AN INVESTIGATION INTO CLINICIAN NUMERACY IN MEDICAL STUDENTS

A thesis submitted to The University of Manchester for the degree of
Doctor of Medicine
in the Faculty of Biology, Medicine and Health

2019

ANNE TAYLOR
Division of Medical Education
School of Medical Sciences

LIST OF CONTENTS

Word Count: 50,850

LIST of TABLES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Appendices

LIST of FIGURES

Chapter 2

Chapter 3

Chapter 4

Chapter 5

LIST of APPENDICES

LIST of ABBREVIATIONS

AIM	Acute Illness Management
A-level	General Certificate of Education - Advanced level
ALS	Advanced Life Support
ANTT	Aseptic Non Touch Technique
ARR	Absolute Risk Reduction
AS-level	General Certificate of Education - Advanced Subsidiary level
ASME	Association for the Study of Medical Education
ATLS	Advanced Trauma Life Support
BBC	British Broadcasting Corporation
BDA	British Dyslexia Association
BPS	British Pharmaceutical Society
CN	Clinician Numeracy
CR	Constructed Response
DEMEC	Developing Excellence in Medical Education Conference
EBM	Evidence-Based Medicine
FT	Foundation Trainee
GCSE	General Certificate of Secondary Education
GMC	General Medical Council
HESA	Higher Education Statistics Agency
HN	Health Numeracy
IV	Intravenous
IQR	Interquartile Range
KSoM	Keele School of Medicine
KU	Keele University
MCQ	Multiple Choice Question
MINT	Medical Interpretation and Numeracy Test
MRSA	Methicillin-Resistant Staphylococcus Aureus
MSC	Medical Schools Council
NACT	National Association of Clinical Tutors
NEWS	National Early Warning System
NICE	National Institute for Health and Care Excellence
NHS	National Health Service
NHSI	National Health Service Improvement
NLQ	Nutritional label questions
NPSA	National Patient Safety Association
NVS	Newest Vital Sign test
OECD	Organisation for Economic Cooperation and Development
OSCE	Objective Structured Clinical Examination
PISA	Programme for International Student Assessment

PSA	Prescribing Safety Assessment
Q	Question number
RCT	Randomised Controlled Trial
RRR	Relative Risk Reduction
SBA	Single Best Answer
SBAR	Situation Background Assessment Recommendation
SD	Standard Deviation
SPLD	Specific Learning Disability
UHNM	University Hospitals of North Midlands
UK	United Kingdom
UKCAT	UK Clinical Aptitude Test
UoM	University of Manchester
UREC	University Research Ethics Committee
US	United States (of America)
VSA	Very Short Answer
WHO	World Health Organisation

ABSTRACT

Background

Numeracy in the general population is low: approximately 50% of adults of working age have numeracy skills at the level expected of an average schoolchild at age 11. There is increasing evidence that many students entering university across the globe have numeracy levels significantly below those required for their chosen courses. Clinician Numeracy (CN) is the ability to understand and use numerical information and quantitative data of all kinds in delivering safe treatment to patients. Research demonstrates that many medical students and doctors worldwide have low clinician numeracy, with implications for safe patient care.

Methods

My research includes the evaluation of various assessments of CN, followed by a comprehensive review of one assessment, the Medical Interpretation and Numeracy Test. I have conducted two randomised controlled trials: the first to assess whether calculators influence CN test results, and the second to investigate whether answer format (multiple choice or constructed response) influences CN test results. Finally, I investigated and classified the types of error being made by test participants.

Results

Following analysis, the Medical Interpretation and Numeracy Test appeared to be the most appropriate measure of CN for medical students and doctors. The Medical Interpretation and Numeracy Test was subjected to an emendation process, and psychometric analysis indicated that it is a reliable and valid test of CN. The revised test was used for my research. Approval for this research was granted by the University of Manchester Research Ethics Committee. All research was carried out on third year medical students in a single institution in England. The first randomised controlled trial showed that calculators did not affect test scores. The second showed that the multiple choice (single best answer) format of the test was equivalent to the constructed response (very short answer) format. The exploration of error demonstrated that most errors related to a failure to set the problem up correctly; furthermore, the type of errors being made were basic mistakes, similar to those made by nursing and biomedical science students.

Conclusion

This research confirms that the level of CN in medical students is variable, and that some have low numeracy. This is consistent with national and global data relating to low numeracy in the general population, in schoolchildren, in university students, and in healthcare professionals. Medical schools must ensure their graduates are competent to provide safe patient care; graduates have a responsibility to identify and remediate areas of weakness that may affect their practice. It is time for Clinician Numeracy to be included in medical curricula.

DECLARATION

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

COPYRIGHT STATEMENT

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and she has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=2442 0), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in The University’s policy on Presentation of Theses.

DEDICATIONS

I would like to thank my supervisors, Lucie and Doug for their help and support throughout this process.

I am very grateful to my colleagues at Keele, for their interest in my research and for their encouragement and support.

I am grateful to HEWM who awarded me a bursary towards course fees for three years of my studies.

I would like to thank Helen in the research department at UHNM for advice regarding statistics.

I am grateful to Len for his enthusiasm and support of my research; I would also like to thank the peer reviewers of my papers for their constructive feedback.

Finally, I could not have embarked on or completed this project without the backing of my wonderful family – go raibh míle maith agaibh go léir, is grá mór.

"Think it over, think it under."
From *Winnie the Pooh*, by A.A. Milne

"…. A hard time we had of it.
At the end we preferred to travel all night,
Sleeping in snatches,
With the voices singing in our ears, saying
That this was all folly."

From *The Journey Of The Magi*, by T.S. Eliot

'"Begin at the beginning," the King said, very gravely, "and go on till you come to the end: then stop."' From *Alice in Wonderland*, by Lewis Carroll

THE AUTHOR

I have been a consultant anaesthetist in the NHS since 1996, initially in Birmingham, and subsequently in Staffordshire. I have been involved in medical education throughout my career, and am an instructor for a range of advanced life support courses (ATLS since 1996, ALS since 2011, AIM since 2012). I have also held formal educational roles, initially in postgraduate education as College Tutor in Anaesthesia from 2001 – 2008, and then in undergraduate education as a Keele University Hospital Dean since 2010.

Qualifications

1986	MB, BCh, BAO	University College Dublin, National University of Ireland
1991	FFARCSI	Royal College of Surgeons of Ireland, Dublin, Ireland
2014	MSc (Med Ed)	University of Manchester, Manchester, England.

Employment

1996	Consultant Anaesthetist	City Hospital NHS Trust Dudley Road, Birmingham
1998	Consultant Anaesthetist	Mid Staffordshire NHS Foundation Trust, Weston Road, Stafford
2014	Consultant Anaesthetist	University Hospitals of North Midlands NHS Trust Stafford and Stoke-on-Trent

Relevant publications

Taylor A.A., & Byrne-Davis, L.M. (2017). Clinician numeracy: use of the medical interpretation and numeracy test in foundation trainee doctors. *Numeracy, 10*(2), 5. Retrieved from http://doi.org/10.5038/1936-4660.10.2.5

Taylor A.A., & Byrne-Davis, L.M. (2016). Clinician numeracy: the development of an assessment measure for doctors. *Numeracy, 9*(1), 5. Retrieved from http://dx.doi.org/10.5038/1936-4660.9.1.5

Taylor, A. (2014). The development, validation and implementation of a test to assess health numeracy in doctors. (Unpublished MSc dissertation). University of Manchester, Manchester, UK.

Taylor, A., & McCarroll, M. (1994). Sedation for non-invasive procedures: a potential hazard. *Irish Journal of Medical Science*, *163*,(4), 217. Retrieved from https://doi.org/10.1007/BF02967232

THESIS PRESENTATION

This thesis is presented in the Journal Format. My supervisors and I discussed the most appropriate format for this thesis before I started my research. We agreed that I should aim to present my thesis in Journal Format, since the nature of my research lends itself to distinct chapters suitable for publication as journal articles. This format was also potentially advantageous, since it meant that I would be able to submit a chapter for publication without having to wait for the full thesis to be completed, thus preventing delay in publishing any findings that are of interest in the current literature. Furthermore, it had the advantage that chapters would be written as papers ready for publication, and so it would avoid the problem of trying to extract a paper from a thesis written in standard format. Finally, presenting the thesis in this way imposed the additional discipline of academic writing for a journal: requiring that chapters were more focussed, and the work was tailored to a restricted word count.

However, a recognised drawback to this format is the necessity to introduce the topic at the beginning of each paper, thereby leading to some degree of repetition in the thesis. While I have made every effort to minimise such repetition, it is an inevitable and accepted limitation of this format, recognised in the University's guidance.

Points of Presentation Style

1. I have followed university guidance suggesting that the thesis be presented in a consistent format, rather than varying font, spacing and referencing style according to the requirements of different journals.
2. After discussion with my supervisors, all references are placed at the end of the thesis, rather than at the end of individual chapters. I have used the APA referencing style throughout. (www.subjects.library.manchester.ac.uk/referencing/referencing-apa)
3. Chapters that may be submitted for publication are written using the first person plural as they will be co-authored with my supervisors.

Contribution of others

I have met my supervisors regularly during the course of my research for this MD, and they have provided useful advice on the structure of the research, the design of individual studies, and all drafts of chapters, and papers included in this thesis.

I have conducted all of the studies, all of the data analysis and written all of the papers and chapters in this thesis.

STRUCTURE OF THESIS

This thesis presents my research into Clinician Numeracy in medical students since 2015, and comprises 6 chapters.

Chapter 1. An introduction to Clinician Numeracy
This chapter starts with an introduction to the concepts of numeracy, health numeracy and Clinician Numeracy (CN). It then discusses patient safety and error in healthcare. It sets the context of my research into CN, and its relevance to safe medical practice and improved patient safety. This chapter then goes on to review the various available assessments of CN, to determine the most appropriate measure to use for my research.

Chapter 2. Revision of the Medical Interpretation and Numeracy Test
This chapter discusses the comprehensive evaluation of the Medical Interpretation and Numeracy Test (MINT) leading to the development of two revised tests, a constructed response format (MINTv2) and a single best answer format (MINTv3). This chapter also presents psychometric analysis of the revised tests, and compares this with data relating to the original test (MINTv1).

Chapter 3. The impact of Calculators on a test of Clinician Numeracy: a randomised controlled trial
This chapter describes a randomised controlled trial conducted to assess whether using a calculator would improve participants' test scores. This chapter has been peer reviewed and accepted for publication in the July edition of the journal *Numeracy*.

Chapter 4. Assessing Clinician Numeracy: Single Best Answer or Very Short Answer? A randomised controlled trial
This chapter considers the relative merits of single best answer and constructed response test formats, and then discusses a randomised controlled trial comparing these test formats. I have submitted this chapter as a "Brief Report" to the *Journal of Experimental Education*.

Chapter 5. Errors made by medical students in a test of Clinician Numeracy
This chapter describes an in-depth exploration into the errors made by medical students in the MINT. It describes a new classification system for the errors made in tests of CN, and considers the kinds of errors made by participants. This chapter has been prepared for submission to the journal *Medical Teacher*.

Chapter 6. Discussion
This chapter reviews my research over the course of the past four years, and considers its relevance to medical education, and its place in the literature.

Blank Page

CHAPTER 1

AN INTRODUCTION TO CLINICIAN NUMERACY

TABLE OF CONTENTS: Chapter 1

TABLES

1.1 BACKGROUND: WHAT IS CLINICIAN NUMERACY?

This thesis is concerned with Clinician Numeracy (CN), but also refers to the concepts numeracy and health numeracy. These concepts are closely related but distinct. “Numeracy” is an essential life skill, encompassing the ability to use and apply numbers and mathematical concepts for problem-solving in all aspects of life, from managing a household budget to understanding quantitative information presented in the media; it “is as much about thinking and reasoning logically as about 'doing sums'” (www.nationalnumeracy.org.uk). This definition is important: numeracy is not only about being able to do maths, it is integral to logical thinking and problem-solving. This is further emphasised by the Organisation for Economic Cooperation and Development (OECD) who define mathematics as “the science of pattern recognition” in data relating to the Programme for International Student Assessment (PISA), (OECD, 2009). This is relevant to clinical medicine, since it is well recognised that for experienced clinicians, decision making and clinical reasoning rely primarily on pattern recognition (Elstein & Schwartz, 2002). However, despite its importance, it is estimated that almost half of adults of working age in the UK have numeracy levels equivalent to those of primary schoolchildren, and thus face challenges in the workplace, in managing their finances and in decision making (www.nationalnumeracy.org.uk). The situation is similar in the US (Reyna *et al* 2009; OECD, 2013).

Low numeracy is associated with adverse outcomes in healthcare, prompting considerable research into numeracy in patients (Weiss *et al* 2005; Rothman *et al* 2006; Huizinga *et al* 2008; Schapira *et al* 2008; Reyna *et al* 2009; Rowlands *et al* 2013; Lag *et al* 2013), and leading to the notion of health numeracy as a discrete entity. People with low health numeracy often struggle to manage chronic medical conditions such as asthma and diabetes, may not comply with medical advice regarding treatment, are at an increased risk of experiencing the side effects of treatments, have an increased incidence of hospital admission, an increased risk of readmission following discharge from hospital, and even increased mortality (Baker *et al* 2002; Gazmararian *et al* 2003; Gazmararian *et al* 2005; Apter *et al* 2006; Baker *et al* 2007; Huizinga *et al* 2008; Reyna *et al* 2009).

Health numeracy (HN) essentially means being able to understand numerical data as it applies to healthcare: Golbeck *et al* (2005) define it as “the degree to which individuals have the capacity to access, process, interpret, communicate, and act on [the] numerical, quantitative, graphical, biostatistical, and probabilistic health information needed to make effective health decisions.” Thus HN is a complex concept, and various frameworks have been developed to consider its components, and to study its application in healthcare (Nutbeam, 2000; Golbeck *et al* 2005; Ancker & Kaufman, 2007; Schapira *et al* 2008). The framework developed by Golbeck *et al* (2005) divides numeracy into four distinct, but overlapping constructs: basic, computational, analytical and statistical numeracy, and provides a useful way of understanding the importance of numeracy for both patients and healthcare practitioners (Table 1.1).

Table 1.1. Health numeracy and Clinician numeracy*

Construct	**Health numeracy** *(Golbeck et al 2005)*	**Clinician numeracy**
Basic numeracy	Number recognition Making sense of numerical data No manipulation of numbers	Prerequisite to formal education
Computational numeracy	Basic mathematical skills Simple manipulation of numbers in a health context e.g. ability to comply with instructions regarding prescribed medicines	Calculation of drug doses, fluid & nutritional regimens Use of medical formulae Accurate disease management Accurate advice to patients regarding disease management
Analytical numeracy	Making sense of numerical data Understanding charts & graphs, proportions, percentages, & frequencies in relation to managing own health	Interpreting medical data presented in different formats, including research results Diagnostic skills Understanding drug pharmacokinetics Estimation of accuracy of calculations Disease management Clinical decision making & treatment selection
Statistical numeracy	Understanding basic biostatistics including concepts such as risk & probability Able to compare treatment options offered by healthcare professionals, & participate in decision making	Understanding data related to risk as presented in various formats Ability to communicate risk Understanding medical research Interpretation of medical statistics Clinical decision making & treatment selection Practicing evidence-based medicine

** Adapted from Taylor & Byrne-Davis (2016)*

While the definition of health numeracy given above would allow it to be used to describe numeracy in healthcare professionals, this term is widely used in the literature to refer to levels of HN in the general population; thus it could be confusing to use it in relation to healthcare professionals. Furthermore, it is implicit in the HN literature that low numeracy is the preserve of patients, with recommendations for doctors on identifying and managing patients with low HN (Schwartz *et al* 1997; Weiss *et al* 2005; Rothman *et al* 2006; Huizinga *et al* 2008; Rowlands *et al* 2013). Finally, since doctors and other healthcare professionals share a responsibility for patient care, they must be competent in interpreting more complex quantitative medical data than is necessary for patients. Thus, the term “clinician numeracy”, defined by Caverly *et al* (2012) as “the ability to use numbers and numeric concepts in the context of taking care of patients”, is more appropriate to use when considering the numeracy of healthcare professionals.

Clinician numeracy (CN), is essential for many everyday tasks in medical practice; it includes the numerical skills required for accurate drug dose calculation (computational numeracy), the ability to interpret medical data (analytical numeracy), clinical reasoning, problem-solving and decision-making skills (analytical numeracy), and understanding and communicating information related to probability and risk (statistical numeracy) (Table 1.1). However, there is evidence from research worldwide that CN in medical students and qualified doctors may be deficient: it has long been recognised that medical students and doctors may find drug dose calculations difficult (Rowe *et al* 1998; Selbst, 1999; Wheeler *et al* 2004a; Wheeler *et al* 2004b; Simpson *et al* 2009; Harries & Botha, 2013). Moreover, many doctors and medical students struggle with data interpretation, particularly in relation to statistics (Sheridan

& Pignone, 2002; Ghosh & Ghosh, 2005; Windish *et al* 2007; Gigerenzer *et al* 2007; Wegwarth *et al* 2012; Johnson *et al* 2014). Nonetheless, the lack of research into CN suggests that this is not considered to be a significant problem, and time is rarely found for CN in medical education curricula despite repeated calls for its inclusion (Wheeler *et al* 2004b; Ghosh & Ghosh, 2005; Gigerenzer *et al* 2007; Johnson *et al* 2014). Gigerenzer *et al* (2007) suggest that the lack of attention paid to CN in medical education may be related to a phenomenon they refer to as "collective statistical illiteracy" i.e. because numeracy is so low generally in the population, a deficiency in medical students and doctors is relatively concealed. However, educators should take note of the emerging recognition of low numeracy in schoolchildren at age 15 (OECD, 2009) and in university students worldwide (McLean *et al* 2011; Tariq, 2002; Tariq, 2008; Sikorskii *et al* 2011; Follette *et al* 2015; National Numeracy, 2019). Thus, low numeracy in medical students and doctors is part of a larger problem affecting the population.

Although there is comprehensive evidence documenting the impact of an individual's low numeracy on their own health, evidence associating a doctor's low numeracy with the quality of their practice or the health of their patients is yet to accrue. However, it is clear that low CN could lead to errors in drug dose calculation, and to poor decision making and treatment selection. Where deficits in knowledge and skills have been identified, steps must be taken to address them. Medical educators have a responsibility to ensure that their graduates are safe practitioners. Medical students and doctors have a responsibility to be aware of their limitations and to address areas of weakness. It is time to pay attention to CN.

1.2 PATIENT SAFETY AND MEDICAL ERROR

The fundamental aim of healthcare services is to make sick people better. People who are unwell seek the help of healthcare professionals, trusting that doctors will be competent to diagnose the cause of their illness, know about the available treatment options, including the benefits and risks associated with treatment, and thus provide informed advice on disease management. However, a wealth of evidence shows that many patients suffer harm because of medical error (Brennan *et al* 1991; Vincent *et al* 2001; Neale *et al* 2001; Kohn *et al* 2002; Leape & Berwick, 2005; Berwick, 2013; Jha *et al* 2013; Vincent *et al* 2014; Malhotra *et al* 2015; O'Hara *et al* 2018). The consequences can be devastating, not only for patients and their relatives, but also for clinicians. Healthcare staff involved in serious errors suffer emotional costs (guilt, shame, depression, loss of confidence and morale), may find their careers ruined, and occasionally may be convicted of manslaughter (Coben & Weeks, 2014; BBC news, 2015; Vaughan, 2018). Healthcare organisations pay the price in terms of financial penalties and reputational damage. All stakeholders, therefore, share a common goal: the provision of safe, effective patient care. Patient safety is now established as the top priority for the National Health Service (NHS) (Department of Health, 2015; General Medical Council (GMC), 2015; National Health Service Improvement (NHSI), 2017), with organisations such as the National Patient Safety Agency (NPSA) and initiatives such as Patient Safety Collaboratives (NHS England, 2014) directed towards the delivery of better quality, safer care for patients. Moreover, patient safety is a priority in medical education, endorsed by the World Health Organisation (WHO) (WHO, 2009) and the GMC (2015). Since deficiencies in CN may lead to medical error and patient harm, medical educators should include CN in undergraduate and postgraduate medical curricula in the interest of patient safety.

1.3 ADVERSE EVENTS IN HEALTHCARE

Despite major efforts to improve the quality of care delivered to patients by the NHS in recent years, approximately 10% of patients continue to suffer adverse events following healthcare intervention (Vincent *et al* 2001; Vincent *et al* 2014). In the NHS, an adverse event is defined as “any unintended or unexpected incident which could have or did lead to harm for one or more patients receiving NHS funded healthcare” (National Patient Safety Association (NPSA), 2011). Adverse events may relate to omissions and failures of various kinds, and can occur at all levels throughout a healthcare organisation: although they vary greatly in their severity, some will cause permanent and catastrophic harm to patients, including death and life-changing disability. Most serious errors are multifactorial, and result from unsafe acts by individuals combined with organisational or system failures (Reason, 2000; Vaughan, 2018). Almost half are preventable (Vincent *et al* 2001; de Vries *et al* 2008; Smits *et al* 2010; Vincent *et al* 2014; Rafter *et al* 2015). A classification of patient harm is shown in Table 1.2, indicating how deficiencies in CN can impact on patient care. In addition to the human cost of healthcare

Table 1.2. Classification of patient harm*

Definition	**Examples**	**Potential impact of deficiencies in CN**
Treatment-specific harm (related to particular treatments or a particular disease) *Not all such harm is preventable*	Adverse drug reactions Adverse effects of chemotherapy Surgical complications Wrong site surgery	Drug dose calculation error Data interpretation error may lead to error in treatment selection
Harm due to over-treatment	Polypharmacy and consequent drug interactions Excessive investigations (blood tests, X-rays) Unnecessary procedures	Drug dose calculation error Data interpretation error may result in inappropriate & unnecessary tests and procedures.
General harm from healthcare (includes failure to recognise the impact of frailties or co-morbidities)	Hospital-acquired infections Falls Delirium Dehydration	Deficiencies in CN may lead to errors in calculating fluid & electrolyte requirements, leading to dehydration
Harm resulting from delayed or inadequate diagnosis	Adverse outcome due to disease progression	Data interpretation error may interfere with timely diagnosis
Harm due to failure to provide appropriate treatment: many patients fail to receive standard evidence-based care	Failure to provide rapid thrombolytic treatment for stroke	Deficiencies in CN may lead to error interpreting research & biostatistical information, thus preventing evidence-based practice

**Adapted from Vincent et al (2014)*

error, the financial consequences are considerable, as illustrated in Table 1.3 (Frontier Economics, 2014). Thus, reducing medical error has the potential to improve patients' lives, and those of their families, as well as improving the use of limited NHS funds.

Table 1.3. Financial costs associated with preventable adverse events in the UK*

Source	Data
National Patient Safety Agency (NPSA) (2007)	Medication errors in 2007 cost £770 million due to: the cost of admissions for adverse drug reactions, and the cost of harm due to medicine during inpatient stays
Cranshaw *et al* (2009)	Drug-related medical errors cost the NHS in England £5 million from litigation between 1995 and 2007
NHS Education for Scotland (2010)	Inadvertent retention of medical equipment inside patients during interventions cost the NHS £9 million in medical negligence compensation over a 5-year period
Briggs (2012)	Infections following total hip or knee replacement can cost £70,000 per patient to treat. If the lowest infection rates could be achieved throughout the NHS, this would save £200-£300 million each year
Parliamentary & Health Service Ombudsman (2013)	Better recognition of sepsis could save the NHS £4,000 per patient in terms of reduced hospital stay This could save £196 million per year

** adapted from Frontier Economics (2014)*

1.4 ADVERSE EVENTS FOR WHICH DOCTORS ARE RESPONSIBLE

As can be seen from Table 1.2, doctors may be responsible for many of the adverse events that occur in hospitals. These can be broadly classified into two groups: medication errors, and errors related to inept clinical decision-making.

Medication Error

Medication error is a common preventable event that causes significant patient morbidity and mortality (Armitage & Knapman, 2003; NHSI, 2017; Wise, 2018; Elliott *et al* 2018). There are approximately 237 million medication errors in the NHS in England annually, of which 66 million are considered to be clinically significant (Elliott *et al* 2018; Wise, 2018). Although it is estimated that only 2% of medication errors are likely to cause severe harm, this results in approximately 700 deaths, and may contribute to a further 22,000 deaths annually (Elliott *et al* 2018; Wise, 2018). The financial cost for the NHS in England is estimated at almost £100 million every year (Elliott *et al* 2018). On a global scale, medication errors are estimated to cost approximately $42 billion worldwide every year (World Health Organisation (WHO) 2017; Donaldson *et al* 2017). Therefore, the WHO has launched a global challenge to reduce the incidence of medication-related harm by 50% over five years (WHO, 2017; Donaldson *et al* 2017).

Data from the NPSA indicates that the most frequent type of medication error in the UK is "wrong dose, strength or frequency of medication" (NPSA, 2007); this may be related to poor clinician numeracy in doctors and nursing staff. Medication error can be broadly classified into prescribing error, and errors related to drug preparation and administration. Prescribing error is extremely common: at least 10-15% of prescriptions contain an error of some kind (Dean *et al* 2002; Coombes *et al* 2008; Dornan *et al* 2009; Franklin *et al* 2011; Tully, 2012), although the incidence of error may be as high as 35 - 44% (Gleason *et al* 2010; Seden *et al* 2013). Prescribing is largely the responsibility of qualified doctors, although some senior nurses are also qualified to prescribe. Prescribing drugs safely requires competence in multiple areas, and is an essential skill for UK doctors (GMC, 2018). Multiple factors contribute to prescribing error, and inadequate training and supervision of junior doctors are significant underlying causes (Table 1.4) (Dean *et al* 2002; Coombes *et al* 2008; Dornan *et al* 2009; Glavin, 2010; Franklin *et al* 2011; Tully, 2012; Avery *et al* 2012; Seden *et al* 2013; Ryan *et al* 2014). To calculate the

Table 1.4. Causes of prescribing error using Reason's model*

Active Failures	**Error-provoking conditions**	**Latent conditions**
Individual unsafe acts	Task and environment	Organisational processes
Knowledge-based mistakes	*Individual*	*General*
Lack of knowledge of drug, including dose & interactions	Hungry, thirsty, tired, low morale, distracted	Long shifts
		Inadequate staffing
Lack of patient information	Inadequate knowledge, skill, experience, training	Reliance on locums
		Reluctance to challenge those with greater authority
Skill-based mistakes		
Slips & lapses: may be due to lack of concentration, multi-tasking, interruptions	*Working environment*	
	Staffing levels inadequate	Need to admit specialist patients out of hours
	New or locum staff	
Memory lapses	Unfamiliar patient	Lack of feedback systems
	High workload, long hours, pressure	Difficulties in storing data
Rule-based mistakes	Lack of access to drug information	Difficult to access specialised expertise
Lack of knowledge of the rule	Lack of access to patient information & computer workstations	
Failure to follow the rule		Poor conflict resolution
Application of the wrong rule		
Inadequate monitoring	*Health-care team*	
	Communication problems e.g. poor handovers, poor hand writing in notes, inadequate records	*Prescribing*
Violations		Lack of training
Deliberate deviation from policy or procedure		Assumption that others will check prescription
	Failure to seek or take advice	
	Inadequate training, knowledge & experience	Simultaneous multiple prescribing tasks
	Inadequate supervision	Pharmacy systems separate from clinical services
	Failure to escalate	
	Difficulty weighing risks & benefits	
	Unclear who is responsible for patient or task	
	Prescribing task	
	Medical chart layout ambiguous	
	Guidelines unavailable	
	Doses & protocols not standardised	
	Doctor unfamiliar with task	
	Patient	
	Acute or Complex problem	
	Complex clinical disease	
	Unhelpful or difficult patient; communication problems	

**Adapted from Dornan et al 2009*

correct dose of a drug for an individual patient, the doctor must take into account information about the drug, including its therapeutic and toxic doses, as well as considering patient factors (e.g. age, renal function). Errors in drug dose calculation may result in a patient receiving a sub-therapeutic dose, or an overdose. A simple error involving one decimal place can be catastrophic, as it will lead to an error of ten times the correct dose. Drug dose calculation errors are particularly devastating in paediatric practice where the margin for error is often very low (Rowe *et al* 1998; Selbst *et al* 1999; Hughes & Edgerton, 2005). Medication error may also result from errors in drug preparation and administration; this may result in the wrong drug or the wrong dose of the drug being given to the patient, or administration of the drug by an incorrect route (e.g. intrathecal rather than intravenous vincristine). Fortunately, the majority of errors are detected before the wrong drug or dose is administered to the patient (Dean *et al* 2002; Coombes *et al* 2008; Dornan *et al* 2009; Seden *et al* 2013; Elliott *et al* 2018). Nevertheless, medication error is a massive problem worldwide, with considerable human and financial costs; therefore, the World Health Organisation (WHO) has recently launched a global initiative to combat medication error worldwide (WHO, 2017; Donaldson *et al* 2017).

Errors in Decision Making

Medical practice is changing all the time. As new procedures become established, old ones become obsolete, and doctors need to keep abreast of new developments and innovations in order to provide the best care for their patients. Furthermore, doctors are expected to practice evidence-based medicine (EBM). This requires them to be competent in evaluating the information relating to new treatments that is published in journals, and presented at conferences. However, this is challenging for many medical students and doctors, who are confused by biostatistical information, particularly in relation to conditional probabilities and screening data (Ben-Shlomo *et al* 2004; Ghosh & Ghosh, 2005; Gigerenzer *et al* 2007; Windish *et al* 2007; Rao & Kanter, 2010; Moyer, 2012; Wegwarth *et al* 2012; Gigerenzer, 2014; Johnson *et al* 2014). Misinterpretation of medical data can lead to poor decision making, and unnecessary interventions, resulting in significant patient harm, and wasting valuable healthcare resources (Tables 1.2 & 1.3).

There are numerous and diverse reasons why errors occur in healthcare, and most serious errors are multifactorial; nonetheless, it is estimated that around 50% of harm in healthcare is preventable (Vincent *et al* 2014; Rafter *et al* 2015). Since poor CN is one possible cause of patient harm, it is important to consider how this can be managed. A review of error management in healthcare generally may provide insight into how to manage error related to poor CN.

1.5 ERROR MANAGEMENT

Where there are humans, there will inevitably be human error, thus error in healthcare cannot be eradicated completely (Reason, 2000). Therefore, it is important to focus on managing error by: a) reducing its incidence; b) developing systems which can deal with and mitigate errors that occur; and c) learning from error (Reason, 2000). Reason's (2000) Swiss cheese model is commonly used to illustrate how a combination of individual and system failures leads to error in healthcare (Table 1.4). Understanding and learning from error in healthcare has been greatly advanced by applying the principles of human factors science to understand the impact of people's behaviour and interactions on outcome in healthcare (Brennan *et al* 2005; NHS England, 2013). The recognition that similar errors occur repeatedly, worldwide, due to the same failures in healthcare systems, has led to improvements in healthcare globally e.g. the introduction of the WHO surgical safety checklist has reduced the incidence of surgical error worldwide (Fudickar *et al* 2012). NHS England advocates the adoption of human factors concepts as a crucial step in delivering culture change, improving patient safety and achieving clinical excellence (NHS England, 2013); thus it is now mandatory to use the WHO surgical safety checklist before all invasive procedures in the NHS. Furthermore, the level of compliance with its use is measured as an indicator of quality in hospitals. Some of the error prevention strategies currently used in the NHS are shown in Table 1.5.

Table 1.5. Healthcare errors and interventions to reduce or prevent error

Error	Intervention
Wrong patient, wrong site, wrong side surgery	WHO surgical safety checklist
Failure to recognise deteriorating patient	Education: AIM, ALS, NEWS
Poor communication	Training in effective communication e.g. SBAR
Increased incidence of surgical complications in small hospitals	Centralisation of services
Drug administration error	Bar-coding in pharmacy
Wound infection	Antibiotic prophylaxis
Cannula site infection	ANTT procedure
Prescribing error	e-prescribing
Inadvertent administration of drug by wrong route (e.g. IV bupivacaine)	Design of new equipment specific to route, incompatible with wrong route
Confusion due to drugs with similar names	Tallman lettering system
Poor handover at change of shift	Targeted training in effective handover
Near miss event	Team training using simulation

Using Reason's model, and Human Factors science, it is apparent that attempts to prevent errors made by clinicians must focus both on the individual and on the overall healthcare system. Gordon *et al* (2013) recommend adopting a Human Factors approach to improving the prescribing skills of doctors. As shown in Table 1.4, errors made by clinicians can be attributed to failures in knowledge, skills or behaviour across all areas of their practice. Therefore, there are diverse ways in which efforts to reduce clinician error can be directed. This thesis addresses one small area - clinician numeracy - where improvements may help to reduce error.

Error Management and Clinician Numeracy: A Systems Approach

Medication error

Initiatives to reduce medication errors have been introduced, following the recognition that similar errors occurred repeatedly either within a single institution, or across many different institutions both nationally and worldwide. Medication errors may relate to prescribing, to drug preparation and/or to drug administration.

Drug presentation and packaging

Drug presentation and packaging is a common preventable cause of medication error (Orser *et al* 2001; Berman, 2004; Momtahan *et al* 2008), and some research suggests that over 50% of medication errors in the US may be due to similarities in drug names, labelling or packaging (Berman, 2004). Drugs from the same ‘family’ often have similar names (e.g. the cephalosporin antibiotics ceftazidime, cefuroxime, cefalexin); however, completely unrelated drugs may also have similar names (e.g. anectine, anexate) with life-threatening results (Taylor & McCarroll, 1994). Furthermore, a drug may be available in different concentrations (e.g. ketamine 10mg/ml and 50mg/ml). The situation is exacerbated by the fact that pharmaceutical companies tend to have their own branding style, and thus present their products in very similar packaging. Finally, many intravenous (IV) drugs are prepared in small vials with minute writing that is difficult to read. These issues pose real risks to patient safety. Using a systems approach to tackle this problem has led to several simple, but effective, solutions e.g. stocking only one concentration of a drug i.e. either ketamine 10mg/ml or ketamine 50mg/l, but not both; or procuring drugs from different pharmaceutical companies in order to reduce the similarities in packaging. An interesting innovation has been the introduction of the Tall Man lettering system to enhance the differences in drug names e.g. the cephalosporin antibiotics are labelled as cefTAZidime and cefUROXime (Filik *et al* 2006; Irwin *et al* 2013). However, although this system has been adopted widely, there is doubt as to its effectiveness in practice (Lambert *et al* 2016; Zhong *et al* 2016).

Drug dose calculation

Drug dose calculation error may initially appear to be a problem that can only be resolved at the individual level, by improving that person’s numeracy. However, a systems approach can be useful, since some of the difficulty related to drug dose calculation, particularly for intravenous (IV) drugs, is caused by variations in drug labelling. Although most IV drugs are labelled as mass per unit volume (e.g. atropine 0.6mg/ml), some are labelled in terms of percentage (e.g. chirocaine 0.5%), and others as proportion (e.g. adrenaline 1:1000). When drugs are labelled in terms of percentage or proportion, medical students and doctors may become confused, as it is not immediately clear how much drug in mg/ml is present in the solution (Wheeler *et al* 2004a; Wheeler *et al* 2007; Harries & Botha 2013). They almost invariably need to convert from the amount given in percentage or proportion to mg/ml in order to calculate the correct volume of drug to administer; introducing this additional step to the process makes the calculation more difficult, and increases the likelihood of error. Standardisation in labelling of IV drugs to mass per unit volume would overcome the need for conversion from percentage or proportion to mg/ml, simplifying the calculation, and reducing the incidence of error (Rolfe & Harper, 1995; Orser *et al* 2001; Wheeler *et al* 2004a; Wheeler *et al* 2007; Momtahan *et al* 2008; Harries & Botha, 2013). Another method of reducing calculation error is to provide tables with doses pre-calculated for patients of different weights; alternatively, Williams & Walker (2014) describe a nomogram to overcome the need for complex calculations in anaesthetic practice.

Electronic prescribing

Electonic prescribing (e-prescribing) is now widely used, and aims to improve patient safety by reducing human error e.g. mistakes due to poor handwriting, and to drug dose calculation error. However, the evidence so far suggests that while many types of error are reduced or eliminated, new problems have arisen (Tully, 2012; Ahmed *et al* 2016). Errors may occur because of incorrect data entry, because patients have complex co-morbidities, and are taking many different drugs, because doctors override the system and ignore warnings designed to prevent error; furthermore, reliance on technology is potentially risky and may lead to deskilling.

Data interpretation

A combined individual and Human Factors approach could also help overcome difficulties related to data interpretation. While better education in medical statistics is required at the individual level, simplification and standardisation of data presentation is also needed. Indeed, Gigerenzer & Edwards (2003) suggest that poor presentation of medical data is the primary cause of difficulty in understanding statistical information, observing that "for each confusing representation [of medical data] there is at least one [good] alternative". An example is the framing of research data as relative risk reduction (RRR) rather than absolute risk reduction (ARR): this enhances the apparent effect of a drug or medical intervention, and is often misleading. The standardisation of presentation to ARR would reduce confusion.

"Choosing Wisely" is an initiative to improve patient care by reducing unnecessary investigations and procedures (Malhotra *et al* 2015). This is relevant because low CN in doctors, and poor understanding of biostatistical information may lead to poor decision making and treatment selection (Gigerenzer *et al* 2007).

Addressing Clinician Numeracy: The Individual Approach

A decline has been observed in the numeracy skills of incoming undergraduates to medical, pharmacy and bioscience degrees (Malcolm & McCoy, 2007; Tariq, 2008; Whittle *et al* 2010; McLean *et al* 2011; Sikorskii *et al* 2011; National Numeracy, 2019). Moreover, some medical students who indicated a lack of confidence and competence in their numeracy skills on entry to university, considered that their numeracy had declined further after a year in medical school (McLean *et al* 2011). This is of concern since numeracy is among the generic transferable skills required to equip doctors for life-long learning and evidence-based practice (Whittle *et al* 2010; McLean *et al* 2011). There is evidence that many UK graduates feel unprepared for clinical decision-making and prescribing (Heaton *et al* 2008; Monrouxe *et al* 2014; Nazar *et al* 2015). Although the aetiology of this is likely to be multifactorial, poor numeracy may be a contributory factor. Medical schools are responsible for producing graduates who are fit to practice (GMC, 2015, 2018), and must address areas where gaps in knowledge or skills have been identified in their graduates. The introduction of educational intervention to improve clinician numeracy in undergraduates may lead to greater competence and confidence in prescribing and decision making by newly graduated doctors; this could also reduce the incidence of error (NPSA, 2007; Gigerenzer *et al* 2007; Wheeler *et al* 2008; Rao & Kanter, 2010; Moyer, 2012; Johnson *et al* 2014).

1.6 THE ASSESSMENT OF CLINICIAN NUMERACY

The evidence demonstrating that many medical errors are caused by inaccurate drug dose calculation or data interpretation suggests that improving CN in medical students and doctors could reduce the incidence of these errors. Therefore, we need a way to assess CN reliably, in order to measure CN in medical students and doctors, and consider whether it is sufficient for safe practice. There are numerous numeracy tests of varying levels of difficulty available in non-healthcare contexts. However, these are not generally suitable for clinicians; this is partly because their content often includes higher mathematical functions such as algebra and trigonometry that are irrelevant to medical practice, but also because the context of the test material is important. Levy *et al* (2014) recruited almost 1000 adults to a study where data on risk and probability were framed as pure mathematics, or in a financial or healthcare context; the numbers and mathematical operations involved in each question were the same, the only variable was the context. Their results demonstrate that performance was influenced by context: in all cases, results were worse when the data was presented in a healthcare context.

Multiple tests of health numeracy have also been developed (Schwartz *et al* 1997; Weiss *et al* 2005; Schapira *et al* 2008; Huizinga *et al* 2008), but because these are aimed to detect low numeracy in patients, they are generally unsuitable (too easy) for healthcare professionals. There is no generally accepted ‘gold standard’ assessment of clinician numeracy. Caverly *et al* (2012) observed that “we cannot currently assess CN because there is no valid measure to test these skills in clinicians. Indeed, measuring the range of skills inherent in the domains of CN may be too much for a single measure.” This is an interesting point, and is borne out by the fact that most of the tests developed to measure numeracy in healthcare professionals focus on a single CN construct e.g. drug dose calculation skills (computational numeracy), or understanding risk and probability data (statistical numeracy). Before discussing the available measures of CN, I will briefly consider two assessments that might be expected to guarantee CN in medical students and doctors.

The UK Clinical Aptitude Test

Applicants to all UK medical schools must take the UK Clinical Aptitude Test (UKCAT) (www.ukcat.ac.uk). This test includes questions testing clinician numeracy, many of which are quite challenging. Therefore, performance on this section of the UKCAT would provide useful information on the CN of applicants to medical school, and potentially this could be used in the selection process or to guide remediation after entry. However, each medical school has its own entry criteria, and many do not use the data available from the UKCAT in selecting prospective students. Indeed, in their study of drug dose calculation ability, Wheeler *et al* (2007) found that there were significant differences in the performance of students from different UK medical schools.

The Prescribing Safety Assessment

The Prescribing Safety Assessment (PSA) for final year medical students has been developed to try to reduce prescribing error (British Pharmaceutical Society (BPS) & Medical Schools Council, (MSC) 2013). All UK final year medical students must pass the PSA in order to start work as foundation trainee doctors, and it is commonly assumed that passing the PSA assures competence in prescribing. However, passing the PSA does not indicate that a doctor has adequate drug dose calculation skills; it is possible to pass the PSA without answering any of the drug dose calculation questions, or having answered them all incorrectly. This is because the PSA has 8 component parts, for which there is a possible total of 200 marks; the pass mark is generally around 62 - 65% (Maxwell *et al* 2017), and the calculation section accounts for only 16 marks (8%) (BPS & MSC, 2013).

1.6.1 Tests of CN in Medical Students and Doctors

Researchers across the world have assessed CN in medical students and doctors, finding evidence that many have low numeracy (Rowe *et al* 1998, Selbst *et al* 1999, Sheridan & Pignone, 2002; Oldridge *et al* 2004; Wheeler *et al* 2004a; Wheeler *et al* 2004b; Ghosh & Ghosh, 2005; Windish *et al* 2007; Gigerenzer *et al* 2007; Simpson *et al* 2009; Rao & Kanter, 2010; Wegwarth *et al* 2012; Harries & Botha, 2013; Johnson *et al* 2014; Taylor & Byrne-Davis 2017). However, most of the tests used in these studies have focussed on a single construct: either computational, in relation to drug dose calculation (Rowe *et al* 1998; Selbst *et al* 1999; Oldridge *et al* 2004; Wheeler *et al* 2004a; Simpson *et al* 2009; Harries & Botha, 2013), or statistical, in relation to interpreting the results of research and screening (Ghosh & Ghosh, 2005; Windish *et al* 2007; Gigerenzer *et al* 2007; Rao & Kanter, 2010; Wegwarth *et al* 2012; Johnson *et al* 2014). The Medical Interpretation and Numeracy Test (MINT) (Taylor & Byrne-Davis, 2016) is the only CN test developed for doctors and medical students that provides a comprehensive assessment of computational, analytical and statistical numeracy.

The construct tested is only one of several important factors to consider when assessing CN tests; other important factors include the participants, the study design, and the length and format of the test. The attributes of twelve tests that have been used to assess CN in medical students and doctors are summarised in Table 1.6, and discussed below. The Critical Risk Interpretation Test described by Caverly *et al* (2012) to assess CN is not included in this analysis, as results of testing with this instrument have yet to be published.

1.6.1.1 Participants

Participants in eight of the studies were qualified doctors, three studies focussed only on medical students, and one included both medical students and trainee doctors.

1.6.1.2 Study design

Most studies were traditional classroom tests, with sample sizes of between 62 and 412 participants. However, one study was an online cross-sectional survey of 2975 doctors (Wheeler *et al* 2004a). Classroom tests are commonly used to assess the performance of medical students and trainee doctors, as they are a practical and feasible means of testing large numbers of participants. However, the validity of classroom tests to assess numeracy in clinicians is debated. Having investigated CN in nurses and student nurses extensively, Wright (2007a) suggests that classroom tests underestimate ability, and that performance is better in the workplace. However, although workplace-based assessments are now commonplace in undergraduate and postgraduate medical education, they are not without problems: an important limitation is that participants know that they are being assessed, and thus tend to modify their practice. Therefore, while clinicians may calculate drug doses more accurately under observation on a ward, this does not necessarily reflect their usual performance.

Table 1.6. Tests of CN in medical students and doctors

Authors Year Country	**Study sample**	**No. items & format of test**	**Construct**	**Outcome**
Rolfe & Harper (1995) UK	150 Hospital doctors of all grades & specialties	5 CR	Drug dose calculation	Difficulty with conversion between units. Seniority improves performance.
Rowe *et al* (1998) US	64 Paediatric trainees	8-10 CR	Drug dose calculation	Need for educational intervention and assessment.
Sheridan & Pignone (2002) US	62 1st year Medical students	6* CR	Risk estimation	Numeracy affects risk comprehension. Framing effect.
Wheeler *et al* (2004a, 2007) UK	2975 Doctors of all grades & specialties	6 CR	Drug dose calculation	Problems with IV drug labelling. Conversion between units causes error. Seniority improves performance; specialty affects performance.
Wheeler *et al* (2004b) UK	168 Medical students	3 CR	Drug dose calculation	As above. Final years did better than first years.
Windish *et al* (2007) US	277 Residents (senior trainee doctors)	20 CR	Biostatistics	Mean score 41% Inverse relationship between score & experience
Simpson *et* a*l* (2009) Australia	141 A&E trainee doctors	12 CR	Drug dose calculation	Seniority improves score Higher score for critical care, anaesthesia, A&E doctors
Anderson *et al* (2011) US	203 Obstetrician-gynaecologists	11* CR	Risk estimation & subjective numeracy	Deficiencies in understanding risk information; confidence unrelated to ability.
Wegwarth *et al* (2012) US	412 Senior doctors	8 CR	Screening statistics.	Screening statistics poorly understood
Harries & Botha (2013) S. Africa	364 3rd & 4th year Medical students	4 CR	Drug dose calculation	Only 23% competent initially; 34% never reached competence.
Johnson *et al* (2014) US	308 Medical students; 50 trainee surgeons	3* CR	Statistical: risk estimation	Results for medical students similar to trainees. Confidence unrelated to ability.
Taylor & Byrne-Davis (2017) UK	135 Foundation trainee doctors	43** SBA	Computation Analytical Statistical	Evidence of poor numeracy; statistical items more difficult than analytical; computational easiest.

** includes the 3 Schwartz et al (1997) probability questions*

*** includes the 3 Schwartz et al (1997) probability questions, but adapted to clinical setting*

Furthermore, Rowe *et al* (1998) and Simpson *et al* (2009) consider that calculation mistakes are more likely on the wards, with the distractions of a busy clinical environment. This opinion is supported by evidence that a stressful working environment is a significant contributory factor in prescribing error (Dean *et al* 2002; Coombes *et al* 2008; Dornan *et al* 2009; Glavin, 2010) (Table 1.4).

Another possible disadvantage of a classroom test is that participants may dismiss it as irrelevant "maths"; furthermore, taking part in research is not a high-stakes event, thus participants may not be bothered about their results. Although Johnson *et al* (2014) expressed such concerns about their research, this does not reflect my own experience: my original study was conducted in examination conditions, and participants took the task seriously. Moreover, when asked to predict how they would perform compared to their peers, most expected to be in the top or middle thirds, indicating their intent to perform to the best of their ability (Taylor, 2014). This is in accordance with the observations of Windish *et al* (2008) and Simpson *et al* (2009) that participants expected their performance to be better than it actually was.

1.6.1.3 Clinical experience

In studies involving doctors from varying backgrounds, senior doctors performed better than their junior colleagues (Rolfe & Harper, 1995; Wheeler *et al* 2004a; Simpson *et al* 2009); similarly, final year students performed better than first years (Wheeler *et al* 2004b). This suggests that CN improves with experience, and familiarity with the drugs involved. However, in their test of biostatistics, Windish *et al* (2007) observed an inverse relationship between seniority and performance; this may reflect deskilling through lack of practice.

When specialty area was considered, doctors from anaesthetic, critical care and emergency medicine backgrounds generally performed better than those from other clinical backgrounds; however, this may be simply because many of the questions used in these tests involved drugs used commonly in emergency situations (Rolfe & Harper, 1995; Wheeler *et al* 2004a; Wheeler *et al* 2004a). Nonetheless, Wheeler *et al* (2004a) observed that doctors who could not calculate the amount of drug in an ampoule often knew how much to administer in an emergency situation, because they remembered or had a rule of thumb about using a full ampoule, or half an ampoule etc.

1.6.1.4 Length of test

Six of the twelve tests were very short: two consisted of only 3 questions, one had four, one five, and two six had questions. Four tests were fairly short, with between 8 and 12 questions, while another was fairly long, with 20 questions. Only one test, the MINT, was lengthy, with 43 questions. The length of the test is important, in terms of reliability and validity: short tests are unlikely to be valid (Schuwirth & Van der Vleuten, 2004). Tests of only 3-6 questions may not be very useful, as the content is very limited, whereas longer tests allow multiple questions of each construct, thus giving a fuller picture of a participant's ability. Furthermore, there is little margin for error in a short test e.g. in the Harries & Botha (2013) study, there were only four questions, and participants needed to answer all four correctly to be deemed competent at drug dose calculation. These authors found that 34% participants never achieved competence, despite educational intervention, with worrying implications for medical educators (Harries & Botha, 2013). However, delivering a longer test may have produced different results.

1.6.1.5 Test format

All studies were constructed response (CR) tests, with the exception of the MINT which was presented in a multiple choice single best answer (SBA) format. The SBA format allows for a longer test, as it is much more feasible to deliver and mark an SBA test. However, performance on SBA tests may be enhanced compared to that on CR tests, as participants can guess at the answer, or they may be prompted to select the correct answer on viewing the available options (cueing) (Downing, 2003; McCoubrie, 2004; DiBattista & Kurzawa, 2011; McAllister & Guidice, 2012). Although these studies are not of CN tests, their findings suggest that scores achieved on the MINT may be enhanced because of its SBA format. The impact of test format on the MINT is the subject of Chapter 4.

1.6.1.6 Constructs tested

Ten of the twelve CN tests assess only one construct: six of these test computational numeracy, and four statistical numeracy. Anderson *et al* (2011) combined a 3-item statistical test (Schwartz *et al* 1997) with the Subjective Numeracy Test (Fagerlin *et al* 2008). Only one assessment, the MINT, tests a range of constructs, with multiple questions assessing computational, analytical and statistical numeracy.

1.6.1.6.1 Test material: computational numeracy

In addition to their limited test content, some CN tests introduce bias by using test material that could influence outcome e.g. basing drug dose calculation questions on drugs used in specific specialty areas, thus conferring an advantage to some participants (Rolfe & Harper, 1995; Wheeler *et al* 2004a). Furthermore, some questions included in these tests are not relevant to clinical practice for doctors e.g. calculations based on counting drop rates in intravenous (IV) infusions (Harries & Botha, 2013). In modern hospital practice, most infusions are delivered using automated systems that regulate flow electronically; therefore, an inability to answer questions related to counting drops may not reflect any risk to clinical practice. Nonetheless, a common and important finding from these studies was that medical students and doctors often found drug dose calculation relating to drugs for IV administration difficult, particularly when they needed to convert numbers between different units of measurement (Rolfe & Harper, 1995; Wheeler *et al* 2004a; Harries & Botha, 2013). These researchers all recommend that labelling of IV drugs should be standardised to mass concentration (mg/ml).

The MINT contains 13 computational questions based on clinical scenarios that are relevant to the workload of a trainee doctor at foundation level (within two years of qualification). Many, but not all, computational questions in the MINT relate to drug dose calculation; none are specialty-specific (Taylor & Byrne-Davis, 2016).

1.6.1.6.2 Test material: statistical numeracy.

Three tests (Sheridan & Pignone, 2002; Anderson *et al* 2011; Johnson *et al* 2014) used the 3-item statistical test developed by Schwartz *et al* (1997) to assess numeracy in patients. These questions are set in a gambling context e.g. chance of winning the lottery. Johnson *et al* (2014) quite rightly acknowledge that these three questions may not constitute a proxy for "numeracy", and that this limits the interpretation of their results. Windish *et al* (2007) developed a 20-item test of biostatistics, while Wegwarth *et al* (2012) used an 8-question test based around two clinical scenarios related to screening for disease; both tests are challenging, and assess constructs relevant to clinical practice.

The MINT contains 13 statistical questions, including the three questions devised by Schwartz *et al* (1997); however, for the MINT they were rewritten to a clinical setting e.g. estimating the likelihood of developing a side effect following treatment. The MINT also uses three statistical questions developed by Sheridan & Pignone (2002). Performance on these three questions is similar for all four studies involving medical students and doctors (Sheridan & Pignone, 2002; Anderson *et al* 2011; Johnson *et al* 2014; Taylor & Byrne-Davis, 2017); furthermore, performance is significantly better than that of the general public (Schwartz *et al* 1997; Taylor & Byrne-Davis, 2017). Interestingly, the outcome of the study by Levy *et al* (2014) would suggest that performance on these questions should have been lower in the MINT, where the questions are set in a healthcare context; however, this did not prove to be the case.

1.6.1.7 Difficulty of test material

The content and level of difficulty of these tests is very variable. Some tests are quite easy, being largely based on the three Schwartz *et al* (1997) questions designed for the general public (Sheridan & Pignone, 2002; Anderson *et al* 2011; Johnson *et al* 2014). Some are straightforward drug dose calculation tests, but are based on a small number of drugs that are commonly used by some doctors but rarely by others (Rolfe & Harper, 1995; Wheeler *et al* 2004a): thus the content is easy for some doctors, but quite difficult for others. Others are more challenging, being based entirely on statistics (Windish *et al* 2007; Wegwarth *et al* 2012). In contrast, questions in the MINT deliberately vary in their level of difficulty (Taylor & Byrne-Davis, 2016). Furthermore, the expected level of performance is also variable e.g. Harries & Botha (2013) required a score of 100% for competence. However, the mean score achieved on the CN test of Windish *et al* (2007) was 41%, and of Taylor & Byrne-Davis (2017) was 76%; in both cases, the authors suggest that their findings indicate deficiencies in CN, although neither suggests what score would be acceptable. Simpson *et al* (2009) report that the mean score on their drug dose calculation test was 72.5%; interestingly, participants suggested that an adequate score would be 91%.

1.6.1.8 Conclusion

Therefore, having reviewed the available tests of CN, there is evidence to suggest that the MINT is the best available measure for research into CN in medical students and doctors, since it is the only assessment that assesses all three numeracy constructs, and the only lengthy test. Psychometric testing indicates that the MINT is a valid and reliable test (Taylor & Byrne-

Davis, 2017). My initial research using the MINT suggested that many trainee doctors have significant deficiencies in CN, and that these could affect safe patient care (Taylor & Byrne-Davis, 2017). However, several factors may affect the interpretation of my initial results, thus additional work is required before conducting further research using the MINT.

1.7 SUMMARY

There is ample evidence that many medical students and doctors have deficiencies in CN (Rowe *et al* 1998; Selbst, 1999; Sheridan & Pignone, 2002; Wheeler *et al* 2004a; Wheeler *et al* 2004b; Ghosh & Ghosh, 2005; Wheeler *et al* 2007; Gigerenzer *et al* 2007; Windish *et al* 2007; Simpson *et al* 2009; Wegwarth *et al* 2012; Harries & Botha, 2013; Johnson *et al* 2014; Taylor & Byrne-Davis, 2017). However, despite this, we know remarkably little about CN in doctors and medical students, since it is under-researched, and because of variations in the tests used to assess CN. The lack of standardisation of level of difficultly together with the variability in test content makes it challenging to assess the overall level of CN in doctors and medical students. No standard of CN has been set for medical graduates; however, determining an acceptable standard of competence would be useful in defining and addressing deficiencies.

It is not clear why medical students and doctors have deficiencies in CN. While it seems counterintuitive, given the academic demands of medical school, there is evidence that numeracy may be low in highly-educated people (Lipkus *et al* 2002; Peters *et al* 2007). However, no research has yet been published on the cause of error in CN tests in medical students and doctors. This area has been extensively researched in nursing (Weeks *et al* 2000, 2001; Johnson & Johnson, 2002; Wright, 2004; Galligan *et al* 2010; Young *et al* 2013; Weeks *et al* 2013a, 2013b, 2013c; Sabin *et al* 2013; McDonald *et al* 2013; Coben & Weeks, 2014; Galligan & Hobohm, 2015), and warrants investigation in medical practice.

Following my initial research, I concluded that many foundation trainee doctors (FTs) had low CN (Taylor & Byrne-Davis, 2017). However, I did not understand why this was, nor did I know what kind of errors were being made. The FTs were a heterogeneous group, including UK and overseas graduates; thus, although participants were likely to be representative of FTs across the UK, it was not clear whether my results were generalisable to UK medical students. Therefore, additional research was needed to assess whether CN in medical students was similar to that of FTs. Furthermore, I considered that it would be important to conduct an investigation into the errors being made.

1.8 AIMS AND OBJECTIVES

My research aims to assess clinician numeracy (CN) in medical students. In order to do this, I need an appropriate assessment measure. Having reviewed twelve tests of CN in medical students and doctors (Table 1.6), it is clear that the MINT offers many advantages over the other available measures. Therefore, it is the most appropriate test to use for my research. The aims of my research are:

1) to evaluate MINT test material to ensure its quality and suitability to assess CN;

2) to investigate external factors that may affect the measurement of CN, specifically:

i) the impact of calculators on test scores; and

ii) the impact of test format on test scores;

3) to investigate CN in medical students using the MINT, and in particular to analyse the errors made by participants when answering MINT questions.

1.9. OVERVIEW OF THESIS

In light of these aims, I will present four research chapters, as outlined below. The first phase of my research involved evaluating the MINT test material, thus no participants were required; however, participants were needed for all later phases. I received approval from the University of Manchester's Research Ethics Committee (Appendix 1) to conduct the various studies described here. Participants in this research were third year medical students from a single institution; participation in each study was optional, and I anonymised all of the data that I collected. Before conducting my research, I consulted statisticians from the research department at the University Hospitals of North Midlands (UHNM), who provided useful advice regarding study methodology.

Chapter 2. Quality of MINT Questions

The first aim of my research is to review the MINT to ensure that all questions are clear and clinically appropriate. In an ideal situation, a candidate's test score should be an accurate reflection of their knowledge and ability; however, in practice several factors may interfere with this relationship. Such factors may be related to the candidate e.g. lack of preparation for the examination, lack of engagement with the test, fatigue, distraction, carelessness and poor examination technique. However, external factors are also important, particularly the quality of the test material: if questions are poorly written or displayed, they may mislead candidates resulting in error. If a test has many such questions, candidates' scores may be low, and may not accurately represent their knowledge and ability. Thus, an essential part of preparing any university examination is the emendation process, in which questions are subjected to detailed scrutiny. This chapter of my thesis concerns my revision of the MINT test material, and its evaluation, including psychometric analysis. This process, as with emendation in university exams, aims to ensure the quality of all test questions, so that results reflect the ability of candidates.

Chapter 3. Effect of calculators on test scores

The second aim of my research is to investigate external factors that may affect test scores. The first factor to be considered was whether using a calculator would have an impact on test score. Calculators are almost universally available in clinical practice, and are commonly used to help in numeracy related tasks, such as drug dose calculation. Nonetheless, there is no consensus on whether healthcare professionals sitting drug dose calculation tests should be given access to calculators: in some tests, calculators are permitted (Coyne *et al* 2013; Fleming *et al* 2014), while in others they are not (McMullan *et al* 2010; Bagnasco *et al* 2016). McMullan *et al* (2010) argue that nurses and student nurses should not be allowed to use calculators, as doing so could lead to deskilling. For my original research, I agreed with that argument, and so did not permit FTs to use calculators (Taylor & Byrne-Davis, 2016, 2017). However, feedback from colleagues and peer reviewers suggested that test results would have been better had FTs been given calculators. Furthermore, calculators are now routinely used in clinical practice. Therefore, I considered that it would be important to determine whether calculators would improve test scores, and conducted a randomised controlled trial (RCT) to test this hypothesis.

Chapter 4. Effect of test format on test scores

The second external factor that could affect test scores, was test format. I had developed the MINT as a multiple choice Single Best Answer (SBA) test. However, SBA tests may enhance performance due to cueing and guessing (Downing, 2003; McCoubrie, 2004; Simkin & Kuechler, 2005; DiBattista & Kurzawa, 2011; Sam *et al* 2016). Therefore, my initial results with the MINT may have overestimated participants' CN. In order to exclude this possibility, I developed a constructed response (CR) version of the test, and then conducted an RCT to investigate the impact of test format.

Chapter 5. Causes of error in the MINT

The third aim of my research was to investigate CN in medical students and doctors, and to explore the causes of error being made in the MINT. I have found no published studies regarding the errors being made by medical students or doctors in CN tests. However, determining the cause of error is a vital first step towards developing an effective educational intervention (Tully, 2012; Wallace, 2019); therefore, my analysis should be useful in considering remediation for deficiencies in CN. This chapter outlines my analysis of the types of error being made by medical students, and considers the implications for medical education.

Chapter 6. Discussion

The final chapter of my thesis summarises my findings, and considers their implications for future research, and for medical education.

Blank Page

CHAPTER 2

REVISION OF THE MINT

This chapter will not be submitted for publication; however, I have given several oral presentations relating to my research, during which I have presented some of the information discussed here. My presentations have included talks and workshops at the following events:

1. Developing Excellence in Medical Education Conference (DEMEC), November 2017
2. National Association of Clinical Tutors (NACT) Walsall, February 2018
3. Association for the Study of Medical Education (ASME) Workshop, Nottingham, April 2018
4. Annual Medical Education Conference, Keele, April 2018
5. Health Education England (HEE) Educators Conference, Birmingham, November 2018
6. Academic Grand Rounds, University of Manchester, February 2019

TABLE OF CONTENTS: CHAPTER 2

LIST OF APPENDICES

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Numerical skills are integral to many everyday clinical tasks, as outlined in Table 2.1. Although a wide range of tests have been used to assess clinician numeracy (CN) in medical students and doctors, all have limitations of some kind. Many tests are short, and therefore their reliability and hence validity is not assured (Schuwirth & van der Vleuten, 2004); others test only one numeracy construct, thus they cannot provide an overall assessment of CN. Having evaluated several assessment measures, as discussed in Chapter 1, I have concluded that the Medical Interpretation and Numeracy Test (MINT) (Taylor & Byrne-Davis, 2016) is the best assessment of CN for medical students and doctors that is currently available. However, experience with the MINT is limited, as it has only been used in one study to date (Taylor & Byrne-Davis, 2017), and data analysis and peer review suggested that some test material could be improved. Furthermore, the MINT cannot be used in its original form for any further research, since some test material is subject to copyright. Therefore, I needed to review and revise the MINT, and to develop new test questions, before conducting any further research into CN.

Table 2.1. Clinician numeracy in practice*

Construct	Clinical application
Computational numeracy	Calculation of drug doses Calculation of IV fluid & electrolyte requirements Calculation/prescription of parenteral nutrition requirements Use of medical formulae e.g. growth charts in paediatrics Managing disease processes e.g. blood glucose control, anticoagulation therapy, weaning from steroids
Analytical numeracy	Interpretation of medical data presented in different formats, including the results of research Diagnostic skills based on the interpretation of test results Problem-solving skills Estimation of accuracy of calculations, and of efficacy of various treatment options Clinical decision making Appropriate treatment selection
Statistical numeracy	Understanding risk information Effective risk communication Interpretation of medical data Clinical decision making & treatment selection Critical analysis of evidence; including the effect of different treatments & results of research published in the literature Practicing evidence-based medicine

**Adapted from Taylor & Byrne-Davis (2016)*

Seven questions used in the MINT were subject to copyright, and permission to use them applied only to our initial research. These questions were from the Newest Vital Sign test (NVS) (Weiss *et al* 2005) and the Programme for International Student Assessment (PISA) test (Organisation for Economic Cooperation and Development (OECD), 2009). An important part of the revision process was to develop new test material to replace these questions. Another important consideration was that the peer reviewers of the paper describing the development of the MINT (Taylor & Byrne-Davis, 2016) indicated that two questions could be improved; in one case, the text was potentially misleading due to the absence of the words “at random”; in the other, a graph was deemed to be of inadequate size to allow precise interpretation. Since I had submitted only a representative sample of 10/43 MINT questions as an appendix to this paper, the identification of problems with two questions suggested that all MINT questions should be reassessed to ensure their quality.

In my original study with the MINT, I found that the mean test score was 32.76/43 (76%) with a median score of 34, a range of 14-42/43 (32-98%), and an interquartile range of 28-38/43 (65-88%); these results suggest that many trainee doctors have low clinician numeracy (Taylor & Byrne-Davis, 2017). However, poor numeracy is only one of several factors that may lead to low test scores: participants in examinations of all kinds may make errors due to lapses in concentration, to poor time management, and due to failure to read questions carefully. These factors should be considered when interpreting test scores.

The quality of test questions is of vital importance to the interpretation of test results (Downing, 2003; McCoubrie, 2004; DiBattista & Kurzawa, 2011; McAllister & Guidice, 2012; DiBattista *et al* 2014). The text of a question may inadvertently mislead participants, prompting incorrect answers; furthermore, the quality of the distractors used in a single best answer (SBA) multiple choice test is important in reducing the potential impact of cueing and guessing (Downing, 2003; DiBattista & Kurzawa, 2011; McAllister & Guidice, 2012). Therefore, it was essential to review all MINT questions to assure the quality of test material. The revision process would also provide an opportunity to recruit colleagues to evaluate test questions, similar to the emendation process performed for university examinations. In addition, all test material was reviewed both by myself, and by those reviewers who were clinicians, to ensure that it remained consistent with current medical practice.

Finally, the revision process was a useful opportunity to reconsider the framework for health numeracy (HN) that I had used when blueprinting MINT questions. This framework was devised by Golbeck *et al* (2005), and describes three key constructs relevant to CN: computational, analytical and statistical numeracy (Table 2.1); however, these constructs often overlap, leading to difficulty in classifying test questions for data analysis. Alternative frameworks for HN have been described (Nutbeam, 2000; Ancker & Kaufman, 2007; Schapira *et al* 2008; Caverly *et al* 2012); therefore, I decided to investigate whether one of these would provide a better structure for considering and analysing MINT content.

The revision of the MINT was an important stage in its development, involving a comprehensive emendation process in which all questions were evaluated. Questions that were subject to copyright were replaced, as were those that had become clinically out of date; in addition, I made suitable adjustments to any questions in which the text or data display might be misleading or confusing.

The revision process had seven distinct aims:

1. To replace test questions that were subject to copyright;
2. To identify and amend any text that might confuse participants;
3. To identify and correct any clinical information that was inaccurate, misleading or obsolete;
4. To review the frameworks for health numeracy, and consider which was most suitable for analysing the MINT;
5. To subject the MINT to external review;
6. To conduct an evaluation study of the revised test paper; and
7. To develop and assess distractors for the SBA version of the MINT.

SECTION 2. METHODS

The five stages of the revision process are summarised below:

Stage 1: Internal review. The initial step of the review involved replacing the seven copyrighted questions, meeting the first aim of the revision process. The revised test paper, MINTv2 was then subjected to detailed scrutiny, meeting the second, third and fourth aims of the revision process.

Stage 2: External Review. Several clinicians and academics evaluated MINTv2 to meet the fifth aim of this research.

Stage 3: Evaluation Study 1. The sixth aim of the revision process was met by recruiting a cohort of third-year students to sit MINTv2.

Stage 4: Development of MINTv3. The answers given by participants in the evaluation study of MINTv2 were analysed, leading to the development of distractors for the SBA version of the test, MINTv3. This met the first part of the seventh aim of the revision process.

Stage 5: Evaluation Study 2. A cohort of third-year students were recruited to sit MINTv3, fulfilling the second part of the seventh aim of the revision process.

2.1 INTERNAL REVIEW

I conducted a comprehensive review of all 43 MINTv1 questions, leading to the development of MINTv2, as shown in Table 2.2. The revised test paper contains 34 MINTv1 questions and 9 new questions; seven new questions replace questions that were subject to copyright, and two replace questions that had become redundant for clinical reasons. Although 34/43 MINTv1 questions are retained in MINTv2, 23/34 of these have been amended in some way, while eleven remain unaltered. Many questions involve hypothetical patients: as I revised questions, I generally gave these “patients” new names to allow me to differentiate between questions used in each test; the new names do not constitute a change for the purposes of data analysis. Additionally, I changed the order of questions in the revised test, MINTv2, thus each question discussed below is identified by its number in the relevant test.

Finally, I reviewed three frameworks for health numeracy, and compared them with that of Golbeck *et al* (2005) to assess which was most appropriate for classifying MINT questions.

Table 2.2. Revision process

	Original MINT (MINTv1)	Revised MINT (MINTv2 and MINTv3)
Original Test Paper	Development of MINTv1* 43 test questions 12 new questions 31 questions from existing sources (7 subject to copyright)	
Stage 1	Internal review Development of MINTv2	7 questions replacing copyrighted questions 2 new questions, replacing obsolete questions 23 modified MINTv1 questions 11 unaltered MINTv1 questions
Stage 2	External Review	Minor amendments to 2 questions
Stage 3	Evaluation Study 1 (MINTv2)	Minor amendment to one question
Stage 4	Development of MINTv3	New distractors provided for 34 questions
Stage 5	Evaluation Study 2 (MINTv3)	

2.1.1 ***Replacement of nutritional label questions***

Four copyrighted questions were based on the interpretation of a nutritional label, and adapted from the Newest Vital Sign test (NVS) (Weiss *et al* 2005). The ability to interpret nutritional labels provides a useful indication of health numeracy (HN) in patients (Weiss *et al* 2005; Rothman *et al* 2006; Huizinga *et al* 2008), and the NVS can be used as a screening tool for health literacy (Weiss *et al* 2005). Therefore, when developing the MINT, I considered that it would be important to assess the ability of medical students and doctors to interpret nutritional labels.

In my original research with MINTv1, I found that only 76/135 (56%) foundation trainees had answered all four NVS questions correctly (Taylor & Byrne-Davis, 2017): a surprising finding in such a highly-educated group. However, my data analysis suggested that the American terminology in one of the NVS questions used in MINTv1 might have confused some participants. This hypothesis is supported by the work of Rowlands *et al* (2013), who conducted a Delphi study to develop a version of the NVS for the UK population: they found that the text of the NVS questions used in the US (Weiss *et al* 2005) was unsuitable for UK participants, and developed new material for the UK NVS. Thus, it would have been necessary to replace the NVS questions, even if there was no issue relating to copyright.

In addition, it was important to include nutritional label questions in the MINT because of the correlation between performance on the NVS questions and overall MINT score in my initial research (Taylor, 2014); this suggested that it might be possible to develop a short test, similar to the NVS, to screen healthcare professionals for CN. However, further research was required to investigate this; my review of the literature on CN has found no published studies relating to nutritional label interpretation in medical students or doctors. Therefore, I reviewed various nutritional label tests (Rothman *et al* 2006; Huizinga *et al* 2008; Rowlands *et al* 2013) to inform the development of new nutritional label questions for MINTv2 that were equivalent to the NVS in terms of content and level of difficulty. The original NVS questions (Weiss *et al* 2005) are based on interpreting the nutritional label of an ice-cream carton. I changed the context of these questions for MINTv1, using the same data, but based on a nutritional drink for a patient; therefore the setting was not subject to copyright, and did not need to be altered for MINTv2. The nutritional label questions used in MINTv1 are shown in Figure 2.1.

Mr Iqbal is recovering from a stroke, and has been prescribed a nutritional supplement drink. The nutritional information available on the drink carton is shown below:

Build-up drink		
Nutrition Facts Serving size Servings per container	 100ml 4	
	Amount per serving	% Recommended daily intake*
Energy	250 Calories	12.5%
Total Fat	13g	20%
of which saturates	9g	40%
cholesterol	28g	12%
Total Carbohydrate	30g	12%
of which sugars	25g	
Dietary Fibre	3g	
Protein	4.2g	8%
Sodium	55mg	2%
* % Recommended daily intake values are based on a 2,000 calorie diet. An individual's daily values may be higher or lower depending on their calorie needs.		

1. If Mr Iqbal drinks he entire container, how many calories will he ingest?
2. If Mr Iqbal is allowed to have 60g of carbohydrate as a snack, how much can he drink?
3. Mr Iqbal usually ingests 42g of saturated fat a day, including one serving of the nutritional drink. If he stops taking the nutritional drink, how much saturated fat would he be consuming each day?
4. Mr Iqbal requires 2500 calories per day. What percentage of his daily value of calories is one serving of the drink? (Q.20-23)

Figure 2.1. Nutritional label questions from MINTv1.

The replacement questions used in MINTv2 are shown in Figure 2.2.

Mrs Doyle is recovering from a stroke, and has been prescribed a nutritional supplement drink. The nutritional information available on the drink carton is shown below:

Build-up drink		
Nutritional information Serving size Servings per carton	 100ml 4	
Typical values	Per 100ml	% Recommended daily intake*
Energy	200 Calories	10%
Total Fat	16g	25%
of which saturates	12.5g	50%
Total Carbohydrate	35g	11%
of which sugars	24g	
Dietary Fibre	7.5g	
Protein	6g	12%
Sodium	75mg	3%
* NOTE % Recommended daily intake values are based on a 2,000 calorie diet. People have different calorie requirements, so an individual's daily values may be higher or lower than those indicated here.		

1. If Mrs Doyle drinks half of the carton, how many calories will she consume?
2. Mrs Doyle needs to increase her protein intake by 9g. What volume of the nutritional drink provides 9g of protein?
3. Mrs Doyle has two servings of the nutritional drink every day. Her total carbohydrate intake is 200g per day. How many grams of her carbohydrate intake comes from sources other than the drink?
4. Mrs Doyle requires 1600 calories per day. What percentage of her daily calorie intake is provided by one serving of the drink? (Q.19 - 22)

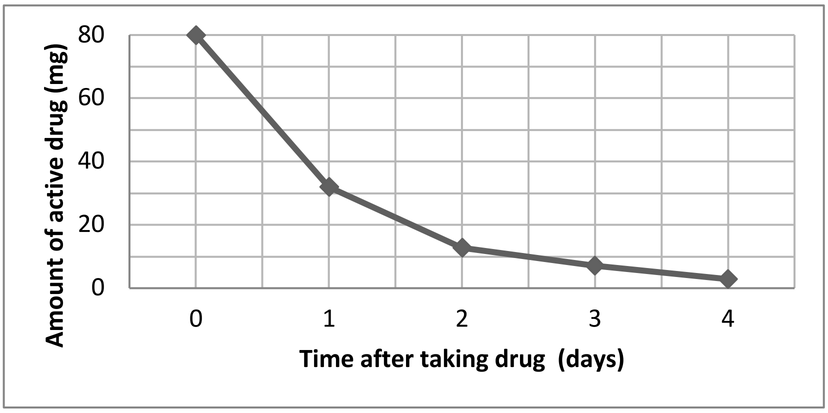
Figure 2.2. Nutritional label questions from MINTv2.

2.1.2 *Replacement of PISA questions*

The remaining three questions subject to copyright came from the PISA test for 15-year olds (OECD, 2009), and were related to drug metabolism. Two of these were based on the interpretation of a line graph showing the changes in the serum concentration of an intravenous (IV) drug over time, and the third was a complex calculation. Analysis of data from MINTv1 indicated that all were useful questions, hence I developed similar material for MINTv2. The questions used in MINTv1 are shown in Figure 2.3 and those used in MINTv2 are shown in Figure 2.4.

In addition to the copyright issue, a peer reviewer of my first paper (Taylor & Byrne-Davis, 2016) considered that the quality of the line graph I had used in MINTv1 was insufficient to allow precise interpretation; therefore, one of my "incorrect" MCQ answer options could be

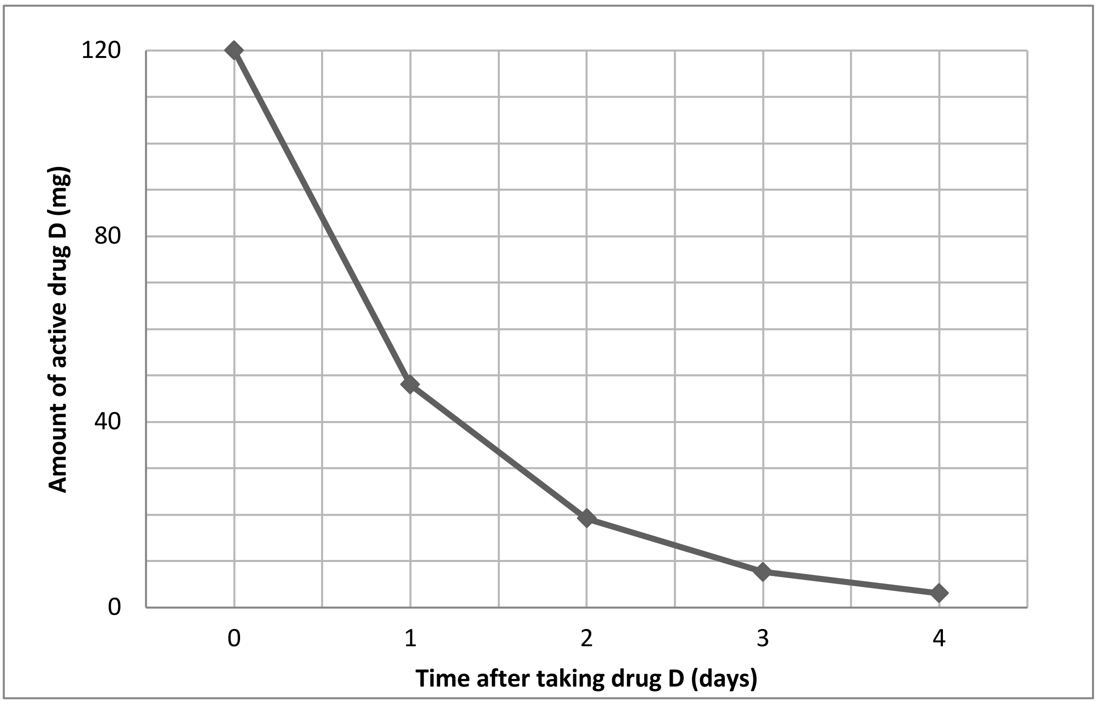
Items 24 – 25. Alex enters a clinical trial, and is given 80mg of the test drug by IV injection. The following graph shows the initial amount of the drug, and the amount that remains active in Alex's blood after one, two, three and four days.



24. Approximately how much of the drug remains active after 36 hours?
25. Each day about the same proportion of the previous day's drug remains active. At the end of each day which of the following is the approximate percentage of the previous day's drug that remains active?

Figure 2.3. PISA drug concentration chart used in MINTv1.

Items 31 - 32. Amy is given 120mg of drug D by intravenous (IV) injection. Blood tests are taken every day for the next four days to check the level of drug D remaining active in her bloodstream. The graph below shows the concentrations of drug D in Amy's blood over the four-day period.



31. Approximately how many mg of drug D remain active after 36 hours?
32. From the graph above, it can be seen that each day about the same proportion of the previous day's drug remains active in Amy's blood. At the end of each day, approximately what percentage of the previous day's drug remains active?

Figure 2.4. Replacement drug concentration chart used in MINTv2.

construed to be correct. This problem occurred because I had reduced the size of the graph for the MINTv1 test paper. Data analysis showed that the facility of this question would increase from 0.42 to 0.60 if both answers were considered to be correct. Therefore, I enlarged the graph for MINTv2 to ensure that it could be interpreted accurately. In addition, as shown in Figures 2.3 and 2.4, although I did not change the first question following the graph, I modified the text of the second question slightly.

The third PISA question to be replaced was based on drug metabolism. Both the original and replacement questions are shown in Figure 2.5.

From MINTv1
Millie receives an injection of IV antibiotic. One hour after the injection, only 60% of the antibiotic will remain active. This pattern continues: at the end of each hour only 60% of the antibiotic that was present at the end of the previous hour remains active. Millie is given a dose of 300 mg of the antibiotic at 0800. Approximately how much antibiotic will remain active at 1100? (Q. 32)

From MINTv2
Clark is admitted to the ward with an infection, and starts on a course of IV antibiotics. One hour after he receives the injection, 70% of the antibiotic remains active. This antibiotic activity continues to decline in this manner: at the end of each hour antibiotic activity is 70% of its value at the end of the previous hour. Clark receives 500mg of the antibiotic at 1200. Approximately how many mg of the antibiotic will still be active at 1600? (Q. 33)

Figure 2.5. Drug metabolism questions.

2.1.3 ***Replacement of clinically redundant questions***

Changes in clinical practice had rendered two questions obsolete. The original question assessed participants' ability to calculate a very small volume of a drug, using decimal places; this was replaced with similar material, but in a different clinical setting (Fig. 2.6).

From MINTv1
Ryan has diabetes, and needs 8 units of Actrapid insulin. Actrapid is prepared in a solution containing 100 units of Actrapid per ml. What volume of solution should Ryan be given? (Q. 30)

From MINTv2
Noah is 5 years old, and weighs 20kg. Following a dose of morphine for postoperative pain relief, he has developed respiratory depression, and now needs reversal with an injection of naloxone. The recommended dose of naloxone is 3 micrograms/kg body weight. Naloxone is prepared in a solution containing 400 micrograms per ml. How many ml of naloxone should Noah be given? (Q. 30)

Figure 2.6. Drug dose calculation questions.

A second MINTv1 question was based on calculating a patient's average daily urine output; however, on review, I considered that this question was unsatisfactory, since more frequent recording of output is necessary if there are concerns about a patient's renal function. The question involved calculating the average of four values, and was replaced with very similar material, as shown in Figure 2.7. These questions are displayed using tables on the test papers (Appendices 2 & 3), but only the relevant numbers are shown here.

From MINTv1
You are asked to review Mr Brown as the ward sister is worried about his urine output. His daily output over the past four days has been: 532, 472, 472, and 364ml respectively. What is his average urine output per day over this 4-day period? (Q.38)

From MINTv2
Mr Price is on the elderly care ward. You are asked to review him regarding his oral fluid intake. His hourly oral fluid intake for the 4 hours from 8am to 12 noon has been: 89, 63, 121 and 63ml. What is his average hourly fluid intake over this 4-hour period? (Q.16)

Figure 2.7. Questions testing calculation of the mean.

2.1.4 *Review and revision of remaining questions*

I amended 23 questions in some way; in most cases, this involved minor modifications of the text. Two questions (Q.2, Q.27) were modified for more than one reason. The reasons for revision are summarised below:

a) to improve the quality of charts and graphs (Q. 2, 23, 24, 34);
b) to improve the clarity of the text (Q. 1, 3, 5, 7, 8, 17, 18, 25, 36, 37, 38, 42, 43);
c) to standardise the terminology (Q. 14, 15, 27);
d) for various clinical reasons (Q. 2, 10, 12, 27, 41).

2.1.4.1 *Quality of charts and graphs*

I made minor alterations to three data displays, affecting four questions. I added white lines between segments of pie in a pie chart, thus making the size of each segment clearer (MINTv1, Q.2; MINTv2, Q.2). I also altered the dates shown on two graphs: a scattergram showing the incidence of lung cancer over time (MINTv1, Q.33 and 34; MINTv2, Q.23 and 24), and a bar chart showing causes of laboratory error over a four-year period (MINTv1, Q.35; MINTv2, Q.34). These changes are insignificant, their only purpose was to ensure that the test paper did not look out of date.

2.1.4.2 *Clarity of text*

I altered the text of thirteen questions to improve their clarity; in seven cases the changes were minor, while a further six required moderate amendments.

2.1.4.2.1 *Minor amendments to text*

I changed the order of the text of four questions as shown in Figures 2.8 – 2.11.

From MINTv1
A patient has diabetes and is planning to exercise in the gym for one hour. She needs to eat 6g of carbohydrate for every 30 mins she exercises. She has some biscuits in her gym bag. Each biscuit contains 8g of carbohydrate. How many biscuits should she eat before she exercises? (Q.1)

From MINTv2
A patient has diabetes and needs to eat 6g of carbohydrate for every 30 minutes of exercise. She is planning to exercise in the gym for one hour. She has some biscuits in her gym bag. Each biscuit contains 8g of carbohydrate. How many biscuits should she eat before she exercises? (Q.1)

Figure 2.8. Questions on diabetes management.

From MINTv1
There is a 2 in 100 chance of living 5 years or longer without treatment for a type of cancer. Drug X increases the chance of living 5 years or longer to 6%. Drug Y increases the chance of living 5 years or longer by 50%. If a patient wants the best chance of living 5 years or longer, which drug should be prescribed? (Q.3)

From MINTv2
There is a 2 in 100 chance of living 5 years or longer without treatment for a type of cancer. Drug G increases the chance of living 5 years or longer by 50%. Drug H increases the chance of living 5 years or longer to 6%. Your patient wants the best chance of living 5 years or longer. Which drug should you prescribe? (Q.3)

Figure 2.9. Questions on risk.

From MINTv1
Mr Perkins is admitted to the gastroenterology ward and is prescribed an omeprazole infusion at a rate of 10ml/hr, to be continued for 72 hours. The solution infused contains 80mg of omeprazole in 100ml of 0.9% NaCl. How much omeprazole does Mr Perkins receive every hour? (Q.5)

From MINTv2
Mr Bradley is admitted to the gastroenterology ward and is prescribed an omeprazole infusion, to be continued for 72 hours. The rate of infusion is 10ml/hr, and the solution infused contains 80mg of omeprazole in 100ml of 0.9% NaCl. How much omeprazole does Mr Bradley receive every hour? (Q.5)

Figure 2.10. Questions relating to an IV drug infusion.

The fourth question for which I changed the order of the text related to local anaesthetic administration; in addition, I amended the question to specify that the answer should be expressed in terms of the volume of bupivacaine to be given to the patient. This was unnecessary for MINTv1 due to its single best answer format; however, it was important for the constructed response format of MINTv2, otherwise two answers would be correct: 40ml and 200mg. In clinical practice, doctors will be required to prepare the correct volume of local anaesthetic, hence I wished to test their ability to perform this calculation.

From MINTv1
Mo weighs 100kg, and presents to A&E with a wound in his thigh. You are asked to suture it, using the local anaesthetic bupivacaine which comes in a solution containing bupivacaine 5mg/ml. The maximum dose of bupivacaine that can be safely given is 2mg/kg. What is the maximum amount of bupivacaine you can use when suturing Mo's wound? (Q.43)

From MINTv2
Mo weighs 100kg, and presents to A&E with a wound in his thigh. You are asked to suture the wound under local anaesthetic. The available local anaesthetic is a solution containing bupivacaine 5mg/ml. The maximum dose of bupivacaine that can be safely given is 2mg/kg. What is the maximum volume of bupivacaine you can use to suture Mo's wound? (Q.43)

Figure 2.11. Drug dose questions.

I altered the text relating to the probability of a tossed coin landing heads up; although the wording seemed clear, 4/135 participants had given incorrect answers in MINTv1. Therefore, I modified it aiming to improve its clarity. These questions are shown in Figure 2.12.

From MINTv1

You are asked to randomise patients for a drug trial by tossing a coin. If the coin lands head up, the patient will receive Drug D, while if it is tails, the patient will be given the placebo. You will be recruiting 1000 patients. Approximately how many are likely to receive Drug D? (Q.16)

From MINTv2

Patients are recruited to a randomised controlled trial of a new drug. Randomisation is done by tossing a coin. If the coin lands head up, the patient will be given the new drug. If the coin lands on tails, the patient will be given a placebo. 1000 patients are recruited to the study. Approximately how many are likely to receive the new drug? (Q.38)

Figure 2.12. Coin toss questions.

I made minor modifications to the text of two further questions to improve their mathematical accuracy. The first was a question based on a bar chart showing possible training rotations for a doctor (MINTv1 and MINTv2, Q.42). Participants were asked to calculate the probability of a trainee being placed in a particular specialty. However, a peer reviewer noted that it was not clear from the text whether training places were allocated at random; this was explicitly stated in the text for MINTv2. The second question changed in the interest of mathematical accuracy required calculation of a patient's peak flow rate, and is shown in Figure 2.13.

From MINTv1

Mrs Cartwright has been admitted with an acute exacerbation of asthma. Clinical guidelines state that she can be safely discharged once her Peak Expiratory Flow Rate (PEFR) is >75% of her normal level. Her normal PEFR is 420 l/min. What is the minimum PEFR that Mrs Cartwright must achieve in order to be allowed home? (Q.27)

From MINTv2

Leanne has been admitted with an acute exacerbation of asthma. Clinical guidelines state that she can be safely discharged once her Peak Expiratory Flow Rate (PEFR) is at least 75% of her normal level. Her normal PEFR is 420 l/min. What is the minimum PEFR that Leanne must achieve before she can go home? (Q.25)

Figure 2.13. Calculation questions.

Although the change from "… (PEFR) is >75%" to "… (PEFR) is at least 75%" is minor, it is particularly important for the MCQ version of the test; the correct answer was intended to be C. 315 l/min, and this answer was selected by 93% of participants. However, 315 is exactly 75% of 420, not greater than 75%; data analysis from MINTv1 showed that 6/135 candidates had selected option D, 345 l/min as their answer – of the available MCQ options, this was the lowest answer that was "greater than" 75%, as specified. If option D was considered to be correct, the facility of this question would rise to 0.97.

2.1.4.2.2 *Moderate amendments to text*

The text of three sets of questions developed for medical students by Sheridan & Pignone (2002), required moderate amendments. These questions are based on comparing the efficacy of two hypothetical treatments that reduce the risk of developing a hypothetical disease; data is provided as relative risk reduction (RRR), absolute risk reduction (ARR) and Number Needed to Treat (NNT). There are six questions in this subset, presented as three sets of two questions: the first question asks participants to select the more effective of the treatments, while the second asks them to calculate the risk after Treatment A. The questions used in MINTv1 are shown in Figure 2.14; those used in MINTv2 are shown in Figure 2.15.

From MINTv1

Imagine that 40 out of 1000 people are likely to develop disease Y over the next 5 years. Treatment A reduces the chance of getting disease Y by 25%. Treatment B reduces the chance of getting disease Y by 10%. Which is better, Treatment A or Treatment B? (Q.7)

Imagine that 40 out of 1000 people are likely to develop disease Y over the next 5 years. Treatment A reduces the chance of getting disease Y by 10 per 1000 people. Treatment B reduces the chance of getting disease Y by 4 per 1000 people. Which is better, Treatment A or Treatment B? (Q.17)

Imagine that 40 out of 1000 people are likely to develop disease Y over the next 5 years. 100 people would have to be treated with Treatment A for 5 years for a benefit against disease Y to be evident in one of them. 250 people would have to be treated with Treatment B for 5 years for a benefit against disease Y to be evident in one of them. Which is better, Treatment A or Treatment B? (Q.36)

Each of these questions is followed by the question:
What is the risk of developing disease Y after receiving Treatment A? (Q.8, 18, 36)

Figure 2.14. Treatment A/B questions from MINTv1.

From MINTv2

Without treatment, 40 out of 1000 people will develop disease Y over the next 5 years. Two treatments are available to reduce the risk of developing disease Y. Treatment A reduces the chance of developing disease Y by 25%. Treatment B reduces the chance of developing disease Y by 10%. Which is better, Treatment A or Treatment B? (Q.7)

Without treatment, 40 out of 1000 people will develop disease Y over the next 5 years. Two treatments are available to reduce the risk of developing disease Y. Treatment A reduces the chance of getting disease Y by 10 per 1000 people. Treatment B reduces the chance of getting disease Y by 4 per 1000 people. Which is better, Treatment A or Treatment B? (Q.17)

Without treatment, 40 out of 1000 people will develop disease Y over the next 5 years. Two treatments are available to reduce the risk of developing disease Y. 100 people must be treated with Treatment A for 5 years to prevent one person developing disease Y. 250 people must be treated with Treatment B for 5 years to prevent one person developing disease Y. Which is better, Treatment A or Treatment B? (Q.36)

Each of these questions is followed by the question:
What is the risk of developing disease Y after receiving Treatment A? (Q.8, 18, 36)

Figure 2.15. Treatment A/B questions from MINTv2.

Written feedback on some test papers from my original study with MINTv1 indicated that some students had difficulty with the concept of these hypothetical treatments, and wanted further clinical information before answering. Therefore, as shown in Figure 2.15, I added an additional sentence to the text in MINTv2, in an attempt to highlight that no further data was needed.

2.1.4.3 *Standardisation of terminology*

My review showed that two MINTv1 questions had significant inconsistencies in their terminology. The first, a question based on mammography devised by Peters *et al* (2007), required significant revision as it used a range of terms interchangeably to refer to cancer. I had not noticed these inconsistencies when developing the MINTv1, and had copied the question verbatim. The original and revised text of this question are shown in Figure 2.16.

> **From MINTv1**
> 100 women attend hospital for a mammogram. 10 of these women have a malignant tumour, while 90 do not. Of the 10 patients with malignancy, the mammogram detects the cancer in 9, but misses the cancer in one patient. Of the 90 women who are disease-free, the mammogram indicates correctly that 81 of them are healthy, but wrongly indicates that 9 of them have cancer. Mrs Jones is told that her mammogram is positive. What are the chances that she actually does have cancer? (Q.26)
>
> **From MINTv2**
> 100 people attend hospital for a cancer screening test. However, the screening test is not completely accurate. 10 of the 100 people have cancer, while 90 do not. Of the 10 people with cancer, the screening test detects the cancer in 9, but misses the cancer in 1 person. Of the 90 people who do not have cancer, the screening test indicates correctly that 81 of them do not have cancer, but indicates incorrectly that 9 of them do have cancer. Mr Iqbal is told that his screening test shows that he has cancer. What is the likelihood that he actually does have cancer? (Q.27)

Figure 2.16. Cancer screening questions.

Although medical students and doctors should not find this variation in terminology confusing, I considered that it was important to standardise the text to use only the term “cancer”. In addition to standardising the terminology, I also changed the clinical context of this question from the specific context of a mammogram to a hypothetical cancer screening test. This was because of a clinical inaccuracy in the question. Following my research with MINTv1, I received feedback from a consultant histopathologist that this question was misleading, as one of the “incorrect” answer options was clinically correct. Therefore, participants with clinical experience in this area might confidently select the correct clinical answer, (which was incorrect for the purposes of the test) rather than using the data given in the question to work out the correct test answer. I had not spotted this anomaly when developing the question, but it clearly needed to be changed. Altering the context so that the question relates to a hypothetical cancer eliminates any impact caused by clinical knowledge.

The second question with significant inconsistencies is that based on a table taken from the National Institute for Health and Care Excellence (NICE) guidance for IV fluid and electrolyte therapy (NICE, 2013). The table shows an individual’s electrolyte requirements in millimoles per kilogram of body weight per day, while glucose requirement is given as grams per day. However, the aim of this question is to exploit and highlight the problems that can arise due to a lack of standardisation of clinical information. Therefore, I did not alter this question.

I made a minor amendment to standardise the text of a set of two questions based on screening for Methicillin-Resistant Staphylococcus Aureus (MRSA). During my review, I noticed that most, but not all, of the numbers given in the question were displayed in digital form. This was a minor discrepancy, nonetheless, I revised it for consistency, as shown in Figure 2.17.

> **From MINTv1**
> Miss Strong, an orthopaedic surgeon, screens 100 patients for MRSA preoperatively. Ten of these patients are actually MRSA carriers. The test used gives a true positive result in 90% of MRSA carriers, and a false positive result in 20% of people who do not carry MRSA.
> How many of these 100 patients are expected to test positive? (Q.14)
> What percentage of those who test positive actually carry MRSA? (Q.15)
>
> **From MINTv2**
> Miss Strong, an orthopaedic surgeon, screens 100 patients for MRSA preoperatively. 10 of these 100 patients are actually MRSA carriers. The test used gives a true positive result in 90% of MRSA carriers, and a false positive result in 20% of people who do not carry MRSA.
> How many of these 100 patients are expected to test positive? (Q.14)
> What percentage of those who test positive actually carry MRSA? (Q.15)

Figure 2.17. MRSA screening questions.

2.1.4.4 *Clinical reasons*

Medicine is changing all the time: new treatments are introduced, old ones abandoned, and clinical guidelines are regularly revised in line with the latest evidence. Therefore, I reviewed the clinical setting of each question, to establish whether the test material remained clinically relevant and accurate. In addition to the two questions that were replaced because they had become clinically obsolete as discussed in section 2.1.3, I identified five questions that needed to be changed for clinical reasons. One of these was inappropriate because of new guidance relating to blood transfusion: this stipulates that a unit of blood must be discarded four hours after removal from storage; a minor amendment was sufficient to bring this question up to date, as shown in Figure 2.18.

> **From MINTv1**
> The volume of a unit of blood is 330ml. If it is infused at a rate of 80ml/hr, approximately what proportion of the blood will have been transfused after 2 hours? (Q.9)
>
> **From MINTv2**
> The volume of a unit of blood is 380ml. If it is infused at a rate of 125ml/hr, approximately what proportion of the blood will have been transfused after 2 hours? (Q.12)

Figure 2.18. Blood transfusion questions.

There were two questions where a participant's clinical knowledge or experience might affect their response: that based on mammography is discussed in section 2.1.4.3. The second question where clinical experience may have had an impact was based on a pie chart with five segments; participants were asked to select the smallest segment. I was surprised that 4/135 MINTv1 participants answered this question incorrectly, and suspected that all had made a careless error, and selected the biggest rather than the smallest segment. However, three different wrong answer options had been selected. I considered the test material: each segment of pie in the chart in MINTv1 relates to a common clinical ward task. I hypothesised that participants who gave incorrect answers may have answered the question based on their own clinical experience, rather than by using the data presented in the question. Although this seemed unlikely, I relabelled the chart so that the segments of pie refer to hypothetical wards. This eliminates the possibility of answering the question based on personal experience.

Finally, I changed the clinical context of two questions to broaden the clinical scope of test material. Six questions in MINTv1 were based on cancer scenarios, and four on screening; this seemed disproportionate, so I changed the setting of two questions. As shown in Figures 2.19 and 2.20, both are so similar to the originals that they cannot be considered as "new" questions.

From MINTv1
The chance of a skin lesion being cancerous is 1%. If 1000 people attend the dermatology clinic with this skin lesion, how many are likely to have cancer? (Q.19)

From MINTv2
The chance of a hip replacement operation being cancelled is 1%. If 1000 people are scheduled for hip replacement operations, how many are likely to have their operation cancelled? (Q.41)

Figure 2.19. Risk questions (percentage to frequency).

From MINTv1
People being screened for a virus are told that the chance of testing positive is 1 in 1000. What percentage of people has the virus? (Q.10)

From MINTv2
People starting on treatment for hypertension are given a 1 in 1000 chance of developing a particular complication. What percentage of people are likely to develop this complication? (Q.10)

Figure 2.20. Risk questions (frequency to percentage).

2.1.5 ***Review of health numeracy frameworks***

I had developed the blueprint for MINTv1 using the framework for health numeracy (HN) devised by Golbeck *et al* (2005). This describes three distinct, but overlapping constructs: computational, analytical and statistical numeracy. I categorised each question as computational, analytical or statistical. In many cases this was straightforward: questions based on analysis of charts and graphs were clearly analytical in construct, while those related to probability concepts were undoubtedly statistical. However, several questions were difficult to categorise because they had significant elements of more than one construct; thus it seemed worthwhile to evaluate alternative frameworks to assess whether another might be more suitable for categorising and analysing the MINT. I reviewed the frameworks developed by Ancker & Kaufman (2007), Schapira *et al* (2008), and Caverly *et al* (2012) and compared them to that of Golbeck *et al* (2005).

2.2 EXTERNAL REVIEW

The internal review process resulted in a revised test paper, MINTv2. This paper was then reviewed by nine independent reviewers involved in medical education: seven clinicians, and two academics. The clinicians had varying levels of training and experience: three were consultants from different specialty backgrounds with significant commitments to either undergraduate or postgraduate medical education at Foundation training level, two were clinical teaching fellows, one was a core medical trainee and one a foundation trainee. The academics were based at Keele School of Medicine, and included a co-lead for year 3, and a lecturer in biology, both of whom are interested in numeracy in medical students.

All external reviewers were given copies of MINTv2 and asked to evaluate test questions. This process was similar to my internal review: each reviewer was asked to assess the clarity of the text of each question, and the quality of all charts and graphs. Additionally, reviewers were asked to consider how difficult they thought test questions were, and to rate the level of difficulty of each question on a scale of 1-5, where 1 is easy and 5 is difficult. All reviewers were also given a copy of the HN framework devised by Golbeck *et al* (2005), and asked to familiarise themselves with the definitions of the different constructs of health numeracy. They were then asked to categorise the principal construct being tested in each MINTv2 question (computational, analytical or statistical) using this framework. Finally, those reviewers who were clinicians were also asked to comment on the clinical relevance of each question.

2.3 EVALUATION STUDY: MINTv2

Once the external review was completed, I finalised the test paper, and obtained approval from the University of Manchester's Research Ethics Committee (UREC) to conduct an evaluation study of the revised test, MINTv2. MINTv2 was developed as a constructed response (CR) test, with room for rough work on the question papers. Implementing the test in the CR format was important to allow analysis of participants' responses; the incorrect answers given to test questions were used to inform Stage 4 of the revision process, the development of distractors for the single best answer (SBA) version of the test, MINTv3.

2.3.1 ***Study Methods***

Participants

Participants were third year medical students at Keele university (KU). The School of Medicine (SoM) at KU evaluates numeracy in medical students at entry, and during the first year of their course, and agreed to include the MINT in formative exams taken by third year students. All third-year students were invited to participate in the study, with no exclusions.

Materials

Test materials included the MINTv2 test paper, an answer sheet, a pencil and eraser, and were provided in a brown A4 envelope. Some of the group were also randomised to receive calculators: the impact of calculator use is discussed in Chapter 3.

Procedure

One month prior to the study, I delivered a teaching session on clinician numeracy and its importance for healthcare professionals to the full year group. I then sent all students an email with a participant information sheet containing further details about the research, and inviting them to participate. I gave a short talk about the MINT at the briefing session prior to the examinations, and emphasised that participation in the research was optional.

All students were given the MINTv2 in its blank answer format. The test was carried out under examination conditions, and students had 90 minutes to complete the test. A few weeks later I delivered a feedback session to the students, discussing both the overall results and each question in detail.

Data Analysis

The aims of this process were: 1) to assess whether participants found any test questions confusing; and 2) to evaluate whether MINTv2 was equivalent to MINTv1, by comparing mean scores and the facility of individual test questions. Data collected was individual test score and facility of test questions; rough work written on the test sheets was also collected for analysis of error. Demographic data collection was limited to student gender and the presence or absence of dyslexia.

Data were analysed using Microsoft EXCEL, and an online statistical tool (MedCalc Software bvba).

2.3.2 ***Study limitations***

The aim of this study was to evaluate whether MINTv2 was equivalent to MINTv1, by comparing mean scores and the facility of individual test questions. However, this comparison has some limitations because of differences in test format, test materials, and participants.

MINTv1 was delivered as an SBA test, while MINTv2 is a CR test. This difference in format is important, and may bias the comparison of results, since scores on SBA tests are often elevated due to cueing and guessing (Simkin & Kuechler, 2005; Betts *et al* 2009; DiBattista & Kurzawa, 2011; Funk & Dickson, 2011).

Participants in the evaluation study with MINTv2 were third year medical students; however, MINTv1 was tested on Foundation Trainee doctors (FTs). This difference may affect test scores, since the questions are set in a clinical context, and FTs have considerably more clinical experience than third-year medical students. However, the MINT is not a test of clinical knowledge, thus additional clinical experience may not affect performance. Furthermore, when developing MINTv1 in 2013, I had conducted a pilot study using a convenience sample of third year medical students: their performance was similar to that of the FTs in the implementation study (Taylor, 2014).

Finally, participants in the evaluation study of MINTv2 were also taking part in a randomised controlled trial (RCT) to assess the impact of calculators. Therefore, approximately half of the participants sitting MINTv2 had calculators, compared to none of those sitting MINTv1. However, the MINT was designed to be done without using calculators, as I considered that calculators would not be helpful for much of the test content, particularly analytical and statistical constructs. The RCT is fully discussed in Chapter 3.

2.4 DEVELOPMENT of MINTv3

2.4.1 ***Overview***

The final stage of the revision process was to develop distractors for MINTv2, to create the SBA version of the test, MINTv3. This stage involved reviewing the distractors used in MINTv1, comparing them to the answers given by participants in the evaluation study of MINTv2, and then writing new distractors based on the most common incorrect answers provided by participants in MINTv2. Therefore, development of the SBA answers for MINTv3 was a separate process to the revision of questions for MINTv2. Furthermore, the evaluation of the SBA answers required a separate testing process; this involved delivering MINTv3 to a full student year group.

2.4.2 ***Rationale***

My analysis of the original data with MINTv1 had suggested that some of the SBA answer options were poor distractors. A question adapted from material developed by Sikorskii *et al* (2011), had a correct answer of '27', with four distractors: '9', ' 72', '80' and '90'. Data analysis showed that 81/135 participants selected the correct answer, while 38/135 selected the answer '9'; the remaining answers were poor distractors (Taylor & Byrne-Davis, 2016). Analysis of rough work provided by participants in MINTv1 showed that many incorrect answers ranged between '20' and '29', hence participants who calculated answers in this range might select '27' in the SBA test, even if they had not calculated the correct answer. Therefore, it was important to ensure that the MINT had plausible distractors, thus I decided to develop evidence-based distractors using data provided by participants in the evaluation study. This approach to developing distractors is advocated by Birenbaum & Tatsuoka (1987) and DiBattista and Kurzawa (2011).

2.4.3 ***Methods***

I analysed all of the answer sheets used by students who participated in the evaluation study of MINTv2. I coded answers given to all 43 questions on every answer sheet with a letter code. Answers that corresponded to the 'A' – 'E' options used in MINTv1 were given the corresponding codes 'A' to 'E', and the absence of an answer was coded 'X'. Incorrect answers that had not been used as distractors in MINTv1 were coded in alphabetical order starting from the letter 'F'. I recorded the answers given to all questions, and also recorded the number of distractor options used in MINTv1 that were not offered as answers by participants in MINTv2. I then developed MINTv3, an SBA test format, using new distractors based on the most common incorrect answers provided for MINTv2. I conducted an evaluation study of MINTv3.

2.5 EVALUATION STUDY of MINTv3

This study was approved by UREC; the study methods were similar to those of the first evaluation study, and are briefly outlined below.

2.5.1 ***Study Methods***

Participants

Participants were third year medical students at Keele university who sat MINTv3 as a formative exam. All third-year students were invited to participate in the study, with no exclusions.

Materials

The test materials included the MINTv3 test paper, an answer sheet, a pencil, an eraser, and a calculator.

Procedure

Approximately one month prior to the study, students attended a teaching session on clinician numeracy; they were then sent a participant information sheet with further details about the research by email, and invited to participate in the study. The test was carried out under examination conditions, with 90 minutes to complete the test. A follow-up session was offered to all students.

Data Analysis

The aim of this process was to assess the performance of MINTv3. In addition to recording test scores, I calculated the facility and item discrimination of each question. I used Cronbach's alpha to measure internal consistency reliability of MINTv3. Demographic data collection included student gender and dyslexia status.

Data were analysed using Microsoft EXCEL, and an online statistical tool (MedCalc Software bvba).

Limitations

This study aimed to evaluate MINTv3 by comparing psychometric data with MINTv1. However, there were two key differences relating to participants: as with the first evaluation study, participants were third year medical students rather than Foundation Trainee doctors (FTs); this may affect test scores. Secondly, all participants in the second evaluation study had calculators, while the FTs who sat MINTv1 did not.

SECTION 3. RESULTS

The revision of the MINT leading to the development of MINTv3 is summarised in Table 2.2 (Section 2). The process involved five stages: 1) an internal review of MINTv1 to produce

MINTv2; 2) an external review of MINTv2; 3) an evaluation study of MINTv2; 4) the development of distractors for MINTv3; and 5) an evaluation study of MINTv3.

3.1 INTERNAL REVIEW

This consisted of a thorough review of MINTv1, leading to the development of a fully revised test, MINTv2. During this process, I replaced 9/43 questions, and made changes to 23/43, although the majority of alterations were minor. The remaining 11/43 questions were unchanged. An outline of the changes made to individual questions during each stage is shown in Table 2.3.

Table 2.3. Development of MINTv2: changes to individual questions

Q. no.	Content	Internal review	External review	Evaluation study MINTv2
1	Calculate, other	Text re-ordered		
2	Pie chart	Clinical context changed; Chart amended		
3	Drug comparison	Text re-ordered		
4	Calculate, other	No change		
5	Drug infusion	Text re-ordered		
6	NICE guidance	No change		
7- 8	A/B (RRR)	Text added		
9	% risk	No change		
10	Conversion	Clinical context changed		
11	Conversion	No change		
12	Proportion	Text amended (clinical)		
13	Calculate, formula	No change		
14 - 15	Screening data	Text standardised		
16	Calculate mean	New question		
17 - 18	A/B (ARR)	Text added		
19 - 22	NLQ 1-4	4 new questions	Text standardised	
23 - 24	Scattergram	Change in dates		
25	Calculate, %	Text amended		
26	Risk	No change		
27	Screening data	Clinical context changed Text standardised		
28	Table & bar chart	No change		
29	Proportion	No change		
30	Drug dose	New question		
31 - 32	Line graph	2 New questions		Text amended
33	Drug dose	New question		
34	Bar chart	Change in dates		
35	Drug dose	No change		
36 - 37	A/B (NNT)	Text added		
38	Risk	Text amended		
39	Table	No change		
40	Drug infusion	No change		
41	Conversion	Clinical context changed		
42	Bar chart	Text amended	Chart amended	
43	Drug dose	Text re-ordered		

3.1.1 ***Development of new MINT questions***

I wrote nine new questions for MINTv2. Seven of these replaced questions that were subject to copyright, and two replaced test material that had become clinically redundant. All new questions were designed to be equivalent to the original material in MINTv1 in terms of construct and level of difficulty. The new questions were evaluated by nine external reviewers, and then tested on medical students.

3.1.2 ***Changes to MINTv1 questions***

I amended 23/43 questions, as outlined below.

3.1.2.1 *Improved clarity of text and data displays*

I revised three charts and graphs (affecting four questions MINTv2 Q.2, 23 and 24, 34); and amended the text of a further 13 questions: for seven of these, the changes were minor (MINTv2 Q.1, 3, 5, 25, 38, 42, 43); but six required more substantial amendments (MINTv2 Q.7 and 8, 17 and 18, 36 and 37).

3.1.2.2 *Standardisation of terms*

I standardised the text of three questions: this required a minor amendment for two questions (MINTv2 Q.14 and 15), but a substantial revision of the terminology in one question (MINTv2 Q.27).

3.1.2.3 *Revision for clinical reasons*

I revised three further questions for clinical reasons. One simply required a minor modification to ensure compliance with current clinical practice (MINTv2 Q.12); and the clinical context of two questions was changed to introduce greater variety into the clinical scope of the test (MINTv2 Q.10, 41). I altered another two questions to eliminate the possible impact of clinical experience on participants' responses; both of these questions had additional alterations, and are mentioned above (MINTv2 Q.2, section 3.1.2.1; MINTv2 Q.27, section 3.1.2.2).

3.1.3 ***Review of numeracy frameworks***

Having used the framework for health numeracy (HN) developed by Golbeck *et al* (2005) to classify questions in MINTv1, I reviewed four alternative frameworks to assess whether they might be more suitable for classifying MINT items (Table 2.4). The framework for health literacy (HL) used by Nutbeam (2000) has three categories: functional, interactive, and critical HL. While useful for assessing and understanding a patient's ability to use healthcare information, it is unsuitable for classifying items in a test of numeracy. The framework described by Ancker & Kaufman (2007) has three categories: basic computation, estimation, and statistical literacy, and is somewhat similar to that of Golbeck *et al* (2005). However, there are two important differences: 1) Ancker & Kaufman (2007) consider "the manipulation of percentages and probabilities" to be "basic computation", whereas this skill is categorised as "statistical" by Golbeck *et al* (2005); and 2) Ancker & Kaufman (2007) consider estimation to be a category in its own right, but do not mention the interpretation of graphs and other data displays as a quantitative skill. Clinicians need to assess medical information provided in charts and graphs of various kinds; therefore, data displays form the basis of several MINT questions. Since Ancker & Kaufman's (2007) framework does not consider data displays, it cannot be used to

Table 2.4. Comparison of health numeracy frameworks

Nutbeam 2000	**Golbeck *et al* 2005**	**Ancker & Kaufman 2007**	**Schapira *et al* 2008**	**Caverly *et al* 2012**

Functional health literacy Understand basic patient education information	*Basic numeracy* Number recognition, understand quantitative data *Computational numeracy* Basic maths skills Simple manipulation of numbers	*Basic computation* Number recognition & comparisons; arithmetic; use of simple formulae; manipulation of percentages & probabilities	*Primary numeric skills* Basic arithmetic, graphs, dates & time	*Primary numeric skills* Counting Basic math functions Calculate ARR, RRR, NNT Scales & graphs
Interactive health literacy Skills needed to act on healthcare information	*Analytical numeracy* Making sense of information Understanding graphs & other data displays Higher functions e.g. inference, estimation, proportions, percentages, frequencies	*Estimation* Generally quicker & simpler than precise calculations; often sufficient for decision making; helps with judging probable correctness of a calculation; important in health contexts.	*Applied health numeracy* Basic: using numbers in healthcare tasks Risk communication: using numbers to communicate probabilistic health information Decision-making: balancing risk and benefit information, assessment of evidence	*Applied health numeracy* Interpret lab values Calculate drug doses Use prognostic or diagnostic tools Disease incidence Risk factor modification Prognosis, survival Information seeking Balancing risks & benefits Assessment of evidence Estimation & sense of magnitude
Critical health literacy Skills in working within a given social & economic context	*Statistical numeracy* Understanding basic biostatistics including probability statements Compare different scales (probability, proportion, percent) Ability to critically analyse quantitative information Understanding concepts such as randomisation	*Statistical literacy* Understanding chance & uncertainty; margins of error; randomisation in clinical trials; ability to evaluate scientific information *Representational fluency* Document literacy Graphical literacy	*Interpretive health numeracy* Understanding the strengths & limitations of numbers to represent disease states, efficacy of an intervention or other healthcare outcome. Includes probability, concept of uncertainty, & principles of scientific methods	*Interpretive health numeracy* Probability & chance Principles of scientific methods Concept of uncertainty Graphic & verbal formats Individual & biologic variation in expected outcomes Estimation & sense of magnitude

classify MINT items. Schapira *et al* (2008) also classify numeracy into three competencies: primary numeric skills, applied HN, and interpretive HN. Caverly *et al* (2012) used this framework as a basis for developing their test of clinician numeracy (CN). However, their assessment of CN, the Critical Risk Interpretation Test, was quite different to the MINT, and aimed to measure “risk gist”, using fuzzy trace theory as a way of understanding medical decision making (Caverly *et al* 2012). Therefore, their framework is more complex, and differs conceptually from the Golbeck *et al* (2005) framework, being designed to assess the thought processes involved in answering questions, rather than classifying the basic mathematical operation involved. Moreover, their framework considers understanding graphs and the concepts of ARR, RRR, NNT to be primary numeric skills, while “applied health numeracy” includes computational, analytical and statistical constructs. Using this framework would entail classifying most MINT items as “applied health numeracy”, and would not be helpful in terms of considering the numeracy skills of medical students and doctors; hence it is unsuitable for classifying the MINT. Therefore, the Golbeck *et al* (2005) framework remained the most appropriate system for categorising MINT items.

3.2 EXTERNAL REVIEW

Nine external reviewers evaluated MINTv2, and advised on the clarity of questions and their level of difficulty, as outlined in section 2.2. Reviewers also classified questions according to numeracy construct. Clinician reviewers considered whether questions were clinically appropriate.

3.2.1 *Review of MINTv2 questions*

Reviewers identified two questions where changes were needed: one observed that the text of the nutritional label questions was inconsistent, as the words “carton” and “container” were used interchangeably. I standardised the terminology to the word “carton”. Another reviewer queried the data given in a bar chart, where the bars added up to 29 places; this was an error, and I adjusted the chart to show the intended 30 places. Changes to the MINTv2 test paper are shown in Table 2.3.

3.2.2 *Allocation of level of difficulty*

Reviewers rated questions on a scale of 1 - 5, where 1 was easy, and 5 difficult: they varied greatly in their assessment of level of difficulty, and were unanimous only in rating the pie chart as easy. For the remainder, there was often little consensus: four of the questions were rated across the full range of 1 to 5, and 15 were rated across four levels (either 1 - 4, or 2 - 5). I ranked the results for each question, and assigned level of difficulty based on the median value of the reviewers’ ratings. I then calculated the mean value of ratings for each question, and cross-checked this with its median value. I found that level of difficulty assigned on the basis of median rating was consistent with the mean rating for 40/43 questions. I reviewed the three questions where the mean and median ratings were inconsistent: two of these had median values of 4, and had been assigned to level 4. However, there was a clear gap between the mean values of these questions (4.1 and 4.2) and the mean values of others assigned to level 4 (3.5 – 3.7). These two questions were the most difficult of the test, and the only ones with mean ratings above 4; therefore, these were re-assigned to level 5. The remaining outlier was a question that had been assigned to level 3 based on its median value; however, since its mean score was in the range for level 4, it was re- assigned to level 4. Allocation of level of difficulty is summarised in Table 2.5.

Table 2.5. Allocation of level of difficulty

Panel rating						**Close *et al* (2008) criteria***	
Mean rating	Median rating	Final rating	Difficulty	No. items	%	No. items	%
1 – 1.6	1	1	Easy	9	21%	23	54
1.7 – 2.3	2	2	Fairly easy	13	30%	4	9
2.4 – 3.4	3	3	Average	15	35%	11	25
3.5 – 3.7	4	4	Fairly difficult	4	9%	2	5
4.1 – 4.2	4	5	Difficult	2	5%	3	7

**See section 3.3.2.3*

The external review process resulted in changes to the assigned level of difficulty of 25 of the original 34 MINTv1 questions: the level of difficulty of 21 questions was reduced and of four was raised (Table 2.6). Thus, for MINTv2, 22 questions were considered by raters to be either easy or fairly easy, 15 to be average, and 6 to be either difficult or fairly difficult. Two of the new MINTv2 questions were rated easy (level 1), three fairly easy (level 2), three average (level 3), and one fairly difficult (level 4). Four of the original nine questions were rated as fairly easy (level 2), two as average (level 3), and three as fairly difficult (level 4).

Table 2.6. Changes in level of difficulty of MINT questions

Item no.	Content	MINTv2	MINTv1	Change in difficulty
1	Calculate, other	2	2	None
2	Pie chart	1	1	None
3	Drug comparison	3	5	Reduced by 2 levels
4	Calculate, other	2	3	Reduced by 1 level
5	Drug infusion	3	3	None
6	NICE guidance	2	4	Reduced by 2 levels
7	A/B RRR 1	1	3	Reduced by 2 levels
8	A/B RRR 2	3	3	None
9	% risk	3	5	Reduced by 2 levels
10	Conversion	1	2	Reduced by 1 level
11	Conversion	3	5	Reduced by 2 levels
12	Proportion	2	1	Increased by 1 level
13	Calculate, formula	3	4	Reduced by 1 level
14	Screening data 1	5	5	None
15	Screening data 2	4	5	Reduced by 1 level
16	Calculate, mean	1	3	New: N/A
17	A/B ARR 1	1	3	Reduced by 2 levels
18	A/B ARR 2	3	3	None
19	NLQ 1	2	2	New: N/A
20	NLQ 2	2	2	New: N/A
21	NLQ 3	3	2	New: N/A
22	NLQ 4	2	2	New: N/A
23	Scattergram 1	2	5	Reduced by 3 levels
24	Scattergram 2	4	5	Reduced by 1 level
25	Calculate %	2	2	None
26	Risk	2	5	Reduced by 3 levels
27	Screening data	5	5	None
28	Table & bar chart	3	1	Increased by 2 levels
29	Proportion	3	5	Reduced by 2 levels
30	Drug dose	4	3	New: N/A
31	Line graph 1	1	4	New: N/A
32	Line graph 2	3	4	New: N/A
33	Drug dose	3	4	New: N/A
34	Bar chart	2	3	Reduced by 1 level
35	Calculate, other	2	3	Reduced by 1 level
36	A/B NNT 1	1	3	Reduced by 2 levels
37	A/B NNT 2	3	5	Reduced by 2 levels
38	Risk	1	2	Reduced by 1 level
39	Table	2	5	Reduced by 3 levels
40	Drug infusion	4	3	Increased by 1 level
41	Conversion	1	2	Reduced by 1 level
42	Bar chart	3	1	Increased by 2 levels
43	Drug dose	3	3	None

3.2.3 *Allocation of numeracy constructs*

Reviewers were asked to assign questions to one of three constructs: computational, analytical or statistical (Table 2.7). The nine new questions were similar in construct to the questions that they replaced, and so all 43 questions are considered together here. Reviewers agreed unanimously on the construct of 7/43 questions, and for the remainder, I assigned construct based on the majority opinion. The construct of 3/43 questions was altered by this process: in MINTv1, 13 questions were considered to be primarily computational, 17 analytical and 13 statistical; for MINTv2, the consensus was that 16 questions were computational, 14 analytical, and 13 statistical (Table 2.7).

Table 2.7. Psychometric data for MINTv2

Q no.	Construct MINTv2	Construct MINTv1 if altered	Facility	Item Discrimination	Cronbach's alpha if question removed
	All Computational		*.77*		
25	Computational		.96	.13	.77
20	Computational		.95	.10	.77
1	Computational		.92	-.04	.76
16	Computational		.88	.17	.76
4	Computational		.85	.23	.76
12	Computational	Analytical	.85	.20	.77
22	Computational		.85	.27	.76
35	Computational		.80	.30	.76
21	Computational		.79	.27	.76
43	Computational		.78	.36	.76
5	Computational		.70	.43	.76
29	Computational	Analytical	.67	.37	.77
30	Computational		.65	.50	.76
19	Computational		.62	.36	.76
40	Computational		.55	.43	.76
33	Computational	Analytical	.54	.50	.76
	All Analytical		*.74*		
2	Analytical		1.0	0	.77
28	Analytical		.91	.14	.76
31	Analytical		.89	.27	.76
34	Analytical		.88	.20	.76
18	Analytical		.83	.33	.76
23	Analytical		.83	.30	.76
39	Analytical		.83	.20	.77
42	Analytical		.83	.37	.76
8	Analytical		.69	.56	.78
32	Analytical		.64	.57	.76
13	Analytical		.60	.23	.77
37	Analytical		.53	.70	.76
6	Analytical		.52	.30	.77
24	Analytical		.39	.50	.76
	All Statistical		*.70*		
38	Statistical		.98	0	.77
41	Statistical		.96	.03	.77
7	Statistical		.93	.10	.77
36	Statistical		.92	.06	.77
17	Statistical		.88	.14	.77
10	Statistical		.87	.23	.77
26	Statistical		.84	.13	.77
9	Statistical		.76	.10	.77
3	Statistical		.58	.53	.76
14	Statistical		.44	.63	.76
11	Statistical		.43	.47	.76
15	Statistical		.27	.43	.77
27	Statistical		.25	.54	.76

3.2.4 *Question order*

Following the internal and external review processes, I changed the order of questions in the paper to create a more balanced spread of questions across the test in terms of level of difficulty and construct.

3.3 EVALUATION STUDY: MINTv2

115 third-year medical students sat the formative year 3 examination, of whom 110/115 consented to participate in the study. There were 59 (54%) female and 36 (33%) male students; 15 (13%) students declined to indicate their gender. Twelve of 110 (11%) students stated that they had been diagnosed with dyslexia, and were given extra time in university examinations. Scores for MINTv2 and MINTv1 are shown in Table 2.8.

Table 2.8. Comparison of scores for MINTv2 and MINTv1

	MINTv2	MINTv1
Participants	Medical students	Foundation trainees
No.	110	135
Test format	Constructed response	SBA
Range	19 - 43	14 – 42
Interquartile range	30 - 35	29 - 38
Mean (SD)	31.8 (5.23)	32.8 (6.64)
Median	33	34
Mode	30, 33	34, 38
Dyslexia	12	0

3.3.1 ***Evaluation of the text of MINTv2 questions***

During the evaluation study, a participant noted a typographical error that had eluded previous review. A line graph showing the metabolism of a hypothetical drug was mislabelled as “Drug Z” rather than “Drug D” (Q. 31 and 32). This error was corrected; no further errors or inaccuracies were identified.

3.3.2 Evaluation of the performance of MINTv2 questions

3.3.2.1 *Overall performance*

Data analysis using Student’s t-test (online calculator, MedCalc Software bvba) showed that there was no difference in mean and median scores of participants in MINTv1 and MINTv2; the overall range and interquartile range of scores were also similar (Table 2.8).

3.3.2.2 *Facility of MINTv2 questions*

The facility of questions in MINTv2 varied from 0.25 to 1.0, and the majority of questions 33/43 (77%) had a facility greater than 0.6 (Table 2.9). There was a significant difference in the facility of ten questions in MINTv2 compared to those in MINTv1: six of these were original MINTv1 questions, while four were new MINTv2 questions. Of the original six MINTv1 questions, 4/6 had a lower facility in MINTv2 (Q.5, 8, 15, 40), and 2/6 had a higher facility (Q.34, 39). Three of the four new questions had a lower facility than the questions they replaced (Q. 19, 30, 33), while one had a higher facility (Q.20). All data relating to facility of test questions is shown in Table 2.9.

Table 2.9. Comparison of facility of test questions in MINTv2 and MINTv1

Question no.	Question Content	MINTv2 N=110	MINTv1 N=135	Diff*	95% CI	χ^2	p value**
1	Calculate, other	.92	.86	.06			ns
2	Pie chart	1.0	.97	.03			ns
3	Drug comparison	.58	.61	-.03			ns
4	Calculate, other	.85	.87	-.02			ns
5	Drug infusion	.70	.88	-.18	7.8-28	12	0.0005
6	NICE guidance	.52	.56	-.04			ns
7	A/B RRR 1	.93	.91	.02			ns
8	A/B RRR 2	.69	.90	-.21	11-31	17	<0.0001
9	% risk	.76	.67	.09			ns
10	Conversion	.87	.86	.01			ns
11	Conversion	.43	.54	-.11			ns
12	Proportion	.85	.95	-.10			ns
13	Calculate, formula	.60	.60	0			ns
14	Screening data 1	.44	.60	-.16			ns
15	Screening data 2	.27	.51	-.24	12-35	14	0.0001
16	Calculate mean***	.88	.79	.09			ns
17	A/B ARR 1	.88	.93	-.05			ns
18	A/B ARR 2	.83	.84	-.01			ns
19	NLQ 1***	.62	.85	-.23	12-34	17	<0.0001
20	NLQ 2***	.95	.76	.19	10-27	17	<0.0001
21	NLQ 3***	.79	.87	-.08			ns
22	NLQ 4***	.85	.92	-.07			ns
23	Scattergram 1	.83	.90	-.07			ns
24	Scattergram 2	.39	.53	-.14			ns
25	Calculate %	.96	.93	.03			ns
26	Risk	.84	.70	.14			ns
27	Screening data	.25	.40	-.15			ns
28	Table & bar chart	.91	.92	-.01			ns
29	Proportion	.67	.58	.09			ns
30	Drug dose***	.65	.85	-.20	9-30	13	0.0003
31	Line graph 1***	.89	.81	.08			ns
32	Line graph 2***	.64	.60	.04			ns
33	Drug dose ***	.54	.76	-.22	10-33	13	0.0003
34	Bar chart	.88	.67	.21	10-30	14	0.0001
35	Drug dose	.80	.83	-.03			ns
36	A/B NNT 1	.92	.84	.08			ns
37	A/B NNT 2	.53	.42	.11			ns
38	Risk	.98	.93	.05			ns
39	Table	.83	.64	.19	8-29	11	0.0009
40	Drug infusion	.55	.91	-.36	25-46	41	<0.0001
41	Conversion	.83	.87	-.04			ns
42	Bar chart	.84	.81	.03			ns
43	Drug dose	.78	.80	-.02			ns

* *A positive value indicates that facility is higher in MINTv2, a negative value, higher facility in MINTv1*

** *Bonferroni correction applied, thus significant at 5% level if p< 0.001*

****New questions, so not a direct comparison*

3.3.2.3 *Facility and allocated level of difficulty of MINT questions*

As discussed in section 3.2.2, the anticipated level of difficulty of questions was determined by nine external reviewers. Following implementation, I analysed the level of difficulty of test questions based on their facility, using the criteria described by Close *et al* (2008). Thus, test items were divided into five groups as follows: easy (facility > 0.8); moderately easy (facility 0.7 – 0.79); average (facility 0.5 – 0.69); moderately difficult (facility 0.40 – 0.49) and difficult (facility <0.40). A comparison of the allocated level of difficulty achieved by the panel and by using facility is shown in Table 2.10. Applying the Close *et al* (2008) criteria to MINTv2, 23 questions were easy; 4 moderately easy; 11 average, 2 moderately difficult and 3 difficult questions (Table 2.5). The panel's allocation of level of difficulty agreed with the Close *et al* (2008) category for 17/43 questions; for 23/43 questions there was a difference of one level, and for 3/43 questions the difference was two levels. In 21/26 cases where there was a difference, the panel suggested a higher level of difficulty than was apparent based on facility (Table 2.10).

Table 2.10. Comparison of allocated level of difficulty by panel and by facility

Item no.	Content	Facility	Level of difficulty by Panel	Level of difficulty by Facility	Difference*
1	Calculate, other	.92	2	1	1
2	Pie chart	1.0	1	1	0
3	Drug comparison	.58	3	3	0
4	Calculate other	.85	2	1	1
5	Drug infusion	.70	3	2	1
6	NICE	.52	2	3	-1
7	A/B RRR 1	.93	1	1	0
8	A/B RRR 2	.69	3	3	0
9	% risk	.76	3	2	1
10	Conversion	.87	1	1	0
11	Conversion	.43	3	4	-1
12	Proportion	.85	2	1	1
13	Calculate, formula	.60	3	3	0
14	Screening data	.44	5	4	1
15	Screening data	.27	4	5	-1
16	Calculate mean	.88	1	1	0
17	A/B ARR 1	.88	1	1	0
18	A/B ARR 2	.83	3	1	2
19	NLQ 1	.62	2	3	-1
20	NLQ 2	.95	2	1	1
21	NLQ 3	.79	3	2	1
22	NLQ 4	.85	2	1	1
23	Scattergram 1	.83	2	1	1
24	Scattergram 2	.39	4	5	-1
25	Calculate %	.96	2	1	1
26	Risk	.84	2	1	1
27	Screening data	.25	5	5	0
28	Table & bar chart	.91	3	1	2
29	Proportion	.67	3	3	0
30	Drug dose	.65	4	3	1
31	Line graph 1	.89	1	1	0
32	Line graph 2	.64	3	3	0
33	Drug dose	.54	3	3	0
34	Bar chart	.88	2	1	1
35	Calculate, other	.80	2	1	1
36	A/B NNT 1	.92	1	1	0
37	A/B NNT 2	.53	3	3	0
38	Risk	.98	1	1	0
39	Table	.83	2	1	1
40	Drug dose	.55	4	3	1
41	Conversion	.96	1	1	0
42	Bar chart	.83	3	1	2
43	Drug dose	.78	3	2	1

**A positive value indicates that the level of difficulty assigned by the panel was greater than that based on facility, a negative value indicates the opposite.*

3.3.2.4 *Discrimination index*

The overall discrimination index of MINTv2 was 0.29, while that of individual test questions ranged from -.04 to 0.7. 30/43 questions had a discrimination index of 0.2 or above, of which 13/43 questions had an index of 0.4 – 0.7.

3.2.4.2 *Reliability*

Internal consistency reliability of the test as measured by Cronbach's alpha was 0.77. I analysed the effect of individual questions by calculating Cronbach's alpha if they were removed: the variation in Cronbach's alpha was small, ranging from 0.76 - 0.78. All psychometric data relating to MINTv2 is shown in Table 2.7.

3.3.2.4 *Performance on nutritional label questions*

52/110 participants (47%) answered all four nutritional label questions (NLQ) correctly, compared to 56% who answered all four NVS questions correctly in MINTv1. Statistical analysis with N-1 Chi-squared test indicated that these proportions were similar (difference 9%, 95% C.I. -3.5 – 21, $\chi^2$1.9, DF 1, ns). The mean scores decreased according to the number of NLQs answered correctly, as shown in Table 2.11. As with MINTv1, there was a strong correlation between number of NLQs answered correctly and mean test score: r = 0.93 for MINTv2 (r = 0.99 for MINTv1). However, the range of scores was broad e.g. participants who answered all four questions correctly had scores varying from 19 – 43 (Table 2.11).

Table 2.11. Comparison of performance on nutritional label questions

No. NLQ answered correctly		MINTv2			MINTv1	
	n (%)	Mean score (%)	Range	n (%)	Mean score	Range
4	52 (47%)	34.3 (79%)	19-43	76 (56%)	34.9 (81%)	19-42
3	38 (35%)	30.3 (70%)	19-38	46 (34%)	31.8 (74%)	16-40
2	11 (10%)	28.6 (66%)	21-36	8 (6%)	25.6 (59%)	18-32
1	9 (8%)	28 (65%)	21-35	2 (1.5%)	23 (53%)	22-24
0	0	N/A	N/A	3 (2.2%)	17.6 (41%)	14-21

3.3.2.5 *Performance on Treatment A v Treatment B questions*

86/110 participants (78%) answered all three versions of the Treatment A v B questions correctly; this was similar to results with MINTv1 (Table 2.12).

Table 2.12. Correct answers to Treatment A v Treatment B comparison questions

	AvB RRR	AvB RRR	AvB RRR	All 3 correct
MINTv2	93%	88%	92%	78%
MINTv1	91%	93%	84%	77%
p value	ns	ns	ns	ns

3.4 DEVELOPMENT OF MINTv3

I analysed the answer sheets of students who had participated in the evaluation study of MINTv2. All answers were coded: those that corresponded to options in MINTv1 were allocated the appropriate letters ‘A’ to ‘E’; unanswered questions were coded ‘X’, and new answers that did not correspond to SBA options in MINTv1 were coded in alphabetical order, starting with the letter ‘F’. The results of this analysis are shown in Table 2.13. The data for the 34 MINTv1 questions and the nine new MINTv2 questions are reported separately.

3.4.1 *Answer analysis relating to 34 MINTv1 questions*

No new answers were provided for 8/34 questions: this includes five questions for which there were a maximum of five possible answers (e.g. Q. 2 pie chart with 5 segments). New answer options were provided for 26/34 MINTv1 questions; the number of new answers provided per question ranged from 1-15, and 11 questions were given 10 or more different new answer options. For 32/34 MINTv1 questions, at least one of the original distractors was not offered as an answer by participants sitting MINTv2; none of the original distractors was used in 4/34 questions.

3.4.2 *Answer analysis relating to 9 new MINTv2 questions*

The new questions were answered correctly by 60/110 – 104/110 participants; thus incorrect answers were provided by 6/110 – 50/110 participants. The number of incorrect answer options for these questions ranged from 2 (Q.19) to 19 (Q.33).

3.4.3 *Development of new distractors for MINTv3*

Following analysis of all incorrect answers, I developed new distractors for the SBA version of the test, MINTv3 (Appendix 4). The number of new distractors created per question is shown in Table 2.14.

Table 2.13. Analysis of answers given in MINTv2*

Q.	Construct	A	B	C	D	E	X	Other answer	Unused option	New answer
1	C	0	0	0	7	**101**	0	2	3	1
2	A	**110**	0	0	0	0	0	0	4	0
3	S	46	**64**	0	0	0	0	0	3	0
4	C	0	0	1	**94**	0	1	14	3	9
5	C	0	8	**77**	0	1	5	19	2	11
6	A	1	13	1	0	**57**	1	37	1	10
7	S	**102**	2	0	0	0	0	6	3	1
8	A	3	**76**	1	3	0	1	26	1	9
9	S	0	**84**	0	16	2	2	6	2	2
10	S	**96**	0	9	3	2	0	0	1	0
11	S	2	10	48	**47**	0	3	0	1	0
12	C	0	**93**	3	0	0	0	14	3	4
13	A	3	0	**66**	18	2	0	21	2	7
14	S	23	0	0	**48**	0	1	38	3	10
15	S	0	1	**30**	17	26	2	34	2	15
16$^{\Psi}$	C	-	**97**	-	-	-	0	13	-	11
17	S	**97**	6	0	1	0	0	6	2	1
18	A	0	0	2	**91**	0	0	17	3	10
19$^{\Psi}$	C	-	**68**	-	-	-	0	42	-	2
20$^{\Psi}$	C	0	0	**104**	-	-	1	5	-	5
21$^{\Psi}$	C	-	-	-	**87**	-	1	12	-	8
22$^{\Psi}$	C	-	-	-	-	**94**	1	15	-	10
23	A	0	0	0	0	**91**	1	18	4	5
24	A	2	1	**43**	15	0	27	22	1	15
25	C	**106**	0	0	0	0	0	4	4	4
26	S	3	**92**	3	2	10	0	0	0	0
27	S	43	7	0	0	**28**	9	23	2	9
28	A	**100**	2	1	3	1	2	1	0	1
29	C	4	0	0	1	**74**	10	21	2	13
30$^{\Psi}$	C	-	**71**	-	-	-	0	39	-	16
31$^{\Psi}$	A	-	-	-	-	**98**	0	12	-	7
32$^{\Psi}$	A	-	-	**71**	-	-	0	39	-	9
33$^{\Psi}$	C	-	**60**	-	-	-	7	43	-	19
34	A	0	8	**97**	4	1	0	0	1	0
35	C	2	4	0	**88**	7	0	9	1	9
36	S	**101**	7	0	1	0	1	0	2	0
37	A	0	0	0	7	**58**	11	34	3	11
38	S	0	**108**	0	0	0	1	1	4	1
39	A	0	0	5	6	**91**	2	6	2	4
40	C	2	0	4	**61**	0	15	28	2	10
41	S	0	0	0	**106**	3	1	0	3	0
42	A	**91**	1	0	0	0	1	17	3	11
43	C	0	2	**86**	0	0	0	22	3	11

* *There were 110 participants in this study; therefore, the numbers in columns A-X + "other answer" add up to 110 (e.g. for question 43, at the bottom of this table, answers A-X = 2 + 86 = 88, plus 22 other answers + 110). The number given in "unused option" refers to the number of A-E options for MINTv1 that were not provided as answers by participants in this study (constructed response rather than SBA). The number of new answer options is shown in the column "New answer"; this number is the same or lower than that in the column "other answer", reflecting the fact that several participants may have provided the same incorrect answer (e.g. for question 43, at the bottom of this table, 22 participants selected "other answers", but the number of new answer options was 11).*

$^{\Psi}$ *These are new questions, therefore the correct answer (A-E) is recorded, all other answers are recorded as "other" since there are no distractors from MINTv1.*

Table 2.14. New distractors for MINTv3

New answers	No. questions	Question numbers
New question: all new answers	9	16, 19 – 22, 30 - 33
No new distractors	9	2, 3, 10, 11, 26, 28, 34, 38, 41
One new distractor	4	1, 7, 17, 36,
Two new distractors	7	9, 12, 13, 15, 27, 29, 39
Three new distractors	9	5, 6, 14, 18, 24, 35, 40, 42, 43
Four new distractors	5	4, 8, 23, 25, 37

3.5 EVALUATION STUDY: MINTv3

I recruited a cohort of 111 third year medical students to sit MINTv3;112 students sat the formative examination, of whom one student declined to participate in the study. There were 63 (57%) female and 45 (40%) male students; 3 (3%) students did not indicate their gender. 13/111 (12%) students stated that they were given extra time in university examinations due to dyslexia. Scores for MINTv3 and MINTv2 are shown in Table 2.15.

The overall facility of the test was 0.77, overall item discrimination was 0.25, and internal consistency reliability as measured by Cronbach's alpha was 0.77. Psychometric data relating to MINTv3 is shown in Table 2.16. I compared psychometric data with that of MINTv1 and MINTv2. Test scores for all three tests are shown in Table 2.17; a comparison of the facility of test questions in MINTv3 and MINTv1 is shown in Table 2.18, and in MINTv3 and MINTv2 is shown in Table 2.19.

Table 2.15 Comparison of scores for MINTv3 and MINTv2

	MINTv3	MINTv2
Participants	Medical students	Medical students
No.	111	110
Test format	SBA	Constructed response
Range	22 - 43	19 - 43
Interquartile range	30 - 37	30 - 35
Mean (SD)	33.2 (4.9)	31.8 (5.23)
Median	33	33
Mode	31, 39	30, 33
Dyslexia	13	12

Table 2.16. Psychometric data for MINTv3

Q. no.	**SBA Answer**						**Facility**	**ID index***	**Cronbach's α if question removed**
	A	**B**	**C**	**D**	**E**	**X**			
Overall							**.77**	**.25**	**.77**
1	0	0	4	0	**107**	0	.96	.04	.77
2	**109**	0	0	0	2	0	.98	-.03	.77
3	57	**53**	1	0	0	0	.48	.63	.76
4	0	2	3	**106**	0	0	.95	.01	.78
5	5	20	**79**	5	2	0	.71	.15	.78
6	14	31	2	11	**53**	0	.48	.20	.78
7	**108**	2	0	0	1	0	.97	0	.77
8	6	**75**	24	4	2	0	.68	.29	.77
9	22	**81**	6	2	0	0	.73	.49	.76
10	**86**	0	4	21	0	0	.77	.45	.76
11	5	6	43	**54**	3	0	.49	.46	.76
12	2	**102**	4	2	1	0	.92	.21	.77
13	4	16	**72**	15	4	0	.65	.09	.78
14	28	10	22	**44**	7	0	.40	.77	.75
15	41	7	**35**	20	8	0	.32	.55	.76
16	0	**108**	1	2	0	0	.97	.07	.77
17	**103**	4	0	4	0	0	.93	.11	.77
18	9	5	6	**87**	4	0	.78	.18	.77
19	29	**77**	4	1	0	0	.69	.49	.76
20	0	5	**105**	0	1	0	.95	.11	.77
21	3	0	7	**101**	0	0	.91	.18	.77
22	0	6	2	2	**101**	0	.91	.14	.77
23	7	4	5	9	**86**	0	.77	.35	.77
24	11	11	**60**	12	17	0	.54	.53	.76
25	**110**	0	0	0	1	0	.99	0	.77
26	4	**94**	1	10	2	0	.85	.25	.77
27	54	1	11	14	31	0	.28	.58	.76
28	**109**	0	1	1	0	0	.98	0	.77
29	9	3	13	20	**66**	0	.59	.62	.76
30	11	**92**	1	1	6	0	.83	.22	.77
31	8	0	2	0	**100**	1	.90	-.09	.78
32	11	6	**83**	0	11	0	.75	.35	.77
33	4	**100**	5	2	0	0	.90	.18	.77
34	2	14	90	4	1	0	.82	.15	.77
35	4	10	2	**95**	0	0	.86	.22	.77
36	**98**	7	2	3	1	0	.88	.28	.76
37	7	13	.27	10	53	1	.48	.53	.76
38	0	**105**	4	1	1	0	.95	.07	.77
39	3	13	3	0	**92**	0	.83	.32	.77
40	5	21	8	**63**	12	2	.57	.56	.76
41	1	6	2	**101**	1	0	.91	.04	.77
42	**97**	5	0	6	3	0	.87	.22	.77
43	1	2	**103**	0	5	0	.93	0	.77

**Item discrimination index*

3.5.1 ***Comparison of test scores***

The mean and median scores are similar for all tests; however, the spread of scores, and thus standard deviation, was greater in MINTv1 (Table 2.17).

Table 2.17. MINT scores for each cohort tested*

Test	n	Format	Mean (%)	SD	Median	Range	IQR
MINTv1	135	SBA	32.8 (76%)	6.64	34	14 - 42	29 - 38
MINTv2	110	VSA	31.8 (74%)	5.23	33	19 - 43	29 – 35
MINTv3	111	SBA	33.2 (77%)	4.9	33	22 - 43	30 - 37

** Candidates did not have calculators for MINTv1; 57/110 had calculators for MINTv2; all had calculators for MINTv3.*

3.5.2 ***Comparison of facility of test questions***

The level of difficulty of test questions is measured by the facility, the fraction of participants who answer the question correctly. The overall facility of the test refers to the mean facility of test items; for MINTv1 this was 0.77, for MINTv2 it was 0.74, and for MINTv3, 0.77. Although the mean facility of the test was higher for the SBA versions (MINTv1 and MINTv3), this difference was not statistically significant. The magnitude of the difference, or its effect size is very small at only 0.03 (3%).

I compared the facility of individual test questions. Since this involved 43 comparisons, I applied the Bonferroni correction; thus any apparent difference was significant only if $p<0.0012$. Comparing MINTv3 and MINTv1, there was a significant difference in the facility of three test questions (Q. 5, 8, 40) (Table 2.18); in each case the facility was lower in MINTv3. Comparing MINTv3 and MINTv2, there was a significant difference in the facility of two test questions (Q. 30, 33); in both cases the facility was higher in MINTv3 (Table 2.19).

3.5.3 ***Comparison of internal consistency reliability***

I measured the Internal Consistency Reliability of the MINT using Cronbach's alpha. To assess the effect of individual questions on the reliability of the test, I calculated Cronbach's alpha if each question was removed (Table 2.16). Cronbach's alpha for MINTv3 was 0.77; there was little variation when any individual question was removed, with the value ranging between 0.75 and 0.78. This data is very similar to the Cronbach's alpha values for MINTv2 (overall value 0.76, range 0.76-0.78). However, these values are lower than those for MINTv1 (overall value 0.868, range 0.860-0.870).

3.5.4 ***Comparison of item discrimination***

I calculated item discrimination of test questions by evaluating the performance of the top and bottom cohorts of participants. I found that 12/43 questions had an item discrimination of >0.4, and a further 10/43 questions had a discrimination of > 0.2. These figures are almost identical to those for MINTv2, for which 13/43 questions had an item discrimination of >0.4, and a further 17/43 questions had a discrimination between 0.2 and 0.4. However, for MINTv1, no question a discrimination index above 0.4, and only 3/43 questions had a discrimination index of 0.2 or higher.

Table 2.18. Comparison of facility of test questions in MINTv3 and MINTv1

Question no.	Question Content	MINTv3 N=111	MINTv1 N=135	Diff	95% CI	χ^2	p value
1	Calculate, other	.96	.86	.10			ns
2	Pie chart	.98	.97	.01			ns
3	Drug comparison	.48	.61	-.13			ns
4	Calculate, other	.95	.87	.08			ns
5	Drug infusion	.71	.88	-.17	7-27	11	0.0009*
6	NICE guidance	.48	.56	-.08			ns
7	A/B RRR 1	.97	.91	.06			ns
8	A/B RRR 2	.68	.90	-.22	12-32	18	<0.0001*
9	% risk	.73	.67	.06			ns
10	Conversion	.77	.86	-.09			ns
11	Conversion	.49	.54	-.05			ns
12	Proportion	.92	.95	-.03			-
13	Calculate, formula	.65	.60	.05			ns
14	Screening data 1	.40	.60	-.20	7-32	9.7	ns 0.0018
15	Screening data 2	.32	.51	-.19	6-30	8.9	ns 0.0028
16	Calculate mean**	.97	.79	.18			-
17	A/B ARR 1	.93	.93	0			ns
18	A/B ARR 2	.78	.84	-.06			ns
19	NLQ 1**	.69	.85	-.16			-
20	NLQ 2**	.95	.76	.19			-
21	NLQ 3**	.91	.87	.04			-
22	NLQ 4**	.91	.92	-.01			-
23	Scattergram 1	.77	.90	-.13			ns
24	Scattergram 2	.54	.53	.01			ns
25	Calculate, other	.99	.93	.06			ns
26	Risk	.85	.70	.15	4-25	7.6	ns
27	Screening data	.28	.40	-.12			ns
28	Table & bar chart	.98	.92	.06			ns
29	Proportion	.59	.58	.01			ns
30	Drug dose**	.83	.85	-.02			-
31	Line graph 1**	.90	.81	.09			-
32	Line graph 2**	.75	.60	.15			-
33	Drug dose **	.90	.76	.14			-
34	Bar chart	.82	.67	.15			ns
35	Calculate, other	.86	.83	.03			ns
36	A/B NNT 1	.88	.84	.04			ns
37	A/B NNT 2	.48	.42	.06			ns
38	Risk	.95	.93	.02			ns
39	Table	.83	.64	.19			ns
40	Drug infusion	.57	.91	-.34	23-44	38	<0.0001*
41	Conversion	.91	.87	.04			ns
42	Bar chart	.87	.81	.06			ns
43	Drug dose	.93	.80	.13	4.4-21	8.4	ns

** Bonferroni correction applied, significance p<0.05/43; so, significant at 5% level if p< 0.0012*

*** New questions, so not a direct comparison*

Table 2.19. Comparison of facility of test questions in MINTv3 and MINTv2*

Question no.	Question Content	MINTv3 2019	MINTv2 2017	Diff	95% C.I.	χ^2	p
1	Calculate, other	.96	.92	4			ns
2	Pie chart	.98	1.0	2			ns
3	Drug comparison	.48	.58	10			ns
4	Calculate, other	.95	.85	10			ns
5	Drug infusion	.71	.70	1			ns
6	NICE guidance	.48	.52	4			ns
7	A/B RRR 1	.97	.93	4			ns
8	A/B RRR 2	.68	.69	1			ns
9	% risk	.73	.76	3			ns
10	Conversion	.77	.87	10			ns
11	Conversion	.49	.43	6			ns
12	Proportion	.92	.85	7			ns
13	Calculate, formula	.65	.60	5			ns
14	Screening data 1	.40	.44	4			ns
15	Screening data 2	.32	.27	5			ns
16	Calculate mean	.97	.88	9			ns
17	A/B ARR 1	.93	.88	5			ns
18	A/B ARR 2	.78	.83	5			ns
19	NLQ 1	.69	.62	7			ns
20	NLQ 2	.95	.95	0			ns
21	NLQ 3	.91	.79	12	2.6-21	6.2	ns
22	NLQ 4	.91	.85	6			ns
23	Scattergram 1	.77	.83	6			ns
24	Scattergram 2	.54	.39	15	1.9-27	4.9	ns
25	Calculate %	.99	.96	3			ns
26	Risk	.85	.84	1			ns
27	Screening data	.28	.25	3			ns
28	Table & bar chart	.98	.91	7			ns
29	Proportion	.59	.67	8			ns
30	Drug dose	.83	.65	18	6.5-29	9.3	0.0023
31	Line graph 1	.90	.89	1			ns
32	Line graph 2	.75	.64	11	-1-22	3	ns
33	Drug dose	.90	.54	36	24-46	35	<0.0001
34	Bar chart	.82	.88	6			ns
35	Calculate, other	.86	.80	6			ns
36	A/B NNT 1	.88	.92	4			ns
37	A/B NNT 2	.48	.53	5			ns
38	Risk	.95	.98	3			ns
39	Table	.83	.83	0			ns
40	Drug infusion	.57	.55	2			ns
41	Conversion	.91	.96	5			ns
42	Bar chart	.87	.83	4			ns
43	Drug dose	.93	.78	15	6-24	10	ns

**Bonferroni correction applied, significance p<0.05/43; significant at 5% level if p< 0.0012*

SECTION 4. DISCUSSION

The revision of the original MINT paper (MINTv1) was a complex process with seven aims: 1) to replace copyrighted questions; 2) to ensure that questions were clearly written and that data displays were of sufficient quality to allow accurate assessment; 3) to identify and correct any clinical information that was inaccurate, misleading or obsolete; 4) to review various frameworks for health numeracy and decide which was most suitable for analysing the MINT; 5) to subject the revised test (MINTv2) to external review; 6) to conduct an evaluation study of MINTv2 and 7) to develop and assess distractors for the SBA version of the test paper, MINTv3. Each of these aims is discussed in detail below.

Aim 1: Replacement of questions that were subject to copyright

Seven questions in MINTv1 needed to be replaced because they were subject to copyright: four nutritional label questions taken from the Newest Vital Sign (NVS) test (Weiss *et al* 2005), and three questions based on the metabolism of IV drugs adapted from the PISA test (OECD, 2009).

Performance of medical students and doctors on a nutritional label test has not previously been researched. This is an interesting area to study: questions based on the interpretation of nutritional labels are used as a screening test for health literacy in the general public (Weiss *et al* 2005; Rowlands *et al* 2013), and so it is possible that similar questions could be used to screen healthcare professionals for CN. Performance on the NVS questions in MINTv1 was lower than expected, and I considered that this might have been due to the American terminology of the NVS. However, results with the new nutritional label questions (NLQs) used in MINTv2 were similar to results for MINTv1, with 47% and 56% respectively answering all four questions correctly (Table 2.11), suggesting that the American terminology did not affect performance.

I did not find convincing evidence to support the use of the NLQs as a screening tool for CN, even though these questions are not difficult, and only 47% of participants answered all four correctly. Although the facility of the first NLQ was relatively low at 0.62, the facility of the remaining three was above 0.75. My data analysis suggests that the low facility of the first question was due to a careless error by participants. The question was *“If Mrs Doyle drinks half of the carton, how many calories will she consume?”* Forty-two participants answered incorrectly, of whom 41/42 gave an answer corresponding to the calories in half a serving of the drink, rather than half of the carton; this suggests that they had not read the question carefully. Overall, 90/110 participants (82%) answered three or four NLQs correctly, hence the careless error on the first question - rather than poor numeracy - may be responsible for the apparently poor performance on the NLQs overall. Although there was a strong correlation between performance on the NLQs and mean MINTv2 score, the range of scores associated with answering different numbers of NLQs correctly was broad (Table 2.11). Five of the nine students who answered only one NLQ correctly were in the lowest quartile for MINT score, while four were in the interquartile range. Thus, evidence for using NLQs to screen for CN is inconclusive; however, the number of participants in our studies is relatively small (135 in

MINTv1 and 110 in MINTv2), therefore further research is needed to establish whether a short test based on NLQs could be used as an initial assessment of clinician numeracy.

Three new questions on drug metabolism replaced questions taken from the PISA test for 15-year olds (OECD, 2009). Two were based on the interpretation of a line graph, and their facility was similar in both tests (MINTv1, 0.81 and 0.60; MINTv2 0.89 and 0.64 respectively). However, the facility of the third new question, a complex calculation, was significantly lower than that of the original - 0.54 compared to 0.76. Since the new question, although similar to the original, requires four steps rather than three, the difference in facility may be related to the additional complexity. However, the difference may also relate to the difference in test format, as the SBA format of MINTv1 permits cueing and guessing.

Aim 2: Clarity of test questions

I modified 23/43 questions to improve their clarity; nonetheless, test results suggest that this did not affect the outcome of any questions. The set of Treatment A/Treatment B questions devised by Sheridan & Pignone (2002), includes three simple comparisons with data presented as ARR, RRR, and NNT (Section 2.1.4.2.2). Although these questions were deemed easy by the external reviewers, and by using the criteria suggested by Close *et al* (2008), 22% of medical students were unable to answer all three questions correctly. This result was similar to the performance of Foundation Trainees in MINTv1 (Table 2.12), confirming that a large proportion of medical students and doctors appear to have difficulty with questions based on the interpretation of uncomplicated data. This is both intriguing, and a matter of concern, since doctors are required to make considerably more complex treatment comparisons in clinical practice (GMC, 2018). This finding merits further research.

I altered a question on screening mammography devised by Peters *et al* (2007) substantially. I changed the clinical context from the specific “mammogram” to a generic, hypothetical, “cancer screening”. Furthermore, I standardised all terminology to “cancer”, removing alternative terms such as “malignant tumour” and “malignancy”. Nevertheless, this remained one of the most difficult questions in the test, and its facility dropped from 0.4 to 0.25, although this was not statistically significant. Since the changes in wording did not affect facility, the variation in terminology in MINTv1 was probably unimportant; similarly, the potential confounding effect of clinical knowledge on answering this question, discussed in section 2.1.4.3, does not appear to be an issue in practice. Therefore, poor performance on this question is likely to be related to its content.

I modified the text of four questions (Q. 1, 3, 5, 43) by changing the order of their wording, in an attempt to make them clearer (Table 2.3). The facility of three of these questions (Q. 1, 3, 43) was similar in MINTv2 to that of MINTv1, probably because the changes were minor; however, the facility of one question (Q. 5) was lower following revision (0.70 in MINTv2 compared to 0.88 in MINTv1) (Table 2.9). It is possible, but unlikely, that the minor amendments to the text made the question less clear; however, it is more likely that the difference in facility is related to test format, since the SBA format of MINTv1 allows guessing and cueing. Nonetheless, the difference in test participants may be the most important factor: the question relates to an IV drug infusion, with which the Foundation Trainees who sat

MINTv1 would be very familiar, unlike third-year medical students. This hypothesis is supported by research demonstrating that a lack of familiarity with the clinical environment and equipment used in preparation of medicines made drug dose calculation challenging for nursing students (Weeks *et al* 2001; Johnson & Johnson, 2002; Wright, 2005; Weeks *et al* 2013). Furthermore, my subsequent research suggests that the difference in performance that I observed is related to participants/clinical experience: the facility of this question in MINTv3 was almost identical to that of MINTv2 (0.71 and 0.70 respectively) (Table 2.19).

Two questions (Q. 14 and 15) taken from Sikorskii *et al* (2011) relate to pre-operative screening; I modified the text slightly so that all numbers were written in digital form. Although I did not anticipate that this minor change would affect performance, I expected the facility of these questions to fall due to the change in test format. I had been unhappy with the distractors used in MINTv1 because they were numerically diverse; thus I considered that cueing/guessing would allow participants to select the correct answers. As expected, the facility of both questions was lower in MINTv2 than in MINTv1; this was statistically significant for question 15 (Table 2.9). However, my later study with MINTv3 showed that the facility of both questions was similar in MINTv2 and MINTv3 (Table 2.19); this result supports my hypothesis that the poor distractors in the original test allowed participants to rule out several options, and that cueing/guessing played a role in selecting the correct answers.

I made minor amendments to the text of four further questions in an attempt to improve their clarity (Q. 25, 35, 38, 42). While the questions may have been clearer, the amendments had no impact on facility (Table 2.9).

Aim 3: Questions amended for clinical reasons

Three questions used in MINTv1 had become clinically redundant by 2016. Two of these questions (Q. 16, 30) were replaced with new material, while the third (Q. 12) simply required a minor amendment. The facility of all three questions was unchanged in MINTv2 (Table 2.9).

Changes in clinical practice meant that a drug dose calculation question related to the drug ‘Actrapid’ in MINTv1 had become obsolete; this was replaced with a question based on the drug ‘Naloxone’ in MINTv2 (Q. 30). The new question was more difficult, with a facility of 0.65 compared to a facility of 0.85 for the original question, and this was statistically significant. The new question was slightly more complex, comprising a three-stage rather than a two-stage calculation; thus the reduction in facility is not unexpected. However, it is possible that the differences in participants and/or test format may also be contributory factors. Analysis of MINTv3 data demonstrates that the facility of this question increased from 0.65 to 0.83; this was statistically significant. The participants in both tests are medical students, although the test format is different; moreover, all participants in MINTv3 had calculators, compared to only half of those in MINTv2. The impact of calculators is discussed in Chapter 3, and of test format in Chapter 4.

Data analysis suggested that the clinical content of a question based on a pie chart (Q. 2) should be changed to prevent participants from answering based on their own experience. I changed the context of this question from bleep calls for different (real) clinical tasks to bleep calls to hypothetical wards. Although the facility of this question increased from 0.97 in MINTv1, to 1.0 in MINTv2 (Table 2.9), this difference does not reach statistical significance. It is possible that the change in context may have been unnecessary given the difference in participants: third-year medical students do not carry bleeps, and so would not have been biased by their clinical experience. Interestingly, the facility of this question was 0.98 in MINTv3 (Table 2.19).

Finally, I altered the clinical context of two questions (Q. 10, 41) to reduce the number of cancer scenarios in MINTv2. Both new questions are otherwise identical to the originals; unsurprisingly, the change in clinical setting did not affect facility.

Performance of unrevised questions

Three questions had such minimal amendments that they are considered here with unrevised material. I altered the timelines of two graphs involving three questions (Q. 23, 24, 34) so that the test would not appear dated. Since the revision was negligible, I did not expect a change in facility. However, the facility of one question (Q.34) increased, reaching statistical significance (Table 2.9). This question was rated fairly easy by the external reviewers, and the facility of 0.88 in MINTv2 seems more appropriate than that of 0.67 in MINTv1. Thus, the difference in facility is more likely to reflect poor performance in MINTv1, although the reason for this is uncertain. Data analysis of MINTv3 showed that this question had similar facility (0.82) to that of MINTv2 (Table 2.19).

Eleven questions were completely unaltered, and for eight of these the facility in MINTv2 was similar to that in MINTv1 (Q. 4, 6, 9, 11, 13, 26, 28, 29) (Table 2.9). However, there was a statistically significant difference in the facility of two of the unrevised questions. The facility of a question based on a table showing data on binge drinking (Q. 39) was higher in MINTv2 (0.83), than in MINTv1 (0.64), although I can find no plausible reason for this difference. The second unaltered question for which performance was different is based on the infusion of an intravenous drug (Q. 40): the facility of this question was 0.55 in MINTv2, compared to 0.91 in MINTv1. This difference in performance may be due to the difference in test format; however, as with Q. 5 discussed above, clinical experience may have given foundation trainees an advantage over third-year medical students. Results with MINTv3 suggest that the difference in participants is important; the facility of this question remained low (0.57) in MINTv3 (Table 2.19).

Aim 4: Review of Numeracy frameworks

I had categorised MINTv1 questions using the framework for health numeracy (HN) devised by Golbeck *et al* (2005) (Table 2.1); however, due to the overlap between numeracy constructs, some questions were difficult to classify. Therefore, I reviewed some other HN/CN frameworks, to assess whether a better classification system might be available. The HN frameworks developed by Nutbeam (2000) and Ancker & Kaufman (2007) were both unsuitable; however, that of Schapira *et al* (2008) looked promising, and had been used by Caverly *et al* (2012), when developing their CN test for doctors (Table 2.4). Nonetheless, this framework was also unsuitable for classifying MINT items, and offered no advantage over the Golbeck *et al* (2005) model.

Aim 5: External review

Numeracy constructs

External reviewers were asked to classify MINTv2 questions as primarily computational, analytical or statistical in construct using the Golbeck *et al* (2005) framework. Not surprisingly, due to the overlap between constructs, there was some variation between external reviewers regarding the principal construct of each question, and they were unanimous for only 7/43 questions. This demonstrates the subjective nature of assessing test content, even when clear definitions are provided, and raises questions about the value of using a framework to classify MINT questions.

Level of difficulty

MINT questions deliberately vary in their level of difficulty to allow assessment of an individual's ability. The majority of questions (31/43) were based on questions originally developed for various non-medical populations, including schoolchildren, the general public, and university students, with 12/43 new questions written for medical students/trainee doctors (Taylor & Byrne-Davis 2016). For MINTv1, I had categorised the level of difficulty of these questions on a 5-point scale according to intended participants e.g. questions developed for primary schoolchildren were rated easy (level 1), while those aimed at university students were rated difficult (level 5). An exception to this was the set of 18 questions developed for medical students and doctors, which varied greatly in difficulty, and thus were reviewed separately and then rated between 1-5. This rating strategy had seemed appropriate and worked well: data analysis with MINTv1 showed a strong inverse correlation between facility and allocated level of difficulty (r = -0.751, p<0.01) (Taylor & Byrne-Davis, 2016). However, there were some discrepancies, and according to this system, only 4/43 questions were rated as easy, while 12/43 were rated difficult: this did not seem to be an accurate reflection of the test, either in our opinion, or that of participants in the original study (Taylor, 2014). Therefore, a review by independent experts was warranted. However, there was little consensus among the reviewers regarding the level of difficulty of test questions, and for 17/43 (39.5%) questions, the range of difficulty varied across four or five levels. Nonetheless, analysis of the mean and median ratings given by reviewers led to consensus on level of difficulty. Furthermore, the process showed that the reviewers did not consider the test to be difficult: they rated 22/43 (51%) questions as either easy or fairly easy, 15/43 (35%) as average, and the remaining 6/43 (14%) as either fairly difficult or difficult. This spread of ratings seemed more appropriate for the MINT than that achieved with the original method. Furthermore, this process showed that the level of difficulty of the new questions developed for MINTv2 was comparable to that of the originals.

The variation in reviewers' ratings made me reconsider the value of assigning level of difficulty in advance, and I wondered whether a retrospective process based on item facility might be more appropriate. Review of the literature revealed that Close *et al* (2008) had used a retrospective analysis to rate maths questions, grading difficulty according to facility as follows: Easy (>80% correct); Moderately easy (70 – 79%); Average (50 – 69%); Moderately difficult (40 – 49%) and Difficult (<40%). Applying these criteria to MINTv2, I found that 23/43 (53%) questions would be classified as easy; 4/43 (9%) as moderately easy; 11/43 (25%) as average; 2/43 (5%) as moderately difficult and 3/43 (7%) as difficult. Interestingly, this is very similar to the classification based on reviewers' ratings; hence for future research using new test questions, I will use the Close *et al* (2008) criteria to classify questions, rather than needing to convene a panel of expert reviewers. Moreover, this analysis confirms my opinion that the MINT is not a difficult test.

Aim 6: Evaluation study

I recruited 110 medical students to participate in the evaluation of MINTv2. I found no difference in their test scores compared to those of foundation trainees (FTs) who sat the MINTv1. This indicates that the two tests are equivalent, and supports the validity and reliability of the MINT as an assessment of CN. Furthermore, the similarity of results suggests that the new material developed for MINTv2 is appropriate, and performs as well as the original questions used in MINTv1.

However, in assessing these results, there are three important differences between MINTv1 and MINTv2 that should be taken into consideration, as discussed in Study Limitations (section 2.3.2, p.55). Firstly, the FTs who participated in MINTv1 had more clinical experience than the medical students who sat MINTv2; as discussed above, this may have had an impact on performance in some test questions. Secondly, MINTv1 was an SBA multiple choice test, while MINTv2 was a CR test; therefore, participants in MINTv1 may have selected some correct answers based on cueing or guessing. Finally, all MINTv2 participants were also taking part in a randomised controlled trial (RCT) to assess the impact of calculators. Although there was no difference between participants in MINTv1 and MINTv2 in terms of overall test scores, I found that there was a significant difference in performance on 10/43 questions; however, four of these were new questions, thus the comparison is only of value in assessing their equivalence to the originals. Of the remaining six questions, the facility of four was lower in MINTv2, suggesting that access to calculators did not confer an advantage. The RCT is fully discussed in chapter 3.

Aim 7: Development and assessment of MINTv3

The questions used in MINTv3 are identical to those of MINTv2, the only difference is in test format: MINTv3 is an SBA test. I created new distractors for 25/34 of the original MINTv1 questions used in MINTv2 and MINTv3 (Table 2.14). These distractors were based on analysis of the incorrect answers provided by medical students who participated in the evaluation study of MINTv2.

Analysis of data relating to the two SBA formats of the test is important, as it eliminates the possible confounding effect of test format affecting the comparison of MINTv2 with MINTv1. Nevertheless, two important potential confounding factors remain: participants in MINTv3 are medical students rather than FTs; and all participants in MINTv3 had calculators, while those in MINTv1 did not. Comparison of the facility of the 34 questions from MINTv1 that were used in MINTv3 shows a significant difference in relation to three questions; two of these (Q.5 & Q.40) relate to IV drug infusions, thus the difference is likely to relate to the difference in clinical experience of participants. The third question relates to the calculation of risk when comparing treatments with information given in relative risk reduction (RRR) format (Q.8); the facility of this question was 0.68 in MINTv2, compared to 0.90 in MINTv1. This question is not difficult; furthermore, it is one of a set of treatment comparison questions, and is identical to Q.18 except that the information in Q.18 is given as absolute risk reduction (ARR). Facility of Q.18 is the same for both groups (0.83 in MINTv2, 0.84 in MINTv1); therefore, the observed difference in Q.8 is evidence of a framing effect, i.e. a difference related to the way in which the data is presented. Framing is an important phenomenon in medicine, and has been shown to influence decision making by doctors (Perneger & Agoritsas, 2011).

Evaluation of data comparing MINTv3 with MINTv2 is also important, since this eliminates bias relating to participants; however, potential confounding effects include the difference in test format, and the impact of calculators. The facility of 41/43 questions was similar in MINTv3 and MINTv2; the two questions where there was a significant difference were complex calculations, and in both cases the facility was higher in MINTv3. This difference may be explained by cueing/guessing in the SBA format, or by the availability of calculators. The latter is discussed fully in Chapter 3, and the former in Chapter 4.

Psychometric analysis of the revised tests

Analysis of psychometric data showed that the overall facility of MINTv3 (0.77) was similar to that of both MINTv1 (0.77) and MINTv2 (0.74); interestingly the facility of both SBA versions of the test is exactly the same, the slightly lower value for MINTv2 may be related to its constructed response (CR) format. This will be explored in Chapter 4.

Item discrimination compares the performance of the top and bottom group of participants on each question, and is calculated by comparing the facility of the question for the upper and lower 27% of students; the facility of the bottom group is subtracted from that of the top group, and the range of possible values is from -1 to 1 (University of Oxford Medical Sciences Division, n.d.). Although it is generally agreed that questions should have a discrimination index of at least 0.2, items with a facility above 0.9 or below 0.3 tend not to discriminate well (DiBattista & Kurzawa, 2011). The overall discrimination index of MINTv3 (0.25) was lower than that of MINTv2 (0.29); this may be related to test format. However, both tests were much more discriminating than the MINTv1, for which the discrimination index was 0.10. The greater discrimination of the revised test is likely to be due to the elimination of cueing and guessing in MINTv2, and the use of evidence-based distractors in MINTv3.

Finally, the internal consistency reliability of the tests, as measured by Cronbach's alpha is similar for MINTv3 (0.77) and for MINTv2 (0.76). This is within the acceptable range (Tavakol & Dennick, 2011), although lower than that of MINTv1 (0.87). Participants in the revised tests were a more homogenous group (cohorts of third year students from a single medical school) than participants in MINTv1 (FTs from a diverse range of backgrounds). This may explain the narrower spread of test scores, and lower standard deviation for the revised tests; this in turn will result in a lower value of Cronbach's alpha).

SECTION 5. CONCLUSION

The revision of MINTv1 leading to the development of a CR version of the test, MINTv2, and an SBA version, MINTv3 was a comprehensive process. In addition to my own evaluation of questions, I recruited nine external reviewers to assess test content, advising on the quality of text and data displays, the principal numeracy construct, and the level of difficulty of individual questions. This analysis indicates that the MINT is not a difficult test, with over 50% questions rated as easy or fairly easy; furthermore, there is a fairly even spread of questions across the three constructs (computational, analytical and statistical numeracy). The first revised test, MINTv2 consists of nine new questions in addition to 34 original MINTv1 questions; however, 23/34 of the remaining MINTv1questions have been amended in some way.

I recruited a cohort of third-year medical students to participate in an evaluation study of MINTv2. Following the evaluation study of MINTv2, I analysed the incorrect answers given by participants, using them to develop evidence-based distractors for the SBA version of the test, MINTv3. I then conducted a second evaluation study, this time to assess MINTv3.

Overall test scores of medical students who sat MINTv2 and MINTv3 were almost identical to those of FTs who sat MINTv1, demonstrating that CN in medical students is at a similar level to that of FTs, and indicating that some medical students have low CN.

Psychometric data from MINTv3 was very similar to that of MINTv2. Additionally, data analysis indicated that both tests were comparable to MINTv1 in terms of overall facility and the facility of individual test questions. This confirmed that the new test material was equivalent to the original, and allayed any concerns that test questions in MINTv1 might have been confusing or misleading. The overall discrimination index of both MINTv2 and MINTv3, and that of individual questions in both tests, was significantly higher than that of MINTv1; this highlights the effectiveness of using evidence-based distractors for the SBA version of the test.

The revision process has confirmed the quality of the MINT as an assessment measure of CN for medical students and doctors. Furthermore, all copyrighted questions have been replaced. Therefore, it is suitable to use for my ongoing research into CN.

Blank Page

CHAPTER 3

THE IMPACT OF CALCULATORS ON A TEST OF CLINICIAN NUMERACY: A RANDOMISED CONTROLLED TRIAL

I have submitted this study to the journal *Numeracy*. The manuscript was reviewed by three reviewers, who suggested some revisions prior to publication. The revised manuscript has been accepted for publication in the July 2019 edition of the journal. This chapter is identical in content to the manuscript in press; however, for consistency, it is presented in the same style as the rest of the thesis.

The results of this study have also been presented at various conferences:

1. Developing Excellence in Medical Education Conference (DEMEC), November 2017
2. National Association of Clinical Tutors (NACT) Walsall, February 2018
3. Association for the Study of Medical Education (ASME) Workshop, Nottingham, April 2018
4. Annual Medical Education Conference, Keele, April 2018
5. Health Education England Educators Conference, Birmingham, November 2018.
6. Grand Rounds, University of Manchester, February 2019

ABSTRACT

Clinician numeracy (CN), the ability to use and understand quantitative data in patient care, is an important skill for healthcare professionals. Nonetheless, it is recognised that many healthcare professionals, including doctors, have deficiencies in CN, and that this may affect patient safety. In our previous research using the Medical Interpretation and Numeracy Test (MINT), we found that many doctors in training in the UK had low CN. However, participants were not permitted to use calculators when taking the MINT, even though staff have access to calculators in clinical practice. Therefore, our original study may have underestimated doctors' CN, compared to their ability in clinical practice. Thus, we designed a randomised controlled trial to assess the impact of calculators on MINT score. We recruited 110 third-year medical students to participate in this study. Our results show that having access to a calculator had no impact on test scores. We consider that this is due to two factors: first, CN is a complex construct that involves problem-solving and analysis, skills that are not improved by using calculators; second, the lack of impact of calculators suggests that the errors made by participants in our study are predominantly errors of understanding rather than mathematical errors. We suggest that participants taking CN tests should have access to calculators as they would do in the workplace. We recognise that further research in this area is needed, but suggest that educational interventions to improve CN should primarily be directed at improving understanding rather than mathematical skills.

INTRODUCTION

Clinician numeracy (CN) is the ability of healthcare professionals "to use numbers and numeric concepts in the context of taking care of patients" (Caverly *et al* 2012). CN is important across the spectrum of clinical work for doctors, from routine tasks such as calculating drug doses to medical decision making; therefore, it is essential to patient safety (Lesar *et al*. 1997; Hughes & Edgerton 2005; Gigerenzer *et al* 2007; Coben & Weeks 2014; Williams and Walker 2014). However, there is evidence that many medical students and doctors have difficulty in calculating drug doses (Rowe *et al* 1998; Selbst *et al*. 1999; Wheeler *et al* 2004a; Wheeler *et al* 2004b; Simpson *et al* 2009; Harries *et al* 2013) and struggle to understand medical data underpinning clinical treatment options (Gigerenzer *et al* 2007; Windish *et al* 2007; Rao & Kanter 2010; Gigerenzer & Gray, 2011; Moyer, 2012; Johnson *et al* 2014; Malhotra *et al* 2015). This is important, since medication errors are common, and a significant cause of morbidity and mortality worldwide; it is estimated that there are approximately 240 million medication errors annually in the NHS in England (Elliott *et al* 2018), while adverse drug events are estimated to cost almost $20 billion annually in the US (da Silva & Krishnamurthy, 2016). The World Health Organisation (WHO) has launched a global challenge to reduce the incidence of medication-related harm by 50% over five years (WHO, 2017). Drug dose calculation errors are a cause of medication error, and although this area has been extensively researched in the nursing literature (Johnson & Johnson, 2002; Wright, 2004, 2005; Hutton *et al* 2010; McMullan *et al* 2010; Sabin *et al* 2013; McDonald *et al* 2013; Weeks *et al* 2013 a,b,c; Young *et al* 2013; Coben & Weeks, 2014; Fleming *et al* 2014; Bagnasco *et al* 2016), there has been little research on drug dose calculation skills in medical students and doctors, perhaps due to the assumption that entry to medical school assures good numeracy (Rowe *et al* 1998; Simpson *et al* 2009; Harries & Botha, 2013).

The assessment of CN in medical students would have many potential uses, including selection, formative assessment to identify areas of learning difficulty, and summative assessment for progression decisions; this would be particularly salient if low CN was associated with difficulties in clinical practice. In order to measure CN in medical students and doctors, we previously developed an assessment of CN, the Medical Interpretation and Numeracy Test (MINT). The MINT is a 43-item assessment with questions testing computational, analytical, and statistical numeracy (Taylor & Byrne-Davis 2016). Our research adds to the evidence demonstrating that medical students and doctors may have deficiencies in CN (Taylor & Byrne-Davis 2017). However, the participants in our study did not have access to calculators, so it is possible that our finding of low CN on the MINT might not translate to difficulties in clinical practice, where calculators are readily available.

Calculators are not helpful for all numeracy questions. Close *et al* (2008) classified numeracy questions as calculator appropriate (complex calculations), calculator optional (where it is unnecessary but not unreasonable to use a calculator), and calculator inappropriate (simple calculations that can be answered readily either mentally or with pen and paper). Questions that are important in determining overall CN, such as data interpretation questions, would also be classified as calculator inappropriate. In our previous research with the MINT, we considered that calculators would be unnecessary, as its content was largely calculator inappropriate. Calculators would not help with analytical questions, involving the interpretation of data presented in charts and graphs, or with statistical questions, testing clinical mathematical reasoning. Furthermore, most computational questions in the MINT were straightforward, and based on numbers that would be easy to manipulate either mentally or using pen and paper. On this basis, we originally classified only one of our 43 questions, a complex calculation, as calculator appropriate. However, we recognise that there may be a significant overlap between numeracy constructs, and that questions classified as primarily “analytical” or “statistical” may also have significant computational elements (Golbeck *et al* 2005). Therefore, our original classification of questions as calculator appropriate or not may have been inaccurate: many MINT questions, whether computational, analytical or statistical, involve multiple steps and calculations, and so could be considered to be either calculator optional or calculator appropriate. We reviewed our test material, classifying all 18 computational questions, along with three analytical and two statistical questions as either calculator appropriate or calculator optional; we considered the remaining 20/43 questions to be calculator inappropriate. Therefore, lack of access to calculators in our initial research with the MINT may have underestimated CN in doctors compared to the real-life clinical situation where calculators are readily available. If this is the case, our previous finding of low CN in doctors would be less relevant to clinical practice, and could also mean that the MINT had lower construct validity. Therefore, to test the hypothesis that using calculators would improve MINT scores, we conducted a randomized controlled trial of the effect of calculators on clinician numeracy, comparing MINT scores in medical students randomly allocated to having or not having a calculator.

METHODS

Study design

The study was a randomised controlled trial. Participants were randomly allocated into one of two groups: group C, who received calculators, and group N, who did not. Ethical approval for the study was obtained from the University of Manchester (UoM) Research Ethics Committee.

Participants

Participants were third year medical students studying at a single institution in England. The MINT was incorporated into the formative mid-year assessments for these students. All students in the year group were eligible for entry to the study. One month prior to the formative assessment, these students attended a teaching session on clinician numeracy and its importance for healthcare professionals and were given preliminary information about the study. Further information and an invitation to participate in the research were sent by email.

Interventions

The Medical Interpretation and Numeracy Test (MINT) is an assessment of clinician numeracy, consisting of 43 questions, testing computational, analytical and statistical constructs; it has high internal consistency reliability as measured by a Cronbach's alpha score of 0.868 (Taylor & Byrne-Davis, 2016). The MINT is available as a multiple choice test, and in a short answer (constructed response) format. For this study, we used the constructed response format.

Outcomes

The outcome measure was the mean score of participants in groups N and C. We also measured the facility of each test item for participants in groups N and C.

Sample size

In order to calculate the sample size required for the trial, we considered the previous mean and standard deviation of the MINT in similar participant groups. The mean MINT score achieved by participants in a previous study was 32.76/43 with a standard deviation of 6.64. We considered that a change of up to 2 marks might represent normal variation (a "good" or "bad" day for an individual), but that a change in score of 4 marks (almost 10%) would demonstrate that an intervention had had a positive effect. With a minimum difference to be detected of 4 marks, and a standard deviation of 6.64, a type 1 error rate of 0.05, and a type 2 error of 0.2, we calculated that 88 participants (44 in each group) would be required (online tool for sample size calculation: Brant, n.d.).

Randomisation

116 students were invited to participate in the study; therefore, a list of potential participants was made, with study identification (ID) numbers from 1 – 116. A table of random numbers was used to allocate the ID numbers 1-116 into study groups C (calculator) and N (no calculator). Test answer sheets were prepared, and recorded the study ID number and group allocation code "C" or "N".

Allocation concealment

The test answer sheets were placed in a brown A4 envelope, alongside the MINT paper, a pencil and an eraser. Basic pocket calculators were added to test envelopes for test papers coded "C". All envelopes were sealed. Since the calculators were small and flat, envelopes containing calculators appeared similar to those containing only a pencil and eraser. The study envelopes were randomly distributed on desks in the examination room, and participants were allowed to select their own seats. Therefore, neither the researcher nor the participants were aware of group allocation until the test commenced, and participants opened their envelopes.

Implementation

The test was carried out under examination conditions, with 90 minutes to complete the test. Once the test was completed, participants returned all test materials to the study envelopes. Participants were aware of the hypothesis that using a calculator would improve test score; those allocated to group N were given the option to request a calculator. When students opted to change their allocation, the coding on their answer sheets was changed accordingly, and this was reported.

Statistical Methods

Data were analysed in Microsoft EXCEL, and an online statistical tool (MedCalc Software bvba (BE) a, b). We described the distribution of scores for each group, and then used Student's t-test to compare the means of participants in the two main study groups (N and C); the primary analysis relates to the intent-to-treat group allocation. We also analysed data relating to the final (per-protocol) group allocations.

We assessed the magnitude of the difference associated with use of a calculator by calculating the effect size. Since the comparison is of mean test scores, it is more appropriate to calculate the absolute effect size rather than using an effect size index (Sullivan & Feinn, 2012).

Subgroup analyses

Participants were asked to indicate their gender, as there is evidence from a study investigating quantitative literacy in US university students that female gender may be associated with lower numeracy (Sikorskii *et al* 2011). Furthermore, we asked participants whether they had dyslexia because there is some overlap between dyslexia and dyscalculia (Gibson & Leinster 2011; British Dyslexia Association (2017). We recorded this data to ascertain whether these attributes were evenly distributed across groups, and if not, to ensure that any effects did not confound observed differences between calculator and non-calculator groups. We performed a logistic regression analysis to assess any apparent effect relating to these characteristics.

RESULTS

Participants

Of 116 third-year students, 110 (95%) consented to participate in the study. 52/110 (47%) of students were allocated to Group C (calculators), while 58/110 (53%) participants were allocated to Group N (no calculators). Five students who had been allocated to Group N requested calculators, and so were reassigned to Group C. Recruitment of participants and allocation to study groups is shown in Figure 3.1.

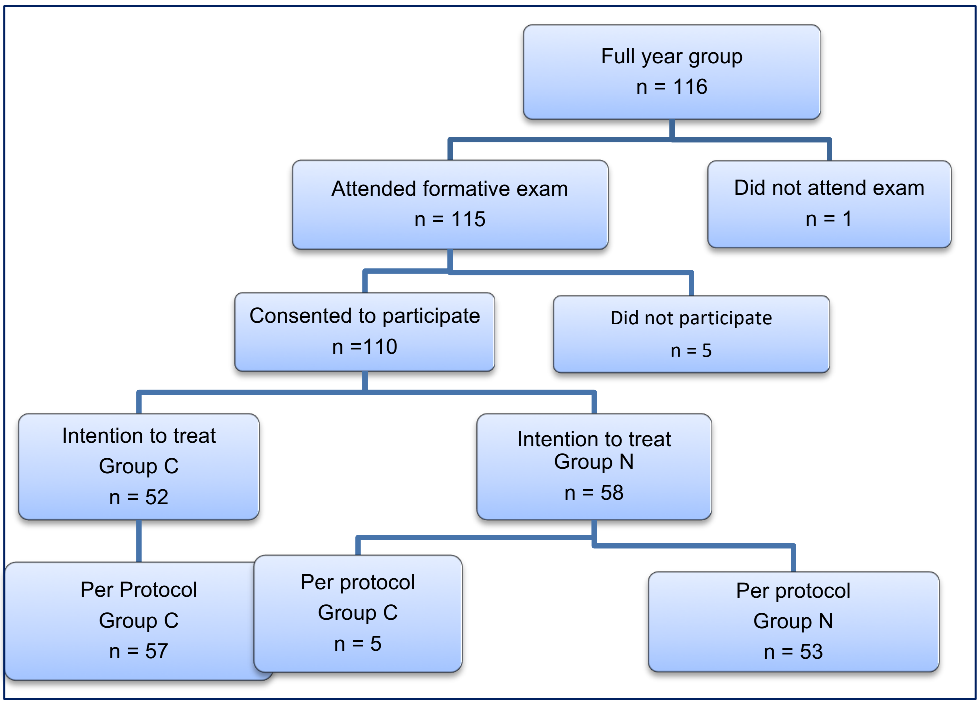


Figure 3.1. Recruitment and allocation of study participants

Demographic data

Of the 110 participants, 59 (54%) were female and 36 (33%) were male, and 15 (13%) did not declare their gender. Twelve students (11%) declared a diagnosis of dyslexia, 75 (68%) stated that they were not dyslexic, and 23 (21%) did not report their dyslexia status (Table 3.1). (The 23 students who did not comment on their dyslexia status includes all 15 who did not indicate their gender).

Table 3.1. Demographic data of study groups

	N	**Male n (%)**	**Female n (%)**	**Unknown Gender n (%)**	**Dyslexia n (%)**	**No dyslexia n (%)**	**Unknown dyslexia n (%)**
Total	110	36 (33%)	59 (53%)	15 (14%)	12 (11%)	75 (68%)	23 (21%)
Group C*	52	14 (27%)	30 (58%)	8 (15%)	4 (8%)	36 (69%)	12 (23%)
Group N*	58	22 (38%)	29 (50%)	7 (12%)	8 (14%)	39 (67%)	11 (19%)

**intent to treat*

Mean Scores

Test scores for all study groups are shown in Table 3.2, and includes scores for the full cohort of 110 participants, as well as scores of participants in different groups. Although the performance of all groups was similar, the mean scores of those who had calculators were higher than mean scores of those without calculators. However, statistical analysis using Student's t-test to compare the mean scores of participants in different groups indicated that the apparent difference in scores was not significant. The primary analysis is based on intention to treat, and thus represents participants whose original allocations were to groups C (n=52) and to group N (n=58). There was no difference in performance of participants in these groups as evidenced by statistical analysis (difference = 1.9; SE = 0.98; 95% CI = -0.05 – 3.8; t = 01.9; DF = 108; ns).

Table 3.1 also provides data relating to the per-protocol group allocations: five students allocated to group N requested calculators and so were re-allocated to group C. Statistical analysis shows no difference between these groups (difference = 1.4; SE = 1.0; 95% CI = -0.58 – 3.38; t = 1.39; DF = 108; ns).

There was no difference in performance of those whose original and final allocations were to group C (difference = 0.3; SE = 0.98; 95% CI = -1.66-2.25; t = 0.3; DF = 107; ns); or to group N (difference = 0.2; SE = 0.99; 95% CI = -1.77-2.2; t = 0.201; DF = 109; ns).

Table 3.2. Test score: all groups

	N	Mean (SD)	Median	Range	IQR
All	110	31.8 (5.2)	33	19-43	29-35
Group C (intent to treat)	52	32.8 (5.1)	33	19–43	30-36
Group N (intent to treat)	58	30.9 (5.2)	33	19-41	28-34
Group C (per-protocol)	57	32.5 (5.2)	33	19-43	29-36
Group N (per-protocol)	53	31.1 (5.3)	33	19-41	28-35

**interquartile range*

Effect Size

The absolute effect size is the difference in the mean scores of Groups N and C. Data was analysed using the intent-to-treat groups, thus absolute effect size was 1.9.

Subgroup Analyses

The mean score of male participants was 34.4/43, while that of females was 30.5/43; this is an effect size of 3.9/43. The mean score of participants with dyslexia was 29.3/43, and of those who were not dyslexic was 32.4/43; the effect size is 3.1/43. Logistic regression analysis indicated that the difference related to gender was significant (Table 3.3).

Table 3.3. Logistic regression analysis of subgroups

Independent variable	b	SE	T	Prob
Calculator	1.7	1.01	1.7	.091
Dyslexia	-3.2	1.638	-1.94	0.56
Gender	-3.9	1.045	-3.75	0.000

Facility of test items

In addition to analysing the mean scores of participants in groups C and N, we compared performance on individual MINT items, to assess whether use of a calculator conferred an advantage for individual questions. Raw data shows that facility was the same for 4/43 questions, was higher in Group C for 29/43 questions and higher in Group N for 10/43 questions (Table 3.4). We used the N-1 Chi-squared test to assess whether these differences were significant. Because this involved conducting 43 individual tests, it was necessary to apply the Bonferroni correction (Perneger, 1998); therefore, significance p<0.05/43, i.e. a difference was only significant at the 5% level if p< 0.001. We found a statistically significant difference in performance in 2/43 questions: in both cases, participants in group C performed better than those in group N. Both questions were computational.

Table 3.4. Facility of test items: Group C v Group N *(intent to treat)*

Q no	Primary	Facility	Group	Group	Diff$^{\Psi}$	95% CI	χ^2	Sig
	construct	n=110	C	N	%			
			n=52	n=58				
25	Computational	.96	.96	.97	-1			ns
20	Computational	.95	.96	.93	3			ns
1	Computational	.92	.96	.88	8			ns
16	Computational	.88	1.0	.78	22	11 - 34	12.6	0.0004*
4	Computational	.85	.90	.81	9			ns
22	Computational	.85	.88	.83	5			ns
12	Computational	.84	.88	.81	7			ns
35	Computational	.81	.87	.74	13			ns
21	Computational	.79	.79	.79	0			ns
43	Computational	.78	.83	.74	9			ns
5	Computational	.70	.65	.74	-9			ns
29	Computational	.67	.65	.69	-4			ns
30	Computational	.64	.77	.53	24	6-40	6.8	0.009**
19	Computational	.62	.62	.62	0			ns
40	Computational	.57	.56	.55	1			ns
33	Computational	.54	.73	.36	37	18 - 52	15	0.0001*
								ns
2	Analytical	1.0	1.0	1.0	0			ns
28	Analytical	.91	.92	.90	2			ns
31	Analytical	.89	.90	.88	2			ns
34	Analytical	.88	.90	.86	4			ns
42	Analytical	.84	.85	.81	4			ns
18	Analytical	.83	.85	.81	4			ns
23	Analytical	.83	.81	.84	-3			ns
39	Analytical	.83	.88	.78	10			ns
8	Analytical	.69	.69	.69	0			ns
32	Analytical	.64	.73	.55	18			ns
13	Analytical	.60	.63	.57	6			ns
37	Analytical	.54	.56	.50	6			ns
6	Analytical	.52	.48	.55	-7			ns
24	Analytical	.39	.42	.36	6			ns
								ns
38	Statistical	.98	.96	1.0	-4			ns
41	Statistical	.96	.96	.97	-1			ns
7	Statistical	.93	.94	.91	3			ns
36	Statistical	.93	.92	.91	1			ns
17	Statistical	.88	.90	.86	4			ns
10	Statistical	.87	.87	.88	-1			ns
26	Statistical	.84	.85	.83	2			ns
9	Statistical	.77	.83	.71	12			ns
3	Statistical	.58	.52	.64	-12			ns
11	Statistical	.44	.44	.41	3			ns
14	Statistical	.44	.42	.45	-3			ns
15	Statistical	.27	.31	.24	7			ns
27	Statistical	.25	.27	.24	3			ns

** p< 0.001, therefore, significant at 5% level when the Bonferroni correction is applied*

*** p> 0.001, therefore, not significant at 5% level when the Bonferroni correction is applied*

$^{\Psi}$ *a positive value in this column indicate that the facility was higher in Group C, while a negative value indicates that the score was higher in Group N*

DISCUSSION

We found that having a calculator did not affect overall scores on our test of clinician numeracy, the MINT. The mean score of Group C was slightly higher than that of Group N, although this difference was not statistically significant. The absolute effect size was 1.9, i.e. participants in Group C had a mean score of 1.9/43 (4%) higher than participants in Group N. We do not think that this is clinically important; the study was designed to detect an effect size of 4/43 (9%).

We found little difference in terms of performance on individual questions. We had considered that using a calculator might improve performance on the 16 computational questions in the test; in addition, we classified three analytical and three statistical questions as either calculator appropriate or calculator optional, thus there were 22/43 questions where having a calculator might prove beneficial. However, participants who had calculators performed better on only two questions: one of these was a complex calculation, the other simply required calculating the mean of four values.

Since research on the use of calculators in tests of CN is limited to tests of drug dose calculation in nursing, with small study samples, it is difficult to compare our results to the existing literature. However, the evidence from nursing studies is conflicting: some researchers found that using calculators improved performance (Shockley *et al* 1989; Bliss-Holtz, 1994), while others observed little or no impact (Murphy & Graveley, 1990; Tarnow & Werst, 2000). Interestingly, there is some debate in the nursing literature about whether to permit the use of calculators in drug dose calculation tests e.g. McMullan *et al* (2010) argue that calculators should not be allowed as they would constitute “a substitute for arithmetical knowledge and skills”. However, we consider that medical students, doctors, and other healthcare staff taking drug dose calculation tests and other tests of CN should be allowed to use calculators, since these are readily available in clinical practice. Furthermore, our results suggest that using calculators will not conceal evidence of low CN.

Our finding that using calculators did not have a positive impact on test scores supports the observation that CN is a complex construct that entails more than the ability to perform simple mathematical operations. This is highlighted by Coben & Weeks (2014), who note that numeracy in nursing practice requires being “competent, confident, and comfortable with one’s judgments on whether to use mathematics in a particular situation and if so, what mathematics to use, how to do it, what degree of accuracy is appropriate, and what the answer means in relation to the context.” Another nursing study describes four distinct areas of competence necessary for accurate drug dose calculation (“the 4 Cs”): computation, conceptualisation, conversion, and critical analysis (Johnson & Johnson, 2002). Therefore, multiple skills are necessary for competence in CN and safe clinical practice: this applies not just to drug dose calculation, but also to clinical tasks involving data interpretation, including basic statistical analysis. Clearly, these skills are required by medical students and doctors as well as nursing students and nurses.

Our results may provide some insight into the type of errors being made by doctors and medical students in the MINT. We consider that these errors may relate to one or more of the “4 Cs”. Research in nursing practice has shown that using calculators reduces the incidence of computational errors, but has no impact on conceptual errors (Murphy & Graveley, 1990; Bliss-Holtz ,1994). Thus our finding that the two questions for which calculators improved performance were computational is in accordance with this literature. Similarly, our finding that calculators did not influence mean test scores may indicate that participants are primarily making conceptual rather than computational errors. Furthermore, errors may occur when converting between different units of measurement: Wheeler *et al* (2007) note that doctors commonly make such errors in drug dose calculation. Finally, participants in our study may not have critically analysed their answers to assess whether they were likely to be correct: there is evidence that errors made by bioscience students (Tariq, 2008) and nursing students (Galligan & Hobohm, 2015) in numeracy tests are often due to failure to cross-check their answers; therefore it is likely that medical students also make this type of error. Determining the type of error being made is an important step in developing appropriate educational intervention, since successful remediation requires that the intervention is targeted at the area of weakness. Further research is needed in this area.

The lack of impact of calculators on MINT scores in this study reinforces our original observation that some doctors have low CN. This is important in relation to patient safety, as errors in drug dose calculation and in data interpretation may lead to serious patient harm (Lesar *et al* 1997; Hughes & Edgerton, 2005; Gleason *et al* 2010; Gigerenzer & Gray, 2011; Abramson *et al* 2012; Seden *et al* 2013; Vincent *et al* 2014; Williams & Walker, 2014; Malhotra *et al* 2015). Moreover, the finding that calculators do not overcome apparent deficiencies in CN is supported by the observation that the introduction of electronic prescribing has had less impact on the prevalence of medication errors than was initially anticipated (Tully, 2012; Ahmed *et al* 2016). Further work is required to elucidate how and why doctors and medical students make errors in tests of CN, as this may have implications for their clinical practice and their education.

We asked participants to report on gender because there is evidence that female gender may be associated with lower numeracy (Sikorskii *et al* 2011; Stoet & Geary, 2013; Bagnasco *et al* 2016). However, a study by Bridgeman *et al* (1992) and a large meta-analysis by Lindberg *et al* (2010) found no difference in mathematical ability related to gender; nonetheless, Lindberg *et al* (2010) found strong evidence of stereotyping girls and women as being inferior at mathematics. In our study, participants identifying as male performed better than those identifying as female. We consider that further research into the association of gender and CN in medical students and doctors could help tease apart different CN constructs to see if some of the different findings are related to gender effects on different aspects of numeracy.

We recorded dyslexia because of the overlap between dyslexia and dyscalculia (Gibson & Leinster, 2011; British Dyslexia Association, 2017); however, we found no statistically significant difference in performance of participants with dyslexia, compared to non-dyslexic participants.

LIMITATIONS

All participants in this study were from a single medical school, and so effects might be related to the context of the course itself, although this is unlikely due to the random allocation between groups. Furthermore, drug dose calculation is a complex task, for which several distinct competencies are required, and we have explored only one small area. Nonetheless, we consider that our findings provide insight into the type of numeracy errors made by doctors and medical students, and may be valuable in terms of determining the direction of educational intervention to remediate drug calculation error.

CONCLUSION

Using calculators did not affect overall MINT score. We consider that this outcome may be related to two key factors: first, since a large proportion of the test material can be classified as either calculator inappropriate or calculator optional, a calculator would not be expected to confer any benefit; and second, our findings suggest that the errors being made in the MINT are not remediable by using calculators, i.e. the errors are conceptual rather than arithmetical. This has implications for educational intervention to reduce drug calculation errors in doctors and medical students.

Blank page

CHAPTER 4

ASSESSING CLINICIAN NUMERACY: VERY SHORT ANSWER OR SINGLE BEST ANSWER? A RANDOMISED CONTROLLED TRIAL

The study presented in this chapter has been prepared for submission as a Brief Report to the Journal of Experimental Education. This chapter is identical in content to the submitted manuscript; however, for consistency, it is presented in the same style as the rest of the thesis.

The results of this study have also been presented at various conferences:

1. Developing Excellence in Medical Education Conference (DEMEC), November 2017
2. National Association of Clinical Tutors (NACT) Walsall, February 2018
3. Association for the Study of Medical Education (ASME) Workshop, Nottingham, April 2018
4. Annual Medical Education Conference, Keele, April 2018
5. Health Education England Educators Conference, Birmingham, November 2018.
6. Grand Rounds, University of Manchester, February 2019

ABSTRACT

Logistical advantages, including rapid objective computer marking, make multiple choice single best answer (SBA) tests cost-effective and easy to deliver to large numbers of candidates, hence their widespread use in undergraduate and postgraduate assessment. However, SBA tests are subject to examination technique including cueing and guessing, and so may overestimate candidates' knowledge and ability. Open-ended or constructed response (CR) tests eliminate measurement error due to examination technique, and are often considered superior to SBA tests; however, they are time-consuming to deliver and mark, and scoring is subjective. While accepting that no single form of assessment is perfect, researchers have attempted to improve the performance of both SBA and CR tests. The Very Short Answer (VSA) test is a promising development of the CR format since it allows rapid, objective, computerised marking. Initial research suggests that the VSA is superior to SBA tests. Our research has focussed on improving the performance of an SBA test by enhancing the quality of the test material. We conducted a randomised controlled trial comparing the SBA and VSA formats of our assessment, the Medical Interpretation and Numeracy Test (MINT). We compared mean test scores, and the facility of individual questions for each assessment format. We found no difference associated with test format. Our study demonstrates that the SBA test format can match the performance of the VSA format, when the SBA test is well constructed, and that SBAs should continue to hold a valuable place in assessments.

INTRODUCTION

Multiple choice single best answer (SBA) tests are popular in undergraduate and postgraduate assessment since they are an efficient and cost-effective method of testing large numbers of candidates, they allow examiners to assess a broad range of topics, test scoring is objective, and accurate results can be produced rapidly (Simkin & Kuechler, 2005; DiBattista & Kurzawa, 2011). In contrast, constructed response (CR) tests, which require candidates to create their own answers, can only test a limited number of topics, are labour-intensive and expensive to deliver and mark, and scoring is slow, subjective, and unreliable (Schuwirth and van der Vleuten 2004; Simkin and Kuechler 2005; Funk & Dickson, 2011). Moreover, they may discriminate against individuals with poor writing skills (Kastner & Stangl, 2011). However, the validity of SBA tests is debated, due to concerns that, unlike CR tests, they do not assess genuine understanding of a topic or higher-level cognitive processing (Downing, 2003; McCoubrie, 2004; Simkin & Kuechler, 2005; DiBattista & Kurzawa, 2011; Jordan, 2013; DiBattista *et al* 2014). However, Hall *et al* (2018) consider cognitive function in terms of "System 1" and "System 2" thinking, as described by Kahneman (2011): System 1 thinking is an intuitive, rapid response, while System 2 thinking is a slower and more reflective process. They note that both types of thinking are used in answering SBA questions, therefore, well-crafted SBA assessments can test higher cognition (Hall *et al* 2018). Nonetheless, a consistent finding is that SBA tests tend to overestimate candidates' ability compared to CR tests (Simkin

& Kuechler, 2005; DiBattista & Kurzawa, 2011; McAllister & Guidice, 2012; DiBattista *et al* 2014; Sam *et al* 2016).

The inflation of test scores in SBA tests is proposed to be due to two factors: guessing and cueing (Downing, 2003; McCoubrie, 2004; Simkin & Kuechler, 2005; DiBattista & Kurzawa, 2011; McAllister & Guidice 2012; DiBattista *et al* 2014; Sam *et al* 2016). Since most SBA tests offer five answer options (one correct answer and four distractors), random guessing will give a student a 1:5 chance of selecting the correct answer, while an informed guess may increase the odds to 1:2; thus, students who should fail a test based on their knowledge and understanding may pass by virtue of guessing. Cueing occurs when candidates can recognise the correct answer from among the available options, even though they would be unable to supply the correct answer unprompted. Cueing is dependent on construct: Traub & Fisher (1977) found that the SBA format led to higher scores when assessing verbal comprehension, but conferred no advantage when testing mathematical reasoning. Furthermore, the quality of the SBA test material must be considered. A consistent finding on analysis of SBA tests is that the quality of test questions and answers is often poor, and it is recognised that is difficult to create realistic distractors (Downing, 2003; McCoubrie, 2004; DiBattista & Kurzawa, 2011; Jordan, 2013; DiBattista *et al* 2014; Sam *et al* 2016). DiBattista & Kurzawa (2011) found that the quality of the distractors used in SBA questions was often poor, and that performance of SBA tests improved when the quality of distractors was raised. Nonetheless, since CR tests are subject to neither cueing nor guessing, they are often considered to be superior to SBA tests (Sam *et al* 2018).

Whatever their perceived disadvantages, the unquestioned advantage of SBA tests has been their feasibility (Downing, 2003; McCoubrie, 2004; Simkin & Kuechler, 2005; DiBattista & Kurzawa, 2011; McAllister & Guidice 2012; DiBattista *et al* 2014; Sam *et al* 2016). However, the Very Short Answer (VSA) test (Sam *et al* 2016), a CR format that limits responses to a maximum of three words, can be marked using computer software, making it logistically comparable to an SBA test. Additionally, comparisons of the VSA and SBA formats suggest that the VSA has greater validity, since it is not subject to cueing or guessing (Sam *et al* 2018). An initial study comparing VSA and SBA formats involved 266 medical students who answered 15 questions in a VSA test, followed by the same 15 questions in SBA format; results showed that the facility of all 15 questions was higher in the SBA format, ascribed to cueing and guessing (Sam *et al* 2016). A further study of 299 medical students comparing a 60-question test in VSA and SBA formats also demonstrated a significant cueing effect associated with the SBA format (Sam *et al* 2018).

In researching clinician numeracy in medical students and doctors, we have developed the Medical Interpretation and Numeracy Test (MINT) (Taylor & Byrne-Davis 2016, 2017). Clinician Numeracy (CN) is the ability to use and apply numerical information in patient care, and is important for safe prescribing and accurate data interpretation. We have delivered the MINT in both SBA and CR test formats; as the answers to all MINT questions require no more than three words, the CR format is by definition a VSA test. The distractors used in the SBA version of the MINT are evidence-based, having been developed using incorrect responses made in a VSA version of the test.

In order to explore the context specificity of the finding that SBA would overestimate abilities compared with VSA, and to determine whether our test of CN would lack validity by the overestimation of abilities due to cueing or guessing, we conducted a randomised controlled trial comparing the SBA and VSA formats of the MINT.

METHODS

Trial design

We designed a randomised controlled trial to compare test scores on the SBA and VSA formats of the MINT. The MINT is a test of clinician numeracy, with 43 questions testing computational, analytical and statistical numeracy; all are posed in a clinical setting, and are designed to assess cognitive skills in analysing and interpreting quantitative data. Both MINT formats contain the same test questions, and the SBA format has five answer options.

Methods

The MINT was developed as an SBA test, and includes 31/43 questions that were adapted from existing material, and 12/43 new questions; analysis of our original data suggested that some of the distractors were ineffective (Taylor & Byrne-Davis, 2016). Having subsequently implemented the MINT as a constructed response (CR) test, we analysed the incorrect answers provided by participants, and used these to inform development of new distractors. This strategy is recommended by Birenbaum & Tatsuoka (1987) and DiBattista and Kurzawa (2011) as a method of providing credible distractors for SBA tests.

Participants

All third-year students at a medical school in the UK were scheduled to attend a teaching session on clinician numeracy and its importance for healthcare professionals. During this session, students were given information about the research, and invited to participate in the study. In addition, in advance of the study, all students were emailed with further details about the research, including a participant information sheet. All students were eligible to participate in the study, with no exclusion criteria.

Interventions

The study took place one month after the teaching session. Participants were randomised into two study groups, Group S received the SBA format of the test, and Group V, the VSA format. The test was carried out under examination conditions, and participants had 90 minutes in which to complete the test.

Sample size

The MINT has 43 questions, each of which scores one mark with no penalty for incorrect answers, so potential scores range from 0 - 43. The study compared two independent samples: sample size calculation requires the expected (or observed) mean scores of each group, and the standard deviation for the control group. We calculated the sample size using the mean test score and standard deviation achieved in our initial study with the VSA format of the MINT (31.8 and 5.2 respectively) (unpublished data). We took statistical advice to calculate the expected mean score of the SBA group, concluding that a significant difference in score would be an increase in score of 4 marks (approximately 10%), to give a mean of 35.8. This data was entered into an online calculator (Brant, n.d.), which indicated that we would need to recruit 54 participants (27 in each group) to reach a significance of 5%, with a power of 80%.

Randomisation: Random Allocation and Allocation Concealment

We prepared a list of study identification (ID) numbers from 1 to 140, and then used a table of random numbers to allocate these study ID numbers to two groups, S and V. Group S received the SBA version of the test, while Group V received the VSA version.

We numbered 140 windowless brown A4 envelopes from 1 – 140, and put test papers in to each envelope according to the random allocation. Additional test material including an answer sheet, pencil, eraser, and calculator was inserted into each envelope. On the day of the test, each participant was given a test envelope; these were not delivered in any particular order i.e. the first 16 students to sit the test did not receive test papers 1 – 16.

Blinding

All test material was contained in a sealed windowless brown A4 envelope. Although the study ID number was written on each envelope, the writing was in pencil and in very small script on the back of the envelope, and not readily visible. Furthermore, without the key to the random allocation (which was not available at the testing site) the test format could not be determined without opening the envelope. Therefore, neither the researcher nor the participant knew in advance whether the envelope contained the SBA or VSA format of the test.

Outcomes

The aim of this study was to assess whether the scores of participants who sat the SBA format of the test were equivalent to those of participants who sat the VSA format. We recorded the mean test scores of both study groups, and the facility of individual test questions for each group. We recorded the presence of dyslexia, since this may be associated with dyscalculia, and thus affect performance (British Dyslexia Association, 2017). We also recorded student gender as there is some evidence that this may be associated with numeracy (Windish *et al* 2007; Sikorskii *et al* 2011).

Data Analysis

We compared the mean scores of the study groups using Students' *t*-test. We compared the facility of individual test questions using the N-1 Chi square test (Campbell, 2007). Data was analysed using Microsoft EXCEL, and an online statistical tool (MedCalc Software bvba).

Ethics

This research was approved by the University of Manchester Research Ethics Committee.

RESULTS

One hundred and thirty-two students sat the MINT in January 2018, of whom 120/132 (91%) agreed to participate in the study. Analysis of demographic data showed that 60/120 (50%) participants were female, and 47/120 (39%) were male, while 13/120 (11%) did not indicate their gender. The cohort included 20/120 (17%) students with specific learning difficulties for which they were allowed extra time in university exams: 18 of these had dyslexia, and two did not specify their disability, but stated that they were not dyslexic.

Following random allocation, 62/120 (51%) participants were assigned to Group S, and 58/120 (49%) to Group V. There were 43 questions, with a mark of 1 for correct answers, and no score for incorrect or unanswered questions; the maximum possible score was 43. The range of scores was similar for all groups: 20 – 42 for the full cohort of 120 students; 22 – 42 for Group S, and 20 – 40 for Group V; the interquartile range was 29 – 36 for all three groups. The mean scores were almost identical for all groups: that of the full cohort was 32.66 with a standard deviation (SD) of 4.9; the mean (SD) scores for Group S and Group V were 32.7 (4.9) and 32.6 (5.04) respectively. The median scores were also similar, with a value of 33 for both the full cohort and Group S, and 32 for Group V.

In addition to comparing the overall scores of both groups, we analysed performance on individual test questions. The facility of a question refers to the proportion of candidates who answer it correctly, and is recorded as a decimal: a question answered correctly by all candidates has a facility of 1.0, while one answered correctly by 50% of candidates has a facility of 0.5. Although the facility of 20/43 (47%) questions appeared higher in Group S, and of 19/43 (44%) appeared higher in Group V, these differences were not statistically significant (Figure 4.1, Table 4.1).

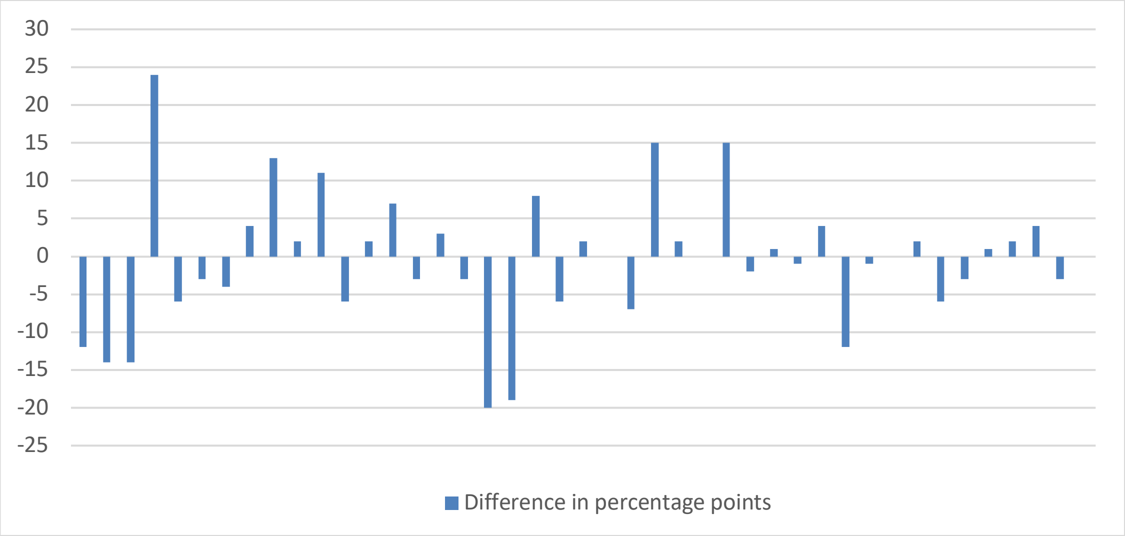


Figure 4.1.
Comparison of facility of test questions.
Questions are ranked in order of facility, from most to least difficult. Positive bars indicate questions where the facility was higher for Group S; negative where the facility was higher for Group V.

Table 4.1. Comparison of facility of individual test questions by test format

Item no.	Primary construct	Overall n=120	Group S n=62	Group V n=58	Difference %	SBA effect	p value*
1	Computational	.88	.90	.86	4	+	ns
2	Analytical	1.0	1.0	1.0	0	0	ns
3	Statistical	.61	.63	.59	4	+	ns
4	Computational	.92	.92	.93	1	-	ns
5	Computational	.70	.68	.71	3	-	ns
6	Analytical	.50	.47	.53	6	-	ns
7	Statistical	.97	.98	.96	2	+	ns
8	Analytical	.68	.69	.67	2	+	ns
9	Statistical	.69	.72	.65	7	+	ns
10	Statistical	.8	.71	.90	19	-	ns
11	Statistical	.53	.52	.55	3	-	ns
12	Computational	.85	.92	.77	15	+	ns
13	Analytical	.63	.64	.62	2	+	ns
14	Statistical	.37	.31	.45	14	-	ns
15	Statistical	.29	.22	.36	14	-	ns
16	Computational	.98	1.0	.96	4	+	ns
17	Statistical	.96	.95	.98	3	-	ns
18	Analytical	.8	.84	.76	8	+	ns
19	Computational	.72	.74	.71	3	+	ns
20	Computational	.98	.97	1.0	3	-	ns
21	Computational	.79	.89	.69	20	-	ns
22	Computational	.87	.87	.88	1	-	ns
23	Analytical	.8	.77	.83	6	-	ns
24	Analytical	.48	.60	.36	24	+	ns
25	Computational	.96	.97	.96	1	+	ns
26	Statistical	.76	.74	.77	3	-	ns
27	Statistical	.28	.22	.34	12	-	ns
28	Analytical	.94	.95	.93	2	+	ns
29	Computational	.66	.63	.69	6	-	ns
30	Computational	.81	.81	.81	0	0	ns
31	Analytical	.82	.79	.86	7	-	ns
32	Analytical	.65	.71	.60	11	+	ns
33	Computational	.84	.85	.83	2	+	ns
34	Analytical	.80	.81	.79	2	+	ns
35	Computational	.84	.84	.84	0	0	ns
36	Statistical	.90	.84	.96	12	-	ns
37	Analytical	.56	.53	.57	4	-	ns
38	Statistical	.93	.93	.93	0	0	ns
39	Analytical	.85	.84	.86	2	-	ns
40	Computational	.61	.66	.53	13	+	ns
41	Statistical	.95	.92	.98	6	-	ns
42	Analytical	.85	.85	.84	1	+	ns
43	Computational	.82	.89	.74	15	+	ns

**With Bonferroni correction, the result is reported as significant if p<0.05/43, i.e. if p < 0.001.*

DISCUSSION

We found no evidence that assessment format influenced test scores on the MINT. Although MINT questions vary in level of difficulty, and in numeracy construct (computational, analytical and statistical), the range of test scores, and both the mean and median scores were similar for both study groups. Furthermore, the facility of individual test questions did not differ significantly between the groups. Thus, we found that in the context of numeracy, with high quality evidence-based distractors, the SBA format was equivalent to the VSA format in terms of test scores.

Our finding that a carefully constructed SBA test can perform just as well as a VSA test is important, and supports previous research indicating that well-designed SBA questions can approximate the performance of CR questions (Bridgeman, 1992; Lin & Singh, 2011). Lin & Singh (2011) observe that developing high quality distractors based on research into the errors made by participants, allows SBA tests to accurately assess participants' cognitive processes in the same way as a CR test. DiBattista & Kurzawa (2011) found that SBA questions were effective and discriminating when credible distractors were provided, and also recommend using the incorrect responses provided by participants in a test, to develop plausible distractors.

The development of the VSA test format (Sam *et al* 2016) has been welcomed as a significant advance, since it combines the logistical advantages of an SBA test with the credibility of the CR format (Sam *et al* 2016; Sam *et al* 2018). However, Sam *et al* (2016) and Sam *et al* (2018) do not discuss how their test questions were developed, so we do not know whether they used evidence-based distractors for the SBA version of their test. If not, it would be interesting to repeat their study comparing the VSA with an evidence-based SBA.

In both previously published studies comparing VSA and SBA, the authors found significant cueing effects associated with the SBA format: in the first study, students sat a test comprising 15 questions, and the facility for all 15 was higher in the SBA format (Sam *et al* 2016). The test used in the second study had 60 questions, and the facility was higher for 56/60 in the SBA format (Sam *et al* 2018). This contrasts with our study, where the facility of questions varied between formats, and was higher for the VSA format (19/43 questions) as often as for the SBA format (20/43), with equal facility for 4/43 questions. Therefore, we did not find any evidence of cueing in our study.

In the context of assessment in clinical medicine, the VSA is considered to have greater validity and authenticity than the SBA, because providing five optional answers to a question is deemed unrealistic, and guessing unscientific (Sam *et al* 2016; Sam *et al* 2018). However, this argument is flawed since doctors are often required to form a differential diagnosis of multiple conditions from which they must select the "best answer", and not uncommonly must resort to making an educated guess as to the most likely diagnosis (Downing, 2003). The MINT test material is based on common medical tasks requiring CN, hence we consider that it is authentic. Furthermore, since participants must solve a numerical problem to answer each question, regardless of format, we believe that the same cognitive process is used for both formats of the MINT. Moreover, the MINT test material includes complex computations which require System 2 thinking; this is the cognition required in difficult clinical situations (Hall *et al* 2018). Therefore, the MINT tests the cognitive processes used by medical students and doctors in daily clinical practice.

Limitations

As with other research demonstrating equivalence between SBA and CR formats (Bridgeman, 1992; Lin & Singh, 2011), our interest is in quantitative reasoning, and it is not clear whether our findings are construct-specific, or are generalizable to other abilities. Indeed, this may explain the difference in outcome between our study and that of Sam *et al* (2016), because although both assessments are set in a clinical context, Sam *et al* (2016) tested medical knowledge (verbal reasoning), whilst we tested clinician numeracy (quantitative reasoning). However, the observations and recommendations of DiBattista & Kurzawa (2011), in relation to improving the quality of SBA distractors are based on their review of almost 1200 SBA questions across a range of disciplines. Furthermore, Schuwirth & van der Vleuten (2004) observe that when the same material is assessed by SBA and CR tests, correlations between test scores are high, demonstrating that the perceived impact of test format on performance may be overestimated.

CONCLUSION

This study shows that a well-constructed SBA test with high-quality, evidence-based distractors, is equivalent to a VSA test. This reassures us that it is appropriate to continue to use the SBA format of the MINT for our ongoing research, including the development of electronic and computer-adaptive versions of the MINT. We have demonstrated the positive impact of developing evidence-based distractors for SBA tests, and consider that this is applicable to a wide variety of disciplines, although we recognise that further research in this area is indicated.

Blank page

CHAPTER 5

ERRORS MADE BY MEDICAL STUDENTS IN A TEST OF CLINICIAN NUMERACY

The study presented in this chapter has been prepared for submission as an article for the journal *Medical Teacher*. It is presented in the same style as the rest of the thesis.

The results of this study have also been presented at the following conferences:

1. Developing Excellence in Medical Education Conference (DEMEC), November 2017
2. National Association of Clinical Tutors (NACT) Walsall, February 2018
3. Association for the Study of Medical Education (ASME) Workshop, Nottingham, April 2018
4. Annual Medical Education Conference, Keele, April 2018
5. Health Education England Educators Conference, Birmingham, November 2018.
6. Grand Rounds, University of Manchester, February 2019

INTRODUCTION

The ability of healthcare professionals to use, manipulate and interpret numerical data to provide safe patient care is called Clinician Numeracy (CN). Although many everyday medical tasks, including drug dose calculation and analysis of test results, require CN, there is increasing evidence that medical students and doctors may have deficiencies in CN (Rowe *et al* 1998; Selbst *et al* 1999; Sheridan & Pignone, 2002; Wheeler *et al* 2004a; Wheeler *et al* 2004b; Simpson *et al* 2009; Rao & Kanter, 2010; Harries & Botha, 2013; Johnson *et al* 2014; Taylor & Byrne-Davis, 2017). The reason for low CN in medical students and doctors is unexplored; however, it may be related to three factors: 1) medical schools are selecting applicants with low numeracy; 2) undergraduate and postgraduate curricula are failing to ensure appropriate standards of numeracy in graduates; and/or 3) flaws in the healthcare system enable errors to occur. Deficiencies in CN may lead to medical error, with significant morbidity and mortality, hence this problem needs to be addressed (Wheeler *et al* 2004b; Gigerenzer *et al* 2007; Harries & Botha, 2013; Taylor & Byrne-Davis, 2017). As with any medical condition, to treat the problem successfully, we need to diagnose the cause and prescribe the correct treatment i.e. we need to investigate the cause of errors being made in tests of CN, so that we can provide appropriate interventions.

Error in healthcare is often considered in terms of individual and system failures (Reason, 2000) (Table 5.1), and there is evidence suggesting that both individual and system factors are involved in errors related to CN in medical students and doctors, although there is little literature available (Wheeler *et al* 2004b; Harries & Botha 2013; Williams & Walker 2014). However, there has been considerable research into error in drug dose calculation tests in nurses and nursing students, where the literature tends to divide errors into two main categories: mathematical and conceptual errors (Bliss-Holtz, 1994; Weeks *et al* 2000; Wright, 2004; Brady *et al* 2009; McMullan *et al* 2010), although a third category, conversion error, is often considered (Blais & Bath, 1992; Zahara-Such, 2013; Fleming *et al* 2014; Koharchik *et al* 2014; Bagnasco *et al* 2016; Hurley, 2017). However, the terminology used in this research is not standardised, and different terms are used to describe the same type of error e.g. errors in basic arithmetic are variously referred to as ‘mathematical’ (Blais & Bath, 1992), ‘computational’ (Weeks *et al* 2000; Koharchik *et al* 2014) and ‘arithmetical operation’ (Bliss-Holtz, 1994) errors. Furthermore, the same term may be used to describe different types of error e.g. ‘arithmetical operation error’ refers to errors in basic arithmetic (Bliss-Holtz, 1994) and also to misunderstanding of mathematical operations (Weeks *et al* 2000). Additionally, the term ‘conceptual error’ is often used to describe situations where the candidate is thought to have misread or misunderstood a question (Blais & Bath, 1992; Wright, 2004; Zahara-Such, 2013; Fleming *et al* 2014), while errors of ‘conceptualisation’ refer to an individual’s lack of understanding of the process of drug preparation and administration (Weeks *et al* 2000; Johnson & Johnson, 2002; Wright, 2007b; Coyne *et al* 2013). (Nursing students with little clinical experience are often unable to conceptualise the clinical equipment involved, and consequently may make errors in drug dose calculation (Weeks *et al* 2000; Johnson & Johnson, 2002; Wright, 2007b; Coyne *et al* 2013)). This variation in terminology highlights the

Table 5.1. Error in healthcare that may be relevant to CN*

Active Failures	Error-provoking conditions	Latent conditions
Individual unsafe acts	Task and environment	Organisational processes
Knowledge-based mistakes Lack of knowledge of drug, including dose and interactions Lack of patient information *Skill-based mistakes* Slips and lapses: may be due to lack of concentration; multi-tasking; Interruptions; memory lapses. *Rule-based mistakes* Lack of knowledge of the rule Failure to follow the rule Application of the wrong rule *Violations* Deliberate deviation from policy or procedure	*Individual* Hungry, thirsty, tired, distracted Inadequate knowledge, skill, experience, training *Working environment* Inadequate staffing; new or locum staff High workload, pressure Lack of access to drug & patient information, and Lack of access to computers *Health-care team* Communication problems Failure to recheck when instructions queried by nursing staff Inadequate training, knowledge & experience In relation to very junior doctors: Assume that others will double-check Difficulty in weighing risks and benefits *Prescribing task* Ambiguous or unavailable guidelines Lack of standardisation	*General* Long hours Inadequate staffing Reluctance to challenge or escalate Lack of feedback systems *Prescribing* Lack of training Low importance attached to task Simultaneous multiple prescribing tasks

**This table is based on Reason's model (2000), and adapted from Dornan et al (2009)*

complexity of the drug dose calculation process (Johnson & Johnson, 2002; Wright, 2005; Coben & Weeks, 2014); however, it also renders these classification systems unfit for general use.

An additional problem with existing classification systems is that the cause of error is often inferred from analysis of students' test papers. While this is logistically the most feasible approach to analysing error, a disadvantage is that it relies on the researcher's interpretation of the data e.g. the nursing studies cited above refer to errors of understanding ('conceptual', 'misinterpretation' and 'misread' errors); however, this interpretation of the data is subjective. Consider an answer showing an incorrectly placed decimal point: this may have occurred because the candidate 1) did not understand the question; 2) did not know how to calculate using decimal places; 3) carried out the mathematical operation inaccurately; 4) performed the wrong mathematical operation; 5) made an error when converting from a different format e.g. percentage; or 6) made a transcribing error. Further information such as evidence from rough work may help determine the cause of the error, otherwise it is impossible to classify this error accurately without discussion with the candidate. An additional problem is that classifying an error as one due to lack of understanding is not specific: a lack of understanding is a cause of error, rather than a definition or class of error; misunderstanding may lead to mathematical, set-up and conversion error. The same applies to the term "careless" error. The difficulty of defining and classifying errors is highlighted by Avery *et al* (2012), who observe that definitions can vary depending on the purpose of classification.

The purpose of defining error in our research was to allow us to identify the type of mistakes being made, and consider their implications in terms of medical education. Therefore, we have designed a study to investigate the cause of error in a CN test in medical students. Our aims were: 1) to develop a classification system that precisely describes the errors that occur; 2) to test this system by using it to reclassify the categories of error documented in nursing studies; 3) to use this system to identify and document the type and frequency of mistakes being made by medical students; and 4) to consider the implications of our findings for medical education in relation to selection, educational intervention and the workplace environment.

METHODS

Participants

All third-year medical students in a single UK medical school were scheduled to take the MINT as a formative examination midway through their third year at medical school. All of these students were eligible to participate in the study, with no exclusions. Information about the study was sent to all students by email in advance of the study, and they were invited to participate. Students were informed that participation in the study was optional, and that they could withdraw at any time until data analysis.

Materials

We have previously developed a reliable and valid test of CN in doctors, the Medical Interpretation and Numeracy Test (MINT) (Taylor & Byrne-Davis, 2016). The MINT is a broad assessment of CN, covering the three key constructs relevant to clinical practice: computational numeracy (16 questions), analytical numeracy (14 questions), and statistical numeracy (13 questions); test material is contextualised to a clinical setting, but no medical knowledge is needed to answer any question. The level of difficulty of MINT questions varies, ranging from material suitable for schoolchildren to questions designed for doctors in their early years of clinical practice.

Procedure

All students sat the test in examination conditions, and were allowed 90 minutes to complete the test. Participants in this study were also taking part in a randomised controlled trial (RCT) comparing single best answer (SBA) and very short answer (VSA) test formats; test papers from those who were randomised to the VSA group were included in this investigation of error. The test papers included large amounts of blank space between questions for rough work. We analysed the data provided as rough work by students and used this as a basis for defining and classifying errors.

Ethics

Ethical approval for this study was received from the University of Manchester Research Ethics Committee.

Data analysis

We reviewed the test papers of all participants who sat the VSA version of the MINT. We conducted a qualitative analysis of the responses provided by participants, and created a database recording every wrong answer given for each of the 43 test questions. The wrong answers were grouped together to assess the frequency of each incorrect response. We then reviewed the rough work provided for each incorrect answer, to explore the strategies used by participants to solve the given mathematical problem. This was an iterative process, where we repeatedly analysed the data to identify patterns in how incorrect answers had been derived. We then developed a coding system based on our findings. Key stages in the process are outlined below.

1. We reviewed the numbers used in the calculation to assess whether the appropriate data needed to answer the question had been extracted from the text.
2. We reviewed the mathematical process shown in the rough work to assess whether the problem had been set up correctly.
3. We documented whether the mathematical process had been performed accurately.
4. For most questions, more than one step was needed to calculate the answer; therefore, we analysed the rough work to assess whether each step had been completed, and whether all steps were performed accurately.
5. We reviewed the various strategies used by participants to answer questions, since different methods were often used to reach the correct answer (e.g. calculating the amount of a drug given in doses of 250mg four times a day: this could be done by multiplying 250mg x 4 = 1000mg; however, the numbers could also be added together in different ways: 250mg + 250mg + 250mg + 250mg = 1000mg; or 250mg + 250mg = 500mg + 250mg = 750mg + 250mg = 1000mg.)
6. We recorded whether an incorrect answer had occurred as a result of converting between different units of measurement.
7. We reviewed incorrect answers to assess their plausibility; the provision of highly implausible answers suggested that participants had not carried out a rough estimation or a cross-check of their answer to establish whether it was likely to be correct.
8. All participants had access to a calculator; we considered whether errors appeared to be related to calculator use.
9. In some cases where no rough work was provided, we were able to determine the type of error that occurred because it was a common and specific error. This is similar to the process of considering error in SBA tests, where the distractor has been developed based on a common error or misconception.

RESULTS

There were two main stages to our research: A, the development of a new system to classify error, and B, an investigation of the errors made in the MINT. Therefore we present our findings in two sections.

A. Classification of error

We reviewed the test papers of all students, recorded all incorrect answers, and analysed the errors as shown in Figure 5.1. We documented five distinct categories of error: basic arithmetical errors, errors in setting up the mathematical problem, measurement errors, transcribing errors and errors in interpreting data displays of different kinds; some classes of error can be further subdivided as shown in Table 5.2. We reviewed the classification systems

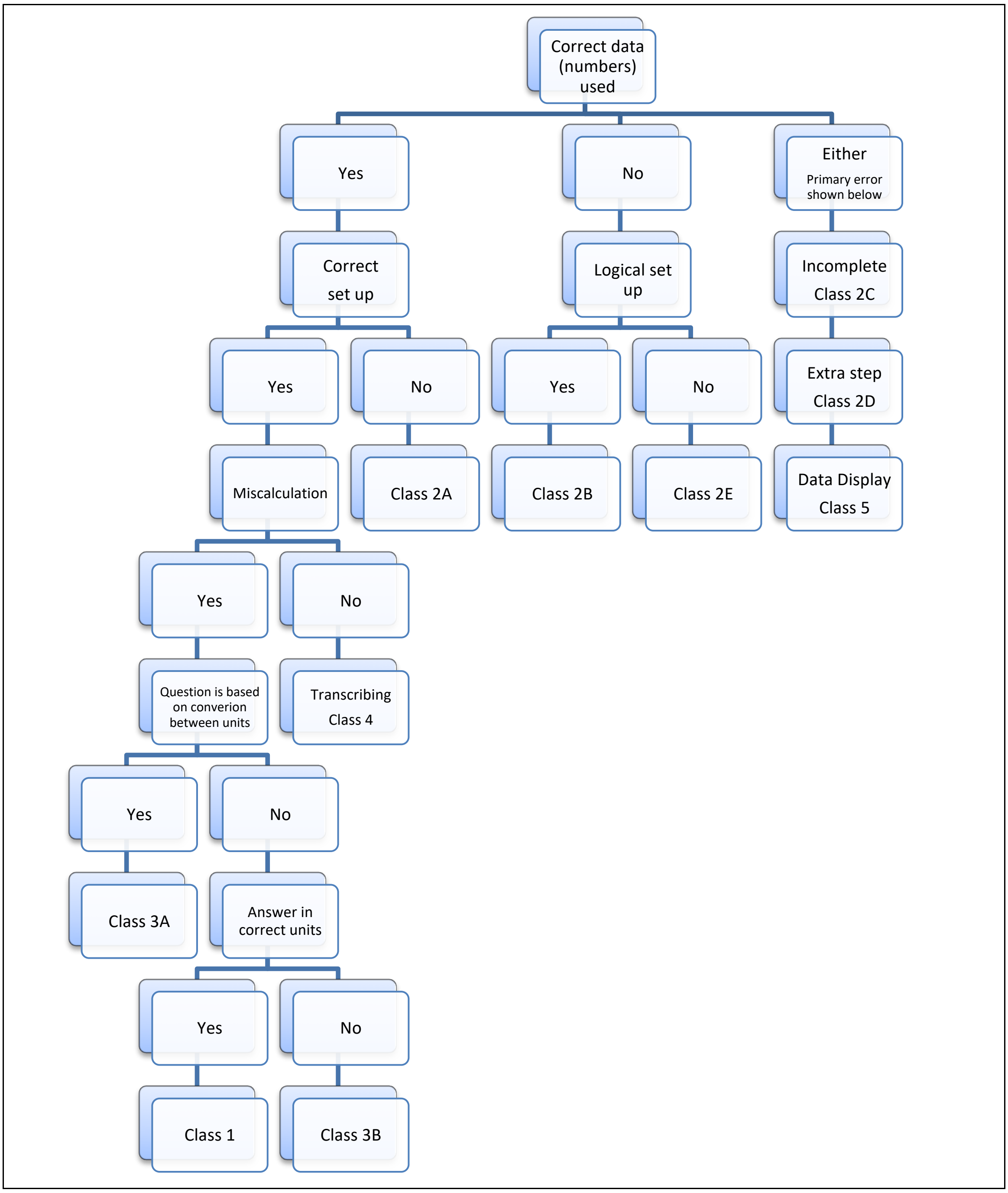


Figure 5.1. Classification of error

NB. Class 2, 3 and 5 errors can be further subdivided depending on whether the calculation is accurate; for clarity these subdivisions are not shown in Figure 5.1.

Table 5.2. Classification of errors in CN tests based on analysis of the MINT

Class of error	Subgroup	Definition
Class 1 Error in basic arithmetic		Failure to add, subtract, divide or multiply accurately Including: whole numbers, decimals, fractions, percentages, numbers expressed in ratio or frequency formats, and failure to use formulae correctly
Class 2 Error in setting up the calculation	A.	The correct data is used, but the wrong process is followed e.g. multiply instead of divide
	B.	Incorrect numbers (data) is used, the correct process may or may not be followed*
	C.	The numbers used and process followed may or may not be correct, but the key error is that the calculation is incomplete*
	D.	The numbers used and process followed may or may not be correct, but the key error is that an unnecessary extra step has been added to the calculation*
	E.	Incorrect data is used, and the process followed is bizarre
Class 3 Error in measurement	A.	Failure to convert between different units accurately e.g. from fraction to percent
	B.	Answer expressed in incorrect units
Class 4 Transcribing error		The correct numbers are used, the correct process is followed, the calculation is accurate, but the wrong answer is entered on the answer sheet
Class 5 Error in interpreting data displays	A.	Error in interpreting data presented in charts
	B.	Error in interpreting data presented in graphs
	C.	Error in interpreting data presented in tables

**Class 2 errors can be further subdivided into subclasses depending on whether there are also miscalculations; for clarity these subdivisions are not shown in Figure 5.1.*

provided in nursing studies, and re-classified them according to our new system (Table 5.3). We report two categories of error not used in nursing studies: transcribing error (Class 4) and error in interpreting data displays (Class 5). However, since the nursing studies are generally based on drug dose calculation, it is not surprising that Class 5 error has not been reported.

B. Error in the MINT

We analysed the test papers of all 58 students who participated in the study. There were 2,494 answers (58 participants x 43 questions) for analysis. We reviewed the rough work relating to all incorrect answers, and categorised them in Classes 1-5 as shown in Table 5.2. We recorded 1892/2494 (76%) correct answers, and 602/2494 (24%) incorrect or unanswered questions. However, not all incorrect answers could be classified, since many participants did not provide any rough work, or provided insufficient data to indicate their problem-solving strategies. Nonetheless, it was sometimes possible to classify incorrect answers in the absence of rough work.

We were able to classify 494/602 (82%) incorrect answers, while 108/602 (18%) of incorrect responses were unclassifiable (the figure 108 comprises 67 incorrect answers and 41 unanswered questions). The majority of incorrect answers 323/494 (66%) were Class 2 errors, involving a failure to set up the calculation properly; 27% consisted of either failure to interpret data displays accurately 68/494 (14%), or inability to convert between different units of

Table 5.3. Classification of errors in drug dose calculation tests.

Researcher	Original classification	Definition	Revised classification
Blais & Bath (1992)	Mathematical error	Failure to add, subtract, divide or multiply accurately Errors in using decimals & fractions	Error class 1: error in basic maths
	Conceptual error	Failure to set up the problem correctly	Error class 2: Set-up error
	Mathematical concept error	Answer expressed in incorrect units	Error class 3: measurement error
	Measurement error	Error when converting between different measurement units	Error class 3
Bliss-Holtz (1994)	Arithmetical operation error	Mistakes in performing basic mathematical procedures	Error class 1
	Mathematical concept error	Mistakes in using formulae	Error class 1
Weeks *et al* (2000)	Computation error	Basic errors of multiplication, division etc	Error class 1
	Arithmetical operation	Misunderstanding of arithmetical operations	Error class 2
Hughes & Edgerton (2005)	Mathematical error	Errors in using decimals, fractions, percentages, and ratios	Error class 1
	Conceptual errors	Failure to understand and or conceptualise the mathematical operation required	Error class 2
Koharchik *et al* 2014	Mathematical computation error	Miscalculation	Error class 1
	Misused ratio, proportion or formula	Mistakes in using formulae	Error class 1
	Incorrect use of Dimensional Analysis	Incorrect use of a specific formula	Error class 1
	Misread or misunderstood question	Misreading or misunderstanding the question	Error class 2
	Incorrect conversion factor	Mistakes in converting between units	Error class 3
	Incorrect rounding	Incorrect rounding	
	Incomplete question	Failure to complete the question	Error class 2
	No math computation shown	No math computation shown	Unclassifiable

measurement 66/494 (13%); and only 35/494 (7%) of errors were due to basic arithmetical errors (Table 5.4). Detailed information regarding the class of error made in each question is

Table 5.4. Frequency of each class of error

	Class 1	Class 2	Class 3	Class 4	Class 5	Total	Unclass-ifiable*	Total
Full test n=43	35	323	66	2	68	494	108	602
%	7	66	13	0	14	100		
Data displays n=16	6	81	35	2	68	226	34	
Text questions n=27	29	242	31	0	0	268	74	

**includes unanswered questions*

shown in Table 5.5, where questions are listed in order of facility, with the easiest questions first. Examples of the different classes of error observed are detailed below.

Table 5.5. MINT items with construct, facility and error analysis

Q no.	Construct	Facility	Class 1	Class 2	Class 3	Class 4	Class 5	U*	Total
Overall		0.82							
2	An	1.0	0	0	0	0	0	0	0
20	Comp	1.0	0	0	0	0	0	0	0
17	Stat	.98	0	0	0	0	0	1	1
41	Stat	.98	0	0	1	0	0	0	1
7	Stat	.96	0	0	0	0	0	2	2
16	Comp	.96	0	2	0	0	0	0	2
25	Comp	.96	1	0	0	0	0	1	2
36	Stat	.96	0	0	0	0	0	2	2
4	Comp	.93	0	3	0	0	0	1	4
28	An	.93	0	0	0	0	1	3	4
38	Stat	.93	0	0	0	0	0	4	4
10	Stat	.90	0	0	6	0	0	0	6
22	Comp	.88	0	5	1	0	0	1	7
1	Comp	.86	2	5	0	0	0	1	8
31	An	.86	0	0	0	0	8	0	8
39	An	.86	1	7	0	0	0	0	8
35	Comp	.84	3	3	0	0	0	3	9
42	An	.84	1	0	1	0	5	2	9
23	An	.83	1	4	0	1	4	0	10
33	Comp	.83	1	7	0	0	0	2	10
30	Comp	.81	1	9	0	0	0	1	11
34	An	.79	0	0	0	0	12	0	12
12	Comp	.77	3	9	0	0	0	1	13
26	Stat	.77	0	1	11	0	0	1	13
18	An	.76	0	9	4	0	0	1	14
43	Comp	.74	5	3	5	0	0	2	15
5	Comp	.71	1	12	0	0	0	4	17
19	Comp	.71	0	17	0	0	0	0	17
21	Comp	.69	2	11	2	0	0	3	18
29	Comp	.69	1	9	4	0	0	4	18
8	An	.67	0	7	4	0	0	8	19
9	Stat	.65	2	16	0	0	0	2	20
13	An	.62	6	16	0	0	0	0	22
32	An	.60	0	6	1	0	15	1	23
3	Stat	.59	0	24	0	0	0	0	24
37	An	.57	1	16	0	0	0	8	25
11	Stat	.55	0	0	26	0	0	0	26
6	An	.53	0	5	0	1	19	2	27
40	Comp	.53	3	16	0	0	0	8	27
14	Stat	.45	0	27	0	0	0	5	32
15	Stat	.36	0	29	0	0	0	8	37
24	An	.36	0	15	0	0	4	18	37
27	Stat	.34	0	30	0	0	0	8	38
	Total		35	334	55	2	68	108	602

**Unclassifiable answers*

Class 1: Basic arithmetical errors

Basic arithmetical errors were the primary error in 7% incorrect answers, and include errors involving whole numbers, decimals, fractions and percentages. There were errors in addition e.g. for the sum 152 + 165 + 13 = 165, incorrect answers included 170, 190, 300, 303, and 317.

There were also errors in subtraction e.g. 200 - 70 = 140; multiplication e.g. 35 x 2 = 75, 65 x 0.1 = 65; and division e.g. 80/10 = 0.8. Some mistakes appeared to result from using a calculator to simplify a fraction: the calculator gives the answer as a decimal, and participants often made errors in recording the answer e.g. adding a % to the decimal, thus 250/380 = 2/3 = 0.66%. Participants often made errors when dealing with proportions e.g. 50 units/50ml = 5 units/1ml, 5 units/1ml = 1.5 units/3.3ml, 6.5 units = 4.3ml. Failure to critically evaluate answers was common, as is evident here.

Class 2: Error in setting up the calculation

The vast majority of errors were in this category, which was subdivided into six groups, 2A – 2F, each of which could be subclassified depending on whether or not the calculations were accurate.

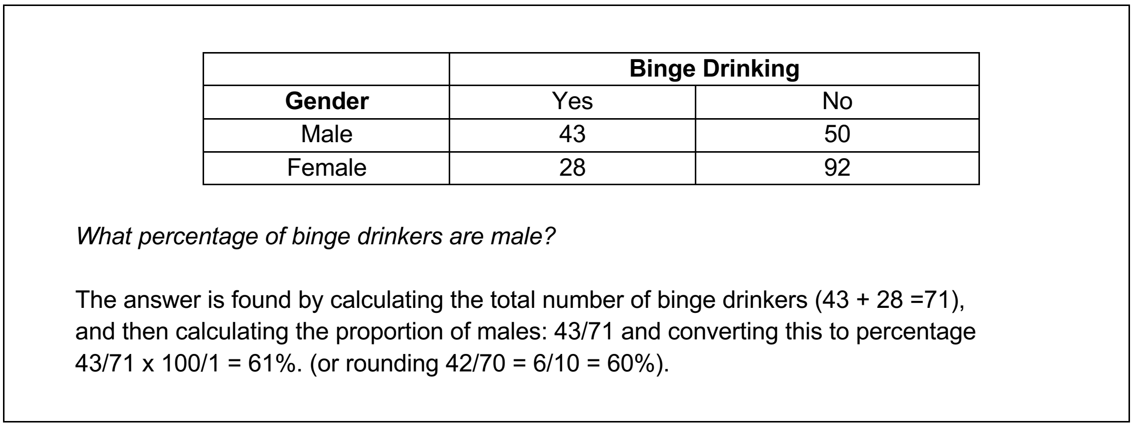
Class 2A errors were those where the correct data was used, but the wrong process was followed e.g. a question asked participants to calculate the rate of an insulin infusion. The information given included the patient's weight (65kg), the clinical guideline for infusion rate (0.1unit/kg/hr), and the standard infusion preparation of a 50ml solution containing 50 units of insulin. Six participants who had correctly calculated that the patient needed 6.5 units of insulin per hour suggested incorrect infusion rates of 7.6 or 7.7 ml/hr; their rough work shows that they divided the volume of infusion (50ml) by the rate required (6.5 units/hr). A further six participants divided 6.5 units/hr by the volume of infusion (50ml) to reach answers of 0.13 ml/hr (or 13 ml/hr if a decimal place error was made). This form of set-up error is likely to be caused by the unfamiliarity of candidates with the clinical task involved, i.e. inability to conceptualise the problem.

Class 2B errors occur when the participant has selected the wrong numbers or data from the text of the question. One question is based on a nutritional drink with four servings per carton; the question was how many calories in half of the carton; 15 participants calculated half the number of calories in one serving, and another participant calculated half the volume of the carton.

Class 2C errors occur when the primary mistake is failure to complete all of the steps required e.g. participants were asked to calculate the proportion of a unit of blood that would be transfused (at a given rate) over a 2-hour period: six participants calculated the volume transfused, but omitted the second step required: calculating the proportion. In contrast, Class 2D errors involve performing an additional extra step e.g. for a question regarding the amount of drug infused per hour, eight participants added a further step and calculated the amount infused over 72 hours.

Class 2E are bizarre errors that occur when the data used is inappropriate and the calculation is set up in a way that is difficult to understand. Two examples are given here, although there are many more. The first question involves a patient with diabetes who needs to eat 6g of carbohydrate (CHO) to support 30 minutes of exercise. She eats biscuits prior to exercising; each biscuit contains 8g CHO. Candidates are asked how many biscuits she should eat before exercising for one hour. One participant's answer was "4.5 biscuits"; rough work showed 6+30 = 36; 36/8 = 4.5 i.e. this person has added 6 (grams CHO) to 30 (minutes), and then divided 36 by 8 to get 4.5. The basic arithmetic has been performed accurately, but the set-up is completely illogical.

Another bizarre answer was provided in relation to the question shown in Figure 5.2.



	Binge Drinking	
Gender	Yes	No
Male	43	50
Female	28	92

What percentage of binge drinkers are male?

The answer is found by calculating the total number of binge drinkers (43 + 28 =71), and then calculating the proportion of males: 43/71 and converting this to percentage 43/71 x 100/1 = 61%. (or rounding 42/70 = 6/10 = 60%).

Figure 5.2. Question on binge drinking.

One candidate set the calculation up as 43/114; 43 is the number of male binge drinkers; 213 people were surveyed, of whom 71 are binge drinkers and 142 are not. This candidate has subtracted 28 (female binge drinkers) from 142 (total non-binge drinkers) to get 114; they then set up the calculation as the number of male binge-drinkers (43) divided by (the number of non-binge drinkers minus the number of female binge drinkers) (114). Once again, the set-up defies logic.

Class 3: measurement error

Class 3A are errors that result from a failure to convert accurately from one format to another e.g. one question stated that the risk of a side effect was 0.3%, and participants were asked how many of 100,000 people exposed to the risk would be expected to get the side effect. Eleven participants gave incorrect answers ranging from 0.0003 to 30000, thus making conversion errors. Class 3B errors occur when the calculation has been done correctly, but the answer is expressed in the wrong units. A question involved calculating the volume of local anaesthetic a patient required: some candidates gave the answer 40mg rather than 40ml.

Class 4: Transcribing error

This was very uncommon: there were two instances where participants had calculated the correct answer in their rough work, but entered incorrect answers on the answer sheet.

Class 5: Error in interpreting data displays

Errors in interpreting tables, charts and graphs were common. The example shown in Figure 5.3 comes from a question based on a table taken from the National Institute for Health and Care Excellence (NICE) clinical guidelines on IV fluid replacement therapy (NICE, 2013).

Normal daily fluid and electrolyte requirements are summarised in the table below.

Water	25-30 ml/kg/day
Sodium, Potassium, Chloride	1 mmol/kg/day
Glucose	50-100g/day

Craig weighs 70kg. What is his approximate daily requirement of water, sodium and glucose?

Figure 5.3. Question based on NICE guidance.

The facility of this question was 0.53, and 19/27 (70%) of the errors occurred because candidates multiplied either 50g or 100g x 70 to reach 3500g or 7000g respectively. This highlights a weakness in the design of the table: while the data for water, sodium, potassium and chloride is presented in units required per kilogram of body weight per day, glucose requirement is shown as the total amount needed, with no need for further calculation. Standardising the table so that the glucose requirement is presented as 1g/kg/day would eliminate confusion.

Unclassifiable errors

Some questions were unclassifiable because they were unanswered, or the answer was unusual, and no rough work was provided. In other cases, different strategies could be used to reach the same answer e.g. in a case where the correct answer was 8mg, several candidates gave the answer 0.8mg; in some cases, rough work showed that they had made arithmetical errors, and miscalculated 80/10 = 0.8, while others had set the calculation up incorrectly as 80/100 = 0.8. Therefore, when no rough work was provided, the answer 0.8mg was unclassifiable.

DISCUSSION

We have investigated low CN in medical students. Our research aims were: 1) to develop a classification system that precisely describes the errors made by participants in our CN test (Table 5.2); 2) to test this system by using it to reclassify the errors documented in nursing studies (Table 5.3); and 3) to use this system to document the type and frequency of mistakes being made by medical students (Tables 5.4 and 5.5); and 4) to consider the implications of our findings for medical education in relation to selection, educational intervention during training, and the workplace environment.

Classification of error

We have developed a classification system that describes the errors made by medical students in the MINT. We found five distinct classes of error, four of which can be sub-divided to allow a precise description of the type of error observed. However, since 108/602 (18%) errors were unclassifiable, it is possible that we may have omitted some categories of error; nonetheless, we consider that our strategy for classifying error, shown in Figure 5.1, is sufficiently comprehensive to detect the different types of error that could be seen in CN tests.

We applied our classification system to the types of error documented in studies of drug dose calculation in nursing, and found that it could be used to classify the observed errors. In most cases this was straightforward, since most of these studies described either two or three categories of error (Table 5.3). However, Koharchik *et al* (2014) documented eight different types of error; nonetheless, these could be assigned to three classes of error in our system. We believe that our classification system is an improvement on previous systems, because it is based on the strategies used by students, rather than defining errors based on the researchers' interpretation of findings e.g. 'misread' or 'misinterpretation' error. Furthermore, defining errors based on precise descriptions of what the participant has done should help in terms of delivering appropriate remediation.

We used our classification system to identify and document the type and frequency of mistakes being made by medical students. Errors in basic arithmetic (Class 1) were uncommon (35/494, 7%). It is somewhat counterintuitive that basic arithmetical errors accounted for such a low percentage of incorrect answers in a numeracy test; however, it is unsurprising that students entering university would have good basic mathematical skills. It was apparent that a range of factors contributed to arithmetical errors e.g. breaking down a sum into several smaller steps introduced additional steps in which errors could occur; furthermore, some errors were related to calculator use. These factors have also been noted to contribute to error in nursing studies (Galligan *et al* 2010).

The majority of incorrect answers were due to Class 2 errors (323/494, 68%) involving a failure to set up the problem correctly. Weeks *et al* (2000) observed that errors made by nursing students in drug dose calculation tests were most frequently due to failure "to grasp the logic of the problem to be solved" leading to set-up errors; Galligan *et al* (2010) report similar findings. Tariq (2008) found that first year bioscience students also made basic errors resulting from a lack of "problem-solving skills rather than mathematical ability". Therefore, our findings are consistent with the literature: medical students make similar errors in numeracy tests to those made by bioscience and nursing students. This may result from lack of practice: Lee *et al* (2010) reported that entrants to university who stopped studying maths at General Certificate of Secondary Education (GCSE) level had retained little of their mathematical knowledge on arrival at university. Evidence from nursing studies also indicates that a lack of practice at calculations leads to deskilling (Hughes & Edgerton, 2005; McMullan *et al* 2010), as does the finding that some specialties outperform others (Rolfe & Harper, 1995; McMullan *et al* 2010). This is important in terms of medical education, as it suggests that medical students may benefit from a refresher course in maths on arrival at university, as well as ongoing relevant tuition in CN. Furthermore, the finding that set-up or problem-solving errors are so common is potentially of concern, since problem-solving is a key skill for doctors.

Measurement errors (Class 3) accounted for 66/494 (13%) observed error, of which 58/66 (88%) were due to mistakes in converting between units (Class 3A) and 8/66 (12%) involved expressing an answer in incorrect units (Class 3B). This class of error is important clinically, because doctors often need to perform complex calculations, including converting between units, to determine the correct volume of an IV drug to administer to a patient. This is because drugs for IV administration may be labelled variously as mass per unit volume (atropine 0.6mg/ml), as percentage (lignocaine 2%), or as a ratio (adrenaline 1:1000), while the drug dose is generally prescribed in mg per kg body weight. Other researchers have observed that medical students and doctors find converting between units difficult (Wheeler *et al* 2004b, Harries & Botha, 2013), and a similar problem is documented in nursing (Weeks *et al* 2000; Johnson & Johnson, 2002; Koharchik *et al* 2014), bioscience (Tariq, 2008), and pharmacy students (Latif & Grillo, 2002; Batchelor, 2007; Malcolm & McCoy, 2007; Hegener *et al* 2013). A smaller number of measurement errors resulted from expressing units in the wrong format; this also has important implications for clinical practice, as a measurement error involving a drug available in a solution of 10mg/2ml could result in the patient being given 5ml (25mg) rather than 5mg (1ml).

There were only two Class 4 (transcribing errors), thus this category may be superfluous. However, Class 5 errors, those relating to interpreting data displays were the second most common type of error in the MINT. This class of error has not been reported in nursing studies, although Galligan *et al* (2010) identify the ability to interpret charts and graphs among seven key mathematical skills required by nurses, suggesting that this category would be included in research involving a comprehensive test of CN in nurses. However, published CN tests in nursing appear to be restricted to studies of drug dose calculation ability. The prevalence of Class 5 errors is important since doctors are frequently required to interpret and act on data presented in tables, charts and graphs. The data interpretation questions in the MINT are straightforward, thus students' difficulty is likely to represent a lack of practice at this skill, due to a lack of appropriate training. The need to include numeracy in curricula teaching Evidence-Based Medicine is already well recognized (Ghosh & Ghosh, 2005; Rao & Kanter, 2010; Johnson *et al* 2014), and supported by our findings.

We found evidence that some students appeared to find the clinical context of some questions confusing. This supports the findings of Tariq (2008), who observed that although staff consider that contextualising a mathematical problem will help students, by making it more meaningful, students find contextualised problems more difficult. We found that students were challenged by a question regarding a patient with Diabetic Ketoacidosis, a common medical condition. The facility of this question was 0.53, compared to a facility of 0.90 in our study with qualified doctors (Taylor & Byrne-Davis, 2017); this difference is statistically significant (p<0.01 with Bonferroni correction). The difference in performance is likely to be due to the students' lack of clinical experience, compared to that of the doctors, who would all have had experience of managing such a case in clinical practice. This finding highlights the importance of being able to conceptualise the data given in CN questions, as documented in nursing research (Johnson & Johnson, 2002; Wright, 2007; Coyne *et al* 2013; Weeks *et al* 2013b). Therefore, medical educators must ensure not only that students are competent in the maths required to calculate drug doses, but also that they understand the clinical context and equipment involved. Moreover, teaching drug dose calculation may be more meaningful to students with clinical experience, hence it may be more effective if delivered in the clinical years.

A further issue highlighted by our study is that many answers were so implausible that participants should never have suggested them e.g. an answer of 126kg as the daily requirement of glucose for a 70kg patient; an answer of 26,000 years as the time it would take for a reduction in death rate by one, where graph shows a decline of 10 deaths over a 30 year-period. There were numerous cases like these, where an error would have been apparent had participants checked their answers. The failure of students to check that their answers make sense is well recognised (Johnson & Johnson, 2002; Tariq, 2008).

Our findings shed some light on the questions posed in the first paragraph of this paper. In relation to selection for medical school, our results are consistent with the literature: medical students make similar errors in numeracy tests to those made by bioscience and nursing students. This suggests that the level of numeracy in medical students entering university is comparable to that of students entering other healthcare/science disciplines. However, this does not mean that their numeracy is appropriate or sufficient for third level studies; indeed the concern about levels of numeracy in entrants to university suggests the opposite (Batchelor, 2007; Malcolm & McCoy, 2007; Tariq, 2008; Lee *et al* 2010; Sikorskii *et al* 2011; Young *et al* 2013; Hegener *et al* 2013; Roohr *et al* 2014; Galligan & Hobohm, 2015). Therefore, securing a place in medical school does not ensure an entrant has good numeracy: this should be assessed, and remediated if necessary, during undergraduate studies.

Our results also add to the evidence that CN should be included in medical curricula, as advocated by many researchers to ensure patient safety (Ghosh & Ghosh, 2005; Wheeler *et al* 2007; Gigerenzer *et al* 2007; Rao & Kanter, 2010; Harries & Botha, 2013; Johnson *et al* 2014). Our results also provide some insight into the type of education needed: Johnson & Johnson (2002) reported that four key skills were required for successful drug dose calculation: computation, conceptualisation, conversion and critical analysis of the process and the answer. They have used this framework successfully to achieve proficiency in drug dose calculation in nursing students (Johnson & Johnson, 2002). Our results suggest that this framework, particularly emphasising the importance of critical analysis of both the process and the answer, would be useful to help medical students and doctors achieve competence in CN. Furthermore, we note that Galligan (2013), emphasises the difference between school maths and “academic numeracy” relevant to a particular third level discipline. They suggest that moving the focus away from tedious revision of school maths to a skill necessary for the workplace will help students engage with the topic; we consider that this recommendation would be helpful in medical education.

The third question posed was whether the healthcare system contributed to errors related to CN. We have found some evidence to suggest that factors in the workplace may contribute to error. These relate to the lack of standardisation of the labelling of IV drugs, an area previously well-researched (Wheeler *et al* 2004b; Harries & Botha, 2013), but also to the lack of standardisation of information provided in clinical guidelines. Both areas are amenable to standardisation.

Finally, we have reviewed the different classes of errors we observed in the MINT, and compared them with the categories of error described by Reason (2000) (Table 5.1). We consider Class 1 (basic arithmetical errors) to be knowledge-based errors, while Class 2 (set-up errors) and Class 3A (conversion errors) are skill-based errors, and Class 3B (answer expressed in incorrect format) are rule-based errors. Although none of the errors made in the MINT are likely to be deliberate violations, Class 2E errors could perhaps be considered a form of violation. However, further research is required to determine whether this is helpful i.e. whether human factors science could help further understand and address CN errors.

Limitations

This research was restricted to medical students in a single institution in the UK, and involves a cohort of 58 participants; therefore, our results must be interpreted with caution. However, MINT scores have been consistent in four separate studies comprising a total of 480 medical students and doctors (Taylor & Byrne-Davis, 2017; Taylor *et al*, *in press*; unpublished data from this thesis), thus our results are likely to be generalisable to UK medical students and doctors.

Our investigation into error was conducted by analysing the rough work (RW) relating to incorrect answers in a classroom test of CN. Not all respondents provided useful RW, thus we were unable to determine the cause of error in 108/602 (18%) cases; this is a significant proportion, and may introduce bias into our results. This could be addressed by conducting a "Think Aloud" (TA) study (Cotton & Gresty, 2006) in which participants discuss their thinking at each step of the process of formulating their answer; however, there is debate about the TA method, with some suggestion that it facilitates clearer thinking, and thus introduces bias. Finally, we cannot tell whether errors made in a classroom test such as the MINT are generalisable to performance in clinical practice.

CONCLUSION

We have explored the causes of error in our test of CN in medical students and doctors, and identified five distinct classes of error. Our results show that the majority of errors made in our CN test were due to an inability to set up the calculation correctly; these errors were similar to those made by nursing and bioscience students. Our results support the evidence that students entering university have become deskilled in mathematics through lack of practice. Therefore, we suggest that numeracy should be included in medical curricula.

Blank Page

CHAPTER 6

DISCUSSION

TABLE OF CONTENTS Chapter 6

SECTION 1. INTRODUCTION

I have presented the results of my research into CN in medical students in detail in the preceding chapters of this thesis. I will now summarise my findings, and consider their relevance to the literature on CN, and to medical education and patient safety. Since chapters 3-5 of this thesis were prepared for publication in different journals, I was limited by word counts in terms of reporting my findings; furthermore, there were some areas of potential interest that I was unable to explore due to the need to focus the discussion on the specific research question posed in the paper. As a result, some observations including the impact of gender on performance, and the effect of dyslexia on test scores have not been discussed. These are reported in Appendices 5 & 6.

Throughout this chapter, I will refer to the original version of the test as MINTv1, to the revised CR version, as MINTv2, and to the revised SBA format as MINTv3; where the discussion centres on the test in general rather than on a specific format, I refer to the test simply as the MINT.

SECTION 2. CN IN HEALTHCARE PROFESSIONALS

There is growing awareness that the level of numeracy in the general public both in the UK and the US is often very low (www.nationalnumeracy.org.uk; Reyna *et al* 2009; OECD, 2013). This has been extensively investigated in healthcare because of the association between low numeracy in patients and adverse outcomes from various disease processes (Gazmararian *et al* 2003; Gazmararian *et al* 2005; Apter *et al* 2007; Weiss *et al* 2005; Reyna *et al* 2009; Rowlands *et al* 2013). The Programme for International Student Assessment (PISA) (OECD, 2018) indicates that 15-year olds in the UK lag behind their contemporaries worldwide in terms of their numeracy. Recent research has shown that university students in the UK often have lower numeracy than is required for their planned undergraduate courses (National Numeracy, 2019). Thus it is timely to consider the numeracy of UK medical students.

There has been an awareness for many years that numeracy among those in the healthcare professions may be low, although the problem has often been seen to relate primarily to nurses and nursing students. Certainly the issue of low numeracy in nursing is well documented and researched (Weeks *et al* 2001; Johnson & Johnson, 2002; Wright 2007). However, despite concerted efforts to address it (Johnson & Johnson, 2002; Young *et al* 2013 Weeks *et al* 2013 a, b, c; Sabin *et al* 2013; McDonald *et al* 2013), the problem appears to persist (Fleming *et al* 2014; Bagnasco *et al* 2016; Hurley, 2017). There is evidence that numeracy in pharmacy students and pharmacists may also be low (Latif & Grillo, 2002; Batchelor, 2007; Malcolm & McCoy, 2007; Hegener *et al* 2013). Research has shown that numeracy in medical students worldwide is often lower than expected (Sheridan & Pignone, 2002; Wheeler *et al* 2004b; Harries & Botha, 2013), and similar concerns have been raised in relation to qualified doctors (Rowe *et al* 1998; Selbst *et al* 1999; Windish *et al* 2007; Simpson *et al* 2009; Wegwarth *et al* 2012; Johnson *et al* 2014; Taylor & Byrne-Davis, 2017). However, despite these findings, little was understood about CN in medical students and doctors, and no research had investigated why it might be low, or what kind of errors were occurring.

Nonetheless, it was recognised that low CN in doctors was a potential cause of patient harm, and many researchers had recommended that CN be introduced into undergraduate and postgraduate medical curricula (Wheeler *et al* 2004b; Harries & Botha, 2013; Johnson *et al* 2014; Taylor & Byrne-Davis, 2017).

During the course of my research for this thesis, I have assessed CN in three separate cohorts, comprising a total of 341 third year medical students. Test scores have been consistent across all groups of students, and they are also comparable to the results of my original research with MINTv1 (Table 2.17) (Taylor & Byrne-Davis, 2017). My results confirm that many medical students have deficiencies in CN, supporting the findings of other researchers worldwide (Sheridan & Pignone, 2002; Wheeler *et al* 2004b; Harries & Botha, 2013). Furthermore, these findings are consistent with recent evidence regarding low numeracy in UK university students (National Numeracy, 2019). Therefore, my research adds to the evidence that CN should be addressed within medical curricula, as suggested by Wheeler *et al* (2004b), Harries & Botha (2013) and Johnson *et al* (2014).

My research supports the evidence that CN is a complex construct. Numeracy is a complex skill, involving the ability to use quantitative information to interpret data, make decisions and solve problems in everyday life (www.nationalnumeracy.org.uk); similarly, CN is a complex construct involving the ability to use quantitative information in healthcare (Caverly *et al* 2012). Johnson & Johnson (2002) observed that accurate drug dose calculation requires not merely the arithmetical ability to perform the calculation, but also the ability to conceptualise the problem, and where necessary to convert numbers between different units of measurement. Furthermore, they emphasise the importance of critical analysis of the answer to ensure that it is correct (Johnson & Johnson, 2002). My analysis of error suggests that these skills are applicable to analytical and statistical as well as computational constructs. Moreover, deficiencies in any of these skills can lead to errors, and potentially cause patient harm.

That CN is complex should not of itself pose a problem for medical students or doctors: most clinical tasks are complex, but with appropriate training and practice, competence can be achieved. My literature review suggests that the problem of low CN is not generally acknowledged in medical education. There has been relatively little research in this area, and when I have presented the results of my research, medical educators tend to recognise the problem, but also to be surprised that this is an issue of concern worldwide. Perhaps the approach recommended by Galligan & Hobohm (2015) should be adopted in medicine: they have introduced the concept of “academic numeracy” to refer to the numerical competence required for professional practice, and to distinguish it from school maths. The term “academic numeracy” may help achieve the engagement of both students and educators in recognising that this is an important skill that warrants a place on the curriculum.

SECTION 3. THE ASSESSMENT OF CN

CN is not regularly assessed in medical students and doctors, nor has a required standard of competence in CN been set; thus there is no agreed assessment measure. However, medical educators may assume that various other assessments are sufficient, thus rendering a separate assessment of CN unnecessary. Such assessments could include ‘A’ grades in national examinations including General Certificate of Secondary Education (GCSE) and General Certificate of Secondary Education - Advanced level (A-level) mathematics. However, there is evidence that these qualifications are not associated with performance in tests of CN (Ben-Shlomo *et al* 2004; Taylor & Byrne-Davis, 2017). Furthermore, although aspiring medical students must sit the UKCAT (www.ukcat.ac.uk) which contains some challenging numeracy questions, this content is limited. Moreover, many medical schools do not consider UKCAT scores when offering places to students. The Prescribing Safety Assessment (PSA) (BPA & MSC, 2013) is mandatory for final year students, and assesses proficiency in prescribing; however, as discussed in chapter 1, a candidate can pass the PSA without completing any of the drug dose calculation questions. Therefore, none of these national assessments assesses CN in medical students.

My research has included an evaluation of twelve CN assessments used in medical students and doctors; this revealed that the MINT (Taylor & Byrne-Davis, 2016) was the best available assessment measure. Although psychometric analysis of the MINT indicated that it is a reliable and valid test (Taylor & Byrne-Davis, 2016), peer reviewers suggested that some of the test material could be improved; moreover, some questions were subject to copyright, and needed to be replaced. Since the interpretation of my results may have implications for medical education, it is vital that test material in the MINT is of high quality. Therefore, I conducted a comprehensive review of the MINT, leading to the development of a new version, MINTv2, a constructed response (CR) test. The revised test paper was evaluated by nine independent reviewers with experience in clinical medicine and/or medical education. These reviewers advised on the clarity of test material, the construct and level of difficulty of test questions, and where appropriate, the clinical relevance of questions. Having analysed data from participants who sat MINTv2, I developed a single best answer (SBA) version of the test, MINTv3, using evidence-based distractors.

Psychometric analysis of MINTv2 and MINTv3 demonstrated that both of the revised tests are reliable and valid. Despite the comprehensive revision process in which 9/43 new questions were introduced, and 23/34 of the original questions were amended, overall test scores remained similar to those of the original test (Table 2.17). This appeared to be a disappointing outcome initially, given the amount of work involved in the revision process. However, the consistency of test scores supports the quality of the original material in MINTv1, and dispels doubt about potentially misleading text and data displays. Additionally, the similarity in test scores suggests that the new material that I developed for MINTv2 is equivalent to that of MINTv1 in quality and level of difficulty. Furthermore, it suggests that the level of CN in medical students is similar to that of doctors in their first years after qualification, suggesting that additional training and clinical experience does not affect CN.

The importance of assessing the level of difficulty of the MINT is important in terms of interpreting test results: I had not considered the MINT to be a difficult test; thus a mean test score of 76% for MINTv1 was a source of concern. However, when I presented data at medical education conferences, many educators considered a score of 76% to be high, since this would be an excellent result in many university examinations. Therefore, confirmation from both the subjective emendation process, and from using the objective criteria described by Close *et al*

(2008), that the MINT is not a difficult test was important, as it supported my view that a mean score of 76% was unimpressive, and indicated overall low numeracy. Moreover, this score demonstrates that participants are making mistakes in easy questions. Had the assessment of test difficulty indicated that the MINT was challenging, I would have had to revise the interpretation of my research.

Another consideration in terms of interpreting test results was the issue of calculators. When I presented results of my original research to medical educators, the consensus view was that that participants would have achieved higher scores had they been allowed to use calculators. Furthermore, since calculators are readily available in the clinical workplace, it was suggested that it was unrealistic and unfair not to allow participants to use calculators to perform calculations. Researchers vary on whether or not it is appropriate to allow participants in drug dose calculation tests to use calculators: some researchers consider that they will overestimate mathematical ability (McMullan *et al* 2010; Bagnasco *et al* 2016), while others argue that calculators are generally available and so should be allowed (Coyne *et al* 2013; Fleming *et al* 2014). Furthermore, results are variable; although Shockley *et al* (1989) and Bliss-Holtz (1994) found that scores were better when calculators were used, both Murphy and Graveley (1990) and Tarnow and Werst, (2000) found that calculators made no difference.

I conducted a randomised controlled trial (RCT) to assess whether scores on the MINT would be improved if participants used calculators; my results showed that having access made no difference to overall test score. This finding suggests that MINT content is largely calculator inappropriate; furthermore, it suggests that the errors being made are not primarily arithmetical in nature. This result should be helpful to others researching CN in healthcare professionals; allowing calculators may not overestimate performance, or conceal error. In practice, I would suggest that participants should be allowed access to calculators for CN tests.

Previous research suggested that the single best answer (SBA) format inflates test scores because of cueing and guessing (McCoubrie, 2004; Betts *et al* 2009; Simkin & Kuechler, 2005; Jordan, 2013; DiBattista *et al* 2014; Sam *et al* 2016; Sam *et al* 2018). Having developed evidence-based distractors for the MINT, I conducted an RCT to assess the impact of test format, finding no difference in test scores. This is consistent with research demonstrating the value of developing high quality questions and distractors for SBA tests (Downing, 2003; DiBattista & Kurzawa, 2011). My findings suggest that with evidence-based distractors, an SBA test can be equivalent to a VSA format, although it is not clear whether my results are generalisable to SBA tests in all situations. However, this result demonstrates that it is appropriate to use the SBA format for my future research; this is important, since the SBA format is logistically easier to deliver, and is feasible for an online test.

SECTION 4. ERRORS IN TESTS OF CN

Research into error in healthcare shows that errors may be caused by individual or system failures, although very often errors are multifactorial (Reason 2000) (Table 1.4). Identifying the cause of errors in healthcare is essential in order to develop effective strategies to prevent and manage them, and thus to improve patient safety. This is important for all kinds of errors, including those related to deficiencies in CN. There is evidence from the literature of efforts to improve CN at the individual level through educational remediation, and at the systems level with innovations such as Tallman lettering to distinguish drugs with similar names, and the introduction of e-prescribing to reduce medication error.

However, in comparison to nursing practice, where educators have accepted that CN may be low in their students and graduates, and extensive research has been conducted into the causes of error and approaches to remediation (Johnson & Johnson, 2002; Galligan *et al* 2010; Weeks *et al* 2013 a, b, c; Young *et al* 2013; Koharchik *et al* 2014; Simonsen *et al* 2014; Galligan & Hobohm, 2015; Mackie & Bruce, 2016; Hurley, 2017), these areas are relatively unexplored in medical students and doctors (Wheeler *et al* 2006; Freeman *et al* 2008; Wheeler *et al* 2008; Ross & Loke, 2009; Harries & Botha, 2013). Therefore, one of the aims of my research was to investigate the errors being made in the MINT, as this could provide useful information to support the development of appropriate educational intervention to improve CN in medical students and doctors. Nonetheless, I was concerned that this might not be worthwhile, since a study by Harries & Botha (2013) suggested that remediation for CN might be ineffective in medical students. In this study, Harries & Botha (2013) found that 125/364 (34%) students never became competent in drug dose calculation, despite repeated teaching and retesting over a two-year period. This suggested not only that a large proportion of medical students had low CN, but more importantly that they could not improve. However, a review of educational intervention in medical education conducted by Cleland *et al* (2013) found that remediation based on repetition and retesting of the same material was often ineffective. Furthermore, in order to successfully remediate for numeracy, it is best to start by performing a learning needs analysis (Wallace, 2019). The strategy for remediation used by Harries & Botha (2013) involved repetition and retesting drug dose calculations, and thus falls into the category deemed ineffective by Cleland *et al* (2013); furthermore, since they do not document that they carried out a learning needs analysis, it is likely that their interventions were not tailored to their learners' needs. Therefore, the lack of improvement they observed may have been due to using inappropriate educational intervention, rather than students' inability to improve.

Review of the nursing literature was encouraging, as a diverse range of strategies from classroom teaching to e-learning courses had been successfully used to improve CN in nursing practice; these are summarised in Table 6.1. Furthermore, there is some evidence that remediation can be effective in improving CN in medical students and doctors. Freeman *et al* (2008) found that an innovative curriculum including videos and animations improved medical students' engagement with the course and their understanding of statistics. Additionally, in their review of approaches to improving prescribing in medical students, Ross & Loke (2009) found that all interventions improved performance. Therefore, having reviewed the literature, I concluded that it would be worthwhile to conduct an exploration of error, as the weight of evidence suggested that remediation would be effective if developed to meet learners' needs.

My exploration of error involved reviewing the test papers of participants who sat the VSA version of the MINT in 2018. I found that available frameworks for considering error in CN tests were unsuitable for the MINT, and hence developed my own classification system as described in chapter 5. The main finding of this part of my research was that the vast majority

Table 6.1. Educational intervention to improve numeracy

Intervention	Authors (year)	No.	Study group	Area	Outcome
Classroom teaching & workbooks	Hutton (1998)	99	Student nurses	Numeracy	Intervention improves performance & confidence
E-learning	Weeks *et al* (2001)	N/A	Student nurses	Drug dose calculation	Improves conceptualisation of clinical environment
Classroom teaching	Johnson & Johnson (2002)	>100	Student nurses	Drug dose calculation	Improves performance and understanding
Classroom & clinical teaching	Wright (2005)	71	Student nurses	Drug dose calculation	improves performance
E-learning (Literature Review)	Cook *et al* (2008)	201 studies	Healthcare Profess-ionals	Healthcare	All interventions have positive impact; e-learning similar to others
Videos, animations, workbooks	Freeman *et al* (2008)	325	Medical students	Statistics	Innovative curriculum improved performance
Classroom, online & simulation	Wheeler *et al* (2008)	72	Medical students	Drug dose calculation	Improved performance, but long-term effect unknown.
Literature Review	Ross & Loke (2009)	15 studies	Medical students	Prescribing	Intervention improves performance, but most studies had small numbers
Classroom teaching	Coyne *et al* (2013)	156	Student nurses	Drug dose calculation	Improved performance and understanding
Online safeMedicate	Weeks *et al* (2013)		Student nurses		Improved performance
Classroom teaching & workbooks	Harries & Botha (2013)	364	Medical students	Drug dose calculation in medical students	157 (43%) became competent; 125 (34%) did not.
Simulation (review)	Zahara-Such (2013)	15 studies	Student nurses	Drug dose calculation	Improves confidence and problem-solving
Clinical skills workshop	Grugnetti *et al* (2014)	77	Nurses	Drug dose calculation	Improved performance and understanding
Classroom teaching	Koharchik *et al* (2014)	75	Nurses	Drug dose calculation	Improves performance
Classroom v e-learning	Simonsen *et al* (2014)	183	Nurses	Drug dose calculation in nurses	No difference, except poor performers, where classroom better
Literature Review	Stolic (2014)	20 studies	Student nurses	Drug dose calculation	All strategies improve performance
Multimodal curriculum with self-audit &reflection	Galligan *et al* (2010)		Student nurses	Academic numeracy for nursing	Increased confidence and competence
E-learning	Mackie & Bruce (2016)	16	Student nurses	Drug dose calculation	Intervention improves performance
Experiential v classroom teaching	Hurley (2017)	76	Student nurses	Drug dose calculation	Experiential programme better than classroom teaching

of errors made by medical students were set up or problem-solving errors (Class 2), rather than basic arithmetical mistakes (Class 1). Class 2 errors occur when the student has not extracted the correct information from the text of the question to solve the problem, and/or does not understand how to set up the calculation. This is a basic error, and researchers have found that bioscience students (Tariq 2008) and nursing students (Galligan *et al* 2010; Galligan & Hobohm, 2015) make similar errors; moreover, Tariq (2008) observed that these errors are similar to those made by schoolchildren. Since my research has shown that the MINT is not a difficult test, the finding that students are making basic errors suggests that their mathematical skills have deteriorated since leaving school; this has previously been documented in relation to engineering students (Lee *et al* 2010). The evidence of deskilling suggests that medical students and doctors would benefit from practice in CN during their training, in order to maintain their skills.

Another important finding was that medical students appeared to be confused by contextualised questions: this was particularly evident in relation to a question based on a scenario of diabetic ketoacidosis, as discussed in chapter 2 (p. 80). This effect is well-documented in bioscience (Tariq, 2008) and nursing students (Weeks *et al* 2001; Johnson & Johnson, 2002; Wright, 2007b). The effect of context suggests that training in CN should be contextualised, rather than simply focussing on basic maths skills, and revising how to set up mathematical calculations. Medical students need training and/or clinical experience in order to understand how to approach contextualised questions i.e. how to extract the relevant information from the text of the question. This was an unexpected finding: when I submitted my proposal for this research to the UREC, I stated that students who had difficulty with the MINT would be directed to educational resources such as Hegartymaths (www.hegartymaths.com) or BBC bitesize (www.bbc.co.uk/education/subjects/z38pycw) to improve their skills. I was confident that this would be appropriate, since the mathematical component of MINT questions is at or below the level required for GCSE. However, following the results of my exploration of error, I would no longer recommend these resources, except to support students who are primarily making Class 1 (basic arithmetical) errors. An alternative resource that may be more appropriate is sn@p (www.sn@p.org) an online educational tool designed to help healthcare professionals with drug dose calculation.

Although arithmetical errors were relatively uncommon (7%), the errors here were basic. While it is tempting to consider that they arose from carelessness, analysis of students' rough work suggested that students often attempted to break a sum down into multiple smaller steps; this strategy is commonly taught to schoolchildren to make calculations easier. However, I found that students who used this method often made errors; Galligan & Hobohm (2015) observed that nursing students made similar mistakes.

I found that students often had difficulty in converting between different units of measurement (Class 3 errors); this is well-documented and thus supports the existing literature calling for standardisation of drug labelling and presentation (Wheeler *et al* 2004a; Harries & Botha, 2013). Class 5 errors related to inability to accurately interpret data presented in tables, charts and graphs; these accounted for 14% of errors. This finding supports evidence that medical students and doctors may struggle with interpreting biostatistical information (Gigerenzer *et al* 2007; Windish *et al* 2007; Wegwarth *et al* 2012; Johnson *et al* 2014); however, since MINT test material is not difficult, this result suggests that medical students struggle with more basic data than has previously been reported. This suggests the need for appropriate educational intervention.

Another important observation was that students often gave completely implausible answers, indicating a failure to sense-check their calculations; this phenomenon has been observed in other student groups (Tariq, 2008; Galligan *et al* 2010), and Johnson & Johnson (2002) include critical evaluation of calculations as one of four essential skills required for competence in drug dose calculation. Failure to sense-check a calculation has potentially serious implications for patient safety e.g. an error of one decimal place will result in a drug dose ten times higher or lower than required, with potentially catastrophic consequences.

Finally, my investigation into error demonstrates that medical students are making the same kinds of errors as nurses and nursing students; therefore, it is likely that the educational interventions used in nursing curricula would be effective in medical students and doctors. Kalet *et al* (2016) discuss the factors that lead to effective remediation in medical education, highlighting the importance of setting clear goals, of a supportive coaching relationship, and of reflective practice. Furthermore, they emphasise that there should be an agreed competence framework. This is an important consideration, and since there is no standard of CN required for medical students and doctors, I would recommend using the framework for competence described by Johnston & Johnston (2002). I would also support the educational approach suggested by Kalet *et al* (2016), and suggest that medical educators adopt the attitude of Galligan and Hobohm (2015), and consider CN as an academic attribute, similar to many clinical skills.

SECTION 5. LIMITATIONS

I recognise that there are some limitations to my work. In the first place, this research is limited to students in one institution. However, the validation of the original MINT (Taylor & Byrne-Davis, 2016), the consistency of my results in 341 participants over a three-year period, and the psychometric data presented in Chapter 2 of this thesis indicate that the MINT is a reliable and valid assessment of CN. Therefore, the evidence is that the findings of my research do reflect students' actual CN. Since the academic criteria for entry to medical school are similar nationally, my results should be generalisable to UK medical students. This is supported by the similarity of my findings to those of my original study with foundation trainees, which included graduates from 26 different UK medical schools (Taylor & Byrne-Davis, 2017).

A second limitation is that the emendation process that I used in evaluating MINTv2 did not involve a meeting of all panel members; each reviewer worked independently, and there was no round table discussion. This did not matter greatly in terms of assessing text and data displays, as I collated all of the information; however, a discussion and exchange of perspectives on allocating construct and level of difficulty to test questions would have been interesting. Nonetheless, I consider that the process of assigning level of difficulty by using the mean and median ratings of reviewers was robust, particularly since it was supported by the second process based on facility of test questions (Close *et al* 2008).

Another limitation is that the MINT is a classroom test, and therefore test scores may not represent performance in practice; some researchers argue that individuals will perform better in the clinical environment (Wright, 2007a). However, it is remarkably difficult to get a true picture of performance in the workplace: people who know they are being observed may “up their game’ while being watched, demonstrating better performance than is normal when unobserved. I consider that performance in a quiet classroom environment should allow optimum performance, and is likely to be a good reflection of knowledge/ability. All medical schools use classroom tests and artificial environments (e.g. Objective Structured Clinical Examinations (OSCEs)) to assess undergraduates and declare them fit to practice. Therefore, a classroom test is an appropriate way to assess CN.

Finally, my exploration of error was limited by the lack of rough work provided by participants, 18% of whom either gave no rough work, or provided data that was insufficient to demonstrate their approach to the calculation. Thus I may have failed to recognise and document some forms of error.

SECTION 6. IMPLICATIONS FOR FUTURE RESEARCH

The completion of a robust emendation process of the MINT has provided a valid and reliable instrument for further research into CN. The results of my research have demonstrated that there is no advantage in having calculators, and that the SBA format is equivalent to the VSA version of the test. Therefore, for future testing, participants can use calculators, and the test can be delivered in its SBA format. This will allow the development of an online version of the MINT, which will increase the capacity for testing CN more widely amongst medical students, doctors and other healthcare professionals. My research has provided a greater understanding of how and why medical students are making errors in CN tests, demonstrating that the majority of errors are set-up errors rather than errors in basic mathematics. This finding will be helpful in terms of developing appropriate educational intervention to ensure that medical students and doctors have a level of CN that is appropriate for safe clinical practice. However, there are several areas yet to be explored, as outlined below.

Analysis of error

My analysis of error is incomplete, since in 18% of answers, there was insufficient rough work (RW) to assess the strategies used by students to answer questions. Although students were encouraged to provide RW, it may be worth changing the scoring system for the test and offering marks for RW. However, the missing data includes situations where some RW was provided but was inadequate in terms of assessing methodology; thus rewarding RW with marks may not be helpful in practice. Perhaps a better strategy would be to conduct a Think Aloud study (Cotton & Gresty, 2006). Think Aloud is a technique involving one to one interviews between the researcher and individual participants, where the participant explains what they are doing step by step, as they go through the calculation. This would provide detailed information of the strategy used to answer the question, and could shed further light on how and why errors are being made in answering MINT questions. However, there are concerns that Think Aloud may act as an intervention, and that the process of talking through the problem can improve performance. Nonetheless, this process may improve the understanding of error in the MINT.

Evaluation of educational material to improve CN

Various researchers have developed educational material aimed at improving numeracy in nurses (Wright, 2005; Coyne *et al* 2013; Sabin *et al* 2013; Coben & Weeks, 2014; Koharchik *et al* 2014; Hurley, 2017), and in doctors (Wheeler *et al* 2006; Freeman *et al* 2008; Wheeler *et al* 2008; Ross & Loke 2009; Harries & Botha 2013) (Table 6.1). I would like to review this educational material, and to evaluate its usefulness for medical students and doctors, based on the results of my investigation into error. This could inform the development of an educational programme for medical students and doctors in training. Future research would include assessing the impact of this intervention.

To compare performance in tests of numeracy in medical students with performance on curricular tests throughout medical school

There is no research investigating whether there is an association between the performance of medical students in tests of CN, and their performance in other undergraduate tests e.g. the prescribing safety assessment, and end of year progress tests. This would be an interesting area to study; furthermore, it would be worth exploring whether a tendency to make certain kinds of error is associated with different types of behaviour in the workplace.

To compare the performance of doctors with that of other healthcare professionals

In my original research with MINTv1, I compared the performance of doctors with that of various non-medical populations on different subsets of the test; in all cases doctors performed better than other groups. However, I do not know how CN in doctors compares with that of CN in other healthcare professionals such as nurses and pharmacists. Therefore, I would like to use the MINT to investigate CN in other healthcare professionals and student healthcare professionals including pharmacists, physician associates and nurses.

Setting a standard of numeracy for doctors

There is no agreed standard of CN required for doctors; the lack of such a standard limits research and educational intervention to address CN. This area warrants further research.

SECTION 7. IMPLICATIONS FOR PRACTICE

My research has many implications for practice in relation to undergraduate and postgraduate education; these primarily relate to medicine, but some apply to other healthcare disciplines e.g. pharmacy and bioscience, or to education more broadly.

Assessing Clinician Numeracy

My research included an evaluation of twelve different tests used to assess CN in medical students and doctors. This is important, since during the course of my research, I found no literature that compared CN tests developed for medical students or doctors. Furthermore, although Coben & Weeks (2014) outline several important properties of CN tests for nursing practice, no researchers have previously considered the characteristics of an ideal CN test for medical students or doctors.

Having compared test format, length of test, content and difficulty of test material, I found that the MINT was the best of the twelve assessments. My further work in subjecting the MINT to a comprehensive internal and external review process demonstrates that it is well-structured, easy to deliver (and mark), with authentic questions testing competence in clinically important areas. Moreover, psychometric analysis shows that the MINT is a valid and reliable test. Therefore, the MINT can be recommended as a useful assessment of clinician numeracy.

Use of Calculators in Numeracy Tests

My research demonstrated that using a calculator did not improve overall test scores. This finding supports the observation that CN is a complex construct, involving four key areas of competence: calculation, conceptualisation, conversion and critical analysis (Johnson & Johnson, 2002). Therefore, although a calculator may help with a tricky calculation, it will not help with problem-solving, or with conceptualisation, converting between units, or critical analysis. Thus calculators will not mask deficiencies in CN. Clinician assessors are sometimes reluctant to permit the use of calculators in CN tests on the basis that calculators may bias test results (McMullan *et al* 2010; Bagnasco *et al* 2016); however, my results indicate that this concern is unfounded. Moreover, since calculators are readily available in clinical practice, they should be permitted in CN tests in the interest of providing a realistic assessment.

Assessment Format

My research also has implications in relation to assessment format. I developed evidence-based distractors for the SBA version of the MINT, and my randomised controlled trial found that the SBA and VSA versions of the MINT were equivalent. This supports the evidence that using high quality distractors can improve the quality of an SBA test. Therefore, educators can be confident in continuing to use SBA tests when high-quality distractors are used.

Error in CN tests

I have developed a new system for classifying error in CN tests, based on analysis of the observed error. This classification system can be used in CN tests other than the MINT, and will provide insight into the kind of errors being made by participants. Understanding how the error has occurred will help educators determine the most appropriate educational intervention required for individual participants.

CN and medical curricula

My research indicates that CN in medical students is lower than might be expected, and that this applies to all three numeracy constructs; computational, analytical and statistical numeracy. I have also found evidence that the errors made by medical students are similar to those made by undergraduate students from other healthcare disciplines; this suggests that educational interventions that have been successful in nursing are likely to be successful in medical students. Since CN is important for patient safety, and for medical decision making, it is essential that deficiencies in CN in medical students and doctors be acknowledged and addressed. It is time for clinician numeracy to be included in medical curricula.

Blank page

APPENDIX 1. University Research Ethics Committee approval letter

MANCHESTER
1824

The University
of Manchester

Ref: ethics: 16459

Dr Anne Taylor and Dr Lucie Byrne-Davis
Manchester Medical School
Stopford Building
University of Manchester
M13 9PL

Research Governance, Ethics and Integrity
The University of Manchester
Oxford Road
Manchester
M13 9PL

Tel: 0161 275 2206/2674
Email: research.ethics@manchester.ac.uk

17th November 2016

Dear Anne and Lucie

Study title: "An Investigation into Clinician Numeracy"

Thank you for attending the UREC 5 Committee on the 7th November 2016 and I am pleased to confirm a favourable ethical opinion for the above research on the basis described in the application form and supporting documentation as submitted by email on the 17th November 2016. This is now approved by the Chair on behalf of the Committee.
This approval is effective for a period of five years. If the project continues beyond that period an application for amendment must be submitted for review. Likewise, any proposed changes to the way the research is conducted must be approved via the amendment process (see below). Failure to do so could invalidate the insurance and constitute research misconduct.

You are reminded that, in accordance with University policy, any data carrying personal identifiers must be encrypted when not held on a secure university computer or kept securely as a hard copy in a location which is accessible only to those involved with the research.

Reporting Requirements:
You are required to report to us the following:

1. Amendments
2. Breaches and adverse events
3. Notification of Progress/End of the Study

Feedback
It is our aim to provide a timely and efficient service that ensures transparent professional and proportionate ethical review of research with consistent outcomes. In order to assist us with our aim, we would be grateful if you would give your view of the service that you have received from us by completing a feedback sheet https://survey. manchester.ac.uk/pssweb/index.php/779758/lang-en

We hope the research goes well.
Yours sincerely,

Patricia Gorham
Secretary to University Research Ethics Committee 5
Cc: Joanne Hart

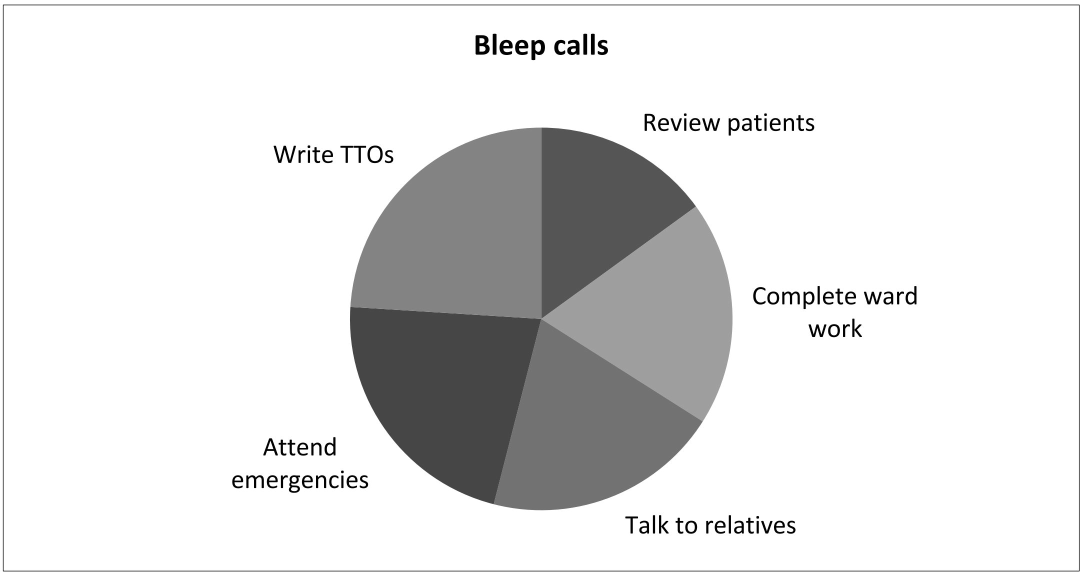
Blank page

APPENDIX 2. MINTv1

1. **Maria has diabetes and is planning to exercise in the gym for one hour. She needs to eat 6g of carbohydrate for every 30 mins she exercises. She has some biscuits in her gym bag. Each biscuit contains 8g of carbohydrate. How many biscuits should she eat before she exercises?**

   A. 1/2 biscuit **B**. 1 biscuit **C**. 3/4 biscuit **D**. 2 biscuits **E**. 1.5 biscuits

Item 2. This pie chart shows the distribution of bleep calls for a trainee on a medical firm.



2. **What is the trainee least frequently called to do?**

   A. Review patients **B**. Complete ward work **C**. Talk to relatives

   D. Attend emergencies **E**. Write TTOs

3. **There is a 2 in 100 chance of living 5 years or longer without treatment for a type of cancer. Drug X increases the chance of living 5 years or longer to 6%. Drug Y increases the chance of living 5 years or longer by 50%. If a patient wants the best chance of living 5 years or longer, which drug should be prescribed?**

   A. Drug Y **B**. Drug X **C**. Either drug, the chance of living longer is the same
   D. Neither drug, the chance of living longer is better without treatment
   E. Don't know

4. **Rose Turner has been referred by her GP with a history of weight loss. Her weight has dropped from 75kg to 67.5kg over the past 5 months. What percentage of her original weight has she lost?**

   A. 7.5% **B**. 20% **C**. 10% **D**. 15% **E**. 5%

5. **Mr Perkins is admitted to the gastroenterology ward and is prescribed an omeprazole infusion at a rate of 10ml/hr, to be continued for 72 hours. The solution infused contains 80mg omeprazole in 100ml NaCl. How much omeprazole does Mr Perkins receive every hour?**

A. 80mg **B**. 576mg **C**. 720mg **D**. 8mg **E**. 10mg

Item 6. Normal daily fluid and electrolyte requirements are summarised in the table below.

Water	25-30 ml/kg/day
Sodium, Potassium, Chloride	1 mmol/kg/day
Glucose	50-100g/day

6. **Cal weighs 70kg. Use the table above to determine which of the following statements regarding his daily requirements is true.**

A. He should have approximately 200ml fluid, 70 mmol sodium and 100g glucose.
B. He should have approximately 2000ml fluid, 70 mmol sodium and 3500g glucose.
C. He should have approximately 2000ml fluid, 70 mmol sodium and 80g glucose.
D. He should have approximately 2000ml fluid, 70 mmol sodium and 350g glucose.
E. He should have approximately 200ml fluid, 70 mmol sodium and 350g glucose.

Items 7 & 8. Imagine that 40 out of 1000 people are likely to develop disease Y over the next 5 years. Treatment A reduces the chance of getting disease Y by 25%. Treatment B reduces the chance of getting disease Y by 10%.

7. **Select the correct answer:**

A. Treatment A and Treatment B are equally effective
B. Treatment A is more effective than Treatment B
C. Treatment B is more effective than Treatment A
D. Don't know
E. Don't know

8. **What is the risk of developing disease Y after receiving Treatment A?**

A. 30:1000 **B**. 15:1000 **C**. 37:1000 **D**. 25:1000 **E**. Don't know

9. **The volume of a unit of blood is 330ml. If it is infused at a rate of 80ml/hr, approximately what proportion of the blood will have been transfused after 2 hours?**

A. 2/3 **B**. 1/2 **C**. 2/5 **D**. 1/3 **E**. ¼

10. **People being screened for a virus are told that the chance of testing positive is 1 in 1000. What percentage of people has the virus?**

A. 0.01% **B**. 1% **C**. 0.001% **D**. 10% **E**. 0.1%

Item 11. Alice and Dave O'Neill undergo genetic screening, and are given the following information regarding the risk of any child they conceive inheriting various diseases.

Inherited disease	**Risk**
Disease A	**0.01%**
Disease B	**0.001**
Disease C	**1:10 000**
Disease D	**0.001%**

11. **Which disease is their child least likely to inherit?**

A. Disease A **B**. Disease B **C**. Disease C **D**. Disease D **E**. Don't know

12. **Jess (age 19, weight 65kg) is admitted to the acute medical ward with a diagnosis of Diabetic Ketoacidosis. Clinical Guidelines state that she should be given insulin at a rate of 0.1 unit/kg/hr. You are asked to prescribe her insulin infusion. The preparation for infusion contains 50 units of insulin in a volume of 50ml saline. What rate of infusion will you prescribe?**

A. 65ml/hr **B**. 1ml/hr **C**. 0.65ml/hr **D**. 10ml/hr **E**. 6.5ml/hr

13. **Adult height for men can be estimated on the basis of parental height, using the following formula:**

$$\text{Adult height (cm)} = \frac{(\text{Mother's height} + \text{Father's height} + 13)}{2}$$

Most boys will reach an adult height within 10cm of this estimation. John is 6 years old. His mother is 152cm tall, and his father is 165cm tall. How tall is John likely to be when he is an adult?

A. 165 – 175cm **B**. 155– 175cm **C**. 160 – 180cm
D. 160 – 170cm **E**. 155 – 165cm

Items 14 – 15. Miss Strong, an orthopaedic surgeon, screens 100 patients for MRSA preoperatively. Ten of these patients are actually MRSA carriers. The test used gives a true positive result in 90% of MRSA carriers, and a false positive result in 20% of people who do not carry MRSA.

14. **How many of these 100 patients are expected to test positive?**

    A. 9 **B**. 80 **C**. 27 **D**. 72 **E**. 90

15. **What percentage of those who test positive actually carry MRSA?**

    A. 72% **B**. 33% **C**. 10% **D**. 90% **E**. 9%

16. **You are asked to randomise patients for a drug trial by tossing a coin. If the coin lands head up, the patient will receive Drug D, while if it is tails, the patient will be given the placebo. You will be recruiting 1000 patients. Approximately how many are likely to receive Drug D?**

    A. 250 **B**. 50 **C**. 25 **D**. 500 **E**. 1000

Items 17 & 18. Imagine that 40 out of 1000 people are likely to develop disease Y over the next 5 years. Treatment A reduces the chance of getting disease Y by 10 per 1000 people. Treatment B reduces the chance of getting disease Y by 4 per 1000 people.

17. **Select the correct answer**

    A. Treatment A is more effective than Treatment B
    B. Treatment B is more effective than Treatment A
    C. Treatment A and Treatment B are equally effective
    D. Don’t know
    E. Don’t know

18. **What is the risk of developing disease Y after receiving Treatment A?**

    A. 36:1000 **B**. 35:1000 **C**. 39:1000 **D**. 30:1000 **E**. Don’t know

19. **The chance of a skin lesion being cancerous is 1%. If 1000 people attend the dermatology clinic with this skin lesion, how many are likely to have cancer?**

    A. 1000 **B**. 1 **C**. 10 **D**. 100 **E**. 0.1

Items 20 – 23. Mr Iqbal is recovering from a stroke, and has been prescribed a nutritional supplement drink. The nutritional information available on the drink carton is shown below:

Build-up drink		
Nutrition Facts		
Serving size	100ml	
Servings per container	4	
	Amount per serving	**% Recommended daily intake***
Energy	250 Calories	12.5%
Total Fat	13g	20%
of which saturates	9g	40%
cholesterol	28g	12%
Total Carbohydrate	30g	12%
of which sugars	25g	
Dietary Fibre	3g	
Protein	4.2g	8%
Sodium	55mg	2%
* % Recommended daily intake values are based on a 2,000 calorie diet. An individual's daily values may be higher or lower depending on their calorie needs.		

20. **If Mr Iqbal drinks the entire container, how many calories will he ingest?**

A. 100 **B**. 250 **C**. 400 **D**. 500 **E**. 1000

21. **If he is allowed to have 60g of carbohydrate as a snack, how much can he drink?**

A. 400ml **B**. 200ml **C**. 100ml **D**. 50ml **E**. 25ml

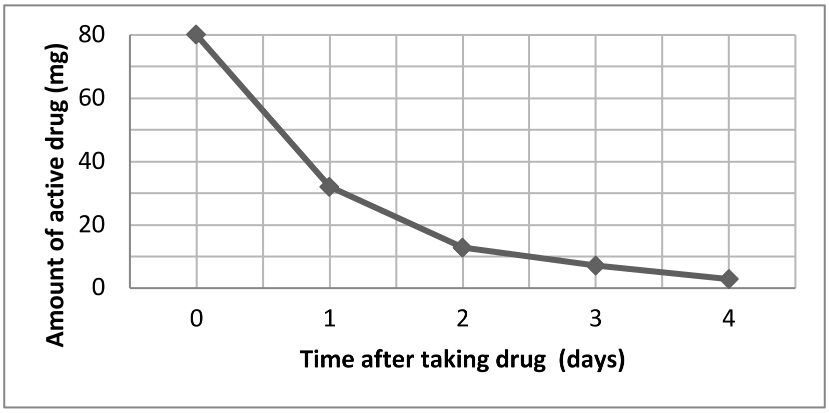
22. **Mr Iqbal usually ingests 42g of saturated fat a day, including one serving of the nutritional drink. If he stops taking the nutritional drink, how much saturated fat would he be consuming each day?**

A. 33g **B**. 29g **C**. 14g **D**.13g **E**. 9g

23. **Mr Iqbal requires 2500 calories per day. What percentage of his daily value of calories is one serving of the drink?**

A. 25% **B**. 12.5% **C**. 10% **D**. 20% **E**. 15%

Items 24 – 25. Alex enters a clinical trial, and is given 80mg of the test drug by IV injection. The following graph shows the initial amount of the drug, and the amount that remains active in Alex's blood after one, two, three and four days.



24. Approximately how much of the drug remains active after 36 hours?

A. 38 mg **B**. 12 mg **C**. 32 mg **D**. 22 mg **E**. 6mg

25. From the graph above, it can be seen that each day about the same proportion of the previous day's drug remains active in Alex's blood. At the end of each day which of the following is the approximate percentage of the previous day's drug that remains active?

A. 50% **B**. 10% **C**. 40% **D**. 20% **E**. 30%

26. 100 women attend hospital for a mammogram. 10 of these women have a malignant tumour, while 90 do not. Of the 10 patients with malignancy, the mammogram detects the cancer in 9, but misses the tumour in one patient. Of the 90 women who are disease-free, the mammogram indicates correctly that 81 of them are healthy, but wrongly indicates that 9 of them have cancer. Mrs Jones is told that her mammogram is positive. What are the chances that she actually does have cancer?

A. 1 in 2 **B**. 1 in 10 **C**. 1 in 9 **D**. 2 in 9 **E**. 9 in 10

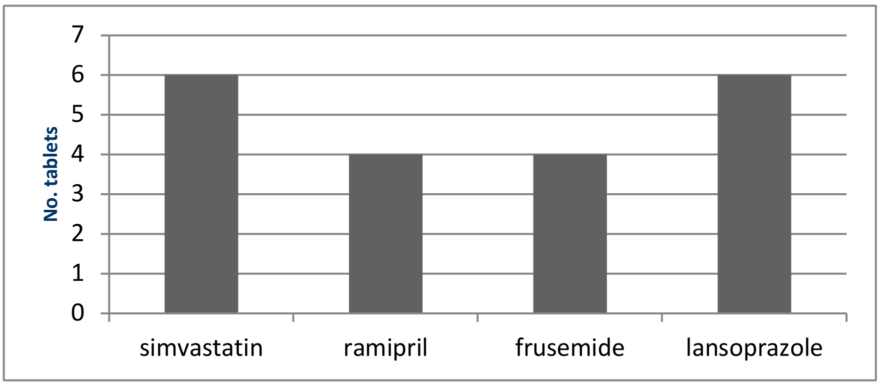
27. Mrs Cartwright has been admitted with an acute exacerbation of asthma. Clinical guidelines state that she can be safely discharged once her Peak Expiratory Flow Rate (PEFR) is >75% of her normal level. Her normal PEFR is 420 l/min. What is the minimum PEFR that Mrs Cartwright must achieve in order to be allowed home?

A. 175 l/min **B**. 255 l/min **C**. 315 l/min **D**. 345 l/min **E**. 495 l/min

Item 28. The table below shows the number of tablets taken daily by five patients.

	Simvastatin	Ramipril	Frusemide	Lansoprazole
Mrs White	1	2	1	0
Mr Brown	2	0	2	1
Miss Scarlet	1	1	0	2
Mr Black	1	2	0	2
Ms Green	2	1	2	1

This graph shows information taken from the table above for four of the five patients.



28. Which patient's information is missing from the graph?

A. Mrs White's **B**. Mr Brown's **C**. Miss Scarlet's **D**. Mr Black's **E**. Ms Green's

29. In a particular university, 1 out of every 49 students is studying medicine. 7 out of 10 medical students are women. What proportion of the university's students are female medical students?

A. 7 out of 10 **B**. 1 out of 59 **C**. 353 out of 490
D. 1 out of 70 **E**. 10 out of 343

30. Ryan has diabetes, and needs 8 units of Actrapid insulin. Actrapid is prepared in a solution containing 100 units of Actrapid per ml. What volume of solution should Ryan be given?

A. 0.008ml **B**. 0.125ml **C**. 8ml **D**. 0.8ml **E**. 0.08ml

31. The chance that an individual gets a certain side effect from a vaccination is 0.3%. If 100,000 people are vaccinated, how many are expected to get the side effect?

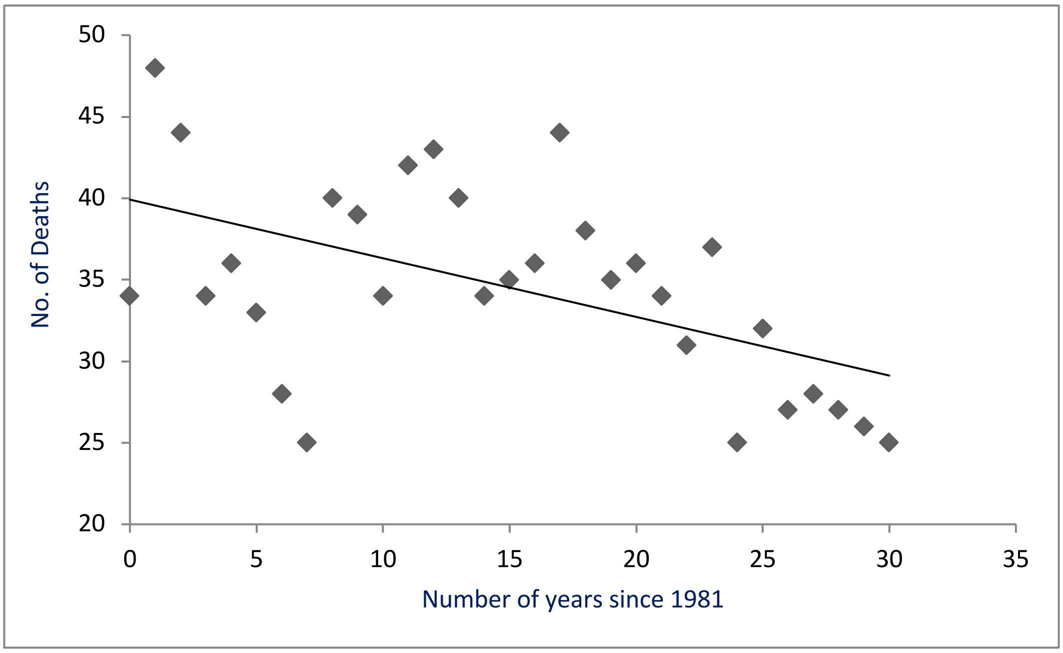
A. 3000 **B**. 300 **C**. 30 **D**. 3 **E**. 30000

32. **Millie receives an injection of IV antibiotic. One hour after the injection, only 60% of the antibiotic will remain active. This pattern continues: at the end of each hour only 60% of the antibiotic that was present at the end of the previous hour remains active. Millie is given a dose of 300 mg of the antibiotic at 0800. Approximately how much antibiotic will remain active at 1100?**

A. 180 mg **B**. 120 mg **C**. 108 mg **D**. 86 mg **E**. 64 mg

<u>Items 33 – 34</u>. The chart below shows the number of deaths from lung cancer per 100,000 smokers for each year since 1981. The best fit is the line drawn on the figure:

Death rate = 39.4 - 0.33 x (number of years since 1981)



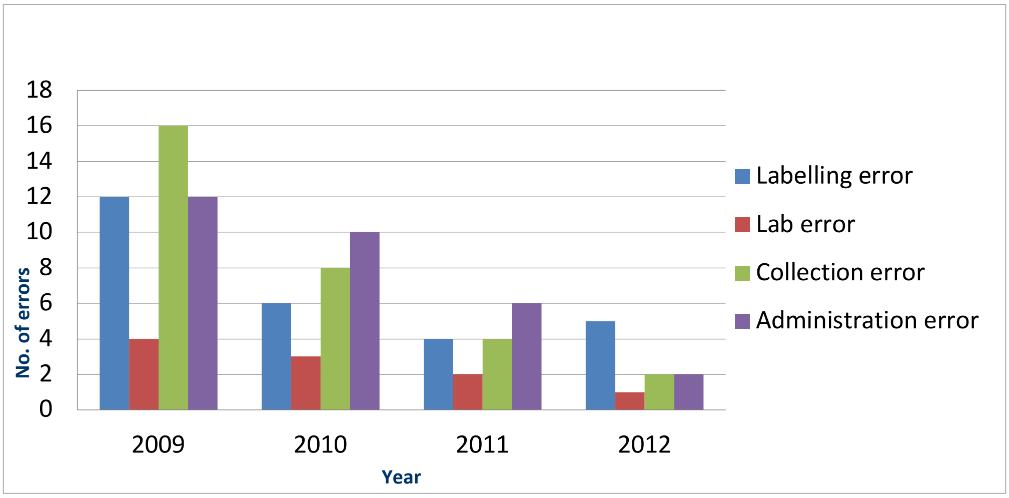
33. **If the trend continues unchanged, approximately how many deaths would be predicted in 2016 from this line?**

A. 28 **B**. 39 **C**. 1 **D**. 11 **E**. 663

34. **Based on the line** *Death rate = 39.4 - 0.33 x (number of years since 1981)* **in the chart above, approximately how many years will it take for the deaths per 100,000 to go down by 1?**

A. 0.33 **B**. 39.4 **C**. 39.07 **D**. 1 **E**. 3

<u>Item 35</u>. The chart below shows a hospital's data regarding the number of errors made in blood transfusion over a four-year period.



35. Regarding the data displayed on this chart, which of the following statements is false:

A. Labelling error is more common than lab error

B. Overall, there have been the same number of collection errors as administration errors

C. Labelling error has decreased steadily since 2009

D. Collection error has decreased by 50% every year

E. Lab error has decreased every year

<u>Items 36 & 37</u>. Imagine that 40 out of 1000 people are likely to develop disease Y over the next 5 years. 100 people would have to be treated with Treatment A for 5 years for a benefit against disease Y to be evident in one of them. 250 people would have to be treated with Treatment B for 5 years for a benefit against disease Y to be evident in one of them.

36. Select the correct answer

A. Treatment A is more effective than Treatment B

B. Treatment B is more effective than Treatment A

C. Treatment A and Treatment B are equally effective

D. Don't know

E. Don't know

37. What is the risk of developing disease Y after receiving Treatment A?

A. 40:900 **B**. 30:1000 **C**. 39:900 **D**. 39:1000 **E**. Don't know

Item 38. You are asked to review Mr Brown as the ward sister is worried about his urine output. The chart below shows Mr Brown's urine output over the past four days:

Day	**Urine output (ml)**
Monday	532
Tuesday	472
Wednesday	472
Thursday	364

38. What is Mr Brown's average urine output per day over this 4-day period?

A. 1460ml **B**. 472m l **C**. 480ml **D**. 460ml **E**. 1840ml

39. About 50% of men and 33% of women will develop cancer at some point. About 20% of cancers in men are found before age 55. What percentage of men are expected to have cancer before age 55?

A. 50% **B**. 10% **C**. 5% **D**. 20% **E**. 1%

40. A patient is on reducing doses of Prednisolone. He starts on 40mg/day, and is advised to reduce the dose by 5mg every third day. Approximately how long will it take him to wean off the Prednisolone?

A. 13 days **B**. 40 days **C**. 24 days **D**. 120 days **E**. 8 days

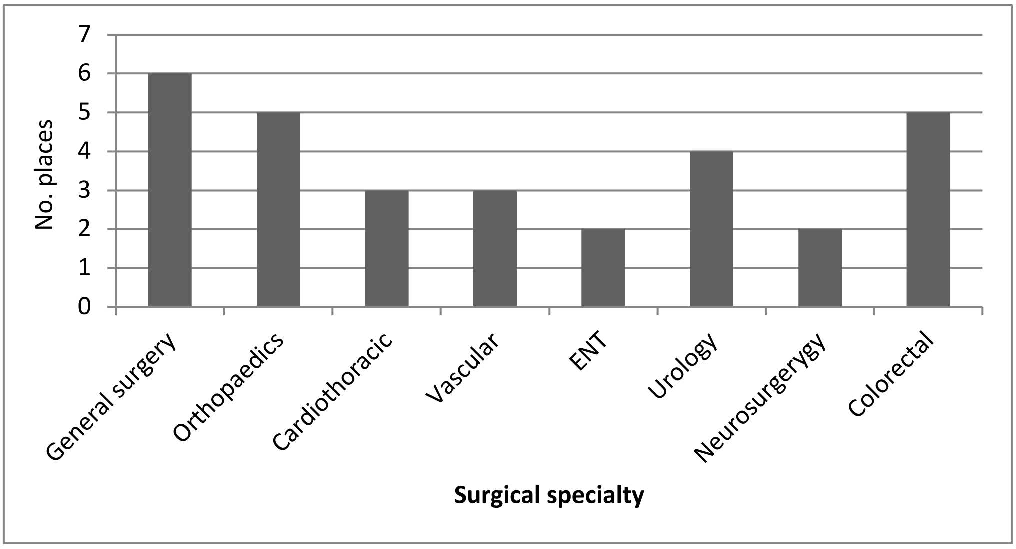
Item 41. Medical students were asked whether they had ever engaged in binge drinking. Their answers, classified by gender, are shown in the table below.

	Binge Drinking	
Gender	**Yes**	**No**
Male	43	50
Female	28	92

41. What percentage of those who reported engaging in binge drinking were male?

A. 86% **B**. 61% **C**. 20% **D**. 46% **E**. 154%

<u>Item 42</u>. The chart below shows the number of training places for FY1 doctors in various surgical specialties in a large teaching hospital.



42. **Sam is an FY1 trainee. How likely is he to be placed in General Surgery?**

A. 50% **B**. 40% **C**. 30% **D**. 20% **E**. 10%

43. **Mo weighs 100kg, and presents to A&E with a wound in his thigh. You are asked to suture it, using the local anaesthetic bupivacaine which comes in a solution containing bupivacaine 5mg/ml. The maximum dose of bupivacaine that can be safely given is 2mg/kg. What is the maximum amount of bupivacaine you can use when suturing Mo's wound?**

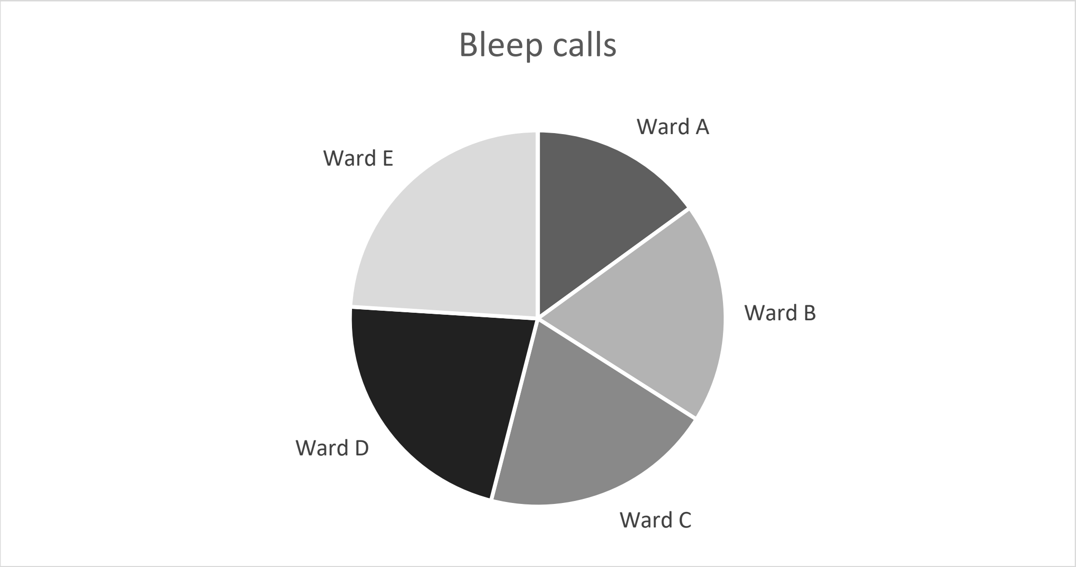
A. 500ml **B**. 20ml **C**. 150ml **D**. 50ml **E**. 40ml

Blank Page

APPENDIX 3. MINTv2

1. **Kerry has diabetes and needs to eat 6g of carbohydrate for every 30 minutes of exercise. She is planning to exercise in the gym for one hour. She has some biscuits in her gym bag. Each biscuit contains 8g of carbohydrate. How many biscuits should she eat before she exercises?**

Item 2. This pie chart shows the distribution of bleep calls for Dr Peters.



2. **To which ward is Dr Peters called least often?**

3. **There is a 2 in 100 chance of living 5 years or longer without treatment for a type of cancer. Drug G increases the chance of living 5 years or longer by 50%. Drug H increases the chance of living 5 years or longer to 6%. Your patient, Bob, wants the best chance of living 5 years or longer. Which drug should you prescribe?**

4. **Katie attends the out-patient clinic with a history of weight loss. Her weight has dropped from 75kg to 67.5kg over the past 5 months. What percentage of her original weight has she lost?**

5. **Mr Bradley is admitted to the gastroenterology ward and is prescribed an omeprazole infusion, to be continued for 72 hours. The rate of infusion is 10ml/hr, and the solution infused contains 80mg of omeprazole in 100ml of 0.9% NaCl. How much omeprazole does Mr Bradley receive every hour?**

<u>Item 6</u>. Normal daily fluid and electrolyte requirements are summarised in the table below.

Water	25-30 ml/kg/day
Sodium, Potassium, Chloride	1 mmol/kg/day
Glucose	50-100g/day

6. **Craig weighs 70kg. Which of the following best represents Craig's approximate daily requirements of water, sodium and glucose?**

<u>Items 7 & 8</u>. Without treatment, 40 out of 1000 people will develop disease Y over the next 5 years. Two treatments are available to reduce the risk of developing disease Y. Treatment A reduces the chance of developing disease Y by 25%. Treatment B reduces the chance of developing disease Y by 10%.

7. **Which treatment is better?**

8. **What is the risk of developing disease Y after receiving Treatment A?**

9. **About 50% of men and 33% of women will develop cancer at some point. About 20% of cancers in men are found before age 55. What percentage of <u>men</u> are expected to have cancer before age 55?**

10. **People starting on treatment for hypertension are given a 1 in 1000 chance of developing a particular complication. What percentage of people are likely to develop this complication?**

A. 0.1% **B**. 1% **C**. 0.01% **D**. 10% **E**. 0.001%

<u>Item 11</u>. Simon and Emma Jones undergo genetic screening, and are given the following information regarding the risk of any child they conceive inheriting various diseases.

Inherited disease	**Risk**
Disease A	**0.01%**
Disease B	**0.001**
Disease C	**1:10 000**
Disease D	**0.001%**
Disease E	**1:1000**

11. **Which disease is their child least likely to inherit?**

12. **The volume of a unit of blood is 380ml. If it is infused at a rate of 125ml/hr, approximately what proportion of the blood will have been transfused after 2 hours?**

13. **Adult height for men can be estimated on the basis of parental height, using the following formula:**

$$\text{Adult height (cm)} = \frac{(\text{Mother's height} + \text{Father's height} + 13)}{2}$$

Most boys will reach an adult height within 10cm of this estimation. Finn is 6 years old. His mother is 152cm tall, and his father is 165cm tall. How tall is Finn likely to be when he is an adult? *Express your answer as a range of heights.*

<u>Items 14 – 15</u>. Miss Strong, an orthopaedic surgeon, screens 100 patients for MRSA preoperatively. 10 of these 100 patients are MRSA carriers. The test used gives a true positive result in 90% of MRSA carriers, and a false positive result in 20% of people who do not carry MRSA.

14. How many of these 100 patients are expected to test positive for MRSA?

15. What percentage of those who test positive actually carry MRSA?

<u>Item 16</u>. Mr Price is on the elderly care ward. You are asked to review him regarding his oral fluid intake. The chart below is an accurate record of Mr Price's oral fluid intake for the 4 hours from 8am to 12 noon.

Time	**08.00 – 09.00**	**09.00 – 10.00**	**10.00 – 11.00**	**11.00 – 12.00**
Volume (ml)	89	63	121	63

16. What is Mr Price's average hourly fluid intake over this 4-hour period?

<u>Items 17 & 18</u>. Without treatment, 40 out of 1000 people will develop disease Y over the next 5 years. Two treatments are available to reduce the risk of developing disease Y. Treatment A reduces the chance of getting disease Y by 10 per 1000 people. Treatment B reduces the chance of getting disease Y by 4 per 1000 people.

17. Which treatment is better?

18. What is the risk of developing disease Y after receiving Treatment A?

<u>Items 19 – 22</u>. Mrs Doyle is recovering from a stroke, and has been prescribed a nutritional supplement drink. The nutritional information available on the drink carton is shown below:

Build-up drink		
Nutritional information Serving size Servings per carton	 100ml 4	
Typical values	**Per 100ml**	**% Recommended daily intake***
Energy	200 Calories	10%
Total Fat	16g	25%
of which saturates	12.5g	50%
Total Carbohydrate	35g	11%
of which sugars	24g	
Dietary Fibre	7.5g	
Protein	6g	12%
Sodium	75mg	3%
* **NOTE** % Recommended daily intake values are based on a 2,000 calorie diet. People have different calorie requirements, so an individual's daily values may be higher or lower than those indicated here.		

19. If Mrs Doyle drinks half of the carton, how many calories will she consume?

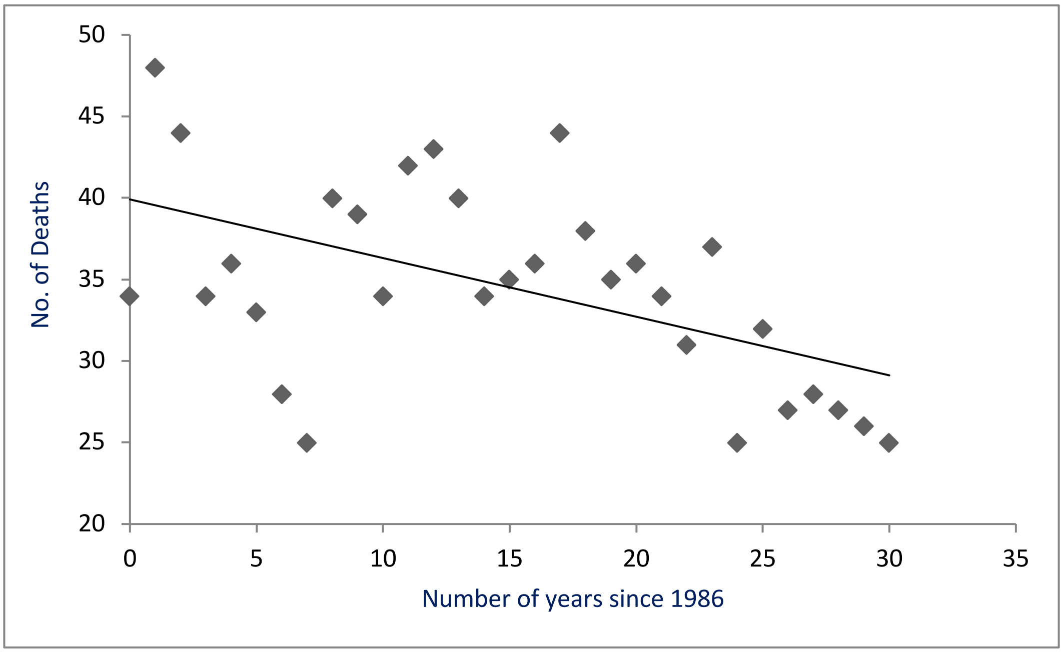
20. Mrs Doyle needs to increase her protein intake by 9g. What volume of the nutritional drink provides 9g of protein?

21. Mrs Doyle has two servings of the nutritional drink every day. Her total carbohydrate intake is 200g per day. How many grams of her carbohydrate intake comes from sources other than the drink?

22. Mrs Doyle requires 1600 calories per day. What percentage of her daily calorie intake is provided by one serving of the drink?

Items 23– 24. The chart below shows the number of deaths from lung cancer per 100,000 smokers for each year since 1986. The best fit is the line drawn on the figure:

Death rate = 39.4 - 0.33 x (number of years since 1986)



23. **If the trend continues unchanged, approximately how many deaths would be predicted in 2021 from this line?**

24. **Based on the line** *Death rate = 39.4 - 0.33 x (number of years since 1986)* **in the chart above, approximately how many years will it take for the deaths per 100,000 to go down by 1?**

25. **Leanne has been admitted with an acute exacerbation of asthma. Clinical guidelines state that she can be safely discharged once her Peak Expiratory Flow Rate (PEFR) is at least 75% of her normal level. Her normal PEFR is 420 l/min. What is the minimum PEFR that Leanne must achieve before she can go home?**

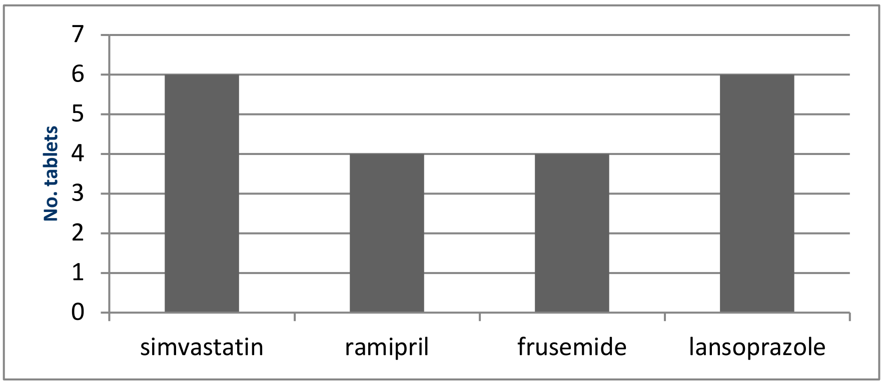
26. **The chance that an individual gets a certain side effect from a vaccination is 0.3%. If 100,000 people are vaccinated, how many are expected to get the side effect?**

27. 100 people attend hospital for a cancer screening test. However, the screening test is not completely accurate. 10 of the 100 people have cancer, while 90 do not. Of the 10 people with cancer, the screening test detects the cancer in 9, but misses the cancer in 1 person. Of the 90 people who do not have cancer, the screening test indicates correctly that 81 of them do not have cancer, but indicates incorrectly that 9 of them do have cancer. Mr Iqbal is told that his screening test shows that he has cancer. What is the likelihood that he actually does have cancer?

<u>Item 28.</u> The table below shows the number of tablets taken daily by five patients.

	Simvastatin	**Ramipril**	**Frusemide**	**Lansoprazole**
Mrs Archer	**1**	**2**	**1**	**0**
Mr Brown	**2**	**0**	**2**	**1**
Miss Cartwright	**1**	**1**	**0**	**2**
Mr Dunne	**1**	**2**	**0**	**2**
Mr Early	**2**	**1**	**2**	**1**

This graph shows information taken from the table above for four of the five patients.

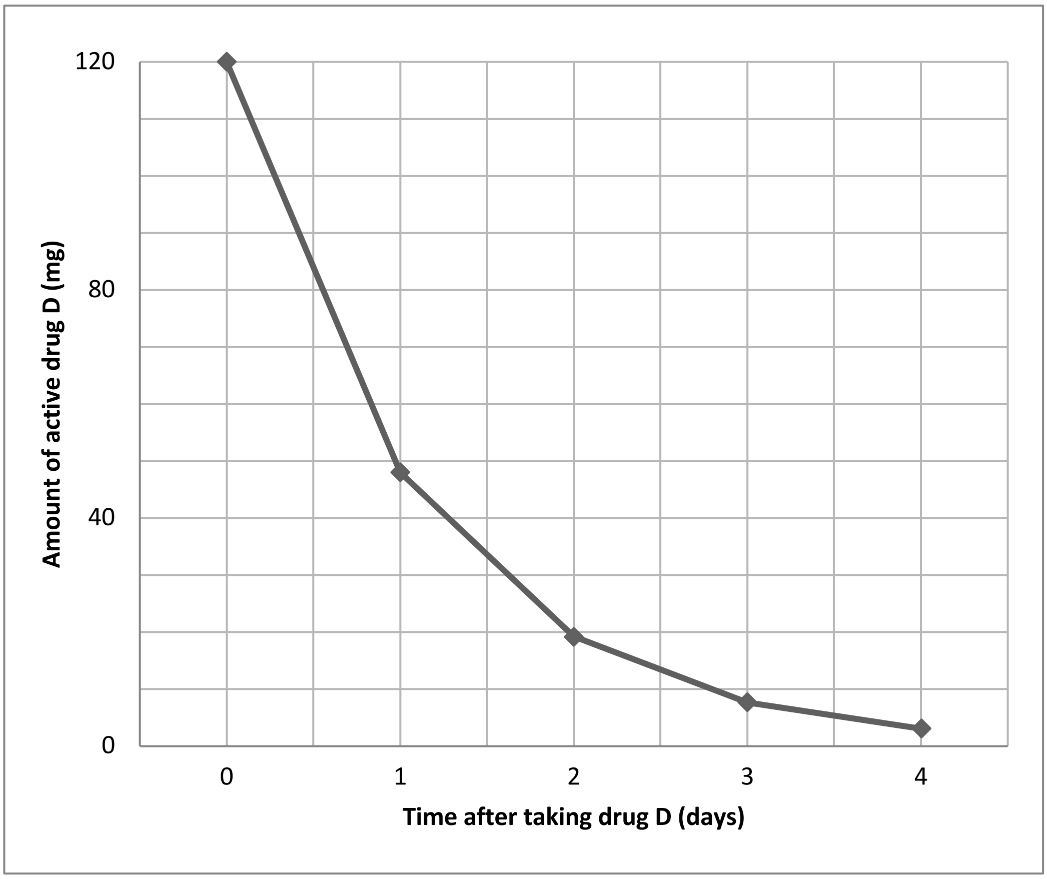


28. Which patient's information is missing from the graph?

29. In a particular university, 1 out of every 49 students is studying medicine. 7 out of 10 medical students are women. What proportion of the university's students are female medical students?

30. **Noah is 5 years old, and weighs 20kg. Following a dose of morphine for postoperative pain relief, he has developed respiratory depression, and now needs reversal with an injection of naloxone. The recommended dose of naloxone is 3 micrograms/kg body weight. Naloxone is prepared in a solution containing 400 micrograms per ml. How many ml of naloxone should Noah be given?**

<u>Items 31-32</u>. Amy is given 120mg of drug D by intravenous (IV) injection. Blood tests are taken every day for the next four days to check the level of drug D remaining active in her bloodstream. The graph below shows the concentrations of drug D in Amy's blood over the four-day period.

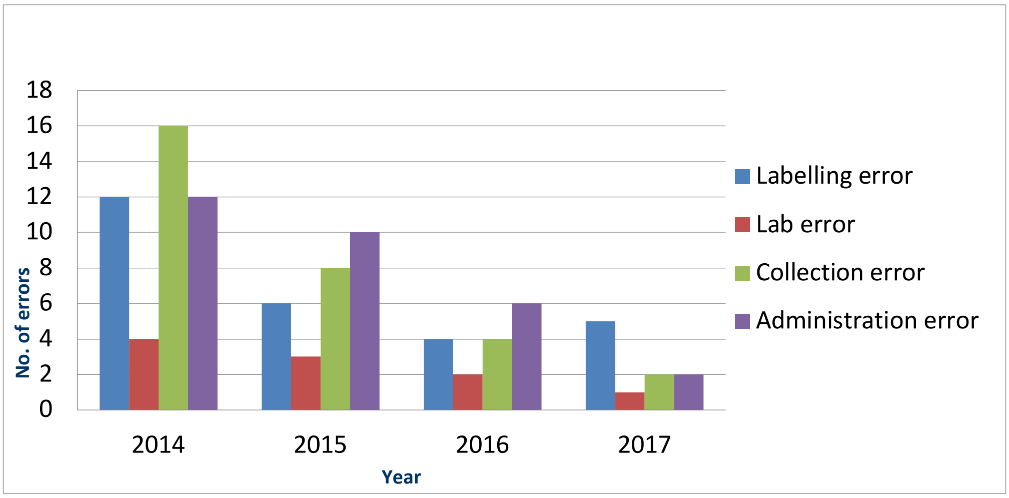


31. **Approximately how many mg of drug D remain active after 36 hours?**

32. **From the graph above, it can be seen that each day about the same proportion of the previous day's drug remains active in Amy's blood. At the end of each day, approximately what <u>percentage</u> of the previous day's drug remains active?**

33. Clark is admitted to the ward with an infection, and starts on a course of IV antibiotics. One hour after he receives the injection, 70% of the antibiotic remains active. The antibiotic activity continues to decline in this manner: at the end of each hour, antibiotic activity is 70% of its value at the end of the previous hour. Clark receives 500 mg of the antibiotic at 1200. Approximately how many mg of the antibiotic will still be active at 1600?

Item 34. The chart below shows a hospital's data regarding the number of errors made in blood transfusion over a four-year period.



34. Regarding the data displayed on this chart, which of the following statements is false:

A. Labelling error is more common than lab error

B. Overall, there have been the same number of collection errors as administration errors

C. Labelling error has decreased steadily since 2014

D. Collection error has decreased by 50% every year

E. Lab error has decreased every year

35. Dave is on reducing doses of Prednisolone. He starts on 40mg/day, and is advised to reduce the dose by 5mg every third day. Approximately how long will it take Dave to wean off the Prednisolone?

Items 36 & 37. Without treatment, 40 out of 1000 people will develop disease Y over the next 5 years. Two treatments are available to reduce the risk of developing disease Y. 100 people must be treated with Treatment A for 5 years to prevent one person developing disease Y. 250 people must be treated with Treatment B for 5 years to prevent one person developing disease Y.

36. Which treatment is better?

37. What is the risk of developing disease Y after receiving Treatment A?

38. Patients are recruited to a randomised controlled trial of a new drug. Randomisation is done by tossing a coin. If the coin lands head up, the patient will be given the new drug. If the coin lands on tails, the patient will be given a placebo. 1000 patients are recruited to the study. Approximately how many are likely to receive the new drug?

Item 39. Medical students were asked whether they had ever engaged in binge drinking. Their answers, classified by gender, are shown in the table below.

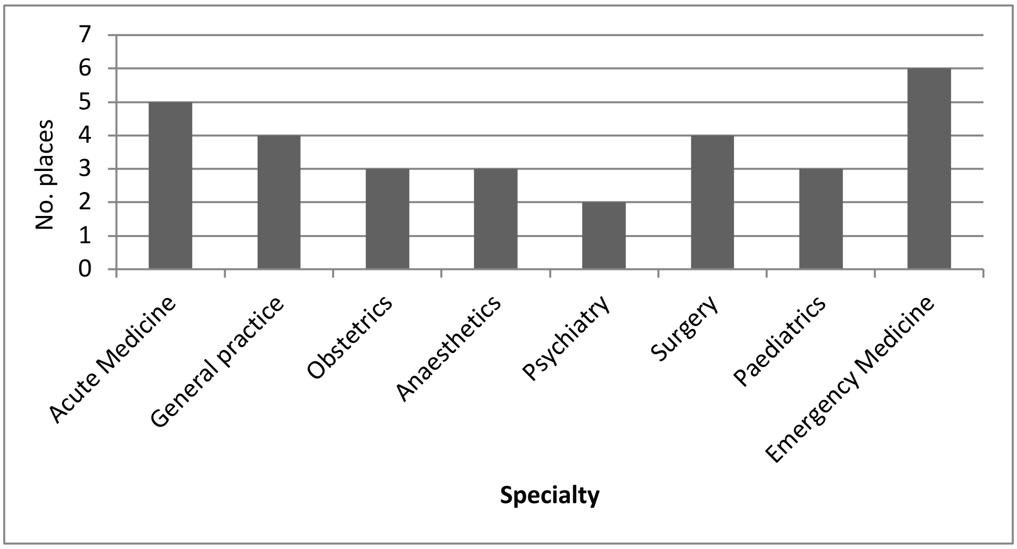
	Binge Drinking	
Gender	**Yes**	**No**
Male	**43**	**50**
Female	**28**	**92**

39. What percentage of those who reported engaging in binge drinking were male?

40. Poppy (age 19, weight 65kg) is admitted to the acute medical ward with a diagnosis of Diabetic Ketoacidosis. Clinical Guidelines state that she should be given insulin at a rate of 0.1 unit/kg/hr. You are asked to prescribe her insulin infusion. The preparation for infusion contains 50 units of insulin in a volume of 50ml 0.9% NaCl. What rate of infusion will you prescribe?

41. The chance of a hip replacement operation being cancelled is 1%. If 1000 people are scheduled for hip replacement operations, how many are likely to have their operation cancelled?

<u>Item 42</u>. The chart below shows the number of training places available in various specialties on a Foundation Year 2 programme.



42. Helen is has applied for a place on this training programme. If places are allocated at random, what is her chance of being placed in Emergency Medicine?

43. Mo weighs 100kg, and presents to A&E with a wound in his thigh. You are asked to suture the wound under local anaesthetic. The available local anaesthetic is a solution containing bupivacaine 5mg/ml. The maximum dose of bupivacaine that can be safely given is 2mg/kg. What is the maximum volume of bupivacaine you can use to suture Mo's wound?

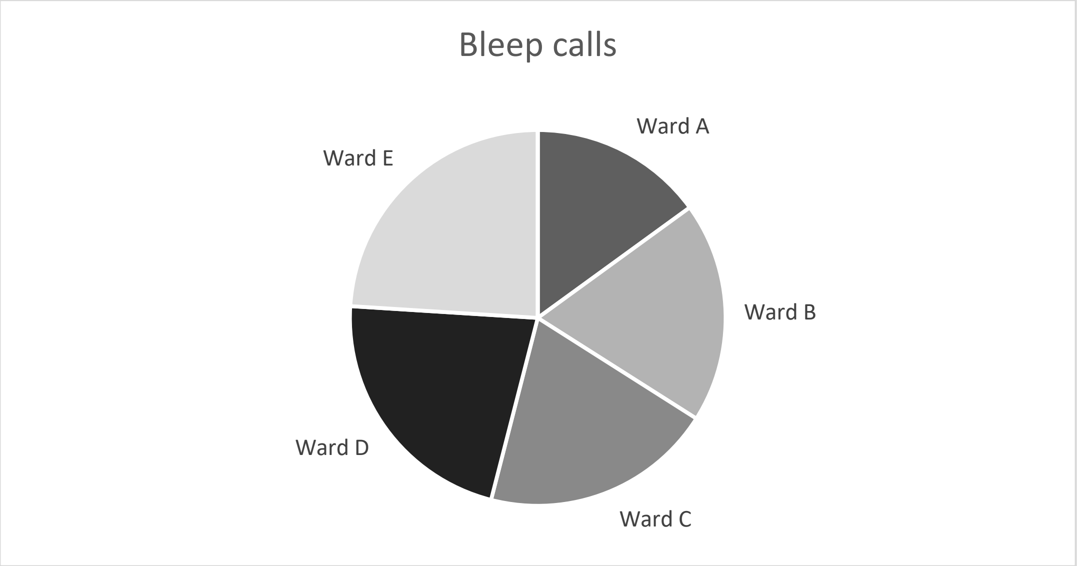
Blank Page

APPENDIX 4. MINTv3

1. **Kerry has diabetes and needs to eat 6g of carbohydrate for every 30 minutes of exercise. She is planning to exercise in the gym for one hour. She has some biscuits in her gym bag. Each biscuit contains 8g of carbohydrate. How many biscuits should she eat before she exercises?**

   A. 2 biscuits **B**. 1 biscuit **C**. 1 2/3 biscuits **D**. 3/4 biscuit **E**. 1½ biscuits

Item 2. This pie chart shows the distribution of bleep calls for Dr Peters.



2. **To which ward is Dr Peters called least often?**

   A. Ward A **B**. Ward B **C**. Ward C **D**. Ward D **E**. Ward E

3. **There is a 2 in 100 chance of living 5 years or longer without treatment for a type of cancer. Drug G increases the chance of living 5 years or longer by 50%. Drug H increases the chance of living 5 years or longer to 6%. Your patient, Bob, wants the best chance of living 5 years or longer. Which drug should you prescribe?**

   A. Drug G **B**. Drug H **C**. Either drug, the chance of living longer is the same

   D. Neither drug, the chance of living longer is better without treatment **E**. Don't know

4. **Katie attends the out-patient clinic with a history of weight loss. Her weight has dropped from 75kg to 67.5kg over the past 5 months. What percentage of her original weight has she lost?**

   A. 1% **B**. 3% **C**. 9% **D**. 10% **E**. 12.5%

5. **Mr Bradley is admitted to the gastroenterology ward and is prescribed an omeprazole infusion, to be continued for 72 hours. The rate of infusion is 10ml/hr, and the solution infused contains 80mg of omeprazole in 100ml of 0.9% NaCl. How much omeprazole does Mr Bradley receive every hour?**

A. 0.8 mg **B**. 576 mg **C**. 8 mg **D**. 1 mg **E**. 48 mg

Item 6. Normal daily fluid and electrolyte requirements are summarised in the table below.

Water	25-30 ml/kg/day
Sodium, Potassium, Chloride	1 mmol/kg/day
Glucose	50-100g/day

6. **Craig weighs 70kg. Which of the following best represents Craig's approximate daily requirements of water, sodium and glucose.**

A. 2000 ml fluid, 70 mmol sodium and 7000g glucose.
B. 2000 ml fluid, 70 mmol sodium and 5250g glucose.
C. 2000 ml fluid, 23 mmol sodium and 80g glucose.
D. 2000 ml fluid, 70 mmol sodium and 3500g glucose.
E. 2000 ml fluid, 70 mmol sodium and 75g glucose.

Items 7 & 8. Without treatment, 40 out of 1000 people will develop disease Y over the next 5 years. Two treatments are available to reduce the risk of developing disease Y. Treatment A reduces the chance of developing disease Y by 25%. Treatment B reduces the chance of developing disease Y by 10%.

7. **Select the correct answer:**
F. Treatment A is more effective than Treatment B
G. Treatment B is more effective than Treatment A
H. Treatments A and B are equally effective
I. Neither treatment is worthwhile
J. Don't know

8. **What is the risk of developing disease Y after receiving Treatment A?**

A. 30:40 **B**. 30:1000 **C**. 10:1000 **D**. 300:1000 **E**. 36:1000

9. **About 50% of men and 33% of women will develop cancer at some point. About 20% of cancers in men are found before age 55. What percentage of men are expected to have cancer before age 55?**

A. 20% **B**. 10% **C**. 40% **D**. 1% **E**. 100%

10. People starting treatment for hypertension are given a 1 in 1000 chance of developing a particular complication. What percentage of people are likely to develop this complication?

A. 0.1% **B**. 1% **C**. 0.01% **D**. 10% **E**. 0.001%

Item 11. Simon and Emma Jones undergo genetic screening, and are given the following information regarding the risk of any child they conceive inheriting various diseases.

Inherited disease	**Risk**
Disease A	**0.01%**
Disease B	**0.001**
Disease C	**1:10 000**
Disease D	**0.001%**
Disease E	**1:1000**

11. Which disease is their child least likely to inherit?

B. Disease A **B**. Disease B **C**. Disease C **D**. Disease D **E**. Disease E

12. The volume of a unit of blood is 380ml. If it is infused at a rate of 125ml/hr, approximately what proportion of the blood will have been transfused after 2 hours?

A. 1/3 **B**. 2/3 **C**. 1/2 **D**. 3/4 **E**. 3/2

13. Adult height for men can be estimated on the basis of parental height, using the following formula:

$$\text{Adult height (cm)} = \frac{\text{(Mother's height + Father's height + 13)}}{2}$$

Most boys will reach an adult height within 10cm of this estimation. Finn is 6 years old. His mother is 152cm tall, and his father is 165cm tall. How tall is Finn likely to be when he is an adult?

A. 150 – 170 cm **B**. 165 cm **C**. 155 – 175 cm
D. 160 – 170 cm **E**. 165 – 175cm

Items 14 – 15. Miss Strong, an orthopaedic surgeon, screens 100 patients for MRSA preoperatively. 10 of these 100 patients are MRSA carriers. The test used gives a true positive result in 90% of MRSA carriers, and a false positive result in 20% of people who do not carry MRSA.

14. How many of these 100 patients are expected to test positive for MRSA?

A. 9 **B**. 10 **C**. 11 **D**. 27 **E**. 29

15. What percentage of those who test positive actually carry MRSA?

A. 90% **B**. 37% **C**. 33% **D**. 80% **E**. 9%

Item 16. Mr Price is on the elderly care ward. You are asked to review him regarding his oral fluid intake. The chart below is an accurate record of Mr Price's oral fluid intake for the 4 hours from 8am to 12 noon.

Time	08.00 – 09.00	09.00 – 10.00	10.00 – 11.00	11.00 – 12.00
Volume (ml)	89	63	121	63

16. What is Mr Price's average hourly fluid intake over this 4-hour period?

A. 74 ml **B**. 84 ml **C**. 79 ml **D**. 63 ml **E**. 88ml

Items 17 & 18. Without treatment, 40 out of 1000 people will develop disease Y over the next 5 years. Two treatments are available to reduce the risk of developing disease Y. Treatment A reduces the chance of getting disease Y by 10 per 1000 people. Treatment B reduces the chance of getting disease Y by 4 per 1000 people.

17. Select the correct answer:

A. Treatment A is more effective than Treatment B
B. Treatment B is more effective than Treatment A
C. Treatments A and B are equally effective
D. Neither treatment is worthwhile
E. Don't know

18. What is the risk of developing disease Y after receiving Treatment A?

A. 39.6:1000 **B**. 10:1000 **C**. 40:1000 **D**. 30:1000 **E**. 39:1000

Items 19 – 22. Mrs Doyle is recovering from a stroke, and has been prescribed a nutritional supplement drink. The nutritional information available on the drink carton is shown below:

Build-up drink		
Nutritional information		
Serving size	100ml	
Servings per carton	4	
Typical values	**Per 100ml**	**% Recommended daily intake***
Energy	200 Calories	10%
Total Fat	16g	25%
of which saturates	12.5g	50%
Total Carbohydrate	35g	11%
of which sugars	24g	
Dietary Fibre	7.5g	
Protein	6g	12%
Sodium	75mg	3%
* **NOTE** % Recommended daily intake values are based on a 2,000 calorie diet. People have different calorie requirements, so an individual's daily values may be higher or lower than those indicated here.		

19. **If Mrs Doyle drinks half of the carton, how many calories will she consume?**

    A. 100 **B**. 400 **C**. 200 **D**. 800 **E**. 150

20. **Mrs Doyle needs to increase her protein intake by 9g. What volume of the nutritional drink provides 9g of protein?**

    A. 100 ml **B**. 133 ml **C**. 150 ml **D**. 175 ml **E**. 50 ml

21. **Mrs Doyle has two servings of the nutritional drink every day. Her total carbohydrate intake is 200g per day. How many grams of her carbohydrate intake comes from sources other than the drink?**

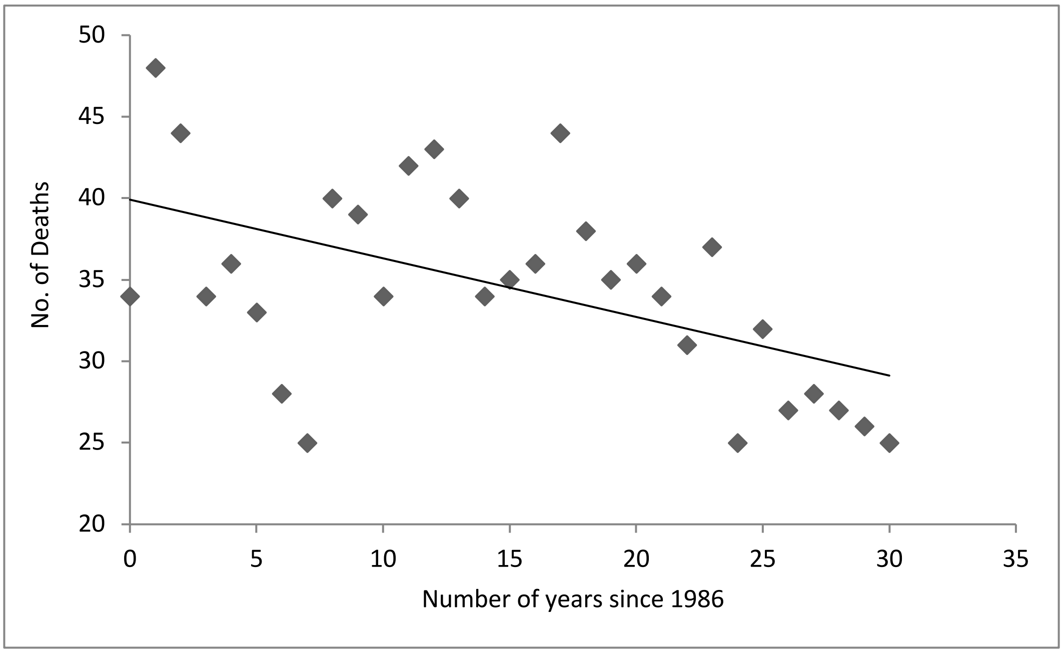
    A. 70g **B**. 152g **C**. 182.5g **D**.130g **E**. 60g

22. **Mrs Doyle requires 1600 calories per day. What percentage of her daily calorie intake is provided by one serving of the drink?**

    A. 10% **B**. 3% **C**. 6.25% **D**. 8% **E**. 12.5%

Items 23– 24. The chart below shows the number of deaths from lung cancer per 100,000 smokers for each year since 1986. The best fit is the line drawn on the figure:

Death rate = 39.4 - 0.33 x (number of years since 1986)



23. If the trend continues unchanged, approximately how many deaths would be predicted in 2021 from this line?

A. 1367 **B**. 25 **C**. 28000 **D**. 32 **E**. 28

24. Based on the line *Death rate = 39.4 - 0.33 x (number of years since 1986)* **in the chart above, approximately how many years will it take for the deaths per 100,000 to go down by 1?**

A. 115 **B**. 10 **C**. 3 **D**. 5 **E**. 1

25. Leanne has been admitted with an acute exacerbation of asthma. Clinical guidelines state that she can be safely discharged once her Peak Expiratory Flow Rate (PEFR) is at least 75% of her normal level. Her normal PEFR is 420 l/min. What is the minimum PEFR that Leanne must achieve before she can go home?

A. 315 l/min **B**. 335 l/min **C**. 520 l/min **D**. 305 l/min **E**. 31.5 l/min

26. The chance that an individual gets a certain side effect from a vaccination is 0.3%. If 100,000 people are vaccinated, how many are expected to get the side effect?

A. 3000 **B**. 300 **C**. 3 **D**. 30000 **E**. 30

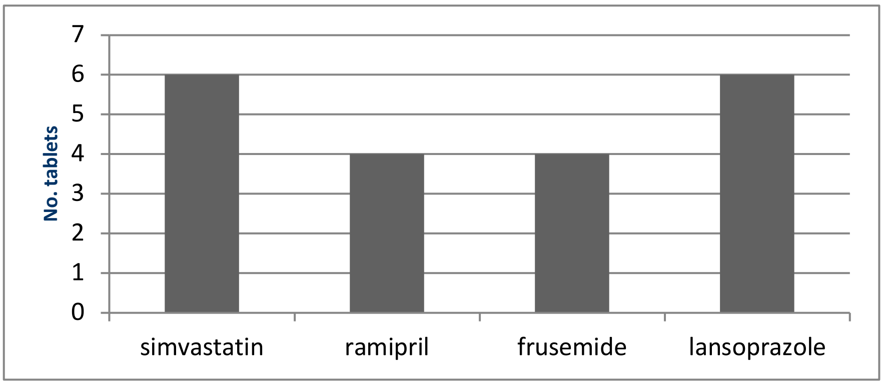
27. 100 people attend hospital for a cancer screening test. However, the screening test is not completely accurate. 10 of the 100 people have cancer, while 90 do not. Of the 10 people with cancer, the screening test detects the cancer in 9, but misses the cancer in 1 person. Of the 90 people who do not have cancer, the screening test indicates correctly that 81 of them do not have cancer, but indicates incorrectly that 9 of them do have cancer. Mr Iqbal is told that his screening test shows that he has cancer. What is the likelihood that he actually does have cancer?

A. 90% **B**. 18% **C**. 10% **D**. 80% **E**. 50%

Item 28. The table below shows the number of tablets taken daily by five patients.

	Simvastatin	**Ramipril**	**Frusemide**	**Lansoprazole**
Mrs Archer	**1**	**2**	**1**	**0**
Mr Brown	**2**	**0**	**2**	**1**
Miss Cartwright	**1**	**1**	**0**	**2**
Mr Dunne	**1**	**2**	**0**	**2**
Mr Early	**2**	**1**	**2**	**1**

This graph shows information taken from the table above for four of the five patients.



28. Which patient's information is missing from the graph?

A. Mrs Archer's **B**. Mr Brown's **C**. Miss Cartwright's **D**. Mr Dunne's **E**. Mr Early's

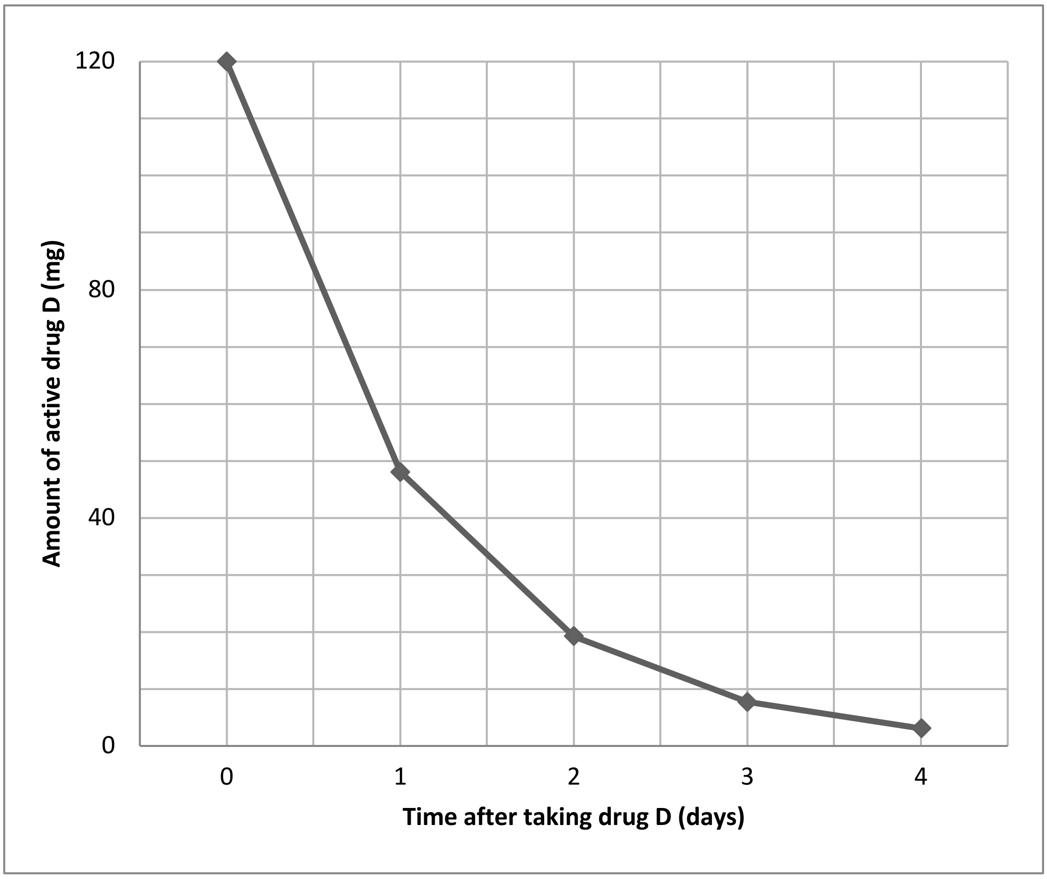
29. In a particular university, 1 out of every 49 students is studying medicine. 7 out of 10 medical students are women. What proportion of the university's students are female medical students?

A. 1 in 7 **B**. 7 in 10 **C**. 10 in 343 **D**. 7 in 343 **E.** 1 in 70

30. **Noah is 5 years old, and weighs 20kg. Following a dose of morphine for postoperative pain relief, he has developed respiratory depression, and now needs reversal with an injection of naloxone. The recommended dose of naloxone is 3 micrograms/kg body weight. Naloxone is prepared in a solution containing 400 micrograms per ml. How many ml of naloxone should Noah be given?**

A. 6.6 ml **B**. 0.15 ml **C**. 1.5 ml **D**. 0.2 ml **E**. 0.6 ml

Items 31-32. Amy is given 120mg of drug D by intravenous (IV) injection. Blood tests are taken every day for the next four days to check the level of drug D remaining active in her bloodstream. The graph below shows the concentrations of drug D in Amy's blood over the four day period.



31. **Approximately how many mg of drug D remain active after 36 hours?**

A. 8 mg **B**. 5 mg **C**. 15 mg **D**. 10 mg **E**. 34 mg

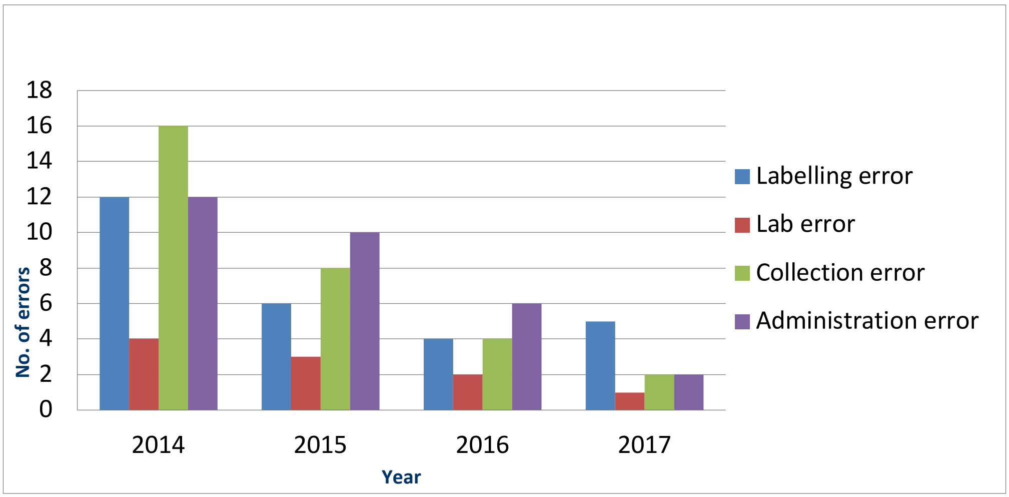
32. **From the graph above, it can be seen that each day about the same proportion of the previous day's drug remains active in Amy's blood. At the end of each day, approximately what percentage of the previous day's drug remains active?**

A. 50% **B**. 30% **C**. 40% **D**. 5% **E**. 60%

33. Clark is admitted to the ward with an infection, and starts on a course of IV antibiotics. One hour after he receives the injection, 70% of the antibiotic remains active. The antibiotic activity continues to decline in this manner: at the end of each hour, antibiotic activity is 70% of its value at the end of the previous hour. Clark receives 500 mg of the antibiotic at 1200. Approximately how many mg of the antibiotic will still be active at 1600?

A. 171.5 mg **B**. 120 mg **C**. 4 mg **D**. 288 mg **E**. 140 mg

Item 34. The chart below shows a hospital's data regarding the number of errors made in blood transfusion over a four-year period.



34. Regarding the data displayed on this chart, which of the following statements is false:

A. Labelling error is more common than lab error
B. Overall, there have been the same number of collection errors as administration errors
C. Labelling error has decreased steadily since 2014
D. Collection error has decreased by 50% every year
E. Lab error has decreased every year

35. Dave is on reducing doses of Prednisolone. He starts on 40mg/day, and is advised to reduce the dose by 5mg every third day. Approximately how long will it take Dave to wean off the Prednisolone?

A. 27 days **B**. 8 days **C**. 17 days **D**. 24 days **E**. 21 days

Items 36 & 37. Without treatment, 40 out of 1000 people will develop disease Y over the next 5 years. Two treatments are available to reduce the risk of developing disease Y. 100 people must be treated with Treatment A for 5 years to prevent one person developing disease Y. 250 people must be treated with Treatment B for 5 years to prevent one person developing disease Y.

36. Select the correct answer:

A. Treatment A is more effective than Treatment B
B. Treatment B is more effective than Treatment A
C. Treatments A and B are equally effective
D. Neither treatment is worthwhile
E. Don't know

37. What is the risk of developing disease Y after receiving Treatment A?

A. 36:1000 **B**. 40:1000 **C**. 39:1000 **D**. 10:1000 **E**. 30:1000

38. Patients are recruited to a randomised controlled trial of a new drug. Randomisation is done by tossing a coin. If the coin lands head up, the patient will be given the new drug. If the coin lands on tails, the patient will be given a placebo. 1000 patients are recruited to the study. Approximately how many are likely to receive the new drug?

A. 1000 **B**. 500 **C**. 50 **D**. 250 **E**. 25

Item 39. Medical students were asked whether they had ever engaged in binge drinking. Their answers, classified by gender, are shown in the table below.

	Binge Drinking	
Gender	**Yes**	**No**
Male	**43**	**50**
Female	**28**	**92**

39. What percentage of those who reported engaging in binge drinking were male?

A. 46% **B**. 20% **C**. 70% **D**. 43% **E**. 61%

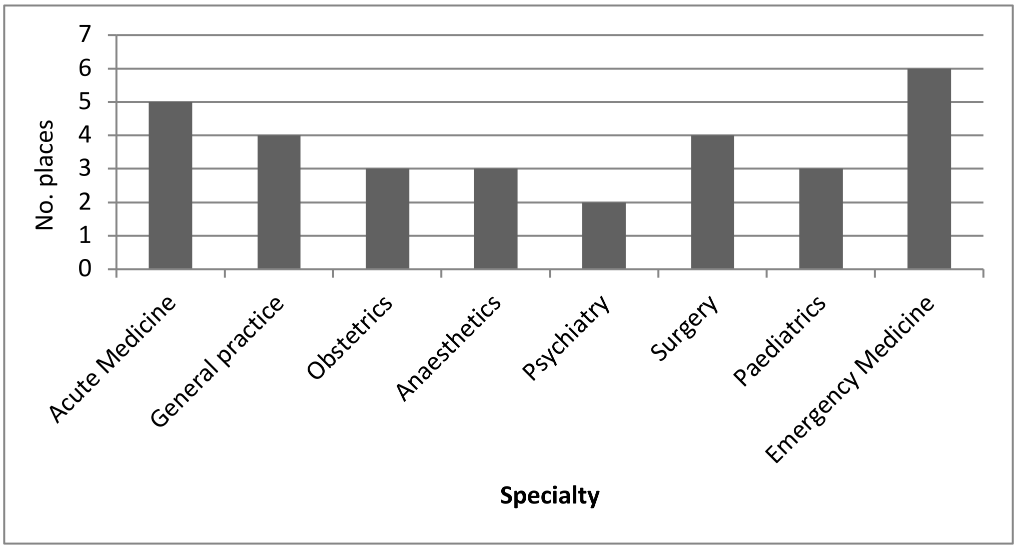
40. Poppy (age 19, weight 65kg) is admitted to the acute medical ward with a diagnosis of Diabetic Ketoacidosis. Clinical Guidelines state that she should be given insulin at a rate of 0.1 unit/kg/hr. You are asked to prescribe her insulin infusion. The preparation for infusion contains 50 units of insulin in a volume of 50ml 0.9% NaCl. What rate of infusion will you prescribe?

A. 13 ml/hr **B**. 7.6 ml/hr **C**. 0.65ml/hr **D**. 6.5 ml/hr **E**. 0.13 ml/hr

41. The chance of a hip replacement operation being cancelled is 1%. If 1000 people are scheduled for hip replacement operations, how many are likely to have their operation cancelled?

A. 1 **B**. 0.1 **C**. 0.01 **D**. 10 **E**. 100

Item 42. The chart below shows the number of training places available in various specialties on a Foundation Year 2 programme.



42. Helen is has applied for a place on this training programme. If places are allocated at random, what is her chance of being placed in Emergency Medicine?

A. 20% **B**. 5% **C**. 40% **D**. 12.5% **E**. 25%

43. Mo weighs 100kg, and presents to A&E with a wound in his thigh. You are asked to suture the wound under local anaesthetic. The available local anaesthetic is a solution containing bupivacaine 5mg/ml. The maximum dose of bupivacaine that can be safely given is 2mg/kg. What is the maximum volume of bupivacaine you can use to suture Mo's wound?

A. 4 ml **B**. 20 ml **C**. 40 ml **D**. 80 ml **E**. 200 ml

Blank Page

APPENDIX 5.

GENDER AND NUMERACY

In my initial study with the MINT, there was some evidence of a gender effect, with male participants outperforming females; although median scores were similar, there were significantly more males in the top decile, and more females in lowest decile. However, there were only 135 participants in the study, so further research would be needed to establish whether this is a consistent finding.

Literature review shows some evidence of a gender effect on performance in mathematics. Assessments of 15-year olds participating in the Programme for International Student Assessment (OECD, 2009) showed no gender difference in the lowest performing students; however, a large gender effect was observed among higher performing students, with boys outperforming girls (Stoet & Geary, 2013). Despite this, the evidence from two large US studies involving university entrants is conflicting: while Bridgeman (1992) found no difference associated with gender, Sikorskii *et al* (2011) found that males outperformed their female counterparts. However, some studies suggest that apparent gender differences in performance may relate to test format rather than to mathematical ability; several studies suggest that females do less well in multiple choice and true/false assessments (Anderson 2002; Simkin & Kuechler 2005; Betts *et al* 2009; Kelly & Dennick, 2009), possibly because they are more risk averse, and less likely to guess at answers. This may be relevant to the research conducted by Stoet & Geary (2013) and Sikorskii *et al* (2011), since both involve multiple choice tests. An investigation into the effect of gender on different types of assessment by Hartley *et al* (2007) was inconclusive; however, although their research involved university students, it did not relate to mathematics. Nonetheless, they present an interesting review of research indicating factors that are considered to contribute towards better performance by men (more confident, more likely to take risks, less anxious, less likely to fear failure) and women (better verbal skills, more conscientious, collaborate with others when revising) (Hartley *et al* 2007).

Gender is rarely discussed in relation to drug dose calculation tests in nurses; however, Bagnasco *et al* (2016) found that men performed better than women on all questions in their drug dose calculation test. Furthermore, Windish *et al* (2007), in their study of 297 doctors in training in the US, reported that men outperformed women on their test of biostatistics; however, they dismiss this finding as an oddity, stating that it is not supported by the literature, since researchers rarely report data according to gender. My data to date suggests that there may be a gender effect associated with the MINT, as male participants have consistently achieved higher scores (Table A.4). However, further research is warranted to determine whether this is a genuine effect.

Table A.5. Gender and MINT score

	n	Gender	Mean (%)	SD	Median	Range	IQR
2017	110	Male (36)	34.4 (80%)	4.4	35	22-43	32-37
		Female (59)	30.5 (71%)	5.3	32	19-41	27-34
		Undeclared (15)					
2018	120	Male (47)					
		Female (60)					
		Undeclared (13)					
2019	111	Male (45)	34 (79%)		35	25 – 43	31 – 37
		Female (63)	32 (74%)		31	22 - 42	29 - 36
		Undeclared (3)					

Blank Page

APPENDIX 6.

DYSLEXIA AND CN

Dyslexia is a specific learning disability, mainly affecting literacy, and affecting approximately 10% the general population (British Dyslexia Association (BDA), 2017). Developmental Dyscalculia is a specific learning disability characterised by impaired acquisition of arithmetic skills, and is estimated to affect approximately 5-6% of schoolchildren (Shalev, 2004). There is some overlap between dsyslexia and dyscalculia (Gibson and Leinster, 2011; BDA, 2017). The incidence of dyscalculia in medical students is not generally recorded at entry to medical school, nor by the Higher Education Statistics Agency (HESA) (HESA, 2017): dyscalculia is coded with dyslexia under the umbrella term of specific learning disability (SPLD). Medical students who have been diagnosed with a SPLD are given increased time for examinations (Gibson and Leinster, 2011; McKendree & Snowling, 2011).

Since medical students with dyslexia/SPLD may not be representative of the overall cohort, I asked students to declare whether they had been diagnosed with dyslexia and were given additional time in university examinations. In addition to analysing their results as part of the overall cohort, I did a subgroup analyses of their results. The incidence of dyslexia was 10 - 12% for participants in my research studies; this includes a small number of students who stated that they were given extra time in examinations for an SPLD other than dyslexia. Although the scores of students with dyslexia/SPLD appeared lower than those of non-dyslexic students, the difference in results was not statistically significant (Table A.5).

Table A.6. Dyslexia and MINT score

	No.	Dyslexia	Mean	SD	Median	Range
2017	12	Yes	29.3		30	21 - 36
2018	20	Yes	32		33	20 – 40
2019	13	Yes	30		30	25 - 36
2017	83*	No	31.8	5.23	33	19 – 43
2018	77*	No	32.6	4.9	33	20 - 42
2019	90*	No	33.2	4.9	33	22 - 43

**Some students did not indicate whether they were dyslexic: the numbers for 2017, 2018 and 2019 are 15, 23 and 7 students respectively.*

Blank Page

REFERENCES

Abramson, E.L., Bates, D.W., Jenter, C., Volk, L.A., Barron, Y., Quaresimo, J., Seger, A.C., Burdick, E., Simon, S. & Kaushal, R. (2012). Ambulatory prescribing errors among community-based providers in two states. *Journal of the American Medical Informatics Association*, *19*, 644–8.

Adams, A., & Duffield, C. (1991). The value of drills in developing and maintaining numeracy skills in an undergraduate nursing programme. *Nurse Education Today*, *11*, 213-219.

Ahmed, Z., Garfield, S., Jani, Y., Jheeta, S. & Franklin, B.D. (2016). Impact of electronic prescribing on patient safety in hospitals: implications for the UK. *Clinical Pharmacist*, *8*(5),1-11.

Ancker, J.S., & Kaufman, D. (2007). Rethinking health numeracy: a multidisciplinary literature review. *Journal of the American Medical Informatics Association*, *14*(6), 713-721.

Anderson, J. (2002). Gender-related differences on open and closed assessment tasks. *International Journal of Mathematical Education in Science and Technology*, *33*(4), 495-503.

Anderson, B.L., Obrecht, N.A., Chapman, G.B., Driscoll, D.A. & Schulkin, J. (2011). Physicians' communication of Down syndrome screening test results: the influence of physician numeracy. *Genetic Medicine, 13*(8)**,** 744-9.

Apter, A.J., Cheng, J., Small, D., Bennett, I.M., Albert, C., Fein, D.G., George, M. & Van Horne, S. (2006). Asthma numeracy skill and health literacy. *Journal of Asthma, 43*(9), 705-710.

Armitage, G., & Knapman, H. (2003). Adverse events in drug administration: a literature review. *Journal of Nursing Management, 11*, 130-140.

Avery, T., Barber, N., Ghaleb, M., Franklin, B.D., Armstrong, S., Crowe, S., Dhillon, S., Freyer, A., Howard, R. & Pezzolesi, C. (2012). *Investigating the prevalence and causes of prescribing errors in general practice. London: The General Medical Council: PRACtICe Study. A report for the GMC*. Nottingham: October 2011. Retrieved from https://www.gmc-uk.org/ Investigating_the_prevalence_and_causes_of_prescribing_errors_ in_general_practice___The_PRACtICe_study_Report_ May_2012_48605085.pdf.

Bagnasco, A., Galaverna, L., Aleo, G., Grugnetti, A.M., Rosa, F., & Sasso, L. (2016). Mathematical calculation skills required for drug administration in undergraduate nursing students to ensure patient safety: A descriptive study. Drug calculation skills in nursing students. *Nurse Education in Practice, 16*, 33-39.

Baker, D.W., Gazmararian, J.A., Williams, M.V., Scott, T., Parker, R.M., Green, D., Ren, J. & Peel, J. (2002). Functional health literacy and the risk of hospital admission among Medicare managed care enrollees. *American Journal of Public Health*, *92*(8),1278-83.

Baker, D.W., Wolf, M.S., Feinglass, J., Thompson, J.A., Gazmararian, J.A., & Huang, J. (2007). Health literacy and mortality among elderly persons. *Archives of Internal Medicine, 167*(14), 1503-9.

Batchelor, H. (2004). The importance of a mathematics diagnostic test for incoming pharmacy undergraduates. *Pharmacy Education*, *4*(2):69-74.

BBC news. (2015). Leicester doctor guilty of manslaughter of Jack Adcock, 6. Retrieved from http://www.bbc.co.uk/news/uk-england-leicestershire-34722885 [accessed March 2, 2019].

Ben-Shlomo, Y., Fallon, U., Sterne, J. & Brookes, S. (2004). Do medical students with A-level mathematics have a better understanding of the principles behind evidence-based medicine? *Medical Teacher, 26*(8)**,** 731-733.

Berman, A. (2004). Reducing medication errors through naming, labeling, and packaging. *Journal of Medical Systems, 28*(1):9-29

Berwick D. (2013). *A promise to learn - a commitment to act. Improving the safety of patients in England*. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/226703/Berwick_Report.pdf

Betts, L.R., Elder, T.J., Hartley, J. & Trueman, M. (2009). Does correction for guessing reduce students' performance on multiple-choice examinations? Yes? No? Sometimes? *Assessment and Evaluation in Higher Education, 34*(1), 1–15.

Birenbaum, M. & Tatsuoka, K.K. (1987). Open-ended versus multiple-choice response formats – it does make a difference for diagnostic purposes. *Applied Psychological Measurement, 11*(4), 385-395.

Blais, K. & Bath, J.B. (1992). Drug calculation errors of baccalaureate nursing students. *Nurse Educator, 17*(1), 12–15.

Bliss-Holtz, J. (1994). Discriminating types of medication calculation errors in nursing practice. *Nursing Research, 43*(6), 373-375.

Brady, A.M., Malone, A.M., & Fleming, S. (2009). A literature review of the individual and systems factors that contribute to medication errors in nursing practice. *Journal of Nursing Management, 17,* 679-697.

Brant, R. (n.d.). Comparing means for two independent samples. Retrieved from http://www.stat.ubc.ca/~rollin/stats/ssize/n2.html [Accessed February 17, 2019]

Brennan, T.A., Leape, L.L., Laird, N.M., Hebert, L., Localio, A.R., Lawthers, A.G., Newhouse, J.P., Weiler, P.C., & Hiatt, H.H. (1991). Incidence of adverse events and negligence in hospitalized patients: results of the Harvard Medical Practice Study I. *New England Journal of Medicine, 324*, 370-376.

Brennan, T.A., Gawande, A., Thomas, E. & Studdert, D. (2005). Accidental deaths, saved lives, and improved quality. *New England Journal of Medicine*, 353(13), 1405-9.

Bridgeman, B. (1992). A Comparison of Quantitative Questions in Open-Ended and Multiple-Choice Formats. *Journal of Educational Measurement, 29*(3), 253-271.

Briggs, T.W.R. (2012). *Getting it right the first time: improving the quality of orthopaedic care within the National Health Service in England*. Retrieved from https://www http://www.gettingitrightfirsttime.com/downloads/BriggsReportA4_FIN.pdf

British Dyslexia Association. (2017). Dyslexic Definitions. Retrieved from: www.bdadyslexia.org.uk/dyslexic/definitions [Accessed August 24, 2017]

British Pharmaceutical Society & Medical Schools Council. (2013). Prescribing Safety Assessment. Retrieved from https://prescribingsafetyassessment.ac.uk/

Campbell, I. (2007). Chi-squared and Fisher-Irwin tests of two-by-two tables with small sample recommendations. *Statistics in Medicine, 26*, 3661-3675.

Caverly, T. J., Prochazka, A., Binswanger, I., Kutner, J. S. & Matlock, D. (2012). Getting the Gist of Health Risks. *Section on Statistical Education – JSM*. Retrieved from http://www.statlit.org/pdf/2012-Caverly-ASA.pdf

Cleland, J., Leggett, H., Sandars, J., Costa, M. J., Patel, R. & Moffat, M. (2013). The remediation challenge: theoretical and methodological insights from a systematic review. *Medical Education, 47*, 242–251.

Close, S., Oldham, E., Surgenor, P., Shiel, G., Dooley, T., & O'Leary, M. (2008). *The effects of calculator use on mathematics in schools and in certificate examinations. Final report on phase 2*. Dublin: St. Patrick's College, Trinity College and Educational Research Centre. Retrieved from http://www.erc.ie/documents/calculator_final_report_phase2.pdf

Coben, D., & Weeks, K. (2014). Meeting the mathematical demands of the safety-critical workplace: medication dosage calculation problem-solving for nursing. *Educational Studies in Mathematics, 86*, 253–270.

Cook D.A., Levinson A.J., Garside S., Dupras, D.M. Erwin, P.J. & Montori, V.M. (2008). Internet-based learning in the health professions: a meta-analysis. *Journal of the American Medical Association, 300*, 1181–96.

Coombes, I. D., Stowasser, D. A., Coombes, J. A. & Mitchell, C. (2008). Why do interns make prescribing errors? A qualitative study. *Medical Journal of Australia, 188*(2), 89.

Cotton, D., & Gresty, K. (2006). Reflecting on the think-aloud method for evaluating e-learning. *British Journal of Educational Technology, 37*(1), 45-54.

Coyne, E., Needham, J. & Rands, H. (2013). Enhancing student nurses' medication calculation knowledge: Integrating theoretical knowledge into practice. *Nurse Education Today, 33*(9), 1014-1019.

Cranshaw, J. Gupta, K.J. & Cook, T.M. (2009). *Litigation related to drug errors in anaesthesia: an analysis of claims against the NHS in England 1995–2007*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2044.2009.06107.x/pdf

da Silva, B. A., & Krishnamurthy, M. (2016). The alarming reality of medication error: a patient case and review of Pennsylvania and National data. *Journal of Community Hospital Internal Medicine Perspectives*, *6*, 31758. Retrieved from http://dx.doi.org/10.3402/jchimp.v6.31758

Dean, B., Schachter, M., Vincent, C. & Barber, N. (2002). Causes of prescribing errors in hospital inpatients: a prospective study. *Lancet, 359*(9315), 1373-1378.

Department of Health. (2015). *Delivering high quality, effective, compassionate care: Developing the right people with the right skills and the right values*. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/203332/29257_2900971_Delivering_Accessible.pdf

de Vries, E. N., Ramrattan, M. A., Smorenburg, S. M., Gouma, D. J. & Boermeester, M. A. (2008). The incidence and nature of in-hospital adverse events: a systematic review. *Quality and Safety in Health Care, 17*(3), 216-223.

DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *Canadian Journal for the Scholarship of Teaching and Learning, 2*(2), 4.

DiBattista, D., Sinnige-Egger, J. & Fortuna, G. (2014). The “None of the Above” Option in Multiple-Choice Testing: An Experimental Study. *Journal of Experimental Education, 82*(2):168-183

Donaldson, L., Kelley, E.T., Dhingra-Kumar, N., Kieny, M, & Sheikh, A. (2017). Medication Without Harm: WHO’s Third Global Patient Safety Challenge. *Lancet, 389*, 1680-1681. Retrieved from https://www.thelancet.com/action/showPdf?pii=S0140-6736%2817%2931047-4

Dornan, T., Ashcroft, D., Heathfield, H., Lewis, P., Miles, J., Taylor, D., Tully, M. & Wass, V. (2009). *An in depth investigation into causes of prescribing errors by foundation trainees in relation to their medical education—EQUIP study*. London: General Medical Council, 1-215. Retrieved from http://www.gmcuk.org/FINAL_Report_prevalence_and_causes_of_prescribing_errors.pdf_28935150.pdf

Downing, S.M. (2003). Guessing on selected response examinations. *Medical Education, 37*(8), 670–671.

Elliott, R.A., Camacho, E., Campbell, F., Jankovic, D., Martyn St James, M., Kaltenthaler, E., Wong, R., Sculpher, M.J. & Faria, R. (2018). Prevalence and economic burden of medication errors in the NHS in England. Policy Research Unit in Economics Evaluation of Health and Care Interventions. Retrieved from www.eepru.org.uk/article/prevalence-andeconomic-burden-of-medication-errors-in-the-nhs-in-england.

Elstein, A. S., & Schwarz, A. (2002). Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *British Medical Journal, 32*, 729–32.

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: Development of the subjective numeracy scale. *Medical Decision Making, 27*(5)**,** 672-680.

Filik, R., Purdy, K., Gale, A. & Gerrett, D. (2006). Labeling of medicines and patient safety: evaluating methods of reducing drug name confusion. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 48*(1), 39-47.

Fleming, S., Brady, A. M. & Malone, A. M. (2014). An evaluation of the drug calculation skills of registered nurses. *Nurse Education in Practice, 14*(1), 55-61.

Follette, K. B., McCarthy, D. W., Dokter, E., Buxner, S. & Prather, E. (2015). The Quantitative Reasoning for College Science (QuaRCS) Assessment, 1: Development and Validation. *Numeracy, 8*(2), 2. Retrieved from http://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=1191andcontext=numeracy

Franklin, B. D., Reynolds, M., Shebl, N. A., Burnett, S. & Jacklin, A. (2011). Prescribing errors in hospital inpatients: a three-centre study of their prevalence, types and causes. *Postgraduate Medical Journal, 87*(1033), 739-745.

Freeman, J.V., Collier, S., Staniforth, D., & Smith, K.J. (2008). Innovations in curriculum design: A multi-disciplinary approach to teaching statistics to undergraduate medical students. *BMC Medical Education, 8*, 28. doi:10.1186/1472-6920-8-28

Frontier economics. (2014). *Exploring the costs of unsafe care in the NHS: A report prepared for the Department of Health*. Retrieved from http://www.frontier-economics.com/documents/2014/10/exploring-the-costs-of-unsafe-care-in-the-nhs-frontier-report-2-2-2-2.pdf

Fudickar, A., Hörle, K., Wiltfang, J. & Bein, B. (2012). The effect of the WHO surgical safety checklist on complication rate and communication. *Deutsches Ärzteblatt International, 109*(42), 695. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3489074/

Funk, S. F., & Dickson, K. L. (2011). Multiple-Choice and Short-Answer Exam Performance in a College Classroom. *Teaching of Psychology, 38*(4) 273-277.

Galligan, L., Loch, B, & Lawrence, J. (2010). *An integrative approach to building professional attributes in a first year nursing course: Description and preliminary analysis of academic numeracy*. Retrieved from https://eprints.usq.edu.au/4330/

Galligan, L. (2013). A systematic approach to embedding academic numeracy at university. *Higher Education Research and Development, 32*(5):734–747.

Galligan, L., & Hobohm, C. (2015). Investigating students' academic numeracy in 1st level university courses. *Mathematics Education Research Journal, 27*, 129–145.

Gazmararian, J.A., Williams, M.V., Peel, J. & Baker, D.W. (2003). Health literacy and knowledge of chronic disease. *Patient Education and Counseling, 51*(3), 267-75.

Gazmararian, J.A., Curran, J.W., Parker, R.M., Bernhardt, J.M. & De Buono, B.A. (2005). Public health literacy in America: an ethical imperative. *American Journal of Preventive Medicine, 28*(3):317-22.

General Medical Council. (2015). *First do no harm: enhancing patient safety teaching in undergraduate medical education*. Retrieved from http://www.gmc-uk.org/First_do_no_harm_patient_safety_in_undergrad_education_FINAL.pdf_62483215.pdf

General Medical Council. (2018). *Outcomes for graduates*. Retrieved from https://www.gmc-uk.org/-/media/documents/dc11326-outcomes-for-graduates-2018_pdf-75040796.pdf

Ghosh, A. K., & Ghosh, K. (2005). Translating evidence-based information into effective risk communication: Current challenges and opportunities. *Journal of Laboratory and Clinical Medicine, 145*(4), 171-180.

Gibson, S., & Leinster, S. (2011). How do students with dyslexia perform in extended matching questions, short answer questions and observed structured clinical examinations? *Advances in Health Sciences Education, 16*, 395–404.

Gigerenzer, G. (2014). *Risk Savvy: How to make good decisions*. London: Allen Lane.

Gigerenzer, G., & Edwards, A. (2003). Simple tools for understanding risks: from innumeracy to insight. *British Medical Journal, 327*(7417), 741.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest, 8*(2), 53-96.

Gigerenzer, G., & Gray, M. (2011). *Better doctors, better patients, better decisions: Envisioning health care 2020*. Cambridge Mass: MIT.

Glavin, R. (2010). Drug errors: consequences, mechanisms, and avoidance. *British Journal of Anaesthesia, 105*(1), 76-82.

Gleason, K.M., McDaniel, M.R., Feinglass, J., Baker, D.W., Lindquist, L., Liss, D., & Noskin, G.A. (2010). Results of the Medications at Transitions and Clinical Handoffs (MATCH) study: an analysis of medication reconciliation errors and risk factors at hospital admission. *Journal of General Internal Medicine, 25*, 441–7.

Golbeck, A.L., Ahlers-Schmidt, C.R., Paschal, A.M., & Dismuke, S.E. (2005). A definition and operational framework for health numeracy. *American Journal of Preventive Medicine, 29*(4), 375-376.

Gordon, M., Catchpole, K. & Baker, P. (2013). Human factors perspective on the prescribing behavior of recent medical graduates: implications for educators. *Advances in Medical Education and Practice, 3*(4)1-9.

Grugnetti, A.M., Bagnasco, A., Rosa, F., & Sasso, L. (2014). Effectiveness of a clinical skills workshop for drug dosage calculation in a nursing program. *Nurse Education Today, 34*(4), 619-624.

Hall, E.T., Weaver, K.W., Perino, A.C., Elder, A., & Verghese, A. (2018). 'A Man Walks into a Bar': Riddles in the Teaching of Medicine. *American Journal of Medicine*, *131*(9), 1000–1002.

Hartley, J., Betts, L., & Murray, W. (2007). Gender and assessment: differences, similarities and implications. *Psychology Teaching Review, 13*(1), 34-47.

Harries, C. S. & Botha J.H. (2013). Can medical students calculate drug doses? *Southern African Journal of Anaesthesia and Analgesia, 19*(5), 248-251.

Heaton, A., Webb, D. J. & Maxwell, S. R. (2008). Undergraduate preparation for prescribing: the views of 2413 UK medical students and recent graduates. *British Journal of Clinical Pharmacology, 66*(1), 128-134.

Hegener, M. A., Buring, S. M., & Papas, E. (2013). Impact of a required pharmaceutical calculations course on mathematics ability and knowledge retention. *American Journal of Pharmaceutical Education, 77*(6), 124.

Higher Education Statistics Agency. (2017). *Higher education student enrolments and qualifications at higher education providers in the United Kingdom 2015/16*. Retrieved from https://www.hesa.ac.uk/news/12-01-2017/sfr242-student-enrolments-and-qualifications [Accessed 22 March, 2019]

Hughes, R. G. & Edgerton, E. A. (2005). Reducing Pediatric Medication Errors: Children are especially at risk for medication errors. *American Journal of Nursing, 105*(5), 79-84.

Huizinga, M. M., Elasy, T. A., Wallston, K. A., Cavanaugh, K., Davis, D., Gregory, R. P., Fuchs, L. S., Malone, R., Cherrington, A. & DeWalt, D. A. (2008). Development and validation of the Diabetes Numeracy Test (DNT). *BMC Health Services Research, 8*(1), 96.

Hurley, T. V. (2017). Experiential teaching increases medication calculation accuracy among baccalaureate nursing students. *Nursing Education Perspectives, 38*(1), 34-36.

Hutton, B.M. (1998). Do school qualifications predict competence in nursing calculations? *Nurse Education Today, 18*, 25-31

Hutton, M., Coben, D., Hall, C., Rowe, D., Sabin, M., Weeks, K.W. & Woolley, N. (2010). Numeracy for nursing, report of a pilot study to compare outcomes of two practical simulation tools. An online medication dosage assessment and practical assessment in the style of objective structured clinical examination. *Nurse Education Today*, *30*, 608-614.

Irwin, A., Mearns, K., Watson, M. & Urquhart, J. (2013). The effect of proximity, Tall Man lettering, and time pressure on accurate visual perception of drug names. *Human Factors, 55*(2), 253-66.

Jha, A.K., Larizgoitia, I., Audera-Lopez, C., Prasopa-Plaizier, N., Waters, H. & Bates, D.W. (2013). The global burden of unsafe medical care: analytic modelling of observational studies. *BMJ Quality & Safety, 25*(4). 22, 809–815.

Johnson, T.V., Abbasi, A., Schoenberg, E.D., Kellum, R., Speake, L.D., Spiker, C., Foust, A., Kreps, A., Ritenour, C.W.M., Brawley, O.W. & Master, V. A. (2014). Numeracy among trainees: are we preparing physicians for evidence-based medicine? *Journal of Surgical Education, 71*(2), 211-215.

Johnson, S. A., & Johnson, L. J. (2002). The 4 Cs: A model for teaching dosage calculations. *Nurse Educator, 27*(2), 79-83.

Jordan, S. (2013). E-assessment: Past, present and future. *New Directions in the Teaching of Physical Sciences, 9*(1), 87-106.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Strauss, and Giroux.

Kalet, A., Guerrasio, J., & C. L. Chou. (2016). Twelve tips for developing and maintaining a remediation program in medical education. *Medical Teacher*, *38*(8), 787-792.

Kastner, M., & Stangl, B. (2011). Multiple Choice and Constructed Response Tests: Do Test Format and Scoring Matter? *Procedia Social and Behavioral Sciences, 12*, 263–273.

Kelly, S., & Dennick, R. (2009). Evidence of gender bias in True-False-Abstain medical examinations. *BMC Medical Education*, *9*, 32.

Koharchik, L., Hardy, E., King, M. & Garibo, Y. (2014). Evidence-based approach to improve nursing student dosage calculation proficiency. *Teaching and Learning in Nursing, 9*, 69-74.

Kohn, L., Corrigan, J. & Donaldson, M. Eds. (2002). *To err is human: building a safer health system*. National Academy of Sciences, Institute of Medicine (US), Committee on Quality of Health Care in America. Washington (DC): National Academies Press.

Lag, T., Bauger, L., Lindberg, M., & Friborg, O. (2013). The Role of Numeracy and Intelligence in Health-Risk Estimation and Medical Data Interpretation. *Journal of Behavioral Decision Making,* [wileyonlinelibrary.com] DOI: 10.1002/bdm.1788

Lambert, B.L., Schroeder, S.R. & Galanter, W.L. (2016). Does Tall Man lettering prevent drug name confusion errors? Incomplete and conflicting evidence suggest need for definitive study. *BMJ Quality & Safety, 25*(4), 213–217.

Latif, D. A., & Grillo, J.A. (2002). Assessing the basic math skills of first-year doctor of pharmacy students. *Journal of Pharmacy Teaching 9*(2),17-25.

Leape, L. L., & Berwick, D. M. (2005). Five years after To Err Is Human: what have we learned? *Journal of the American Medical Association, 293*(19), 2384-2390.

Lee, S., Browne, R., Dudzic, S. & Stripp, C. (2010). *Understanding the UK Mathematics Curriculum Pre-Higher Education: a guide for Academic Members of Staff. Mathematics in Education and Industry*. Retrieved from http://www.mei.org.uk/files/pdf/Pre_Uni_Maths_Guide.pdf

Lesar, T., Briceland, L. & Stein, D. (1997). Factors related to errors in medication prescribing. *Journal of the American Medical Association, 277*, 312–7.

Levy, H., Ubel, P.A., Dillard, A.J., Weir, D.R. & Fagerlin, A. (2014). Health numeracy: the importance of domain in assessing numeracy. *Medical Decision Making, 34*,107-115.

Lin, S-K., & Singh, C. (2013). Can free-response questions be approximated by multiple-choice equivalents? *American Journal of Physics, 81*, 624-629.

Lindberg, S.M., Shibley Hyde, J.S. & Petersen, J.L. (2010). New trends in gender and mathematics performance: a meta-analysis. *Psychological Bulletin*, *136*(6), 1123-1135

Lipkus, I. M., Samsa, G. & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making, 21*(1)**,** 37-44.

Mackie, J.E., & Bruce C. D. (2016). Increasing nursing students' understanding and accuracy with medical dose calculations: A collaborative approach. *Nurse Education Today, 40*,146-153.

Malcolm, R.K., & McCoy, C.P. (2007). Evaluation of numeracy skills in first year pharmacy undergraduates 1999-2005. *Pharmacy Education*, *7*(1), 53-59.

Malhotra A., Maughan, D., Ansell, J., Lehman, R., Henderson, A., Gray, M., Stephenson, T. & Bailey, S. (2015). Choosing Wisely in the UK: the Academy of Medical Royal Colleges' initiative to reduce the harms of too much medicine. *British Medical Journal, 350*, h2308.

Maxwell, S.R.J., Coleman, J.J., Bollington, L., Taylor, C. & Webb, D.J. (2017). Prescribing Safety Assessment 2016: Delivery of a national prescribing assessment to 7343 UK final-year medical students. *British Journal of Clinical Pharmacology, 83*, 2249–2258.

McAllister, D., & Guidice, R.M. (2012). This is only a test: A machine-graded improvement to the multiple-choice and true-false examination. *Teaching in Higher Education, 17*(2), 193–207

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: a literature review. *Medical Teacher, 26*(8), 709-712.

McDonald, K., Weeks, K.W., & Moseley, L. (2013). Safety in numbers 6: Tracking pre-registration nursing students' cognitive and functional competence development in medication dosage calculation problem solving: The role of authentic learning and diagnostic assessment environments. *Nurse Education in Practice*, *13*, e66–77.

McKendree, J., & Snowling, M.J. (2011). Examination results of medical students with dyslexia. *Medical Education*, *45*, 176–182.

McLean, M., Shaban, S., & Murdoch-Eaton, D. (2011). Transferable skills of incoming medical students and their development over the first academic year: The United Arab Emirates experience. *Medical Teacher, 33*(6), e297-e305.

McMullan, M., Jones, R. & Lea, S. (2010). Patient safety: numerical skills and drug calculation abilities of nursing students and registered nurses. *Journal of Advanced Nursing, 66*(4), 891-899.

MedCalc Software bvba (BE)a. Comparison of means calculator. Retrieved from www.medcalc.org/calc/comparison_of_means.php [accessed 4 October 2018]

MedCalc Software bvba (BE)b. Comparison of proportions calculator. Retrieved from www.medcalc.org/calc/comparison_of_proportions.php [accessed 4 February 2017]

Monrouxe, L., Bullock, A., Cole, J., Gormley, G., Kaufhold, K., Kelly, N., Mattick, K., Rees, C., Scheffler, G. & Jefferies, M. C. (2014). *How prepared are UK medical graduates for practice?* Retrieved from: http://www.gmc-uk.org/How_Prepared_are_UK_Medical_Graduates_for_Practice_SUBMITTED_Revised_140614.pdf_58034815.pdf

Momtahan, K., Burns, C. M., Jeon, J., Hyland, S. & Gabriele, S. (2008). Using human factors methods to evaluate the labelling of injectable drugs. *Healthcare Quarterly, 11*(Sp), 122-128.

Moyer, V. A. (2012). What we don't know can hurt our patients: physician innumeracy and overuse of screening tests. *Annals of Internal Medicine, 156*(5), 392-393.

Murphy, M. A., & Graveley, E. A. (1990). Drug Calculation Examinations: Do Calculators Make a Difference? *Nurse Educator, 15*(1), 35-43.

National Health Service Education for Scotland. (2010). *Patient Safety: Cost Implications of Adverse Health Events*. Retrieved from http://www.nes.scot.nhs.uk/media/6472/PS%20Cost%20Briefing%20Paper.pdf

National Health Service England. (2013). *Human Factors in Healthcare: A Concordat from the National Quality Board.* Retrieved from https://www.england.nhs.uk/wp-content/uploads/2013/11/nqb-hum-fact-concord.pdf

National Health Service England. (2014). *Patient Safety Collaboratives*. Retrieved from https://www.england.nhs.uk/patientsafety/collaboratives/

National Health Service Improvement. (2017). *Request under the Freedom of Information Act: Medication Error*. Retrieved from https://improvement.nhs.uk/documents/2184/FOI_Medication_Errors_saMoNx4.pdf

National Institute for Health and Care Excellence (NICE). (2013). *Intravenous fluid therapy in adults in hospital CG174*. Retrieved from http://guidance.nice.org.uk/CG174

National Numeracy. (2019). *Numeracy for Q-Step. Assessing the numerical ability of undergraduate students in order to better support them*. Retrieved from https://www.nationalnumeracy.org.uk/sites/default/files/numeracy_for_q-step_report_february_2019.pdf [Accessed 5 March, 2019]

National Patient Safety Association. (2007). *Safety in doses: medication safety incidents in the NHS*. Retrieved from http://www.nrls.npsa.nhs.uk/EasySiteWeb/getresource.axd?AssetID=61392

National Patient Safety Association. (2011). *Patient Safety*. Retrieved from http://www.nrls.npsa.nhs.uk/report-a-patient-safety-incident/

Nazar, H., Nazar, M., Rothwell, C., Portlock, J., Chaytor, A., & Husband, A. (2015). Teaching safe prescribing to medical students: perspectives in the UK. *Advances in Medical Education and Practice, 6*, 279.

Neale, G., Woloshynowych, M. & Vincent, C. 2001. Exploring the causes of adverse events in NHS hospital practice. *Journal of the Royal Society of Medicine, 94*(7), 322-330.

Nutbeam, D. (2000). Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century. *Health Promotion International, 15*(3), 259-267.

O'Hara, J.K., Reynolds, C., Moore, S., Armitage, G., Sheard, L., Marsh, C., Watt, I., Wright, J., & Lawton, R. (2018). *BMJ Quality and Safety, 27*, 673–682.

Oldridge, G., Gray, K., McDermott, L., & Kirkpatrick, C. (2004). Pilot study to determine the ability of health-care professionals to undertake drug dose calculations. *Internal Medicine Journal, 34*(6), 316-319.

Organisation for Economic Cooperation and Development. (2009). *Take the Test: sample questions from OECD's PISA assessments*. Retrieved from http://www.oecd-ilibrary.org/docserver/download/9809051e.pdf?expires=1381664341andid=idandaccname=ocid177243andchecksum=E8BD943BCB2E202C04E307B6BC33ECAE

Organisation for Economic Cooperation and Development. (2013). *Time for the U.S. to Reskill? What the Survey of Adult Skills Says*. *OECD Skills Studies*. OECD: Paris. Retrieved from https://doi.org/10.1787/9789264204904-en.

Organisation for Economic Cooperation and Development. (2018). *PISA 2012 Results in Focus: What 15 year olds know and what they can do with what they know*. Retrieved from https://www.oecd.org/pisa/keyfindings/pisa-2012-results-overview.pdf

Orser, B. A., Chen, R. J., & Yee, D. A. (2001). Medication errors in anesthetic practice: a survey of 687 practitioners. *Canadian Journal of Anaesthesia, 48*(2), 139-146.

Parliamentary Health Service Ombudsman. (2013). Time to Act: Severe Sepsis: rapid diagnosis and treatment saves lives. Retrieved from https://www.ombudsman.org.uk/publications/time-act-severe-sepsis-rapid-diagnosis-and-treatment-saves-lives-0 [Accessed 22 March, 2019]

Perneger, T.V. (1998). What's wrong with Bonferroni adjustments. *British Medical Journal, 316*, 1236–8.

Perneger, T. V. & Agoritsas, T. (2011). Doctors and Patients' Susceptibility to Framing Bias: A Randomized Trial. *Journal of General Internal Medicine*, 26(12),1411–7.

Peters, E., Dieckmann, N., Dixon, A., Hibbard, J. H., & Mertz, C. (2007) Less is more in presenting quality information to consumers. *Medical Care Research and Review, 64*(2), 169-190.

Rafter, N., Hickey, A., Condell, S., Conroy, R., O'Connor, P., Vaughan, D., & Williams, D. (2015). Adverse events in healthcare: learning from mistakes. *Quarterly Journal of Medicine, 108*(4), 273-277.

Rao, G., & Kanter, S. L. (2010). Physician numeracy as the basis for an evidence-based medicine curriculum. *Academic Medicine, 85*(11), 1794-1799.

Reason, J. (2000). Human error: models and management. *British Medical Journal*, *320*(7237), 768-770.

Reyna, V.F., Nelson, W.L., Han P.K. & Dieckmann NF. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin*, 135(6):943.

Rolfe, S., & Harper, N. (1995). Ability of hospital doctors to calculate drug doses. *British Medical Journal*, *310*(6988), 1173.

Roohr, K. C., Graf, E. A. & Liu, O. L. (2014). *Assessing quantitative literacy in higher education: An overview of existing research and assessments with recommendations for next-generation assessment.* ETS Research Report No. RR-14-22. Princeton, NJ: Educational Testing Service. Retrieved from https://files.eric.ed.gov/fulltext/EJ1109328.pdf

Ross, S., & Loke, Y.K. (2009). Do educational interventions improve prescribing of medical students and junior doctors? A systematic review. *British Journal of Clinical Pharmacology*, *67*(6), 662–670.

Rothman, R. L., Housam, R., Weiss, H., Davis, D., Gregory, R., Gebretsadik, T., Shintani, A. & Elasy, T. A. (2006). Patient understanding of food labels: the role of literacy and numeracy. *American Journal of Preventive Medicine*, *31*(5), 391-398.

Rowe, C., Koren, T. & Koren, G. (1998). Errors by paediatric residents in calculating drug doses. *Archives of Disease in Childhood*, *79*(1), 56-58.

Rowlands, G., Khazaezadeh, N., Oteng-Ntim, E., Seed, P., Barr, S. & Weiss, B. D. (2013). Development and validation of a measure of health literacy in the UK: the newest vital sign. *BMC Public Health*, *13*(1), 116.

Ryan, C., Ross, S., Davey, P., Duncan, E.M., Francis, J.J., Fielding, S., Johnston, M., Ker, J., Lee, A.J. & MacLeod, M. J. (2014). Prevalence and causes of prescribing errors: the PRescribing Outcomes for Trainee doctors Engaged in Clinical Training (PROTECT) study. *PloS ONE, 9*(1), e79802. doi:10.1371/journal.pone.0079802 Retrieved fromhttp://journals.plos.org/plosone/article?id=10.1371/journal.pone.0079802

Sabin M., Weeks K.W., Rowe D.A., Hutton, B.M, Coben, D., Hall, C. & Woolley, N. (2013). Safety in numbers 5: evaluation of computer-based authentic assessment and high fidelity simulated OSCE environments as a framework for articulating a point of registration medication dosage calculation benchmark. *Nurse Education in Practice, 13*, e55–65.

Sam, A.H., Hameed, S., Harris, J. & Meeran, K. (2016). Validity of very short answer versus single best answer questions for undergraduate assessment. *BMC Medical Education, 16*, 266. Retrieved from: https://bmcmededuc.biomedcentral.com/articles/10.1186/s12909-016-0793-z

Sam, A.H., Field, S.M., Collares, C.F., van der Vleuten, C.P.M., Wass, V.J., Melville, C., Harris, J. & Meeran, K. (2018). Very short answer questions: reliability, discrimination and acceptability. *Medical Education, 52*, 447-455.

Schapira, M. M., Fletcher, K. E., Gilligan, M. A., King, T. K., Laud, P. W., Matthews, B. A., Neuner, J. M. & Hayes, E. (2008). A framework for health numeracy: how patients use quantitative skills in health care. *Journal of Health Communication, 13*(5), 501-517.

Schuwirth, L.W., & van der Vleuten, C.P. (2004). Different written assessment methods: what can be said about their strengths and weaknesses? *Medical Education, 38*(9), 974-9.

Schwartz, L. M., Woloshin, S., Black, W. C. & Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine, 127*(11), 966.

Seden, K., Kirkham, J. J., Kennedy, T., Lloyd, M., James, S., Mcmanus, A., Ritchings, A., Simpson, J., Thornton, D. & Gill, A. (2013). Cross-sectional study of prescribing errors in patients admitted to nine hospitals across North West England. *British Medical Journal Open, 3*(1), e002036. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23306005

Selbst, S. M., Fein, J. A., Osterhoudt, K., & Wayne, H. (1999). Medication errors in a pediatric emergency department. *Pediatric Emergency Care, 15*(1), 1-4.

Shalev, R.S. (2004). Developmental Dyscalculia. *Journal of Child Neurology*, *19*(10):765-71.

Sheridan, S., & Pignone, M. (2002). Numeracy and the medical student's ability to interpret data. *Effective Clinical Practice, 5*(1), 35.

Shockley, J. S., McGurn, W. C., Gunning, C., Graveley, E. & Tillotson, D. (1989). Effects of Calculator Use on Arithmetic and Conceptual Skills of Nursing Students. *Journal of Nursing Education, 28*(9), 402-405.

Sikorskii, A., Melfi, V., Gilliland, D., Kaplan, J. & Ahn, S. (2011). Quantitative Literacy at Michigan State University, 1: Development and Initial Evaluation of the Assessment. *Numeracy, 4*(2), 5. Retrieved from https://scholarcommons.usf.edu/numeracy/vol4/iss2/art5/

Simkin, M.G., & Kuechler, W.L. (2005). Multiple-Choice Tests and Student Understanding: What Is the Connection? *Decision Sciences Journal of Innovative Education, 3*, 73-97.

Simonsen, B.O., Daehlin, G.K., Johansson, I., & Farup, P.G. (2014). Improvement of drug dose calculations by classroom teaching or e-learning: a randomised controlled trial in nurses. *BMJ Open*, *4*, e006025. doi:10.1136/bmjopen-2014- 006025

Simpson, C. M., Keijzers, G. B., & Lind, J. F. (2009). A survey of drug-dose calculation skills of Australian tertiary hospital doctors. *Medical Journal of Australia*, *190*(3), 117.

Smits, M., Zegers, M., Groenewegen, P., Timmermans, D., Zwaan, L., Van der Wal, G., & Wagner, C. (2010). Exploring the causes of adverse events in hospitals and potential prevention strategies. *Quality and Safety in Health Care*, *19*(5), 1-7.

Steen, L.A. (2004). Data shapes, symbols: Achieving balance in school mathematics. In B. Madison & L.A. Steen (Eds.), *Quantitative Literacy: Why numeracy matters for schools and colleges* (p. 55). Washington D.C., USA: Mathematical Association of America.

Stoet, G., & Geary, D.C. (2013). Sex differences in mathematics and reading achievement are inversely related: within- and across-nation assessment of 10 years of PISA data. *PLoS ONE 8*(3): e57988. Retrieved from http://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0057988andtype=printable

Stolic, S. (2014). Educational strategies aimed at improving student nurses' medication calculation skills: A review of the research literature*. Nurse Education in Practice, 14*(5), 491-503.

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the p value is not enough. *Journal of Graduate Medical Education, 4*(3), 279 – 282.

Tariq, V. N. (2002). A decline in numeracy skills among bioscience undergraduates. *Journal of Biological Education, 36*(2), 76-83.

Tariq, V.N. (2008). Defining the problem: mathematical errors and misconceptions exhibited by first-year bioscience undergraduates. *International Journal of Mathematical Education in Science and Technology, 39*(7), 889-904.

Tarnow, K.G., & C. L. Werst. (2000). Drug calculation examinations: do calculators make a difference? *Nurse Educator, 25*(5), 213-215.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53-55.

Taylor, A. (2014). The development, validation and implementation of a test to assess health numeracy in doctors. MSc dissertation. University of Manchester.

Taylor A.A., & Byrne-Davis, L.M. (2016). Clinician numeracy: the development of an assessment measure for doctors. *Numeracy, 9*(1), 5. Retrieved from http://dx.doi.org/10.5038/1936-4660.9.1.5

Taylor A.A., & Byrne-Davis, L.M. (2017). Clinician numeracy: use of the medical interpretation and numeracy test in foundation trainee doctors. *Numeracy, 10*(2), 5. Retrieved from http://doi.org/10.5038/1936-4660.10.2.5

Taylor A.A., Corfield, D.R., & Byrne-Davis, L.M. (2019). Clinician numeracy: use of the medical interpretation and numeracy test in foundation trainee doctors. *Numeracy, in press*

Taylor, A., & McCarroll, M. (1994). Sedation for non-invasive procedures: a potential hazard. *Irish Journal of Medical Science*, *163*,(4), 217. Retrieved from https://doi.org/10.1007/BF02967232

Traub, R.E., & Fisher, C.W. (1977). On the Equivalence of Constructed- Response and Multiple-Choice Tests. *Applied Psychological Measurement*, 1, 355-369

Tully, M. (2012). Prescribing errors in hospital practice. *British Journal of Clinical Pharmacology*, *74*(4), 668-675.

University of Oxford Medical Sciences Division. (n.d.). *Online Assessment: What do difficulty, correlation, discrimination, etc. in the question analysis mean?* Retrieved from https://www.medsci.ox.ac.uk/support-services/teams/learning-technologies/faqs/what-do-difficulty-correlation-discrimination-etc-in-the-question-analysis-mean [Accessed 3 March 2019]

Vaughan, J. (2018). The long road to justice for Hadiza Bawa-Garba. *thebmjopinion*, https://blogs.bmj.com/bmj/2018/08/14/jenny-vaughan-the-long-road-to-justice-for-hadiza-bawa-garba/

Vincent, C., Neale, G., & Woloshynowych, M. (2001). Adverse events in British hospitals: preliminary retrospective record review. *British Medical Journal, 322*(7285), 517-519.

Vincent, C., Burnett, S., & Carthey, J. (2014). Safety measurement and monitoring in healthcare: a framework to guide clinical teams and healthcare organisations in maintaining safety. *BMJ Quality and Safety, 23*(8), 670-677.

Wallace, D. (2019). Parts of the Whole: Theories of Pedagogy and Kolb's Learning Cycle. *Numeracy, 12*,(1), 17. Retrieved from https://doi.org/10.5038/1936-4660.12.1.17

Weeks, K., Lyne, P. & Torrance, C. (2000). Written drug dosage errors made by students: the threat to clinical effectiveness and the need for a new approach. *Clinical Effectiveness in Nursing, 4*(1), 20-29.

Weeks, K. W., Lyne, P., Mosely, L. & Torrance, C. (2001). The strive for clinical effectiveness in medication dosage calculation problem-solving skills: the role of constructivist learning theory in the design of a computer-based 'authentic world' learning environment. *Clinical Effectiveness in Nursing, 5*(1), 18-25.

Weeks, K.W., Clochesy, J.M., Hutton, B.M. & Moseley, L. (2013b). Safety in numbers 3: Authenticity, Building knowledge and skills and Competency development and assessment: The ABC of safe medication dosage calculation problem-solving pedagogy. *Nurse Education in Practice, 13*, e33–42.

Weeks, K.W., Hutton, B.M, Coben, D., Clochesy, J.M. & Pontin, D. (2013c). Safety in numbers 4: The relationship between exposure to authentic and didactic environments and Nursing Students' learning of medication dosage calculation problem solving knowledge and skills. *Nurse Education in Practice, 13*, e43–54.

Weeks, K.W., Hutton, B.M., Young, S., Coben, D., Clochesy, J.M. & Pontin, D. (2013a). Safety in numbers 2: Competency modelling and diagnostic error assessment in medication dosage calculation problem-solving. *Nurse Education in Practice, 13*, e23–32.

Wegwarth, O., Schwartz, L. M., Woloshin, S., Gaissmaier, W. & Gigerenzer, G. (2012). Do physicians understand cancer screening statistics? A national survey of primary care physicians in the United States. *Annals of Internal Medicine, 156*(5), 340-349.

Weiss, B. D., Mays, M. Z., Martz, W., Castro, K. M., DeWalt, D. A., Pignone, M. P., Mockbee, J. & Hale, F. A. (2005). Quick assessment of literacy in primary care: the newest vital sign. *The Annals of Family Medicine, 3*(6), 514-522.

Wheeler, D. W., Remoundos, D. D., Whittlestone, K. D., House, T. P. & Menon, D. K. (2004b). Calculation of Doses of Drugs in Solution. *Drug safety, 27*(10), 729-734.

Wheeler, D. W., Remoundos, D. D., Whittlestone, K. D., Palmer, M. I., Wheeler, S. J., Ringrose, T. R. & Menon, D. K. (2004a). Doctors' confusion over ratios and percentages in drug solutions: the case for standard labelling. *Journal of the Royal Society of Medicine, 97*(8), 380-383.

Wheeler, D., Whittlestone, K., Salvador, R., Wood, D., Johnston, A., Smith, H. & Menon, D. (2006). Influence of improved teaching on medical students' acquisition and retention of drug administration skills. *British Journal of Anaesthesia, 96*(1), 48-52.

Wheeler, D. W., Wheeler, S.J. & Ringrose, T.R. (2007). Factors influencing doctors' ability to calculate drug doses correctly. *International Journal of Clinical Practice, 61*(2):189-194.

Wheeler, D., Degnan, B., Murray, L., Dunling, C., Whittlestone, K., Wood, D., Smith, H. & Gupta, A. (2008). Retention of drug administration skills after intensive teaching. *Anaesthesia, 63*(4), 379-384.

Whittle, S. R., Pell, G. & Murdoch-Eaton, D. G. (2010). Recent changes to students' perceptions of their key skills on entry to higher education. *Journal of Further and Higher Education, 34*(4), 557-570.

Williams, D. J., & Walker, J.D. (2014). A nomogram for calculating the maximum dose of local anaesthetic. *Anaesthesia, 69*(8), 847-853.

Windish, D. M., Huot, S. J. & Green, M. L. (2007). Medicine residents' understanding of the biostatistics and results in the medical literature. *Journal of the American Medical Association, 298*(9), 1010-1022.

Wise, J. (2018). Government takes steps to reduce annual burden of medication errors in England. *British Medical Journal*, *360*, 903.

World Health Organisation. (2009). *Patient Safety Curriculum Guide for Medical Schools*. Retrieved from https://www.who.int/patientsafety/education/curriculum_guide_medical_schools/en/

World Health Organisation. (2017). *WHO Global Patient Safety Challenge: Medication Without Harm.* Retrieved from https://www.who.int/patientsafety/medication-safety/medication-without-harm-brochure/en/

Wright, K. (2004). An investigation to find strategies to improve student nurse maths skills. *British Journal of Nursing, 13*(21), 1280-1284.

Wright, K. (2005). An exploration into the most effective way to teach drug calculation skills to nursing students. *Nurse Education Today, 25*, 430-436.

Wright, K. (2007a). A written assessment is an invalid test of numeracy skills. *British Journal of Nursing, 16*(13), 828-831.

Wright, K. (2007b). Student nurses need more than maths to improve their drug calculating skills. *Nurse Education Today, 27*(4), 278-285.

Wright, K. (2010). Do calculation errors by nurses cause medication errors in clinical practice? A literature review. *Nurse Education Today, 30(*1), 85-97.

Young, S., Weeks, K.W. & Hutton, B.M. (2013). Safety in numbers 1: Essential numerical and scientific principles underpinning medication dose calculation. *Nurse Education in Practice, 13*, e11–22.

Zahara-Such, R. M. (2013). Improving medication calculations of nursing students: An integrative review. *Clinical Simulation in Nursing, 9*(9), e379-e383.

Zhong, W., Feinstein, J.A., Patel, N.S., Dai, D. & Feudtner, C. (2016). Tall Man lettering and potential prescription errors: a time series analysis of 42 children's hospitals in the USA over 9 years. *BMJ Quality & Safety, 25*(4), 233–240.