Introduction

The history of psychological science appears short when placed alongside physics, chemistry and biology. Nevertheless, the field has consistently evolved in response to new challenges. For example, one of the crucial features of scientific psychology in the twentieth century was grounding itself in objectivity. By changing the subject to the study of behaviour, psychology could be based on scientific laws of behaviour. In contrast, introspection relied exclusively on an observation of one's mental state. Behaviourism initially helped psychologists better understand learning and behavioural change, but motivations and other mediational processes (e.g., thinking) remained hidden because they were not directly observable (Skinner, 1971). Despite academics arguing that the scientific assessment of behaviour can be yoked to cognition and emotion, behaviourism was never universally accepted, especially in Europe or Canada, because it could only provide a partial account of what it means to be human (Baddeley, 2018; Watson, 1913). These gaps in knowledge became crystallised further following rapid advances in technology during the Second World War, which led to a new approach that brought together computer science, psychology, linguistics and philosophy (Miller, 2003). Psychology became a key player of what has frequently been referred to as the cognitive revolution (Miller, 2003). Some scholars have even attempted to mark the exact date when cognitive science was conceived (11 September 1956). The day included presentations from IBM, which demonstrated how a computer could test a psychological theory, that aimed to explain how the brain might operate (Miller, 2003).

Baddeley (2018) points out that progress was only possible as several disciplines became closer after reaching similar conclusions. Many at the time expected that the events of the 1950s might lead to a new discipline that included linguistics, psychology, anthropology and computer science, but this never happened. Each academic area brought different priorities and while some have become increasingly intertwined with each passing

decade, this was not sufficient at the time to create a new integrated field. While these methodological transitions are often portrayed as a revolution, almost every new development can be described as a gradual evolution of the discipline with the benefit of hindsight. In the present day, interdisciplinary research that brings ideas and methods from multiple fields, while not revolutionary, has become the norm. As with the cognitive evolution, computer science, engineering and psychology have allowed research involving other new technologies to prosper (Lazer et al., 2009; Mazzucato, 2013). The modern-day equivalents of IBM are Apple, Amazon and Google. Similarly, the smartphone has become the latest mass-adopted technology to drive the modern economy.

However, long before industry started to drive the notion of a 'new' digital age, the development of tools to measure behaviour or underlying cognitions often involved modifying or adapting existing technologies. In many cases, these provided new empirical measurements and helped establish entire bodies of research that continue to the present day. Despite a number of journals documenting some of these developments, information is scattered widely across psychology and beyond. At the same time, many of the debates and challenges within psychological science today echo arguments of the past. Therefore, while this introduction is not a complete history of methodological development in psychology, it is important to understand and appreciate our history within and beyond the laboratory in the first instance (Baddeley, 2018).

Analogue to Digital

The development of digital computers provided a new way of thinking, measuring and understanding a variety of psychological processes. Originating with the British mathematician Alan Turing's proof that a 'simple machine' could, in principle, carry out any possible computation, this led to the notion of programmable machines (Turing, 1937). These and related developments were forerunners to the digital age in which we currently reside. However, modern micro-computers that were even remotely recognisable as the ubiquitous computers of today only started to become readily available in universities from the mid 1970s onwards (Weinberg, 2019). As a result, many psychologists throughout the 1960s and early 1970s had limited access to digital technology. The majority of methodological innovation remained grounded in an analogue domain. This included the use of simple machines to measure key press responses.

Those who wanted to go beyond simple behavioural responses and study social or biological processes had to innovate with materials that were readily available. For example, tracking eye movement across a scene involved behavioural observations in the first instance. Early observers noticed that reading does not involve a smooth sweeping of the eyes along text but a series of several stops (called fixations) and quick saccades. The first eye tracker to track such observations automatically and with increased precision used a lens connected to the pupil and an aluminium pointer that moved in response to movement (Huey, 1908/1968). This was extremely unpleasant for participants.

Improving or adapting such processes became important as psychologists sought to enhance the accuracy and reliability of such measures. Daniel Kahneman describes the development of a set-up that could capture pupil dilation. He and colleagues theorised at the time that these changes may reveal different processes involved with cognitive processing. Previous work had suggested that the pupils respond distinctly to mental effort and emotional arousal (Kahneman, 2011). Beatty and Kahneman developed an experimental set-up to measure this behavioural response more accurately. This involved a participant placing their head on a chin-and-forehead rest while staring at a traditional film-based camera. Participants listened to recorded information and answered questions while the audio of a metronome clicked. Each beat triggered an infrared flash, whereby a picture was captured to film. When pictures were developed, the researchers projected images of the pupil onto a large screen and used a ruler to measure the width of the pupil. Modifying existing technologies of the day, Kahneman and colleagues discovered that the pupil varied in response to changing demands of a specific task. Their methods advanced to the point where a video-camera system would project a participant's pupil onto a screen in another room. This removed the need to wait for the film to be developed. Kahneman writes that the development of these methods had a large impact on both his own and Beatty's careers.

Such an approach may still appear clunky and prone to considerable measurement error by today's digital standards, but it paved the way for others to develop eye-tracking systems that would eventually measure both pupil dilation and eye movement using digital computers running specialised software. Portable and head-mounted eye-trackers are now being used routinely within and beyond the laboratory (e.g., Mele and Federici, 2012; MacDonald and Tatler, 2018; Pérez-Edgar et al., 2020) and may transmit data to smartphones in the near future. Alternatively, technology within

inexpensive desktop trackers that allow for basic eye-tracking could be integrated into cameras placed on the front of most smartphones. This might lead to self-contained eye-tracking systems that are synchronised with experiments running on the same device (Wilcockson, 2017).

Like eye-tracking, early versions of existing methods relied on analogue technologies until digital computers allowed the field to expand. Throughout the 1960s and 1970s psychologists developed, often in collaboration with engineers and computer scientists, a number of other new tools. These helped capture a variety of social and physiological responses, including social interactions from audio records captured to analogue tape (Azrin et al., 1961), heart-rate variability using small portable recorders (Holter, 1961) and devices to assess sexual responses (Freund, 1963, 1965; Hanson and Bussière, 1998; Rachman, 1966; Simon and Schouten, 1992). Similarly, slide or overhead projectors that presented stimulus, for example, were eventually replaced by computers, but only when it became possible to present high-resolution graphics or photographs.

Innovative methods of the time became more established when they could be applied to multiple research questions across multiple disciplines. As with social and personality psychology today, cognitive science has benefited enormously from advances in physics, computer science and industry. This includes the development of functional magnetic resonance imaging (fMRI) that can detect changes in blood flow in the brain. As with all new methods, however, it takes time to develop best practices. Misunderstandings can have serious consequences. Specifically, larger volumes of data allow researchers to conduct more statistical tests, and some of these will always produce significant results (the multiple comparisons problem). In fMRI, each brain scan is divided into around 40,000 cubic units (voxels). The majority of analysis treats each voxel independently and compares them accordingly many thousands of times (Uttal, 2001). Many techniques have been developed for dealing with these problems, including Gaussian random field theory to calculate the probability of falsely 'finding' activated areas just by chance and to keep this acceptably low (Chumbley and Friston, 2009). However, the majority of fMRI research in the past did not control for these comparisons. In 2008, a dead salmon was placed in an MRI scanner and shown a series of photographs of humans in social situations and asked to determine the emotional response of each person in the photo (Bennett, Miller and Wolford, 2009). Without correcting for multiple comparisons, it would be possible to conclude that the salmon was in fact reacting to the photographs. To be clear, this study does not show that fMRI is in itself problematic, but it

demonstrates the importance of understanding how to correctly process and interpret data from data-intensive methods. Research in this area now often reports corrected and uncorrected comparisons as a result.

The methodological developments outlined earlier involve tools or devices that are only suitable for use in the laboratory. This includes presentation software and hardware to measure a variety of behavioural (e.g., reaction time) and physiological responses (e.g., heart rate). Today, these often sit within integrated systems that have become easier to use and less susceptible to error. As a result, psychologists have access to a variety of tools that can assist with quantitative and qualitative investigations. However, these developments alone have struggled to address ongoing concerns associated with conducting research in a laboratory environment.

Limitations of the Laboratory

Psychological science has historically held onto the notion that data obtained from strictly controlled laboratory settings, where the effect of single variables can be observed, remain the apex of research practices. As a result, the majority of psychologists run experiments, but these are often not particularly naturalistic. Welcomed into a testing room, participants are asked to perform tasks in return for money or course credit.

Lab-based research remains vulnerable to a range of biases and participant effects can be an artefact of an experiment rather than natural response to a manipulation. These are often referred to as demand characteristics (Hogg and Vaughan, 2005). The dynamic between an experimenter and participant can also impact responses. While these can be controlled statistically to an extent, contextual differences that occur within a lab environment, which is naturally less anonymous, are far reaching and can also change how participants express themselves and respond to psychometric assessments (Joinson, 1999).

Laboratory studies may be ideal when it comes to understanding perceptual or cognitive process, but they are less than perfect when it comes to understanding other everyday behaviours. For example, attempting to examine how people use personal technology in a lab will provide little new knowledge when life largely unfolds around these devices throughout the day. This is also difficult to experimentally manipulate for long periods of time. Many other behaviours or thought patterns that are presented in a lab are simply not reflective of what happens in the real world. For example, laboratory results suggesting conformity to authority go against what we know about how people behave when faced with conformity

outside the laboratory (Francis, 2012). Within developmental psychology, similar challenges have, until recently, faced researchers who wish to study related phenomena (Nielson et al., 2015). Emotional regulation, for example, is a central process that underpins adaptive social and emotional development. However, this area of research has struggled with conceptual and methodological challenges in operationalising and investigating effects over time, in different contexts and with multiple measures (Cole, Martin and Dennis, 2004). Specifically, how someone responds to negative emotional events in their everyday life, especially during development, cannot be simulated ethically in a laboratory.

Another example concerns a renewed interest in how individuals interact in spaces where intergroup contact takes place (McKeown and Dixon, 2017). Specifically, not only is intergroup contact a way to improve attitudes towards outgroups, it has also been argued as equally important when it comes to informing people about how social systems are regulated and can improve problem-solving, enhance cognitive flexibility and generate creativity (Hodson et al., 2018). Methodological barriers have previously made it difficult to examine intergroup contact in real life (Thai and Page-Gould, 2018). As a result, the literature is either reliant on aggregated data across individual contact interactions or uses laboratory environments to study interactions between strangers. Current methods are simply unable to capture the dynamics of group behaviour, and this will almost certainly lead to additional discrepancies between experimental and more ecologically valid approaches in the future (Keil, Koschate-Reis and Levine, 2020). Methodological choices that determine how contact is measured, however, already go beyond issues pertaining to ecological validity. Opposing methods have been shown to change the direction of results, whereby intergroup contact appears to undermine support for social change that leads to greater equality (Hässler et al., 2020; Saguy et al., 2009). Understanding when and why these inconsistencies occur is important and will likely involve the combination of lab-based methods and large-scale secondary data analysis (Ellis, 2012).

While many experimental results across psychology have not replicated or held up to closer inspection, the same issues likely afflict a variety of other methodological practices (Open Science Collaboration, 2015). Any carry-over effects for those who place a high value on more ecologically friendly methods are important to understand because there are many other circumstances across all areas of science where it is not possible to conduct an experiment. Replication alongside methodological pluralism is important, however, if we are to ever establish a general evidence base for

psychological phenomena. Regardless, there remains an ongoing tension between those who wish to maintain ecological validity and those who wish to focus on experimental controls and the removal of extraneous variables. This has continued to fuel a long-standing debate; however, recent developments, largely derived from applied psychology, can help reduce that tension.

Beyond the Laboratory

Methodological advances that attempt to bring the best elements of the scientific method, while avoiding the ecological limitations of the laboratory, have historically been less prominent within psychology. However, a variety of technologies that can help researchers observe people inside and outside the laboratory are both readily available and comparatively inexpensive. These have evolved across multiple disciplines, including medicine, computer science and sociology. For example, early digital systems helped researchers collect observational data using portable devices that have many attributes in common with modern smartphones. However, perhaps due to the disparate nature of early micro-computer systems, many early opportunities were either poorly advertised or never widely available (Farrell, 1991). This remains a challenge in the digital age. Therefore, in the following sections I trace some of these developments, which have eventually provided psychologists with a new set of research tools following the widespread adoption of mobile technology.

Direct Observation

Prior to the advent of computer-assisted systems that helped researchers record observational data, social scientists were limited to using a clipboard and timer to log and record events. Computers here made a researcher's job easier by ensuring data was captured accurately and in real time. Timestamps could be automatically logged with each event (Kahng and Iwata, 1998). In addition, the number of events that could be recorded also allowed for more complex statistical analyses. PROCODER, for example, was a software system that could be used to code video data (Tapp and Walden, 2000). Time codes on VHS tapes allowed researchers to find events or sections of tape very quickly. The enormity of software available for coding and analysing such observations at the end of the 1990s was reviewed by Emerson, Reeves and Felce in 2000. However, these methods only started to gain popularity as portable computing became more readily

available. Indeed, some of the software discussed ran on an Apple Newton, which could be considered a very early, albeit commercially disastrous, incarnation of a smartphone. Palmtop computing, as it was referred to at the time, was a precursor to the development of smartphones, but rather than being used by participants, these were used by researchers to record behaviour directly. While smaller than a laptop, they were functionally identical. To use a description of the Epson HX20 from Emerson et al. (2000):

> The Epson HX20 was one of the first truly portable, battery driven computers. Approximately the size of a closed laptop computer, it was relatively robust, with a small liquid crystal display mounted above a QWERTY keyboard, rather than a hinged screen. It also had a microcassette drive for the storage of programme and data files and a small integral printer similar to those in supermarket cash registers. (Emerson et al., 2000, pp. 48–49)

Many palmtop computers were developed during the 1990s by a variety of manufacturers who unlike Apple never went on to develop smartphones, including Casio, Sharp and Philips. Some of these systems were used in health-care settings to study patient behaviour and improve treatment. This included predicting, for example, self-injury on a mental health ward based on previous context to reduce the risk of self-harm in the future (Emerson et al., 2000). They appeared less frequently in psychological science, demonstrating that not every methodological development or new technology will facilitate subsequent research. Recent failures also include Google Glass, which from research potential may have been extremely useful for observation, but concerns around security meant that as a product it was quickly discontinued (Hofmann, Haustein and Landeweerd, 2017).

Mass Communication (Pre-Smartphone)

The adoption of mass-communication technologies has continued to transform psychological science. Before the internet became popular, psychologists in the 1990s relied on television, radio and print media to investigate if the public could detect lying through different communication channels. For example, a previous body of work had considered what cues people use to detect lying, but the majority of this research had been carried out in a laboratory with small numbers of university students (Wiseman, 1995). In one study, Wiseman wanted to see if similar findings would generalise across the general population. Working closely with the BBC, a well-known celebrity was interviewed twice. In one interview, he

consistently told the truth, but consistently lied in another. Transcripts of these interviews were printed in the *Daily Telegraph* (verbal cues only), broadcast on BBC Radio 1 (verbal and vocal cues) and shown on television (verbal, vocal and visual cues). For each medium the public were asked to report via telephone which of the two interviews they thought contained lies. Over 40,000 members of the public responded. Visual clues appeared to reduce individuals' ability to detect lying. Of course, this could tell us something specific about the listeners of radio versus television, but nevertheless, the results replicate previous laboratory findings. Such research also helped initiate the idea of citizen science, whereby the public are both involved with collection and directly linked back into dissemination activities that take place on a larger scale (Wiseman, 1996, 2007).

Wisemen and others went on to develop similar innovations using books and the internet. This included a real-world experiment that demonstrated how people are often unconsciously influenced by the size of another individual's pupils. Specifically, men tend to rate pictures of women with large pupils as more attractive (Tombs and Silverman, 2004). Using his own book, Wiseman requested that the publisher produce two slightly different front covers (Wiseman and Watt, 2010). Both included a photo of a smiling woman, but the woman's pupils on one cover were digitally enlarged. Those books also had unobtrusive marks on the back cover to distinguish them. A final page of the book asked readers to visit a website where respondents could indicate if they were male or female, the mark on the back of the book, and whether they had purchased the book in a shop or online. Data from high-street purchases confirmed that a significantly greater percentage of readers had chosen the cover with the large pupils. Through the use of online purchasing data, Wiseman and Watt (2010) were also able to demonstrate the absence of an effect, which confirms that their result was not driven by a participant's decision to take part in the experiment. This demonstrates how consumer behaviour can be unconsciously influenced by subtle cues.

Before the work of Wiseman and others, most psychologists initially used the internet to collect small quantities of survey data within university campuses or in specific populations after developing custom web-pages (e.g., Joinson, 1999). However, twenty years later, the internet has become a standard data collection tool for psychological science (Musch and Reips, 2000; Reimers and Stewart, 2015). By avoiding pen and paper, not only can this approach reduce participant and researcher errors, but the internet has allowed psychologists to generate additional statistical power by increasing

sample sizes without the need for additional lab space (Buchanan and Smith, 1999; Sassenberg and Ditrich, 2019).

Crowdsourcing to recruit large numbers of participants to complete surveys online previously involved working with traditional media (e.g., the BBC) that also provided a dissemination platform (Chamorro-Premuzic et al., 2009; Reimers, 2007). For example, Reimers (2007) collected data from around 255,000 participants to investigate sex differences. The impact on individual differences/personality research has been particularly prominent as research designs placed online allow for the presentation of personality assessments alongside other media (Chamorro-Premuzic et al., 2009). In recent years, cognitive experiments can also be developed online using HTML5 or freely available tools, including PsychoPy (Peirce et al., 2019; Reimers and Stewart, 2015, 2016).

Recruitment has also become more researcher-focused as commercial systems like Amazon Mechanical Turk and Prolific Academic now act as integrated participant compensation systems and provide access to large, diverse participant pools (Buhrmester, Kwang and Gosling, 2011). Critics have argued that these may still not be representative of the general population and continue to generate new challenges associated with data quality; however, many solutions have been developed to address these issues (Chandler, Mueller and Paolacci, 2014; Paolacci and Chandler, 2014; Peer, Vosgerau and Acquisti, 2014). Nevertheless, while the 1990s were characterised as a decade of behaviour because these were viewed as societally important (Fowler, Seligman and Koocher, 1999), recent years have witnessed the dominance of survey-based research, which often aligns poorly with actual behaviour (Baumeister, Vohs and Funder, 2007). The internet has, at least in the short term, increased our reliance on self-report, which cannot address every research question.

In contrast, experience sampling can also be conducted using web-pages or via mobile text-messaging systems (Conner and Lehman, 2011; Reimers and Stewart, 2009). Originally appearing in the 1970s, experience sampling allows for 'real time', in situ assessments of behaviour temporally close to the moment of enactment. Early attempts involved participants carrying beepers. These small devices would remind a participant at intermittent intervals to write down something about themselves or their environment. Using the internet, participants can dispense with pen and paper and log on to specific pages (usually daily) and provide information about the previous 24 hours or complete specific tasks (e.g., a cognitive test). Perhaps one of the major discoveries in this area concerns the importance of social