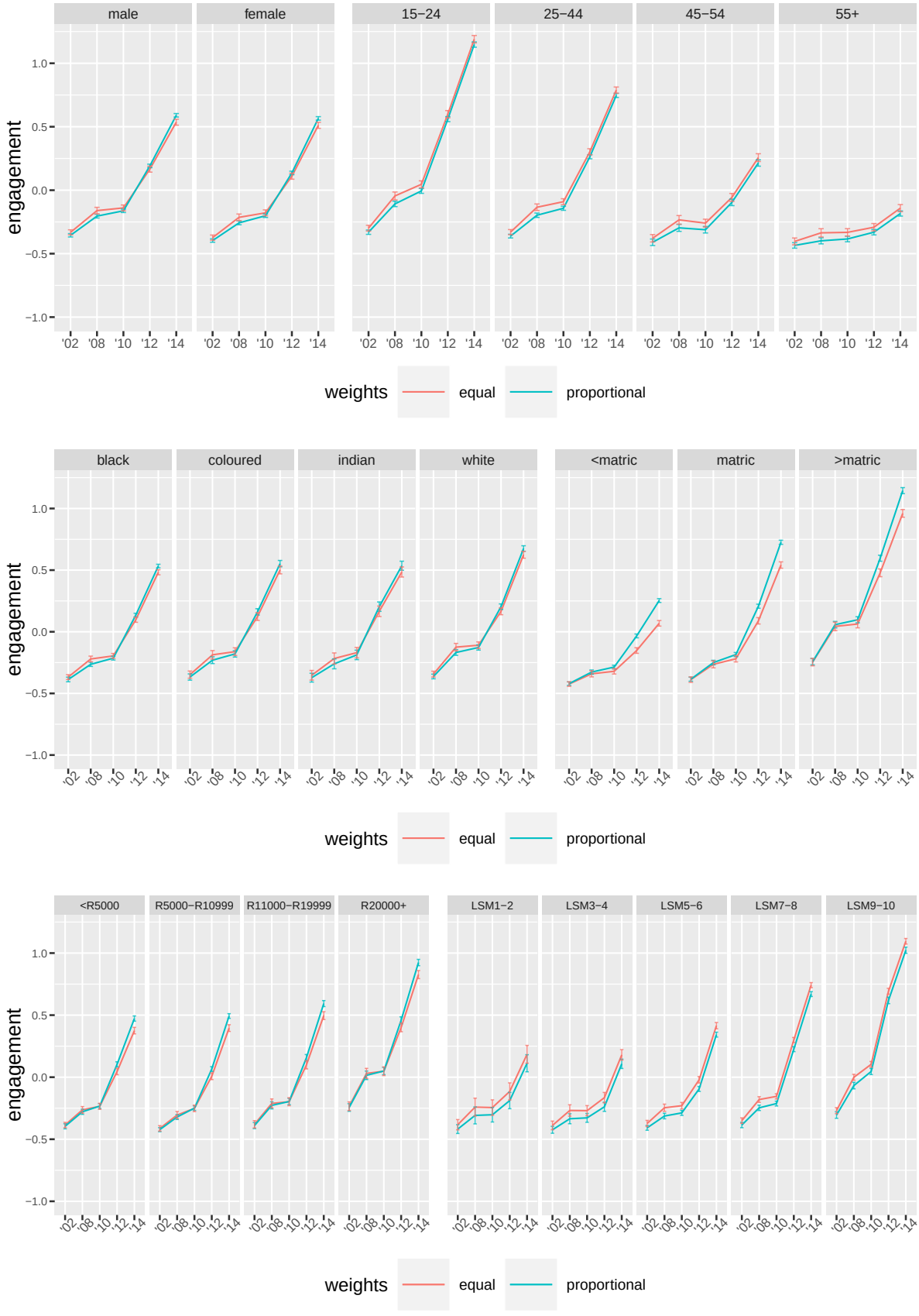
Engagement on internet shows generally consistent and steep increases over the period of this analysis. The only exception appears to be for the 55+ age group, where the adoption of the internet appears to be growing at a markedly lower rate than for other age groups. An interesting feature in the profiles is the apparent slowing down between 2008 and 2010, which could be explained by the economic recession as well as the growth in other media *types* such as radio, tv and print. Furthermore, the full effect of the expansion of smart phones and the extension of wifi may well account for the very steep rises post 2008. While most levels in the various categories start out at similarly low levels, some categories attain considerably higher levels of engagement. Most markedly in this regard are the 15-24 year age group, the post-matric education level and the upper LSM brackets. The profiles and levels for gender and population groupings appear very similar, suggesting that these demographic categories

Figure 6.10: Estimated Marginal Means with error bars showing 95% confidence intervals: Magazines



are not good predictors of internet engagement.

Figure 6.11: Estimated Marginal Means with error bars showing 95% confidence intervals: Internet



6.3.4 Summary Discussion of Results by Demographic Category

It seems clear, and not entirely unexpected, that while engagement in print showed general declines across all demographic categories, internet engagement grew steeply. This summary considers the six categorical variables separately.

6.3.4.1 Gender

Both male and female respondents' engagement in print media and radio declined at similar rates over the period of study. This decline was more apparent for newspapers than for magazines and radio. The levels of engagement between male and female respondents indicate slightly higher levels for males on radio and newspapers, with slightly higher levels for females on magazines and the internet. In contrast, both male and female respondents showed increases in engagement on TV, although female respondents showed steeper growth on TV, and considerably steeper increases on the internet over the period. Both genders showed very similar levels and profiles of engagement on the internet.

6.3.4.2 Age

As for gender, all age groups showed declines in engagement in print and radio. These declines however differed quite significantly by age groups in terms of both engagement levels and profiles of decline. Younger respondents generally showed higher levels of engagement than older ones for both radio and magazines, but the profiles of decline are considerably steeper for younger respondents than for older age groupings, with the 55+ age group showing some stability. Also, similar to gender, all age groupings showed increases in engagement on TV and more extreme increases on the internet. While older age groups started at lower levels for TV, they showed steeper growth over the period under review. With regard to internet engagement, the plots would suggest quite considerable differences in the rates of change for different age brackets. With similar levels in 2002, the younger two age groups (in particular the 15-24 age group) far exceeded the increases shown for the older two age groups.

6.3.4.3 Population Group

With the marked exception of the internet, where there is little differentiation on the basis of race in a steep increase of engagement over the period, different population groups responded quite differently to engagement on other media types over the period under review. For magazines, white respondents showed the highest initial levels of engagement while indian respondents the lowest. Both these groups then showed similar steep declines. In contrast, both black and coloured respondents appear to be reasonably stable in their engagement on magazines. With regard to newspapers, indian respondents showed the highest and white

respondents the lowest initial levels but both showed similar steep declines over the period under review. Black and coloured respondents also showed declines, but remained at higher levels of engagement than white and indian respondents. For radio, black respondents show considerably higher levels of engagement than the other population groups. And all groups with the exception of white respondents showed similar declines over the period. TV engagement among black respondents showed steep increases from a relatively low base. Coloured respondents also increased but less steeply. indian and white respondents remained relatively stable, although the engagement of white respondents was considerably lower than for the other groupings.

6.3.4.4 Education

Respondents from different levels of education showed quite similar profiles of decline on radio since 2008, although the levels for those with lower than matric were also slightly lower than others. These declines since 2008 are also very similar for newspapers and for magazines, with marked differences in levels: higher levels of education generally showed higher levels of engagement on the media *type.* For television, different levels of education showed similar starting levels, all showed quite consistent growth, although post-matric respondents grew slightly slower than the other two levels. For internet, post-matric respondents showed higher initial levels and steeper growth than other levels, while respondents with matric started at similar levels to those with lower levels of education, but exhibited steeper growth profiles.

6.3.4.5 Household Income

All household income levels showed similar profiles of decline on radio, although slight increases in levels of engagement are apparent for higher levels of income. For newspapers and magazines, all levels of income also showed declines, although higher levels of income showed higher engagement levels, they also exhibited steeper declines. For television middle income groups showed similar starting levels and similar growth, the lowest and highest brackets in turn showed lower starting levels and lower rates of growth. For internet engagement those respondents earning household incomes less than R20 000 showed similar starting levels and only slight differences in the steepness of growth, with higher levels slightly higher than lower levels. Respondents above R20 000 showed higher starting levels and also steeper growth profiles than the other levels.

6.3.4.6 Living Standards Measure

For radio all levels showed similar initial growth until 2008, followed by some declines, with the levels increasing by increasing LSM levels. Both newspapers and magazines also exhibited similar profiles for all LSM brackets, but with levels increasing with LSM levels. Newspapers appear to show steeper declines than magazines. For television LSM 3-4 show some declines

until 2010, followed by growth, LSM 5-8 exhibit similar growth patterns, but LSM 7-8 start at higher levels and show slightly steeper growth, and LSM 9-10 start at lower levels than LSM 5-8, but show quite steep growth until 2010, when the growth appears to have slowed down. For internet all groups under LSM 9-10 appeared to start at the same levels, but the steepness of growth appears to be steeper for progressively higher levels of LSM. LSM 9-10 is differentiated by higher starting levels and also steeper growth than other levels.

6.3.5 Code for Chapter 6

The R code for this chapter can be found at the following url:

- *https://raw.githubusercontent.com/hanspeter6/explore_type/master/explore_type_simple_print.R*

Chapter 7

Conclusion

The objective of this chapter is to provide a general conclusion to the project by presenting the most salient features of the results and to consider limitations and suggestions for future research.

This project set out to explore, in predominantly two different but complementary approaches, how the levels of media engagement (as in the *degree* or *intensity* of engagement) in South Africa changed between 2002 and 2014. The two approaches were: first, considering how aggregate degrees of engagement for different demographic categories changed for different media *types* over this period; and secondly, examining how respondents' aggregate degrees of engagement with latent factors of media *vehicles* , spanned various media *types* - identified as *repertoires*, changed over the period of study. This project has achieved the aims and the objectives described in chapter 1.

Xu et al. (2014, citing Gentzkow and Shapiro, 2011; Ksiazek, Malthouse and Webster, 2010) identified that "one of the main research questions" in the media world today, was how audiences responded to an "explosion of choice" in a world characterised by a "proliferation of media". They used as an example: "does it lead to people consuming a steady and consistent diet of their preferred news genre or do they expand their consumption to a wider, more diverse range of sources"? Although this study has provided many contributions to the bigger question of how South African audiences have responded during the period of study, the answer to this specific question appears to be that South Africans are divided on this matter: on the one extreme it would appear that many have remained loyal to a media diet dominated by free-to-air television; while on the other hand, a smaller group have largely swapped their traditional fare for internet as a dominant channel for news.

This study also compares well - in terms of results if not in terms of methods - with that of Edgerly (2015, p. 2), who used a national survey of media use by U.S. adults to identify six distinct news *repertoires*. She found that some were ideologically based, spanning multiple media platforms (*types)*, others functioned at a media level, and others again showed respondents who consumed bipartisan news. This study, which also identified six *repertoires,* found that some were also "ideologically" defined, which in the South African context found

expression in terms of race or culture in the *repertoires african* and *afrikaans.* And that these, like those of Edgerly (2015, p. 2), span multiple media *types.* This study also found support for Edgerley's (2015, p. 2) *repertoires* that function at a media level in identifying *freeTV* and *intnews* which function respectively at predominantly television and internet news levels. Finally, while her study found *repertoires* that showed respondents who consumed bipartisan news, this study found similar indiscriminate media diets in the *repertoire* referred to as *print5.*

The findings of this project however do not support the findings of Schröder (2015, pp. 70-71)'s study into *repertoires* in Denmark, which identified seven types of news consumers who all used a mixture of traditional and new sources of news. This project in contrast found that not all South African media consumers used mixtures of traditional and new sources of news, but that - while the internet has undoubtedly had a disruptive effect on all traditional media - many consumers have remained largely loyal to tradition media, especially to free-to-air television.

Fulton (2017)'s figure 2.1 on page 13 indicates the relative proportions of South Africans engaging with different media *types* over a period that largely spans the period of this study. While his trends showed relative stability or even some growth, this study has made two important contributions: firstly, that while proportions may show relative stability, intensity or the degree of engagements do not. For example, the proportion of South Africans engaging with free-to-air television remained relatively stable in figure 2.1, while the intensity or regularity of that engagement has shown declines since 2010. Another contribution of this study is to show that these changes in engagement can vary quite substantially for different demographic categories and levels.

7.1 Radio

Figure 2.1 reminds readers that by far the largest proportion of South Africans engaged with radio and television (excluding DSTV) over the period 2004 to 2015 (note the dates here refer to release dates and not study period, so for example 2014 in this study would be indicated as 2015 in figure 2.1) and that the proportion of people engaging on radio from a high of around 92% in 2002 saw growth to close to 100% before declining again to around 92% between 2004 and 2009. After 2009 the proportions appeared relatively stable until the end of the period. This study showed increases in radio engagement between 2002 and 2008, coinciding with the growing proportion of radio listeners over the same period and thus providing some explanation for the growth in engagement. That is, the increases shown likely had more to do with the fact that more people were engaging on radio than about higher levels of engagement. But this study stands in contrast to the relative stability shown in radio listenership proportions, by suggesting general declines in degrees of engagement since 2008. Furthermore, the levels of engagement appeared generally higher for younger people,

black respondents, and respondents who are better educated, from higher income household and higher LSM brackets. In contrast to television, the *repertoires* identified in this study did not isolate a factor dominated by radio listenership.

7.2 Television and *freeTV*

With regard to television (excluding DSTV), figure 2.1 shows a steady increase in proportion from around 80% in 2004 to 92% at the end of the period. In this study all television is combined in the television by *type*, but since free-to-air television loaded on its own *repertoire* it was possible to compare this study's findings with the proportions. While television as *type* showed growth, with the exception of indian and white viewers, the engagement on *freeTV* showed declines, particularly since 2010 among younger, upper income bracket and LSM 7-9 respondents. The value of figure 2.1 is that it offers an explanation for this in the steep growth of the proportion respondents watching DSTV. The contrast between the steady increases and the slow declines since around 2010 on *freeTV* would suggest that although the proportion of South Africans watching free-to-air television may have increased somewhat, their intensity of engagement appears to have decreased. Differences in demographic categories are also apparent, with similar profiles of growth on television as *type* for age and education levels, but with regard to population groups, steep increases among black respondents, less steep for coloured, stable and even declining profiles for indian and white respondents. With LSM 5-8 showing reasonably consistent growth, while LSM 9-10 show a flattening off from 2010.

7.3 Newspapers and Magazines with *afrikaans, african* and *print5*

With regard to both newspapers and magazines figure 2.1 shows some growth in the proportions for newspapers and magazines between 2006 and 2009, followed by some marginal declines in the proportion of South Africans reading newspapers and magazines until 2014. This study also shows some initial growth in the intensity of engagement between 2002 and 2008, possibly explained by the growing proportions indicated in figure 2.1 - a reason for which could be the penetration of new tabloids such as the Daily Sun. This study however showed quite steep and consistent profiles of declines since 2008. Therefore, while the proportion of newspaper and magazine readers showed slow declines since 2008, the intensity of engagement appears to have declined also. Also, while the declines are apparent across demographic categories, the levels in engagement increase by increasing levels of income, LSM, education and age. For population groups, while some growth is apparent among black respondents until 2008, declines are consistent among all races post 2008.

The *repertoires* with strong representation from print are: *african, afrikaans* and *print5.* While engagement on *african* showed little change over the period under review, the degree

of engagement varied slightly with more engagement for higher LSM and education levels and slightly higher engagement levels for younger and male respondents. The declines apparent in print as *type* - both in terms of proportion and degrees of engagement - are less apparent in both *african* and *afrikaans,* with both showing reasonable stability over the period under review. The general declines in print appear to have had a greater effect on *print5* where all demographic categories showed some declines, albeit from slightly different levels, with higher levels apparent for higher levels of eduction, income and LSM, while the opposite is true for age.

7.4 The internet with *intnews* and *social*

The internet proportional engagement shown in figure 2.1 indicate proportions from around 5% at the start of the period followed by quite consistent growth to reach a high of 46% at the end of the period. Steep growth on internet as *type* as well as some growth on the two *repertoires* dominated by internet engagement, namely *intnews* and *social,* are also apparent. This would suggest that at least some of the growth on these measures (internet as media *type* as well as *intnews* and *social* as *repertoires)* was due to increasing numbers of people engaging on the medium rather than only due to degrees of engagement. The main growth on *intnews* occurs after 2010 with somewhat different profiles for different demographic categories: steeper for younger, better educated, higher income and higher LSM brackets. A similar picture of the differences between demographic categories is apparent for *social,* but the increase appear consistently steeper, especially post 2010. This may partially be due to the inclusion of DSTV on the *repertoire* referred to as *social.*

7.5 Limitations and Suggestions for Future Research

Although *Datafirst* was invaluable in facilitating access to the datasets that were used here, it did not offer all the sets in the period under review. This was exacerbated by the changing nature of the surveys and - an unavoidable limitation - the changing nature of the media landscape, with new media *vehicles* continually coming into being while others either changed their names or were discontinued.

The national perspective taken in the definition and exploration of *repertoires* ignored many local community newspapers and the local nature of radio listenership in particular. Furthermore, the four-level ordinal scales for broadcast and internet media were defined by recency, while for print the six-level scales were defined by the number of issues. Future surveys into media usage in South Africa would offer more appropriate data for longitudinal analyses if more use could be made of at least six- or seven-point interval ratings scales. It would be particularly useful if the scales could be standardised across different media *types.*

The researcher's initial aim was to focus on the use of media for news and information in

particular. Some attempt was made to achieve this by excluding such reasons for accessing the internet as banking or emails and by the exclusion of obviously partisan magazines and newspaper such as various loyalty magazines and products like *Junk Mail*. Given the importance of such a news-focussed approach to media research, as highlighted in chapter 2, future research could explore ways in which to categorise and isolate news consumption on the various channels provided by the internet; and also on how to isolate news as an activity on traditional media.

With regard to the methods employed in this study, although SEM offered an effective way of estimating both latent factors (*repertoires)* and the effects of demographic variables on these factors - especially in the light of the large sample size; it did not allow for efficient estimation of confidence intervals. Accordingly, processing-intensive bootstrapping procedures had to be used to provide the confidence intervals shown in the relevant plots. The results obtained in this way offered some interesting patterns that have been fully described in the project. The use of *k-means* clustering procedures to consider partitioning of media *type* in the dataset, although more efficient than alternatives that were considered such as hierarchical clustering and self-organising maps, still did not deliver very strong separations, thus weakening the strength of the interpretations that followed. The use of estimated marginal means to aggregate the impact of demographic categories was both appropriate and effective for both *repertoires* and *types*. The fact that population weights provided in the AMPS data were not used in this analysis must be identified as a limitation and also as something to be considered for inclusion in any future research using these datasets.

Finally, in spite of some of the limitations described here, the results from the analyses undertaken in this study offer a base-line for how media engagement has changed over a period renown for the disruptive effects of the internet on the media industry; and could serve as a starting point for researching the revolution that is impacting industry of media and news in South Africa and elsewhere.

Bibliography

A. Abdelmonem, S. May, and V. A. Clark. *Practical multivariate analysis*. CRC Press, 5th edition, 2011.

Hervé Abdi. Factor rotations in factor analyses. *Encyclopedia for Research Methods for the Social Sciences. Sage: Thousand Oaks, CA*, pages 792–795, 2003.

C. W. Anderson, E. Bell, and C. Shirky. Post-industrial journalism: adapting to the present. Technical report, Tow Center for Digital Journalism, 2012.

Anon. Toward economic sustainability of the media in developing countries. Technical report, Center for International Media Assistance, 2007.

B. Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*, 2017. URL `https://CRAN.R-project.org/package=gridExtra`. R package version 2.3.

A. Berndt and D. Petzer, editors. *Marketing research*. Pearson Education South Africa, 2011.

C. Brown. *dummies: Create dummy/indicator variables flexibly and efficiently*, 2012. URL `https://CRAN.R-project.org/package=dummies`. R package version 1.5.6.

C. Brown. *formula.tools: Programmatic Utilities for Manipulating Formulas, Expressions, Calls, Assignments and Other R Objects*, 2018. URL `https://CRAN.R-project.org/package=formula.tools`. R package version 1.7.1.

T. A. Brown and M. T. Moore. Confirmatory factor analysis, 2016. URL `https://www.researchgate.net/profile/Michael_Moore8/publication/251573889_Hoyle_CFA_Chapter_-_Final/links/0deec51f14d2070566000000.pdf`. Accessed: 2 September 2016.

F. B. Bryant and P. R. Yarnold. *Reading and understanding multivariate statistics*, chapter four, pages 99–136. American Psychological Association, 2004.

J. Carifio and R. Perla. Resolving the 50-year debate around using and misusing likert scales. *Medical Education*, (42):1150–1152, 2008.

D. Conrad. Deconstructing the community radio model: applying practice to theory in East Africa. *Journalism*, 15:773–789, 2014.

C. Corder. Pointing the way in media research:, March 2003. URL http://www.saarf.co.za/amps-presentations/2002/2002Bamps.pdf. Accessed: 22 October 2018.

S. Edgerly. Red Media, Blue Media, and Purple Media: News Repertoires in the Colorful Media Landscape. *Journal of Broadcasting & Electronic Media*, 59(1):1–21, January 2015. ISSN 0883-8151, 1550-6878. doi: 10.1080/08838151.2014.998220. URL http://www.tandfonline.com/doi/abs/10.1080/08838151.2014.998220.

A. Fulton. Forty years of amps: insights from a life well lived. Pdf slides, May 2017. URL http://www.samra.co.za/wp-content/uploads/2017/05/40-years-of-AMPS_Perumal.pdf.

J. M. Garbuszus and S. Jeworutzki. *readstata13: Import 'Stata' Data Files*, 2018. URL https://CRAN.R-project.org/package=readstata13. R package version 0.9.2.

D. Gefen, D. Straub, and M. Boudreau. Structural equation modeling and regression: Guidelines for research practice. *Communications of the association for information systems*, 4(1):7, 2000. URL http://aisel.aisnet.org/cgi/viewcontent.cgi?article=2531&context=cais.

R. W. Gregory, B. Ben, L. Thomas, and C. J. Randall. *gmodels: Various R Programming Tools for Model Fitting*, 2018. URL https://CRAN.R-project.org/package=gmodels. R package version 2.18.1 with contributions from Randall C. Johnson are Copyright SAIC-Frederick and Inc. Funded by the Intramural Research Program and of the NIH and National Cancer Institute and Center for Cancer Research under NCI Contract NO1-CO-12400.

L. G. Grimm and P. R. Yarnold. *Reading and Understanding Multivariate Statistics*. American Psychological Association, 2004. URL http://psycnet.apa.org/psycinfo/1995-97110-000.

J. F. Hair, R. E. Anderson, R. L. Tatham, and W. C. Black. *Mulivariate Data Analysis*. Prentice-Hall International Inc., 5th edition, 1998.

J. F. Hair, W. C. Black, R. E. Anderson, B. J. Babin, and R. L. Tatham. SEM Basics: A Supplement to Multivariate Data Analysis, Supplement to Multivariate Data Analysis. Pearson Prentice Hall Publishing, 2006. URL \url{http://www.mvstats.com/downloads/supplements/sem_basics.pdf}.

P. Haupt. The evolution of amps, 2012. URL http://www.saarf.co.za/amps/amps-evolution.asp.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*. Springer, 2013.

S Jarek. *mvnormtest: Normality test for multivariate variables*, 2012. URL `https://CRAN.R-project.org/package=mvnormtest`. R package version 0.1-9.

R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*. Pearson Education International, 5th edition, 2002.

A. Kassambara. *ggcorrplot: Visualization of a Correlation Matrix using 'ggplot2'*, 2016. URL `https://CRAN.R-project.org/package=ggcorrplot`. R package version 0.1.1.

D. Kenny. Measuring model fit, 2015. URL `http://davidakenny.net/cm/fit.htm`. Accessed: 30 May 2018.

P. Kline. *An Easy Guide to Factor Analysis*. Routledge, 1994.

M. Ko. How are estimated marginal means calculated, December 2017. URL `https://www.cscu.cornell.edu/news/statnews/stnews93.pdf`. Accessed: 18 October 2018.

M. Kuhn. *caret: Classification and Regression Training*, r package version 6.0-80 edition, 2018. URL `https://CRAN.R-project.org/package=caret`. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Besty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt.

R. Lenth. *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2018a. URL `https://CRAN.R-project.org/package=emmeans`. R package version 1.2.3.

R. Lenth. *Interaction analysis in emmeans*, September 2018b. URL `https://cran.r-project.org/web/packages/emmeans/vignettes/interactions.html`. Accessed: 18 October 2018.

R. Lenth. *Basics of estimated marginal means*, September 2018c. URL `https://cran.r-project.org/web/packages/emmeans/vignettes/basics.html#motivation`. Accessed: 18 October 2018.

A. Liaw and M. Wiener. Classification and regression by randomForest. *R News*, 2(3):18–22, 2002. URL `https://CRAN.R-project.org/doc/Rnews/`.

D. F. Morrison. *Multivariate Statistical Methods*. Thomson Learning Inc., 4th edition, 2005.

B. Peters. The future of journalism and challenges for media development. *Journalism Practice*, (4):268–273, 2010.

S. E. Radloff. *Mathematical Statistics 3, General Linear Models: Course Notes*. Department of Statistics, Rhodes University, 2015.

G. Raiche. *an R package for parallel analysis and non graphical solutions to the Cattell scree test*, 2010. URL `http://CRAN.R-project.org/package=nFactors`. R package version 2.3.3.

W. Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois, 2018. URL `https://CRAN.R-project.org/package=psych`. R package version 1.8.4.

Y. Rosseel. lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36, 2012. URL `http://www.jstatsoft.org/v48/i02/`.

RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2016. URL `http://www.rstudio.com/`.

SAARF. *The 1995 All Media and Products Survey*. South African Advertising Research Foundation;, Johannesburg, South Africa, 1998. URL `https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/193/download/9060`. Distributor: South African Data Archive.

SAARF. All media and products survey 2002[dataset]. DataFirst[distributor], 2002. URL `https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/525`.

SAARF. All media and products survey 2005[dataset]. DataFirst[distributor], 2005. URL `https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/607`.

SAARF. All media and products survey 2008[dataset]. DataFirst[distributor], 2008. URL `https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/671`.

SAARF. All media and products survey 2010[dataset]. DataFirst[distributor], 2010. URL `https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/524`.

SAARF. All media and products survey 2012[dataset]. DataFirst[distributor], 2012. URL `https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/509`.

SAARF. All media and products survey 2014[dataset]. DataFirst[distributor], 2014a. URL `https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/626`.

SAARF. Universe and sample. techreport, South Arican Audience Research Foundation, 2014b. URL `http://www.saarf.co.za/amps-technicalreport/technicalreport-Jan15-Dec15/data%20files/Technical/13%20-%20Tech%202015B%20~%20Pages%2021-25.pdf`. Accessed: 24 October 2018.

SAARF. Frequenty asked questions, 2018. URL `http://www.saarf.co.za/saarf/Faqs.asp`. Accesed: 22 October 2018.

J. Santana, L. Allison. *The bootstrap: an introductory course*, 1916.

J. B. Schreiber, A. Nora, F. K. Stage, A. Barlow, E, and J. King. Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of educational research*, 99(6):323–338, 2006. URL `http://www.tandfonline.com/doi/abs/10.3200/JOER.99.6.323-338`.

K. Schröder. News media old and new: fluctuating audiences, news repertoires and locations of consumption. *Journalism Studies*, 16(1):60–78, 2015.

T. Susman-Pena. Making media development more effective. Special report, Center for International Media Assistance., 2012.

J. Turcotte, C. York, J. Irving, R. M. Scholl, and R. J. Pingree. News Recommendations from Social Media Opinion Leaders: Effects on Media Trust and Information Seeking. *Journal of Computer-Mediated Communication*, 20(5):520–535, September 2015. ISSN 10836101. doi: 10.1111/jcc4.12127. URL `http://doi.wiley.com/10.1111/jcc4.12127`.

W. N. Venables and B. D. Ripley. *Modern applied statistics with S*. Springer, New York, fourth edition, 2002. URL `http://www.stats.ox.ac.uk/pub/MASS4`. ISBN 0-387-95457-0.

Taiyun Wei and Viliam Simko. *R package "corrplot": Visualization of a Correlation Matrix*, 2017. URL `https://github.com/taiyun/corrplot`. (Version 0.84).

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `http://ggplot2.org`.

H. Wickham. *tidyverse: Easily Install and Load the 'Tidyverse'*, 2017. URL `https://CRAN.R-project.org/package=tidyverse`. R package version 1.2.1.

H. Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations*, 2018. URL `https://CRAN.R-project.org/package=stringr`. R package version 1.3.1.

H. Wickham, R. François, L. Henry, and K. Müller. *dplyr: A Grammar of Data Manipulation*, 2018. URL `https://CRAN.R-project.org/package=dplyr`. R package version 0.7.6.

Hadley Wickham and Lionel Henry. *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*, 2018. URL `https://CRAN.R-project.org/package=tidyr`. R package version 0.8.1.

J. Xu, C. Forman, J. B. Kim, and K. Van Ittersum. News Media Channels: Complements or Substitutes? Evidence from Mobile Phone Usage. *Journal of Marketing*, 78(4):97–112, July 2014. ISSN 0022-2429. doi: 10.1509/jm.13.0198. URL `http://journals.ama.org/doi/abs/10.1509/jm.13.0198`.

Appendix: Dataset AMPS 2002

Introduction

The 2002 dataset used for this exploration consists of responses from 29 791 participants representing a population of 29 583 000 people. The following describes the variables that were considered in this exploration and, where relevant, the transformations that were effected are described. The data was collected in face-to-face interviews over two surveys from January 2002 until December 2002.

The sampling procedure included defining a universe, from which the AMPS sample was drawn, comprising adults older than 15 years in South Africa. Certain geographic areas were excluded for having negligible numbers of people from particular population groups. A multi-stage, stratified, quasi-probability design was employed. This study is based on a full annual sample.

The first two variables signify respondent unique identifiers questionnaire numbers *(qn)* and population weights (*pwgt*). These are followed by 11 demographic variables, of which only six (sex, age, education, race, household income, and LSM) are used, and 6 media *type* variables. With regard to media *vehicles*, the set consisted of 39 newspapers, 13 magazines, 14 radio stations, and 6 TV channels, each including an *other* variable. Internet consisted of 5 purposes for using the internet.

Demographic Variables

Demographic variables indicating category codes are described below:

- Demographic Variables:
  - ***age*:** age brackets

    Aggregated levels:

    1 = 15 - 24

    2 = 25 - 44

    3 = 45 - 54

    4 = 55+

- ***sex*:** gender

  1 = male

  2 = female

- ***edu*:** education level

  Aggregated levels:

  1= <matric (original codes: 1,2,3,4)

  2 = matric (original codes: 5)

  3 = >matric (original codes: 6,7,8)

- ***hh_inc:*** household income

  Aggregated levels:

  1 = <R2 500 (original codes: 1,2,3,4)

  2 = R2 500 - R6 999 (original codes: 5,6)

  3 = R7 000-R11 999 (original codes: 7)

  4 = >=R12 000 (original codes: 8)

- ***race*:** population group

  1 = black

  2 = coloured

  3 = indian

  4 = white

- ***lsm*:** living standards measure

  Aggregated groups

  1 = Groups 1& 2

  2 = Groups 3 & 4

  3 = Groups 5 & 6

  4 = Groups 7 & 8

  5 = Groups 9 & 10

Media Vehicles

Variables reflecting relative engagement values per media vehicle were created. The details are described in the sections below.

Newspapers and Magazines

- The scale variable for these media *types* was based on the response to a survey question about how many issues of a particular print *vehicle* the respondent read in a given issue

period. For example, dailies would be between zero and five and monthly magazines would be considered over a six week period (ie 0-6). The scale will differ depending on the issue periods.

- Club or loyalty magazines such as *Vodaworld*, and *Edgars Club* were excluded as well as tv guides such as *Magic* and *Tv Plus*. Purely advertising products such as *Junk Mail* and *Cape Ads* were also excluded.
- Additional notes on Newspapers:
  - The *Daily Sun*, *Isolezwe* and *Son* - all of which became a key newspaper products later, were launched after this period and therefore do not appear in this set.
  - This dataset contained information on local, community newspapers. These were excluded since later sets did not include them.
  - An *other* variable was created to include: Herald on Sat
- Additional notes on Magazines:
  - *The Motorist* later became *AA Traveller*. For the sake of consistency, it was changed to *AA Traveller* in this dataset
  - All magazine titles with fewer than 5% response were delegated to an *"other.mags"* variable

Television and Radio

- Relative engagement was determined by how recently a respondent personally watched a particular channel or listened to a particular radio station. Resulting in:
  - 0 = "not at all"
  - 1 = "In the past four weeks"
  - 2 = "in the past 7 days"
  - 3 = "yesterday"
- Additional notes on Radio:
  - *P4 Cape Town* became *Heart FM*
  - *P4 KZN* became *Gagasi FM*
  - *CKI FM* became *Tru FM*
  - All radio stations with fewer than 5% response were delegated to an *"other.*radio" variable

- Additional notes on TV:
  - *MNet CSN* and *MNet* Main will not be included, since open time ended in 2007 and access through *DSTV* is represented already
  - *Bop TV* was ended in 2003, so it was included in "*Other TV*" for 2002

Internet

- Analogous to Radio and TV, a first level of engagement was determined by considering the recency of accessing the internet on either computer or mobile. In particular:
  - 0 = "not"
  - 1 = "in the past four weeks"
  - 2 = "in the past 7 days"
  - 3 = "yesterday"
- A second aspect of internet engagement relates to what the internet was used for. Since the focus of this project was on media for the purposes of news, information and entertainment, the use of a computer or cellphone for example accessing email, banking, dating or shopping was excluded. Accessing the internet for purposes of search, social networking, print media sites, listening to the radio, or accessing news were extracted.
- To create a relative engagement level for each of these purposes, the recency values (0-3) were multiplied by the bivariate values (0/1) for each of the purposes.

Media Types

- To create a variable signifying relative engagement on media *type* (newspapers, magazines, tv, radio, internet), a total engagement over all the vehicles described above was created by summing the *vehicle* engagement values.
- A variable *all* was added to reflect engagement over all media *types* by summing the the *type* variables after first standardising to mean = 0 and standard deviation = 1.

Appendix: Dataset AMPS 2008

Introduction

The 2008 dataset used for this exploration consists of responses from 21 083 participants representing a population of 31 305 000 people in South Africa. The data was collected in face-to-face interviews over two surveys during 2008. The drop in sample size from previous years was a result of budget costs at the SAARF.

The sampling procedure included defining a universe, from which the AMPS sample was drawn, comprising adults older than 15 years in South Africa. Certain geographic areas were excluded for having negligible numbers of people from particular population groups. A multi-stage, stratified, quasi-probability design was employed. This study is based on a full annual sample.

The first two variables signify respondent unique identifiers questionnaire numbers *(qn)* and population weights (*pwgt*). These are followed by 13 demographic variables, of which only six (sex, age, education, race, household income, and LSM) are used, and 6 media-type variables. With regard to media vehicles, the set consisted of 85 media *vehicle* variables, comprising of 47 newspapers, 14 magazines, 17 radio and 7 TV channels. All the categories included an *other* variable. Internet engagement was represented by 6 purposes for using the internet.

Demographic Variables

- Demographic Variables:
  - ***age*:** age brackets

    Aggregated levels:

    1 = 16 - 24

    2 = 25 - 44

    3 = 45 - 54

    4: = 55+
  - ***sex*:** gender

    1 = male

2 = female

- ***edu*:** education level

  Aggregated levels:

  1 = <matric (original codes: 1,2,3,4)

  2 = matric (original codes: 5)

  3 = >matric (original codes: 6,7,8)

- ***hh_inc:*** household income level (original codes: ordered factors)

  Aggregated levels:

  1 = <R5 000(original codes: 1,2,3,4)

  2 = R5 000 - R10 999 (original codes: 5,6)

  3 = R11 000 - 19 999 (original codes: 7)

  4 = R20 000+ (original codes: 8)

- ***race*:** population group

  1 = black

  2 = coloured

  3 = indian

  4 = white

- ***lsm*:** living standards measure

  Aggregated groups:

  1 = Groups 1& 2

  2 = Groups 3 & 4

  3 = Groups 5 & 6

  4 = Groups 7 & 8

  5 = Groups 9 & 10

Media Vehicles

Variables reflecting relative engagement values per media vehicle were created.

Newspapers and Magazines

- The scale variable for these media *types* was based on the response to a survey question about how many issues of a particular print *vehicle* the respondent read in a given issue period. For example, dailies would be between zero and five and monthly magazines would be considered over a six week period (ie 0-6). The scale will differ depending on the issue periods.

- Club or loyalty magazines such as *Vodaworld*, and *Edgars Club* were excluded as well as tv guides such as *Magic* and *Tv Plus*. Purely advertising products such as *Junk Mail* and *Cape Ads* were also excluded.

- Additional notes on Newspapers:

  - Additional newspapers included in 2008: *Daily Voice*, *The Times*, *Um Afrika*, *Ilanga Sunday*
  - Included in *other*: *The Weekender*, *Sondag*
  - *Son Ooskaap* was coded as *Son* in this dataset

- Additional notes on Magazines:

  - *Finweek* combined previous *Finance Week* and *Finansies and Tegniek*
  - *Ideas Idees* was previously *Womans' Value*
  - All magazine titles with fewer than 5% response were delegated to an *"other.mags"* variable

Television and Radio

- Relative engagement was determined by how recently a respondent personally watched a particular channel or listened to a particular radio station. Resulting in:

  - 0 = "not at all"
  - 1 = "In the past four weeks"
  - 2 = "in the past 7 days"
  - 3 = "yesterday"

- Additional notes on Radio:

  - Several new radio stations that appear to be community stations were added
  - All radio stations with fewer than 5% response were delegated to an *"other.*radio" variable

- Additional notes on TV:

  - *Soweto TV* was added

Internet

- Analogous to Radio and TV, a first level of engagement was determined by considering the recency of accessing the internet on either computer or mobile. Note, this set included a level for "the past 12 months", which was excluded in order to align with earlier surveys. In particular:
  - 0 = "not"
  - 1 = "in the past 4 weeks"
  - 2 = "in the past 7 days"
  - 3 = "yesterday"
- A second aspect of internet engagement relates to what the internet was used for. Since the focus of this project was on media for the purposes of news, information and entertainment, the use of a computer or cellphone for example accessing email, banking, dating or shopping was excluded. Accessing the internet for purposes of search, social networking, print media sites, listening to the radio, accessing news, or watching tv were extracted.
- To create a relative engagement level for each of these purposes, the recency values (0-3) were multiplied by the bivariate values (0/1) for each of the purposes.

Media Types

- To create a variable signifying relative engagement on media *type* (newspapers, magazines, tv, radio, internet), an aggregate over all the *vehicles* was created by summing the *vehicle* engagement values.
- A variable *all* was added to reflect engagement over all media *types* by summing the the *type* variables after first standardising to mean = 0 and standard deviation = 1.

Appendix: Dataset AMPS 2010

Introduction

The 2010 dataset used for this exploration consists of responses from 25 160 participants representing a population of 34 020 000 people in South Africa. The following describes the variables that were considered in this exploration and, where relevant, the transformations that were effected are described. The data was collected in face-to-face interviews over two surveys during 2010.

The sampling procedure included defining a universe, from which the AMPS sample was drawn, comprising adults older than 15 years in South Africa. Certain geographic areas were excluded for having negligible numbers of people from particular population groups. A multi-stage, stratified, quasi-probability design was employed. This study is based on a full annual sample.

The first two variables that are common in all datasets signify respondent unique identifiers questionnaire numbers *(qn)* and population weights (*pwgt*). These are followed by 13 demographic variables, of which only six (sex, age, education, race, household income, and LSM) are used, and 6 media-type variables. With regard to media *vehicles* the set included 87 media *vehicle* variables, comprising of 48 newspapers, 15 magazines, 17 radio and 7 TV channels. With the exception of TV, all the categories included an *other* variable. Internet engagement was represented by 6 purposes.

Demographic Variables

- Demographic Variables:
  - ***age*:** age brackets

    Aggregated levels:

    1 = 15 - 24

    2 = 25 - 44

    3 = 45 - 54

    4: = 55+

- ***sex*:** gender

  1 = male

  2 = female

- ***edu*:** education level

  Aggregated levels:

  1= <matric (original codes: 1,2,3,4)

  2 = matric (original codes: 5)

  3 = >matric (original codes: 6,7,8)

- ***hh_inc:*** household income level

  Aggregated levels:

  1 = <R5 000 (original codes: 1,2,3,4)

  2 = R5 000 - R10 999 (original codes: 5,6)

  3 = R11 000-19 999 (original codes: 7)

  4 = R20 000+ (original codes: 8)

- ***race*:** population group

  1 = black

  2 = coloured

  3 = indian

  4 = white

- ***lsm:*** living standards measure

  Aggregated groups

  1 = Groups 1& 2

  2 = Groups 3 & 4

  3 = Groups 5 & 6

  4 = Groups 7 & 8

  5 = Groups 9 & 10

Media Vehicles

Variables reflecting relative engagement values per media vehicle were created.

Newspapers and Magazines

- The scale variable for these media *types* was based on the response to a survey question about how many issues of a particular print *vehicle* the respondent read in a given issue period. For example, dailies would be between zero and five and monthly magazines

would be considered over a six week period (ie 0-6). The scale will differ depending on the issue periods.

- Club or loyalty magazines such as *Vodaworld*, and *Edgars Club* were excluded as well as tv guides such as *Magic* and *Tv Plus*. Purely advertising products such as *Junk Mail* and *Cape Ads* were also excluded.

- Additional notes on Newspapers:

  - *Pretoria News Weekend* is missing
  - *Isolezwe Sunday* and *Sondag Son* are new products
  - The *Zimbabwean*, *Soccer Week* and *Sondag* are included in the variable "*other.news*"

- Additional notes on Magazines:

  - All magazine titles with fewer than 5% response were delegated to an *"other.mags*" variable

Television and Radio

- Relative engagement was determined by how recently a respondent personally watched a particular channel or listened to a particular radio station. Resulting in:

  - 0 = "not at all"
  - 1 = "In the past four weeks"
  - 2 = "in the past 7 days"
  - 3 = "yesterday"

- Additional notes on Radio:

  - All radio stations with fewer than 5% response were delegated to an *"other*.radio" variable

- Additional notes on TV:

  - *Cape Town TV* was added

Internet

- Analogous to Radio and TV, a first level of engagement was determined by considering the recency of accessing the internet on either computer or mobile. Note, this set included a level for "the past 12 months, but it was excluded from the datasets". In particular:
  - 0 = "not"
  - 1 = "in the past 4 weeks"
  - 2 = "in the past 7 days"
  - 3 = "yesterday"
- A second aspect of internet engagement relates to what the internet was used for. Since the focus of this project was on media for the purposes of news, information and entertainment, the use of a computer or cellphone for example accessing email, banking, dating or shopping was excluded. Accessing the internet for purposes of search, social networking, print media sites, listening to the radio, accessing news, or watching tv were extracted.
- To create a relative engagement level for each of these purposes, the recency values (0-3) were multiplied by the bivariate values (0/1) for each of the purposes.

Media Types

- To create a variable signifying relative engagement on media *type*, an aggregate over all the *vehicles* was created by summing the *vehicle* engagement values.
- A variable *all* was created to reflect engagement over all media *types* by summing the *type* variables, after first standardising them to means = 0 and standard deviations = 1.

Appendix: Dataset AMPS 2012

Introduction

The 2012 dataset used for this exploration consists of responses from 25 108 participants representing a population of 34 935 000 people in South Africa. The data was collected in face-to-face interviews over two surveys during 2012.

The sampling procedure included defining a universe, from which the AMPS sample was drawn, comprising adults older than 15 years in South Africa. Certain geographic areas were excluded for having negligible numbers of people from particular population groups. A multi-stage, stratified, quasi-probability design was employed. This study is based on a full annual sample.

The first two variables that are common in all datasets signify respondent unique identifiers questionnaire numbers *(qn)* and population weights (*pwgt*). These are followed by 13 demographic variables, of which only six (sex, age, education, race, household income, and LSM) were used, and 6 media *type* variables. With regard to media *vehicles*, the set included 94 media *vehicle* variables, comprising of 51 newspapers, 15 magazines, 17 radio and 11 TV channels. All the categories included an *other* variable. Internet engagement was represented by 6 purposes.

Demographic Variables

- Demographic Variables:
  - ***age*:** age brackets

    Aggregated levels:

    1 = 15 - 24

    2 = 25 - 44

    3 = 45 - 54

    4: = 55+
  - ***sex*:** gender

    1 = male

2 = female

- ***edu*:** education level

  Aggregated levels:

  1= <matric (original codes: 1,2,3,4)

  2 = matric (original codes: 5)

  3 = >matric (original codes: 6,7,8)

- ***hh_inc:*** household income level

  Aggregated levels:

  1 = <R5 000 (original codes: 1,2,3,4)

  2 = R5 000 - R10 999 (original codes: 5,6)

  3 = R11 000-19 999 (original codes: 7)

  4 = R20 000+ (original codes: 8)

- ***race*:** population group

  1 = black

  2 = coloured

  3 = indian

  4 = white

- ***lsm*:** living standards measure

  Aggregated groups:

  1 = Groups 1& 2

  2 = Groups 3 & 4

  3 = Groups 5 & 6

  4 = Groups 7 & 8

  5 = Groups 9 & 10

Media Vehicles

Variables reflecting relative engagement values per media vehicle were created.

Newspapers and Magazines

- The scale variable for these media *types* was based on the response to a survey question about how many issues of a particular print *vehicle* the respondent read in a given issue period. For example, dailies would be between zero and five and monthly magazines would be considered over a six week period (ie 0-6). The scale will differ depending on the issue periods.

- Club or loyalty magazines such as *Vodaworld*, and *Edgars Club* were excluded as well as tv guides such as *Magic* and *Tv Plus*. Purely advertising products such as *Junk Mail* and *Cape Ads* were also excluded.
- Additional notes on Newspapers:
  - Important new product: *The New Age*, *Isolezwe Saturday*
  - *Sondag*, and *The Zimbawean* were included in "*other.news*"
- Additional notes on Magazines:
  - All magazine titles with fewer than 5% response were delegated to an *"other.mags*" variable

Television and Radio

- Relative engagement was determined by how recently a respondent personally watched a particular channel or listened to a particular radio station. Resulting in:
  - 0 = "not at all"
  - 1 = "In the past four weeks"
  - 2 = "in the past 7 days"
  - 3 = "yesterday"
- Additional notes on Radio:
  - All radio stations with fewer than 5% response were delegated to an *"other.*radio" variable
- Additional notes on TV:
  - *IKZN TV*, *Bay TV* and *Top TV* were added

Internet

- Analogous to Radio and TV, a first level of engagement was determined by considering the recency of accessing the internet on either computer or mobile. Note, this set included a level for "the past 12 months, which was excluded in order to align with earlier surveys. In particular:
  - 0 = "not"
  - 2 = "in the past 4 weeks"

  - 2 = "in the past 7 days"
  - 3 = "yesterday"

- A second aspect of internet engagement relates to what the internet was used for. Since the focus of this project was on media for the purposes of news, information and entertainment, the use of a computer or cellphone for example accessing email, banking, dating or shopping was excluded. Accessing the internet for purposes of search, social networking, print media sites, listening to the radio, accessing news, or watching tv were extracted.

- To create a relative engagement level for each of these purposes, the recency values (0-3) were multiplied by the bivariate values (0/1) for each of the purposes.

Media Types

- To create a variable signifying relative engagement on media *type*, an aggregate over all the *vehicles* was created by summing the *vehicle* engagement values.

- A variable *all* was created to reflect engagement over all media *types* by summing the *type* variables, after first standardising to means = 0 and standard deviations = 1.

Appendix: Dataset AMPS 2014

Introduction

The 2014 dataset used for this exploration consists of responses from 25 584 participants representing a population of 37 665 000 people in South Africa. The data was collected in face-to-face interviews over two surveys during 2014.

The sampling procedure included defining a universe, from which the AMPS sample was drawn, comprising adults older than 15 years in South Africa. Certain geographic areas were excluded for having negligible numbers of people from particular population groups. A multi-stage, stratified, quasi-probability design was employed. This study is based on a full annual sample.

The first two variables that are common in all datasets signify respondent unique identifiers questionnaire numbers *(qn)* and population weights (*pwgt*). These are followed by 13 demographic variables, of which only six (sex, age, education, race, household income, and LSM) were used, and 6 media-type variables. With regard to media *vehicles*, the set included 92 media *vehicle* variables, comprising of 49 newspapers, 14 magazines, 17 radio and 12 TV channels. All the categories included an *other* variable. Internet engagement was represented by 6 purposes.

Demographic Variables

- Demographic Variables:
  - ***age*:** age brackets

    Aggregated levels:

    1 = 15 - 24

    2 = 25 - 44

    3 = 45 - 54

    4: = 55+
  - ***sex*:** gender

    1 = male

2 = female

- ***edu*:** education level

  Aggregated levels:

  1= <matric (original codes: 1,2,3,4)

  2 = matric (original codes: 5)

  3 = >matric (original codes: 6,7,8)

- ***hh_inc:*** household income level

  Aggregated levels:

  1 = <R5 000 (original codes: 1,2,3,4)

  2 = R5 000 - R10 999 (original codes: 5,6)

  3 = R11 000-19 999 (original codes: 7)

  4 = R20 000+ (original codes: 8)

- ***race*:** population group

  1 = black

  2 = coloured

  3 = indian

  4 = white

- ***lsm*:** living standards measure

  Aggregated groups:

  1 = Groups 1& 2

  2 = Groups 3 & 4

  3 = Groups 5 & 6

  4 = Groups 7 & 8

  5 = Groups 9 & 10

Media Vehicles

Variables reflecting relative engagement values per media vehicle were created.

newspapers and magazines

- The scale variable for these media *types* was based on the response to a survey question about how many issues of a particular print *vehicle* the respondent read in a given issue period. For example, dailies would be between zero and five and monthly magazines would be considered over a six week period (ie 0-6). The scale will differ depending on the issue periods.

- Club or loyalty magazines such as *Vodaworld*, and *Edgars Club* were excluded as well as tv guides such as *Magic* and *Tv Plus*. Purely advertising products such as *Junk Mail* and *Cape Ads* were also excluded.
- Additional notes on Newspapers:
  - *The Zimbabwean* was included in "*other.news*"
  - Um Afrika not included in this dataset
- Additional notes on Magazines:
  - All magazine titles with fewer than 5% response were delegated to an *"other.mags*" variable

Television and Radio

- Relative engagement was determined by how recently a respondent personally watched a particular channel or listened to a particular radio station. Resulting in:
  - 0 = "not at all"
  - 1 = "In the past four weeks"
  - 2 = "in the past 7 days"
  - 3 = "yesterday"
- Additional notes on Radio:
  - All radio stations with fewer than 5% response were delegated to an *"other.*radio" variable
- Additional notes on TV:
  - *Tswane TV* was added

Internet

- Analogous to Radio and TV, a first level of engagement was determined by considering the recency of accessing the internet on either computer or mobile. Note, this set included a level for "the past 12 months". In particular:
  - 0 = "not"
  - 1 = "in the past 4 weeks"
  - 2 = "in the past 7 days"

  - 3 = "yesterday"

- A second aspect of internet engagement relates to what the internet was used for. Since the focus of this project was on media for the purposes of news, information and entertainment, the use of a computer or cellphone for example accessing email, banking, dating or shopping was excluded. Accessing the internet for purposes of search, social networking, print media sites, listening to the radio, accessing news, or watching tv were extracted.

- To create a relative engagement level for each of these purposes, the recency values (0-3) were multiplied by the bivariate values (0/1) for each of the purposes.

Media Types

- To create a variable signifying relative engagement on media *type*, an aggregate over all the vehicles was created by summing the vehicle engagement values.

- A variable *all* was created to reflect engagement over all media *types* by summing the standardised (to means = 0 and standard deviations = 1) *type* variables.