BritishRedCross

A Literature Review of Potential Outcome Measures for Effective First Aid Education

Lucy Ellis

November 2017

Contents

Executive summary

The aim of this literature review is to identify and explore possibilities for a global measure of effective first aid education. This work was commissioned by the International Federation of Red Cross and Red Crescent Societies (IFRC) and was conducted by the research team in the Education Directorate of the British Red Cross. The search approach was not restricted to any specific discipline, date or format in order to provide a broad scope of possibilities. This enabled the review to cover current practice, theoretical backdrops and the insights that can be gained from other professions.

Part One: Current thinking

For the purpose of this review the IFRC (2016) definition of effective first aid education was chosen. This definition is underpinned by theories of health behaviour including the Theory of Reasoned Action (TRA), Theory of Planned Behaviour (TPB) and the Integrative Model of Behavioural Prediction (IMBP). The definition suggests that effective first aid education increases learners' knowledge, skills, confidence and willingness. This is consistent with current practice in the research literature as these are some of the main constructs emerging as measurements of effective first aid education.

Table 1: Overview of constructs included in the IFRC definition

Construct		Measurement approach	Considerations
Knowledge	Increases in knowledge are most frequently used as evidence of effective education.	Knowledge is predominantly assessed through formal multiple choice tests.	There are some issues with knowledge self-assessments as they are influenced by a wider range of affective factors such as learner motivation and attitudes.
Skills	The focus for skills is applying knowledge. Skills are often associated with practical performance.	Skills are mainly assessed through observations of simulated situations and skill demonstration.	For the purpose of this measurement tool there is a need to consider practical skill measurement feasibility in terms of time, cost and resources.
Confidence / self-efficacy[1]	Self-efficacy can determine how people think and behave.	Self-efficacy is mainly assessed through a self-assessment of confidence on a scale of 0-100.	It is important to make a clear distinction between measuring confidence and competence.
Willingness	This is portrayed as	This is mainly	Although there are clear

[1] Self-efficacy can be defined as an individual's beliefs about how capable they are of carrying out a behaviour (Bandura, 1994).

/ intention	the most important construct in behavioural theory models. It concerns the amount of effort one is willing to exert to achieve an outcome.	measured by directly asking for strength of intention to carry out specific behaviours.	links between willingness and intention there is also a suggestion that they are different concepts.

Part Two : Widening the approach

The IFRC definition used for the purpose of this literature review emerged based on the models outlined for health behaviour. These models are not without criticisms. Within this review the models face three main criticisms: their causal nature due to the structure of the models; their perceived western influences; and the charge that they fail to provide sufficient explanation of variance in behaviour for all.

The purpose of this literature review is to provide suggestions for a global measure and therefore issues of western dominance are particularly salient. This review aims to broaden the discussion of effective first aid education and provide additional suggestions of constructs that could be relevant.

Table 2: Overview of alternative constructs

Construct	Consideration	Measurement approach
Attitudes	The literature suggests that attitudes are key to assessing and predicting behaviour.	Attitudes are most often measured through direct measures. Respondents position themselves on a scale between two bipolar adjectives.
Behavioural intention vs behavioural expectation	There is some suggestion that behavioural expectation may be a better predictor of behaviour.	Behavioural expectations are most often measured directly asking 'I expect' as opposed to 'I intend' for behavioural intentions.
Critical Thinking	The importance of this skill is well established in the field of education.	There are a range of existing standardised formal tests that are used in education settings.
Resilience	There is some evidence that first aid education can increase resilience.	There are a range of existing standardised measures; these most often use self-assessments on a five point scale.
Motivation	The IFRC guidelines refer to the importance of learner motivation for learning.	This is a difficult construct to measure, but evidence shows some crossover with other constructs which may be more measurable.
Helping	Theories of helping behaviour	These can be measured in terms of

behaviour	aim to explain why people help. This is an important element to consider if the overall aim of first aid education is to increase the chances of a learner helping.	helping generally or in specific situations. Direct measures are often in the form of deception experiments. There are some potential suggestions for indirect measures.

This literature review also draws on evidence from other professions

- Firefighters (National Firefighter Questionnaire) measuring confidence and resilience
- Paramedics (training) including non-cognitive skill such as emotional resilience, critical thinking, communication and situational awareness
- National Health Service (Patient Activation Measure) measure of knowledge, skills and confidence to measure own healthcare
- Workplace Assessments
  - Situational Judgement Tests
  - Personality Tests

Recommendations

This review has provided suggestions of constructs that could be relevant for a measure of effective first aid education. Some of these are already being used to evaluate first aid education, but not in a consistent way, such as knowledge, skills and confidence. Some of these are established within education, such as critical thinking and resilience and some provide more innovative suggestions such as indirect measures of helping behaviour.

This literature review demonstrates the need for a global measurement tool for effective first aid education. Currently there is a lack of consistency across research studies. A challenge highlighted by this review is the inherent difficulty in creating a global measure. Any tool developed will need to be generalizable and relevant across cultures.

Emerging from the review are two repeating themes in relation to multiple constructs. Firstly the importance of aggregation, having multiple ways to measure a construct provides more accurate information. Secondly context, evidence shows that more accurate measures are situated within a context. This allows learners to take into consideration all the elements necessary for that situation and provides a more realistic measure.

Limitations

Selected constructs should be informed by further desk research before they are tested. This review is an overview of possible ideas and does not offer the depth of insight into each construct which would be required for further development.

The search for this literature review was conducted in English. This may have limited the review.

Key definitions

Construct: within this review the elements or concepts discussed are referred to as constructs, this is in line with the field of psychology

Likert scale: ordinal scales that measure levels of agreement with a statement. Often a five point scale for example, strongly agree / agree / don't know / disagree / strongly disagree.

Bipolar scale: a scale including two polar opposites, for example 'satisfied' to 'dissatisfied'.

Systematic review: a review of all the available evidence using a clear inclusion criteria and explicit search methods

Standardised assessment: a test taken in a consistent way by all participants which makes it possible to compare performance.

Introduction

The aim of this literature review is to explore existing models for measuring educational effectiveness. This includes considering the outcomes identified by the models as important and the tools that could be used to measure these outcomes.

The first part of this literature review examines the contribution that the field of education effectiveness research can make to a measure for first aid education. For the purpose of this review the definition of effective first aid education has been taken from the International Federation of the Red Cross and Red Crescent guidelines (2016). The IFRC National Societies are some of the leading providers of first aid education worldwide (IFRC, 2016). To the author's knowledge there is no other published agreed definition of effective first aid education. Each outcome identified by the definition is considered.

The second part of this literature review explores other potential avenues that have not necessarily been considered within the IFRC definition of effective first aid education, or within first aid education outcome measures. Each of these concepts is briefly explained including suggestions of approaches to measurements.

Methodology

A search was conducted using a range of online libraries including, Wiley, JSTOR Science Direct and Google Scholar. The search incorporated terms including but not limited to; 'education effectiveness' 'first aid education' 'measuring effective education'. This was later expanded on in part two when considering other measurement outcomes, for example searches included 'measuring critical thinking', 'resilience in education'. The intention of this literature review was to scope potential measures and so there were few restrictions on the inclusion criteria. The search was not restricted to any specific discipline to avoid narrowing any insights gained. Reports and other sources of information were also included. No date restrictions were placed on the search, but attempts to focus on more recent literature where appropriate were made. Although the search was not a systematic review, extrapolated information from previous systematic reviews was included when appropriate. The searches were all conducted in the English language.

1. Part One : Current thinking

1.1 Education Effectiveness Research

The aim of this literature review is to inform the design and development of a global measure of effective first aid education. The first approach taken was to consult the

current literature in the field of education effectiveness research (EER). Education effectiveness research (EER) has rapidly increased recently in both quantity and quality (Reynolds, et al., 2014). There is a significant demand for research within this field due to its potential impact on learner outcomes.

Over time EER has progressed through multiple stages. Attention increased on this area as a result of Coleman's (1966) study, which suggested schools have little effect on learner outcomes. This proposed that education received in the school environment made no difference to learner achievement (Reynolds, et al., 2014). The traditional approach to EER focused on an input / output relationship. For example, whether increasing the amount of funding per student increased the learner outcomes. However evidence showed that the relationship between such variables is not necessarily linear as previously thought (Creemers & Kyriakides, 2006). Today the fields of school and teacher effectiveness have merged to create an overall field of education effectiveness. Education effectiveness is currently seen as multilevel with the different contexts such as schools, classroom and students, having an influence on each other. Looking at education in this way provides a more comprehensive understanding of effective education.

There is a current argument for developing a dynamic model for EER. The dynamic model uses a multilevel approach to education effectiveness, demonstrating that a range of factors impact on effectiveness and that the relationships between these factors may not be linear (Creemers & Kyriakides, 2006). The model highlights the complexity of education effectiveness. The example outlined by Creemers and Kyriakides (2006) is an assessment policy. Such a policy may not just directly impact on when and which assessments are carried out in school, but also the design of assessments and the approach to data monitoring.

The dynamic model (Creemers & Kyriakides, 2006) proposes that factors can be measured through five aspects; frequency, focus, stage, quality and differentiation. Factors include teacher behaviours like questioning techniques and lesson structure. This model provides valuable insights into how education effectiveness can be measured through the five aspects and how this information can be used to improve future teaching and learning. The focus for this model, as well as other models of EER, is on the best practice at the teacher, school, classroom and system level.

EER offers details for how to improve a first aid education intervention for learner achievement, but does not provide much insight into how we can measure learner's outcomes. Student achievement is the main outcome measurement for EER (Reynolds, et al., 2014) and frequently this is through formal assessment data, for example statutory tests or international studies of student learning such as the Programme for International Student Assessment (PISA). Although more formal tests of knowledge might be useful for learner outcomes, and will be discussed in detail later, there is an argument that this would not entirely capture understanding of the learner outcomes in terms of first aid education effectiveness. Therefore EER, although insightful, does not provide a

complete answer to the questions being addressed within this literature review. The field of EER does highlight that focusing solely on learner outcomes may not provide a complete picture of educational effectiveness, other approaches such as measuring trainer behaviours, are also important to include. This is something to consider for the future.

1.2 IFRC definition of effective first aid

The International Federation of the Red Cross (IFRC) define effective first aid education as:

> "interventions that increase the learner's knowledge, skills, confidence and their willingness to apply first aid competencies. These elements have the potential to influence behaviour change." (IFRC, 2016)

Based upon this definition of effective first aid education, the focus is to increase the following four constructs;

- Knowledge
- Skills
- Confidence
- Willingness

The IFRC summarise theories of survival behaviours that contribute to helping a learner act in an emergency (p.29). These appear to have provided a theoretical groundwork for the development of the definition.

These theories predominantly focus on the concept of behaviour change. If the ultimate aim of first aid education is 'change the behaviour of the rescuer to meet the first aid objective' (p.40) then this step seems logical. The first model outlined is Fishbein and Yzer's Integrative Model of Behaviour Prediction (IMBP, figure 2). This model is a development on Azjen's Theory of Reasoned Action (TRA) and Theory of Planned Behaviour (TPB) (figure 1). These models are widely used in relation to health behaviours (Armitage & Conner, 2001). They provide frameworks outlining the key factors that influence behaviours (Montano & Kasprzyk, 2015).

1.3 The Theory of Reasoned Action (TRA) and the Theory of Planned Behaviour (TPB)

The Theory of Reasoned Action (TRA), Theory of Planned Behaviour (TPB) and the Integrative Model of Behaviour Prediction (IMBP) are all social cognitive theories of behaviour change. These models were developed sequentially, each building on the previous. The initial TRA was based on the idea that intention to act is the best predictor of behaviour (Ajzen & Fishbein, 1980). The TRA aimed to describe behaviours that were

under a person's control (Sheppard, Hartwick, & Warshaw, 1988) based on the assumption that behaviour was voluntary. Behaviour is not always voluntary and so the additional construct of perceived behavioural control was added to the model, developing the model into the TPB (Sheppard, Hartwick, & Warshaw, 1988). Perceived behavioural control links to the concept of self-efficacy and refers to a person's belief about their ability to perform a behaviour. An integral aspect of the TPB, like the TRA, is the idea that the stronger the intention of a controlled behaviour, the more likely a behaviour will happen (Ajzen, 1991). The TPB is still one of the most frequently used and influential models for behaviour change (Ajzen, 2011).

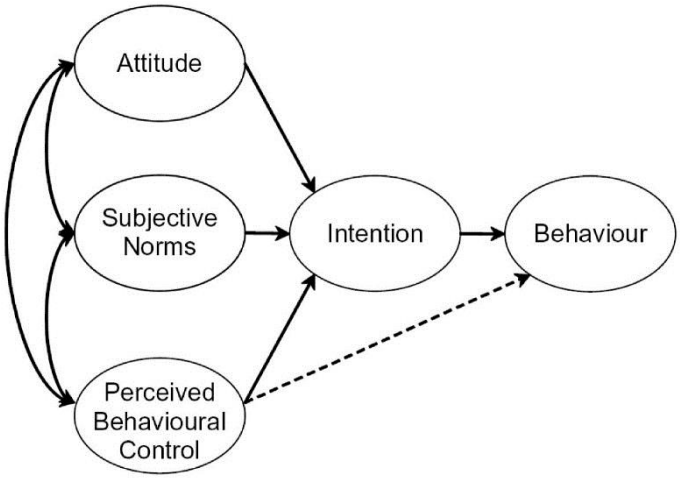

Figure 1: the theory of planned behaviour model

1.4 The Integrative Model of Behaviour Prediction

The IMBP is a development of the TRA and TPB. It extends the previous models and includes the constructs of skills and environmental barriers as influencing intent. It also expands upon the idea of normative beliefs and their role in determining behaviour (Yzer, 2012). The model accounts for both rational and irrational behaviours as it proposes that all behaviour is reasoned from the beliefs that person holds about said behaviour.

The main concept of the IMBP in line with the previous model iterations, is that intention to perform behaviour is the biggest predictor of actual behaviour (Yzer, 2012). This intention is informed by beliefs about the behaviour, the individual's attitude, perceived norms and self-efficacy.

Attitudes within this model can be defined as 'a person's evaluation of how favourable or unfavourable his or her performing a particular behaviour would be' (p. 24) (Yzer, 2012). Perceived norms can be defined as 'the social pressure one expects regarding the behaviour' (p.24) (Yzer, 2012). Perceived norms are seen as incorporating two aspects, injunctive norms and descriptive norms. Injunctive norms include one's social networks' *expectations* about the behaviour and descriptive norms include the extent to which one's social networks *perform* the behaviours (author's emphasis). Self-efficacy can be defined as 'the extent to which a person feels capable of effectively performing the behaviour'

(p.24) (Yzer, 2012). This was previously defined in the TPB as perceived behavioural control.

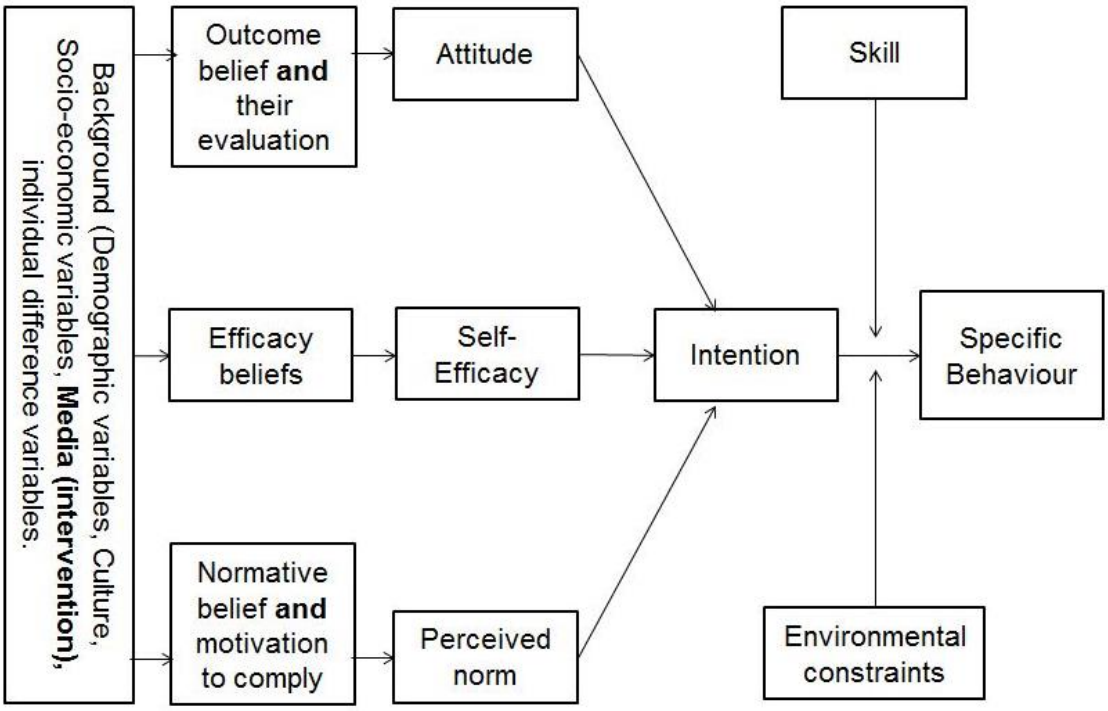

Figure 2: the Integrated Model of Behavioural Prediction

1.5 What has previous research used to measure the four constructs?

The theories of behaviour change propose constructs such as attitude, self-efficacy and perceived behavioural control, that can be operationalised and measured. There is no standard published tool to measure behaviour change given to accompany these models. Icek Ajzen, who proposed the TPB, provides detailed support and guidelines of how to develop questionnaires to measure the main constructs of the Theory of Planned Behaviour (see bibliography). Ajzen advises using self-evaluation with a seven point bipolar adjective scale. To increase reliability, multiple questions for each construct should be used.

The IFRC guidelines (2016) give examples of best practice from the British, German and Swiss Red Cross (see p. 39-40) but no direct instruction on how to measure the outcomes of effective first aid.

The examples from the German and Swiss Red Cross are training evaluations using the Kirkpatrick model. This is a hierarchical model for training to impact on reactions, learning, behaviour and results (Kirkpatrick & Kirkpatrick, 2006). The first two levels operate at the individual level, considering whether learners felt positive or negative about the training and then whether the learner has met the learning objectives which may be

achieving the intended knowledge, skills and attitudes. The second two levels operate at a more institutional level, considering whether they apply the behaviours learned to the workplace and the extent to which the targeted overall outcomes occur. To date, the model seems to have been applied more often to the workplace or organisational training setting. The IFRC guidelines state that currently the German and Swiss Red Cross only use this model at the first level, considering learner reactions to training.

To the author's knowledge, only two systematic reviews have been conducted to explore the effectiveness of first aid education. These reviews have been explored, extrapolating where possible how the studies have measured effective first aid. Alongside these, additional studies are outlined to show the outcome measures used in more recent research studies. Table 5 provided in the Appendix provides an overview of the outcome measures used in all the studies outlined in this section.

Van de Velde et al (2008) conducted a systematic review of the literature to review evidence on the effectiveness of first aid education. They found four studies that met their criteria for inclusion. These studies all explored the effectiveness of a first aid intervention. The approaches to measuring outcomes varied by study. Moore (1997) assessed specific first aid knowledge and skills through a written test. Kelly et al (2003) used a baseline questionnaire and a follow up interview to assess knowledge about first aid specifically of poisoning. Both Shotland and Heinold (1985) and Hawks and Egan (1998) used deception experiments to observe first aid skills and helping responses. Hawks and Egan (1998) also assessed knowledge and skills through a multiple choice question test based on published test banks as well as directly observing skills during the intervention.

A more recent systematic review that incorporated a wider range of studies as a result of a broader inclusion criteria was carried out by He, Wynn and Kendrick (2013). They included 23 studies. Of the 23 studies, 16 included outcome measures on knowledge, 13 included measures of skills, 5 included measures of helping behaviours and 3 included measures of confidence in undertaking first aid. Two of the studies are outlined below as examples, see table 5 for additional details of the measurements used in the other studies included.

Table 3: Overview of two studies included in the systematic review by He, Wynn and Kendrick (2013)	
Bollig (2009)	Conducted a first aid teaching programme for six to seven year old children in Norway. They explored whether participation in the programme affected performance in a first aid scenario. The experimental group (n=117) received a series of five lessons in first aid. Children in the experimental group and control group (n=111) were tested through observations of performance in a practical scenario. Children were measured on their performance of six predefined tasks, including assessment of the situation, knowledge of steps to take and placing the victim in the correct recovery position.

Lippmann (2010)	Compared a traditional two day classroom first aid intervention to a combination approach of a classroom and online learning first aid intervention. They assessed first aid knowledge through a 24 item multiple choice test. Practical skills were assessed through performance of simulated scenarios on a predefined criteria grading performance in each situation, from not competent to very good. Assessors also used objective feedback from the manikin's display. These assessments were carried out immediately after training and then three months later.

The inclusion criteria for this systematic review included whether the studies reported outcomes relating to first aid knowledge, skills, behaviours and confidence. It could be argued that this limits the range of studies from the outset by restricting the possible outcome measures. There is no comment about whether any studies were excluded from the review based on this criterion. Both systematic reviews also focused on the layperson and excluded any studies that used individuals who had received formal healthcare education. More recent studies have been carried out since the systematic reviews evaluating the effectiveness of first aid education and are outlined below. All of the studies in this section have also been included in the overview in table 5 (Appendix 1).

Table 4: Overview of how more recent studies have measured first aid education effectiveness	
Eze et al (2015)	Evaluated a sample of 226 trainee teachers' knowledge, attitude and first aid skills pre and post an epilepsy intervention. They used self-administered questionnaires with closed questions. The knowledge section included 28 questions about teachers' understanding of epilepsy and where they had gained it, for example if they had witnessed a seizure. Attitudes were measured in a separate section, with nine questions exploring how positive the attitudes of teachers were towards epilepsy. The final section of the questionnaire explored the teachers' first aid management of epilepsy, assessing whether teachers knew the correct actions to take in a given scenario. Responses were scored 1 or 0, with 1 being the correct response. A higher total score indicated more positive knowledge, attitude and first aid management.
Schumann et al (2012)	Assessed the retention of wilderness first aid knowledge, self-efficacy beliefs and skills in a sample of volunteer course participants. First aid knowledge was measured using a multiple choice test consisting of 25 questions. Self-efficacy was assessed using an adapted self-efficacy scale where learners self-assessed their perceived ability to perform specific wilderness first aid skills on an 11 point scale ranging from, 'I cannot do it at all' to 'I am highly certain I can do it'. Learners skills were assessed through skill based scenario simulations, where they were observed by multiple examiners using an objective checklist. This study focused on assessing retention of learning and so learners were assessed at 4 months, 8 months or 12 months after training.

Nhu Lam et al (2017)	Conducted a survey among high risk workplace groups to evaluate first aid knowledge specifically of burns. The questionnaire consisted of 12 multiple choice questions about first aid management for burns. The questionnaire also asked what to do in specific fire situations and about where they thought they could find more information about burns if needed.
Mahony, Griffiths, Larsen and Powell (2008)	Evaluated airline cabin crew's knowledge and retention of CPR and AED skills, first aid knowledge and perceived levels of confidence. All participants had completed a two day training course 12 months previously. Mahony et al used a simulated test scenario asking participants to manage the situation and their actions were assessed using a checklist by an observer. Participants also completed a self-assessment of performance and knowledge based on three first aid situations. These were open questions for which they were asked to describe in detail their first aid management and actions. Participants self-confidence of first aid knowledge and skills was assessed through a self-assessment using a five point Likert scale.
Everett-Thomas et al (2016)	Assessed first responders cardiopulmonary resuscitation (CPR) skills using 40 simulated sudden cardiac arrest scenarios after a two hour training lecture. Responder's performance was evaluated using a 19 item assessment observation form. Responders were scored on their time to initiate response, the skills required to perform CPR and performance using the AED.

These study outlines show how effective first aid education can be measured in a variety of ways. It is important to note that not all of the studies were evaluating the effectiveness of first aid education, some were just assessing general first aid knowledge, but this does provide important insights into how first aid knowledge is measured. The majority of the studies outlined focused on specific first aid skills, for example burns or CPR. Mahony et al (2008) explored first aid knowledge generally but this was within a specific unique context of the airplane. Participants in this study were cabin crew workers and an expectation to be able to perform some first aid is part of their role.

The IFRC guidelines state that most measurement currently comes in the form of knowledge acquisition and limited skill demonstration (p.31). These studies support this statement, with knowledge and skills being assessed the most frequently, see table 5. In these examples knowledge is most often measured through formal assessments such as multiple choice tests. Skills are most frequently measured through observations. Some of the studies also measured attitudes such as self-efficacy and confidence, highlighting the importance of these skills within first aid. The definition of effective first aid education includes improving knowledge, skills, confidence and willingness, but none of the outlined studies directly observed willingness. There is an argument that the studies of helping behaviours did encompass an element of participant's willingness to act, but only three of the studies outlined measured helping behaviours. Measures of helping behaviours will be discussed more in section 2.7. Most importantly these studies demonstrate an

inconsistency in the approaches taken to measuring outcomes and the need for an objective universal measure to evaluate first aid education effectiveness.

1.6 The value of knowledge

Knowledge is defined as 'facts, information and skills acquired through experience or education; the theoretical or practical understanding of a subject' (Oxford Dictionary). In the studies outlined above knowledge was most frequently measured through formal multiple choice assessments.

Standardised assessments could enable comparisons across contexts. Formal tests are widely used to measure learner outcomes. Global assessments such as the Organisation for Economic Co-operation and Development (OECD) PISA tests provide international comparisons between countries performance through a formal assessment using both multiple choice and extended questions. It is recognised as an established global comparison of education systems.

Most of the studies outlined used multiple choice tests for knowledge. Multiple choice tests are widely used in training and assessment in the medical profession (Epstein, 2007; Considine, Botti, & Thomas, 2005). They can be efficient and objective (Brady, 2005) and enable assessment of a range of content areas (Epstein, 2007). If multiple choice questions are to be used, it is important to ensure that they are reliable and valid (Considine, Botti, & Thomas, 2005). There are many factors to consider when designing multiple choice questions including but not limited to, readability, response options, inclusive alternatives, difficulty, discrimination, formatting, stability, internal consistency and construct validity (Considine, Botti, & Thomas, 2005). In a comparison of clinical scenario multiple choice questions (CSMCQ) and multiple choice questions (MCQ), Vuma and Sa (2017) found that CSMCQ were more effective. Therefore it might be that multiple choice questions embedded within a scenario provide more effective insights into learner knowledge.

For a more practical skill such as first aid, Miller (1990) argues that knowledge based tests fail to demonstrate what the learner will do in future situations. Willet et al (2009) compared standard measures in training with performance measures and found that standard evaluation measure scores did not correlate to actual clinical performance. This suggests knowing how to complete an action is not always an effective predictor of carrying out the action.

Self-assessments are widely used in research as they allow learners to estimate how much they know or have learned in a relatively quick and efficient way (Sitzmann, Ely, Brown, & Bauer, 2010). In a meta-analysis, Sitzmann et al (2010) found that nearly a third of studies used self-assessed knowledge as evidence of learning. There is some evidence suggesting that self-assessments of knowledge are not as accurate as other forms of assessment (Witt & Wheeles, 2001; Kruger & Dunning, 1999) as people often over-estimate their abilities (Kruger & Dunning, 1999). Evidence from the medical profession suggests that professionals are unable to successfully, accurately self-assess

knowledge (Tracey, Arroll, Barham, & Richmond, 1997; Jansen, Tan, Van der Vleuten, Luijk, Rethans, & Grol, 1995; Speechley, Weston, & Dickie, 1994). A systematic review comparing physician's self-assessment to observations of competence concluded that physicians are often poor at self-assessing knowledge (Davis, Mazmanian, Fordis, Van Harrison, Thorpe, & Perrier, 2006).

Self-assessments can be influenced by additional external factors such as learner motivation and attitudes which can introduce bias (Sitzmann, Ely, Brown, & Bauer, 2010; Stewart, O'Halloran, Barton, Singleton, Harrigan, & Spencer, 2000). Self-assessments of knowledge were found to be more strongly correlated with motivation and satisfaction rather than cognitive learning (Sitzmann, Ely, Brown, & Bauer, 2010). This suggests that if self-assessment is used to assess purely cognitive knowledge acquisition then they might not be the most appropriate method. Multiple choice test of knowledge require participants to select an answer, demonstrating actual knowledge rather than perceived knowledge and so are not as influenced by bias.

Affective measures could be as insightful as cognitive knowledge when evaluating first aid education effectiveness. Self-assessment may then be considered a more reliable measure. A meta-analysis examining the TPB found additional support for self-assessments in this context. When the behaviour measures used were self-reports, 11 per cent more variance of behaviour was explained compared to when the measures were objective or observed (Armitage & Conner, 2001). Interpreting the findings from self-assessment in relation to knowledge may not be simple, but they can provide important insights (Speechley, Weston, & Dickie, 1994) useful for first aid education.

1.7 The value of skills

In the majority of the studies outlined in section 1.5, skills were assessed through practical observed simulated situations and only in a few studies were skills assessed through a written test, sometimes in combination with observations. All studies either compared skills to a control group, an experimental group receiving a different mode of delivery, a pre and post training assessment or a combination of these approaches. Skills are defined as 'a particular ability' (Oxford Dictionary). Having the skills to perform a first aid action is a logical desired learner outcome of effective first aid education.

There is some overlap between knowledge and skills and for many they are similar concepts. The PISA guidance uses the phrase 'knowledge and skills' without differentiating between the two. Skills can be defined as *applying* knowledge to situations (Boulet) (author's emphasis)

There are some examples of written tests of skills. PISA uses written tests to assess skills such as problem solving and communication. The World Bank Skills Towards Employability and Productivity (STEP) programme measures cognitive, socio-emotional and job relevant skills through written surveys in developing countries.

Simulations can provide the link between theoretical knowledge and clinical skills and practice (Durham & Alden, 2008). Simulations provide a representation of reality but in an educational context. They are widely used to assess clinical competence and can include a variety of approaches, from paper and computer based to practical performances (Munshi, Lababidi, & Alyousef, 2015). It is important to provide opportunities to develop skills through practice (Boulet) and simulations provide the opportunity for this, both in training and assessment. Simulations are widely used for teaching healthcare professionals clinical skills (Crea, 2011; Munshi, Lababidi, & Alyousef, 2015; Durham & Alden, 2008; Nehring, Ellis, & Lashley, 2001). For example SimMan® is a realistic simulation of a patient that is often used as an effective tool to assess skills (Everett-Thomas, et al., 2016).

If skills are defined as applying knowledge then a valid measure must provide an opportunity for this. The IFRC state that skill demonstration is the most common current practice to assess skills. Medical students are assessed through an objective structured clinical examination (OSCE). Students are observed by an examiner using a checklist to evaluate performance on different scenarios. The OSCE is an established method of clinical skills assessment (Mitchell, Henderson, Groves, Dalton, & Nulty, 2009; Ward & Barratt, 2005) and is seen as an objective and reliable measuring tool (Major, 2005).

The IFRC guidelines (2016) reinforce the importance of effective measurements in terms of time, costs and resources (p.32). A measurement approach similar to the OSCE may not be feasible in terms of these three factors. Assessing each learner on a course in a simulated scenario environment using observations with an objective checklist could be time consuming and expensive. Equally it would raise considerations such as observer bias which adds additional costs and resources.

1.8 The value of confidence and self-efficacy

In the studies outlined in section 1.5, confidence was most often assessed through the construct of self-efficacy. Self-efficacy can be defined as an individual's beliefs about how capable they are of carrying out a behaviour (Bandura, 1994). It is important to emphasise that this is perceived ability, not necessarily actual (Yzer, 2012). Self-efficacy can determine how people think and behave. Individuals with a strong sense of self-efficacy are confident about their ability to control and act in a situation. Those with low self-efficacy are more likely to have low aspirations and focus on the barriers to performing behaviours. Evidence suggests that self-efficacy is a predictor of positive behaviour change (Burke, Dunbar-Jacob, Sereika, & Ewart, 2003). Within the Integrative Model of Behavioural Prediction (Yzer, 2012), self-efficacy is proposed as one of the three determinants of intention and so behaviour.

There is limited research evidence regarding the link between self-efficacy and responding to a general first aid situation. An area which has been researched related to first aid is the link between self-efficacy and performing CPR. A lack of confidence has been identified as a common barrier to an individual performing CPR (Dwyer, 2008). Dwyer (2008) conducted surveys gathering information about experiences and attitudes

towards CPR. Thirty-one per cent of respondents said that they were not confident to perform CPR. Those who rated themselves as least confident reported concerns over performing CPR incorrectly or fears of failing. This suggests a need for a level of confidence or self-efficacy alongside the basic cognitive knowledge and skills in order to perform CPR if needed (Hernandez-Padilla, Suthers, Fernandez-Sola, & Granero-Molina, 2016).

There are a range of existing measures of self-efficacy. These are usually self-assessments where individuals rate their confidence on a scale. Bandura (2006) recommends a 0 to 100 scale with 10 unit intervals as it gives a higher level of sensitivity. The emergency response confidence tool by Arnold et al (2009) contains 17 items that have been designed to evaluate participant's confidence in reacting to an emergency situation. Participants self-rate their confidence on a scale of 0 to 100 per cent. Another measure is the General Self-Efficacy scale (Schwarzer & Jerusalem, 1995). This asks individuals to rate their response to 12 statements on a scale from one to four from 'not true at all' to 'exactly true'. The statements are more general for example, 'I can always manage to solve difficult problems if I try hard enough'.

Bandura (2006) argues for the importance of measuring self-efficacy in a way that is both multifaceted and tailored to the specific behaviour. This helps to reduce any ambiguity of the situation and lets the respondent take into account the situational demands. The studies outlined in table 5 that measured self-efficacy as an outcome, used unique self-efficacy scales developed for each individual study. This enables the scales to provide more accurate insights about respondent's actual self-efficacy beliefs for the first aid context. Situating the statements enables the self-assessment to take into account the specific demands of that situation and provide a more valid and reliable measure.

Confidence vs competence

It is important to define the difference between confidence and competence as although they are linked and both important, they are not the same. Confidence can be defined as the self-assurance of ones abilities. Competence can be defined as the ability to do something successfully. Within some research studies the terms are used synonymously. For example Speechley et al (1994) used self-assessed competence as a measure of confidence. However they are different concepts and not intrinsically linked. Research evidence has found that a positive expression of confidence does not always correlate to an expression of competence (Stewart, O'Halloran, Barton, Singleton, Harrigan, & Spencer, 2000). In a study using confidence surveys and competency assessments, Clanton et al (2014) found that medical students confidence was only associated with competence post intervention.

First aid education is seeking to equip people to help someone when they are in a real first aid situation. Learners need both the confidence and competence to act. Competence can be measured through practical skill demonstration, but it is not clear whether this also sufficiently encompasses a measure of confidence.

1.9 The value of willingness and intention

Intention can be defined as the amount of effort one is willing to exert to achieve an outcome (Sheeran, 2002). Theories of behavioural change argue that behavioural intention is the most important construct to determine whether someone will act (Gibbons & Gerrard, 1997). Behavioural intention demonstrates how motivated an individual is to perform a behaviour and so how hard they will try to achieve it (Ajzen, 1991).

Recent research evidence supports the argument for intention predicting first aid behaviour. Yap and Jorm (2012) conducted a national survey and follow up interviews of young people in Australia. Their methodological approach incorporated the use of vignettes. Vignettes are widely used in social research (Hughes & Huby, 2002). They involve asking participants to respond to a created stimulus such as a text or image (Schoenberg & Ravdal, 2000; Hughes & Huby, 2002). Vignettes are an important tool for gaining insights into participants attitudes and beliefs by depersonalising a situation (Schoenberg & Ravdal, 2000; Hughes & Huby, 2002). Within this study vignettes were used to establish what participants would do in a mental health situation. Participants were asked to rate whether first aid interventions were helpful, harmful or neither. Yap and Jorm found that young people's first aid intentions and beliefs in the initial survey predicted later actions. Similarly Rossetto et al (2016) found that mental health first aid intentions predicted first aid behaviour six months later. They also assessed intentions and behaviours using hypothetical questions using a vignette. Interestingly Rossetto et al (2016) found that past behaviours and experiences were also a strong indicator for future behaviours. These two studies do provide evidence to support the link between intentions and behaviours, but it is important to note that they were both carried out in the field of mental health first aid, not physical first aid. There is a lack of research specifically exploring behavioural intentions in relation to later physical first aid actions.

There are existing behavioural intention measures available (Gibbons & Gerrard, 1997). Ajzen (1991) suggests that intention is measured through direct questions for example, 'I intend to…' where individuals are then asked to rate their agreement on a Likert scale to indicate the strength of their intention for each behaviour. In order to increase reliability and validity of behavioural intention measures it is important to use aggregation and include a number of measures to target the same behaviour (Ajzen & Fishbein, 2005). There are often multiple intentions that result in a behaviour. As important is the idea of compatibility, the question wording needs to be carefully constructed to include a specific context and time (Ajzen & Fishbein, 2005).

Meta-analyses exploring behavioural intention suggest that intent accounts for 28 per cent of the variance of behaviour (Sheeran, 2002). This suggests that although intention is important, it does not explain fully why an individual acts or does not act.

For first aid education specifically, White and McNulty (2011) found that willingness as a measure was not an outcome of first aid training. None of the studies outlined in table 5 measured willingness or intention directly. Some groups are more likely to be willing to act, research has shown that factors such as knowing the person affects whether a

person is likely to help them (Swor, Khan, Domeier, Honeycutt, Chu, & Compton, 2006). If willingness is affected by the specific situation context in which a person is acting then it questions whether it is the most appropriate construct to measure to assess the effectiveness of first aid education. The models of behaviour change outlined in section 1.3 and 1.4 argue that intention is the most important influence on behaviour, and so this questions the suitability of such models for first aid education.

Concerns about the assumption linking willingness to intention

Although there are clear links between the concepts of willingness and intention, there is a suggestion that they are different concepts. The prototype / willingness model (Gibbons & Gerrard, 1997) is a dual process theory that is an extension to the previous models of behavioural change (figure 3). It is based on the idea that behaviour is not always intentional (Kalebic-Maglica & Martinac-Dorcic, 2016; Todd, Kothe, Mullan, & Monds, 2014) and so there is a difference between measuring intention and willingness. Prototypes are social images about the type of person that one believes is likely to engage in behaviour. Willingness can be measured by asking participants how they would respond in imaginary situations, and rate their willingness on a scale from not at all likely to extremely likely (Gibbons, Gerrard, Blanton, & Russell, 1998).

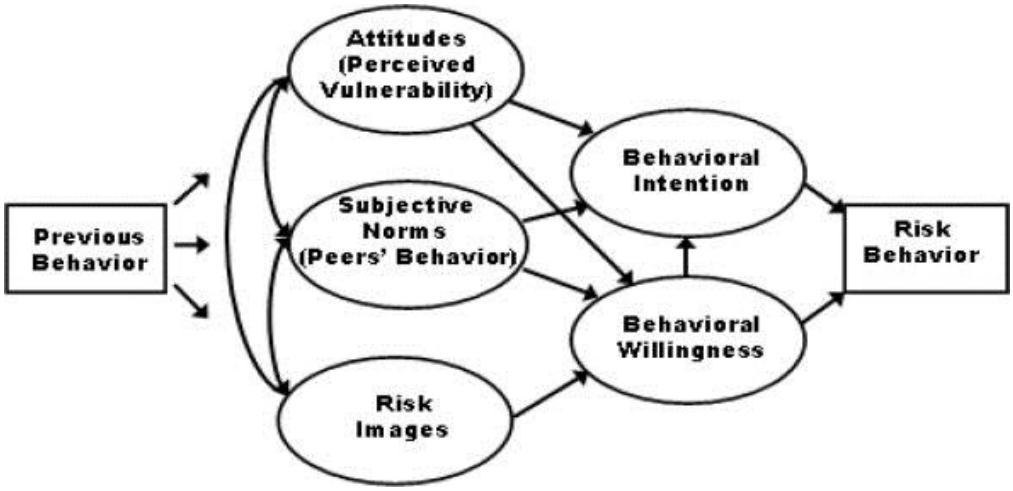

Figure 3: The prototype / willingness model (Gibbons & Gerrard, 1997)

In a meta-analysis of 81 studies using the prototype / willingness model, willingness explained five per cent more variance than intention (Todd, Kothe, Mullan, & Monds, 2014). This provides evidence to suggest that the addition of a prototype / willingness branch to the behavioural models helps to explain behaviour better.

This is not necessarily a consistent finding. Ajzen and Fishbein (2010) measured intention and willingness and found that adding a measure of willingness did not improve the prediction of behaviours. Equally, Matterne, Diepgen and Weisshaar (2011) compared the TPB with the prototype / willingness model and found that intention was a better predictor of behaviour.

The prototype / willingness model was initially created to explain adolescents and young adult's behaviours in social, risky situations. The model assumes that there is a clear social image associated with health risk behaviours and that this explains in part why young people are more likely to engage in said behaviours even if they did not intend to. Although a first aid emergency might be a risk situation, it is different from the types of risk behaviours that have been measured using the prototype / willingness model. It is not clear whether this is an appropriate model to use for the first aid situation. However it is important to consider the evidence that the link between intention and willingness can not necessarily be assumed.

2. Part Two : Widening the approach

In order to explore how to measure effective first aid education it is important to first agree a definition of effective first aid. This review used the definition published in the IFRC guidelines (2016). This definition comes with some potential limitations.

The development of the definition was heavily influenced by the behavioural models outlined in part one. The TRA / TPB and IMBP models have been the leading theoretical approaches to guide research around health related behaviours (Sniehotta, Presseau, & Araujo-Soares, 2014). They have been used to develop effective behaviour interventions predicting a variety of health behaviours including for example smoking and substance use. A systematic review of studies using the TPB found that it explains for 19.3 per cent of variability in behaviour, with intention being the strongest predictor (McEachan, Conner, Taylor, & Lawton, 2011). Evidence suggests that changing the key constructs outlined in the models can lead to changes in behaviour (Montano & Kasprzyk, 2015). Literature tends to recommend the use of the IMBP above the two previous models as it encompasses the detail of these alongside other behavioural theories (Montano & Kasprzyk, 2015). There is a lot of existing support on how to develop effective health behavioural change interventions based on the model (see for example, Yzer, 2012; Montano & Kaspryzyk 2015), but less so on the most effective approach for evaluation purposes.

These models of behaviour change are not without criticism. It has been questioned whether the level of explanation of variance in behaviour is enough (Sniehotta, Presseau, & Araujo-Soares, 2014). Many of the studies providing evidence for the models are correlational (Weinstein, 2007). In a systematic review of 24 experimental studies evaluating the TPB, Hardeman et al (2002) found no robust evidence in support of the models. The systematic review carried out by McEachen et al (2011) found that the model was most effective at predicting behaviours for young, healthy, affluent people. The evidence suggests that the model is less predictive when considering longitudinal behaviour or measuring outcomes with objective methods as opposed to self-assessments (Sniehotta, Presseau, & Araujo-Soares, 2014).

The models are all causal in nature, assuming that if the constructs exist then behaviour will occur. They assume that a limited number of variables can determine any given behaviour (Fishbein, 2000). There is an argument that human behaviour is more complicated than this. The models are constructed using arrows indicating direction of effect, however it is known that this is not always the case. There is often interplay between the variables that works both ways, for example, Rossetto et al (2016) found that past behaviours were the most significant indicator of future behaviours. Equally within the model, environmental constraints are only said to influence intention at the last stage, without being linked to previous variables, but the argument could be made that environmental constraints are influenced by background factors such as demographics and culture. Cognitive models are criticised for not capturing the wider structural, economic, environmental and social factors that affect behaviour change (Morris, Maizano, Dandy, & O'Brien, 2012). Therefore the structure could be suggested to limit the model.

The models have been criticised as being Western constructs and not applicable across cultures (Airhihenbuwa & Obregon, 2000). The models are based on the idea that an individual is an actor and empowered to make a decision about their actions. However this is not necessarily applicable globally. Self-efficacy is often argued to be a Western construct. As a construct, self-efficacy developed in Western cultures and most of the research evaluating it has also been conducted in a Western context (Klassen, 2004). A review of studies showed that self-efficacy beliefs tended to be lower in non-Western cultures suggesting that there might be cultural differences in how self-efficacy is understood and measured (Klassen, 2004). In response to such criticisms, Fishbein and Ajzen highlight the importance of carrying out elicitation interviews in order to make the model context specific (Fishbein & Ajzen, 2010). Considering the beliefs that underlie the specific behaviour in the target population enables the model to be applicable across cultures (Middlestadt, 2012).

Most of the studies included in the systematic review of effective first aid come from high income countries (He, Wynn, & Kendrick, 2013). He et al (2013) suggest that these findings therefore would be unlikely to be generalised to other countries. It is important that any global measure is applicable to the context it is designed for.

The purpose of this literature review is to lead into the development of a global measure of effective first aid education and so it is an important to consider how culturally appropriate the definition is for a global audience. To widen the thinking in this area the second part of this review will consider other constructs that may be deemed important learner outcomes from effective first aid education. This is based on the consideration that the overarching aim of effective first aid education is that the individual steps out the classroom more able and likely to help than before. This is a broader definition of effective first aid education than the IFRC definition. Different constructs will briefly be outlined alongside any existing measurement tools. It is important to note that the current literature review relied solely on English language searches, leading to a lack of

accessibility to research not published in English. This does create some potential limitations.

2.1 Attitudes

The TRA was initially developed to gain a better understanding of the relationship between attitudes and behaviours (Fishbein, 1967). Attitudes are included as one of the three vital constructs in the IMBP influencing intent and behaviour. Many theoretical frameworks including the TRA and TPB include the relationship between attitudes and behaviour. Attitudes are not included directly in the IFRC definition of effective first aid education.

In table 5, four studies measured attitudes towards first aid. This was done in a variety of ways. Engeland et al (2002) measured attitudes towards specific first aid and intended behaviours using a 1 to 100 semantic differential scale. Frederick (2000) used a 'draw and write' (McWhirter & Wetton, 1994) method to measure attitudes and hypothetical behaviour. For this participants watched a video and then developed the story further through either writing or drawing. Responses were evaluated by an independent observer. Eze et al (2015) measured attitudes specifically towards epilepsy using 9 questions, positive attitudes were scored 1 and negative attitudes scored 0. It is not clear whether this was through a closed or open question format. Azeredo and Stephens-Stidham (2003) similarly do not give full details about the measures used, but did use a mixture of true / false and multiple choice and simple written answers.

The IMBP separates attitudes into experiential and instrumental attitudes. Experiential attitudes involve individuals emotional ideas about carrying out a behaviour, instrumental attitude is based on beliefs about the outcomes of carrying out a behaviour (Montano & Kasprzyk, 2015). These two types of attitude are defined by many other theorists as affective and cognitive attitudes respectively (Conner, Gaston, Sheeran, & Germain, 2013).

Measuring attitudes are key to assessing and predicting behaviour (Kroenung & Eckhardt, 2015). There are two main approaches to attitude measurement, direct and indirect measurement. Direct measurement includes tools such as questionnaires with Likert scales and semantic differential scales (Montano & Kasprzyk, 2015). Indirect measurement includes projective techniques (McLeod, 2009) and asking how beliefs are associated with feelings and outcomes (Montano & Kasprzyk, 2015).

Direct measures have been found to be strongly associated with intentions (Montano & Kasprzyk, 2015). Semantic differential scales are a rating scale that is used to measure attitudes towards aspects such as a behaviour. They are widely used in the history of attitudinal research (Montano & Kasprzyk, 2015) and across a wide range of disciplines (Lavrakas, 2008). Chraska and Chraskova (2017) compared semantic differential scales across different cultures and found inaccurate results when comparing attitudes. This raises a concern about the cross cultural applicability of such measures. Ajzen and

Sexton (1999) acknowledge this concern, stating that different contexts may produce different attitudes. Concepts could be interpreted differently and so it is important that the measure used is tested for reliability and validity in the specific context with the target population. In the TPB measurement tool guidance (see bibliography), Ajzen (1991) suggests that it is important to use specific situations when assessing attitudes. General attitudes surveys are not be able to predict specific behaviour (Ajzen, 1991).

Direct measures can be influenced by social desirability (McLeod, 2009). Indirect measures aim to overcome this disadvantage. Projective tests ask participants to interpret an ambiguous stimulus. Such tests include the Thematic Apperception Task, Rorschach Inkblot test and the Draw a Person Task. The third task is generally used for younger participants. Indirect measures are criticised for lacking objectivity (McLeod, 2009).

2.2 Behavioural intentions vs behavioural expectations

Intentions and their influence over behaviour were discussed in section 1.9. There is a suggestion that in some literature behavioural intention has been confounded with behavioural expectation (Warshaw & Davis, 1984) even though there is a difference between the two constructs. Behavioural intentions can be defined as how hard people are willing to perform a particular behaviour (Ajzen, 1991) whereas behavioural expectations are the individuals' prediction of their future performance and the likelihood that they will perform the behaviour (Warshaw & Davis, 1984). The gap outlined above between intent and behaviour has been labelled as the intention-behaviour gap (Moghavvemi, Salleh, Sulaiman, & Abessi, 2015). This gap suggests that although intentions are important, they are insufficient in determining whether an action is performed (Bhattacherjee & Sanford, 2009).

Some evidence suggests that behavioural expectation is a better predictor of behaviour than behavioural intention. Warshaw and Davis (1984) compared two versions of the same questionnaire, a behavioural intent version and a behavioural expectation version. For example, a question assessing behavioural intent might ask to 'indicate whether you intend to perform the given behaviour sometime next weekend'. A question assessing behavioural expectation might ask 'how likely is it that you will actually perform the given behaviour sometime next weekend' (Warshaw & Davis, 1984). They found that behavioural expectation was a better predictor of 18 common behaviours. Warshaw and Davis (1984) argue that behavioural expectation considers more behavioural determinants and so is a broader more inclusive construct leading to it being a better predictor of behaviour. A more recent study has also found evidence that behavioural expectations are more effective predictors of behaviour than behavioural intentions. Rothschild and Wolfers (2011) compared behavioural intentions and expectations on poll surveys and found that in 78 per cent of cases expectations were more accurate predictors. In this study, participant's intention was measured by asking 'who do you think you will vote for in the election'. Participant's expectation was measured by asking 'who do you think will be elected' (p.4) (Rothschild & Wolfers, 2011).

The addition of self-efficacy in the IMBP was assumed to resolve the intention-behaviour gap (Armitage, Norman, Alganem, & Conner, 2015). Recent evidence shows that behavioural expectation still may be a better predictor of behaviour than behavioural intention, even in the instances when self- efficacy is controlled for. Armitage et al (2015) assessed participant's behavioural intentions, expectations and self-efficacy in questionnaires on alcohol consumption and weight loss. They found that behavioural expectations were still more predictive of behaviour than behavioural intentions. However, this particular study is not without limitations and its applicability to first aid situations stands to be evidenced and so further research is needed.

2.3 Critical thinking

Critical thinking is an important skill for decision making (Papathanasiou, Kleisiaris, Fradelos, Kakou, & Kourkouta, 2014) especially in a clinical setting (Jenkins, 1985; Malek, 1986; Miller M. , 1992; Papathanasiou, Kleisiaris, Fradelos, Kakou, & Kourkouta, 2014; Macpherson & Owen, 2010). Whilst there is no universally agreed definition of critical thinking, it is can be thought of as the process of deciding what to believe and do (Ennis & Millman, 1985), including the skills and strategies used for making decisions (Rudinow & Barry, 1999).

There are many existing assessments of critical thinking including standardised, formalised tests such as those listed below. Most of the tests use a selected response format such as multiple choice or Likert scale items (Liu, Frankel, & Roohr, 2014).

- The Watson Glaser Critical Thinking (CT) Appraisal was initially developed in the 1920's. It is one of the most common critical thinking tests. It has continued to be refined with the most recent version published in 2011. It assesses the individual's ability to infer, assume, deduct, come to conclusions and interpret and evaluate arguments.
- The Cornell Critical Thinking Test (CT) evaluates student's skills in induction, deduction, credibility, identification of assumptions, semantics, definition and prediction in planning.
- California Critical Thinking Disposition Inventory focuses on assessing open mindedness, analytic skills, cognitive skills, maturity, truth seeking, systematicity, inquisitiveness and self-confidence.
- California Critical Thinking (CT) Skills Tests measures drawing conclusions in terms of analysis, inference, evaluation, deductive reasoning and inductive reasoning. It was initially designed for college students.
- Health Services Reasoning Test (HSRT) was designed specifically for health science students and practitioners. The items are based in clinical contexts. Participants are expected to infer, interpret, analyse, identify claims and reasons and evaluate the quality of arguments.

In a comprehensive review of current critical thinking assessments, Liu, Frankel and Roohr (2014) found some issues. They argue that many of the existing assessments are limited and do not capture the higher level of sophisticated critical thinking skills. There is

an argument that assessments of critical thinking should use open ended task to enable learners to move beyond fact recall to applying knowledge to realistic problem contexts (Lai, 2011). Liu, Frankel and Roohr (2014) recommend using a mixture of both open and closed questions to provide an authentic and sound assessment.

A large proportion of research into critical thinking has been conducted in the educational context, with improvements in critical thinking seen as key for quality education (UNESCO, 2000; OECD, 2012). Carter et al (2015) carried out a systematic review of critical thinking tools in a nursing and midwifery context. They reviewed 34 studies, 21 of which used one of four commercially available standardised critical thinking assessments, the California CT disposition inventory, the California CT skills test, the Watson-Glaser CT appraisal and the Health Services Reasoning Test. The other studies measured critical thinking through non standardised tools. Carter et al (2015) concluded that for the nursing and midwife context the results were inconsistent and that more work is needed to be done to establish a successful measure of critical thinking in this context. This suggests that although the idea of critical thinking in education is established, more work needs to be done to ensure it is applicable and appropriate to contexts outside of education specifically.

The majority of critical thinking assessments treat the concept of critical thinking as a general skill (Liu, Frankel, & Roohr, 2014). This may be due to the cost involved in developing a standardised measurement tool, as there may be more of a demand for a generic measure than for each individual subject (Liu, Frankel, & Roohr, 2014). There is an argument that the specific context and subject knowledge skills are important for critical thinking (Daly, 2001; Merrell, 2003). Situating the assessment within a context enables other factors to be considered, such as cultural considerations, and could provide a more accurate measure of critical thinking (Gelerstein, del Rio, Nussbaum, Chiuminato, & Lopez, 2016). Current research evidence does not confirm whether critical thinking is a general or specific skill (Nicholas & Labig, 2013).

Many of the studies included in Carter et al's systematic review were conducted in Western contexts. There is an argument that the concept of critical thinking is a Western construct (Daly, 2001). This needs to be considered and further literature and evidence consulted if critical thinking is to be included in a global assessment.

2.4 Resilience

White and McNulty (2011) found that first aid training appeared to be positively correlated to an individual's resilience. Resilience can be defined as the process of adapting to cope in potentially traumatic and stressful events (American Psychological Association, 2017). A resilient individual may be more likely to cope and act in a first aid crisis and so there could be many benefits to increasing and measuring resilience through first aid education.

There are a range of existing measures of resilience. Five are briefly described below.

- The Connor-Davidson Resilience scale contains 25 items that are rated on a zero to five scale.
- The Resilience Scale for Adults is a 37 item, five point rating scale focusing on five factors, personal competence, social competence, family competence, social support and family structure.
- The Brief Resilience Coping Scale is a four item questionnaire with five point rating scales designed to measure tendencies to cope and adapt with stress.
- The Resilience Scale is a 25 item questionnaire with a seven point rating scale focusing on personal competence and acceptance of self and life.
- The Baruth Protective Factors Inventory is a 16 item five point rating scale measuring four protective factors to the construct of resilience, adaptable personality, supportive environment, fewer stressors and compensating experiences.

Windle, Bennet and Noyes (2011) carried out a comprehensive review of 15 existing measures of resilience, evaluating them against a clear set of criteria. Measures were scored for quality on content validity, internal consistency, criterion validity, construct validity, reproducibility, agreement, reliability, responsiveness, floor and ceiling effects and interpretability. Whilst they found no current 'gold standard' resilience measure, they found that the Connor-Davidson Resilience Scale, The Resilience Scale for Adults and the Brief Resilience Scale received the best psychometric rating based on their criteria.

The majority of resilience measuring tools use rating scales, most of them a five point rating scale. They all make use of multiple items to generate a total resilience score. An important consideration and something to consider in a global effectiveness measure, is the suggestion that a measure of resilience could be culturally and contextually dependent (Ungar, et al., 2008).

2.5 Empathy

Empathy can be defined as the 'ability to experience the same feelings as those of another person in response to a particular situation' (p.624, Nesdale et al, 2005). It is a complicated and multifaceted construct (Yu & Kirk, 2008) that is shown to promote prosocial behaviour (Spreng, McKinnon, Mar, & Levine, 2009). It is seen as an important construct for caring (Reynolds, Scott, & Jessiman, 1999) and an important trait for medical professionals (Hemmerdinger, Stoddart, & Lilford, 2007). There are existing tools available for measuring empathy, but little consensus on the most effective approach (Yu & Kirk, 2008). Much of the research into empathy has taken place in Western countries (Yu & Kirk, 2008).

Existing measuring tools include the following:

- Toronto Empathy Questionnaire (Spreng, McKinnon, Mar, & Levine, 2009) is a self-assessment of 16 items.
- Empathy Components Questionnaire (Batchelder, Brosnan, & Ashwin, 2017) is a self-assessment tool of 27 items.

> Interpersonal Reactivity Index (Davis, 1980) is a self-assessment of 28 items on a 5 point Likert scale.

2.6 Motivation

Motivation can be defined as the psychological force that enables action (Toure-Tillery & Fishbach, 2014). The IFRC 2016 guidelines highlight the importance of learner motivation on intervention effectiveness and so it is an important construct to consider. Motivation of learners to attend training impacts on the effectiveness of the training (Tai, 2006). However, motivation is difficult as it cannot be observed or recorded directly (Toure-Tillery & Fishbach, 2014). Without motivation behaviour is not likely to happen (Montano & Kasprzyk, 2015) but equally behaviour is not the sole outcome of motivation (Toure-Tillery & Fishbach, 2014).

Toure-Tillery and Fishback (2014) suggest three approaches to measuring motivation, self-reports, cognitive measures and behavioural measures. Self-assessments asking participants to rate their motivation, but this can be limited by interpretation and social desirability concerns (Toure-Tillery & Fishbach, 2014) alongside the previous limitations with self-assessment outlined in section 1.6. Cognitive measures could be used to measure motivation by considering the degree to which an objective is evaluated positively using an explicit measure such as willingness to pay. Behavioural measures of motivation consider the strength of motivation in terms of whether actions correspond with the goal in focus. However there are many other factors that may affect these measurement approaches and so it is important to use multiple measures (Toure-Tillery & Fishbach, 2014).

There are also links between motivation and other constructs outlined in this review. For example if one was more motivated then it may be logical that they are also more willing. Another area where the links are clear is personality testing (see section 2.11). Personality type is suggested to determine the level and type of motivation an individual has (Stocker, 2016). For example, someone with an extroverted personality type is more likely to be motivated by recognition than someone with an introverted personality type.

2.7 Theories of helping behaviour

Helping behaviour involves providing aid or benefit to another person. The term is broad (Levine, 2003) and the concept has a long history. Modern thinking about helping behaviours increased as a result of the Kitty Genovese murder case in 1964 where people failed to help despite witnessing an attack (Latane & Darley, 1969).

Existing measures of helping behaviour can be classified into two categories, either global measures or specific situation measures (Carlo & Randall, 2002). Global measures assess behaviours across contexts and motives, ‘volunteering to help someone is valuable’. Specific context measures are observations of opportunities designed for specific situations. One example of specific context measures are those used in Levine’s (2003) study where researchers created five scenarios and in the field measured how

many people helped, for example helping someone with a hurt leg trying to pick something up. Global measures are argued to be limited because helping behaviours in specific situations are influenced by additional internal norms and extrinsic motivators. Observed measures of behaviours in specific situations may have more ecological validity, but they are not without their disadvantages as they are resource intensive and may be susceptible to observer bias.

Studies that measure helping behaviours often use deception tests because they are high in ecological validity. For examples, see table 5. Deception tests have potential ethical issues as participants are not necessarily aware of all the conditions of the experiment and full informed consent will not have been provided.

Two of the studies outlined in table 5 used what could be described as indirect measures of helping behaviour, Campbell (2001) and Kendrick (2012).

Campbell (2001) asked whether participants in their study had a first aid kit at home and what was in the kit. One year later, participants were asked whether they had since implemented any extra first aid / home safety practices. An increase in having a first aid kit at home could be seen to be a shift in intentions as a result of the interventions, with participants more likely to help as a result of being more prepared to help. It is important to note though that this would be an indirect measure and it could not be claimed that as they have purchased a first aid kit, they are definitely going to help in an emergency.

Kendrick (2012) measured participant's attendance to accident and emergency (A&E) using patient records to see whether attendance had decreased as a result of being involved in the first aid intervention. The focus of this study was to encourage parents to implement better home safety. First aid knowledge and confidence specifically was measured through a questionnaire, but it could provide a suggestion for how increased helping behaviour could be indirectly measured through reduced medical admissions.

Neither of these studies claims to be measuring helping behaviour indirectly through the outcome measures that they have used. However, they offer ideas for potential avenues to indirectly measure increases in helping behaviours. It is important to note though that these two measures were carried out longitudinally, comparing measures taken at the time of the intervention to a later date which is more resource intensive and at risk of attrition.

2.8 First responder training

Another avenue to consider is the training assessments of other types of first responders. Their jobs require them to respond first to emergencies and decide on an appropriate action. Interesting insights could come from considering how they are assessed to be ready to respond.

Fire fighters

Alongside physical training, firefighters undertake psychological tests. This includes the National Firefighter Questionnaire (NFQ) which assesses motivation and attitudes based on the fire service personal qualities and attributes. Learners are required to self-assess the extent to which they agree with statements on a formal written paper and pencil test. One of these personal qualities and attributes assessed is confidence and resilience. This stands out as a quality that could be applicable to a first responder in a first aid emergency.

Paramedics

Paramedic training includes sessions on non-cognitive skills such as emotional resilience, critical thinking, communication and situational awareness. These form part of both their formative and summative assessments throughout the course and in their final paramedic assessment. There does not currently exist a standard test for paramedics as there is for firefighters, this varies by training provider.

2.9 Patient Activation Measure

The Patient Activation Measure (PAM) was developed initially in the USA but has been validated in the UK and is promoted by the National Health Service (NHS). It is an assessment of a patient's ability to manage their own health and care. The assessment is based on constructs including self-efficacy and readiness to change but adopts a broader approach. It consists of 13 statements that assess the patient's activation by asking them about their beliefs, confidence in health related tasks and self-assessed knowledge. Participants are asked the extent to which they agree or disagree with set statements. PAM scores have been found to predict a number of health behaviours. The measures have been found to be valid and reliable over time and across cultures (Hibbard & Gilburt, 2014).

2.10 Situational Judgement Tests

Situational Judgement Tests (SJTs) have been used as far back as the 1940's (Rahman, 2007; Whetzel & McDaniel, 2009) and are popular in many professions, including the training and assessment of medical students and civil servants. A SJT assesses a person's judgement in a workplace situation (Weekley & Ployhart, 2006). A range of options are given for each situation and the participant has to rank them from most effective / appropriate to least effective / appropriate (Rahman, 2007; Whetzel & McDaniel, 2009). They are often used in conjunction with a knowledge based test (Patterson, Kerrin, Edwards, Ashworth, & Baron, 2015; Rahman, 2007) but they reflect a move from the value of knowledge as facts to applied knowledge (Rahman, 2007). Research evaluating SJTs indicate that they are effective (Taylor, Mehra, Elley, Patterson, & Cousands, 2016) and frequently used as selection tools both in the US and Europe (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). They have been shown to be reliable and valid assessment tools for non-academic skills in healthcare professions (Taylor, Mehra, Elley, Patterson, & Cousands, 2016).

2.11 Personality tests

Personality characteristics play a role in behaviour (Stocker, 2016). There are many commercially available personality tests for example;

- Colour works. This is an online questionnaire consisting of 25 questions which then generates a personality profile. It is accredited by the British Psychological Society and based on the work of the psychologist Carl Jung.
- The Myers Briggs Type Indicator. This is a structured questionnaire of 88 items where the participant is asked to choose one of two responses which is either a word or short phrase to generate a personality profile.

Personality research is generally agreed to be described through the five factor model (Grumm & Collani, 2007). This includes Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness (McCrae & John, 1992). The Big Five Aspects Scale (DeYoung, Quilty, & Peterson, 2007) based on these concepts contains 44 items that are rated on a five step scale from disagree strongly to agree strongly. It has been shown to be reliable and valid (DeYoung, Quilty, & Peterson, 2007). More recently there is a demand for shorter, quicker personality tests (Rammstedt & John, 2007). This has led to the development of a shorter version of the Big Five Aspects Scale, the BFI-10. This shortened version has been shown to still retain significant reliability and validity (Rammstedt & John, 2007).

Personality tests tend to use self-assessment measures. There are some concerns about the quality of self-assessment personality measures (Pomerance & Converse, 2014). They require each individual to interpret the personality items in the same way. One suggestion to improve this is to provide a frame of reference for the measures. This provides a context and enables an individual to respond to items more accurately. Frame of reference measures have been found to be better at predicting behaviours than non-frame of reference measures (Bowling & Burns, 2010).

Personality tests are widely used in both the workplace and for job interview assessments. It is questionable whether this is applicable to first aid situations and whether it appropriately measures whether an individual is motivated to help.

3. Conclusions and considerations moving forward

This review of the literature has highlighted a range of potential outcome measures that could be relevant to evaluate the effectiveness of first aid education. It offers suggestions for how some constructs could be measured and evidence from research studies of how they have been measured.

It highlighted the need for a consistent approach to measuring educational effectiveness generally as well in the specific first aid environment. A global measure would enable a level of consistency. Currently most studies evaluate first aid effectiveness by creating their own instruments, having an available measure could support future research. A recognised measure would enable comparisons to be made and ultimately lead to an improvement in the quality of first aid education.

Two repeating themes emerged across many of the constructs discussed. This was the importance of context and aggregation. More accurate measures situate assessments in a context. This allows respondents to draw on all the influencing factors and respond more accurately. In relation to aggregation, having multiple ways of measuring the same construct increases the accuracy of the measurement. These are two key considerations to take forward regardless of the measures selected.

A main challenge that is expected is ensuring any measure can be generalised to a global population. The generalisability of measuring instruments was a repeated concern throughout. Many of the constructs explored could be argued to be Western in nature and the current review is limited by an English language only search.

It is important to note that this review provides only a brief overview of the suggested constructs and measures. If any of the measures are chosen to be taken forward then additional desk research would need to be carried out to explore that measure in detail.

There is a gap in research evidence showing how people are actually helping. This is not necessarily evidence to show that those receiving first aid education do help. Responding to a first aid emergency requires more than the constructs outlined, it also relies on having an opportunity to be in the right situation at the right time. Further insights into the actual current situation and how people help could inform the development of an effective measure.

Bibliography

Airhihenbuwa, C., & Obregon, R. (2000). A critical assessment of theories/models used in health communication for HIV/AIDS. *Journal of Health Communication* , 5-15.

Ajzen, I. (1991). The Theory of Planned Behavior. *Organisational Behavior and Human Decision Processes* , 179-211 .

Ajzen, I. (2011). The theory of planned behaviour: Reactions and reflections . *Psychology and Health* , 1113-1127.

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior.* Englewood Cliffs, NJ: Prentice-Hall.

Ajzen, I., & Fishbein, M. (2005). The influence of attitudes on behavior. In D. Albarracin, B. Johnson, & M. Zanna, *The handbook of attitudes* (pp. 173-221). Mahwah, NJ: Erlbaum .

Ajzen, I., & Fishbein, M. (2010). *Predicting and Changing Behavior: The Reasoned Action Approach.* New York: Psychology Press (Taylor and Francis) .

Ajzen, I., & Sexton, J. (1999). Depth of processing, belief congruence and attitude-behaviour correspondence . In S. Chaiken, & Y. Trope, *Dual-process theories in social psychology* (pp. 117-138). New York, NY: Guilford .

American Psychological Association. (2017). *What is resilience* . Retrieved from psychcentral: https://psychcentral.com/lib/what-is-resilience/

Anderson, G., Gaetz, M., & Masse, J. (2011). First aid skill retention of first responders within the workplace . *Scandinavian Journal of trauma, resuscitation and emergency medicine* .

Armitage, C., & Conner, M. (2001). Efficacy of the theory of planned behaviour: A meta-analytic review. *British Journal of Social Psychology*, 471-499.

Armitage, C., Norman, P., Alganem, S., & Conner, M. (2015). Expectations are more predictive of behavior than behavioral intentions: Evidence from two prospective studies . *Annals of Behavioral Medicine*, 239-246 .

Arnold, J., Johnson, L., Tucker, S., Malec, J., Henrickson, S., & Dunn, W. (2009). Evaluation Tools in Simulation Learning: Performance and Self-Efficacy in Emergency Response . *Cliniacl Simulation in Nursing* , 35-43.

Azeredo, R., & Stephens-Stidham, S. (2003). Design and implementation of injury prevention curricula for elementary schools: lessons learned . *Injury prevention* , 274-278.

Bandura, A. (1994). Self-efficacy. In V. S. Ramachaudran, *Encyclopedia of human behaviour* (pp. 71-81). New York: Academic Press .

Bandura, A. (2006). Guide for constructing self-efficacy scales . In T. Urdan, & F. Pajares, *Self-efficacy beliefs of adolescents* (pp. 307-337).

Batchelder, L., Brosnan, M., & Ashwin, C. (2017). The Development and Validation of the Empathy Components Questionnaire (ECQ). *PLoS ONE*.

Bhattacherjee, A., & Sanford, C. (2009). The intention behaviour gap in technology usage: the moderating role of attitude strength. *Behaviour and Information Technology*, 389-401.

Bollig, G., Wahl, H., & Svendsen, M. (2009). Primary school children are able to perform basic life-saving first aid measures. *Resuscitation*, 689-692 .

Boulet, G. (n.d.). *The difference between knowledge and skills: knowing does not make you skilled* . Retrieved 09 2017, from elearning inudstry : https://elearningindustry.com/difference-between-knowledge-and-skills-knowing-not-make-skilled

Bowling, N., & Burns, G. (2010). A comparison of work specific and general personality measures as predictors of work and non work criteria. *Personality and Individual differences* , 95-101.

Brady, A. (2005). Assessment of learning with multiple-choice questions . *Nurse Education in Practice* , 238-242.

Burke, L., Dunbar-Jacob, J., Sereika, S., & Ewart, C. (2003). Development and Testing of the Chloesterol Lowering Diet Self Efficacy Scale . *European Journal of Cardiovascular Nursing*.

Campbell, N., Ayala, G., Litrownik, A., Slymen, D., Zavala, F., & Elder, J. (2001). Evaluation of a first aid and home safety program for hispanic migrant adolescents . *American Journal of Preventive Medicine*, 258-265.

Carlo, G., & Randall, B. (2002). The Development of a Measure of Prosocial Behaviors for Late Adolescents . *Faculty Publications, Department of Psychology* .

Carter, G., Creedy, D., & Sidebotham, M. (2015). Evaluation of tools used to measure critical thinking development in nursing and midwifery undergraduate students: A systematic review . *Nurse Education Today*, 864-874.

Chraska, M., & Chraskova, M. (2017). Semantic Differential and its Risks in the Measurement of Students' Attitudes. *Procedia- Social and Behavioral Sciences* , 820-829.

Clanton, J., Gardner, A., Cheung, M., Mellert, L., Evancho-Chapman, M., & George, R. (2014). The relationship between confidence and competence in the development of surgical skills. *Journal of Surgical Education* , 405-412.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., et al. (1966). *Equality of educational opportunity.* Washington, DC: US Government Printing Office .

Communities and Local Government . (2009, December). *National Firefighter Selection Process National Firefighter Questionnaire (NFQ).* Retrieved from www.gov.uk.

Conner, M., Gaston, G., Sheeran, P., & Germain, M. (2013). Some feelings are more important: Cognitive attitudes, affective attitudes, anticipated affect and blood donation. *Health Psychology*, 264-272.

Considine, J., Botti, M., & Thomas, S. (2005). Design, format, validity and reliability of multiple choice questions for use in nursing research and education . *Collegian*, 19-24 .

Crea, K. (2011). Practice Skill Development Through the Use of Human Patient Simulation . *American Journal of Pharmaceutical Education*.

Creemers, B. P., & Kyriakides, L. (2006). Critical analysis of the current approaches to modelling educational effectiveness: The importance of establishing a dynamic model. *School Effectiveness and School Improvement*, 347-366.

Daly, W. (2001). The development of an alternative method in the assessment of critical thinking as an outcome of nursing education. *Journal of Advanced Nursing* , 120-130.

Davis, D., Mazmanian, P., Fordis, M., Van Harrison, R., Thorpe, K., & Perrier, L. (2006). Accuracy of Physician Self-assessment Compared With Observed Measures of Competence . *Clinicians corner*, 1094-1102.

Davis, M. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology*.

DeYoung, C., Quilty, L., & Peterson, J. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology*, 880-896.

Durham, C., & Alden, K. (2008). Enhancing Patient Safety in Nursing Eduaction Through Patient Simulation. In R. Hughes, *Patient Safety and Quality: An evidence based handbook for nurses* (p. Chapter 51). Rockville, US .

Dwyer, T. (2008). Psychological factors inhibit family members' confidence to initiate CPR. *Prehospital Emergency Care*, 157-161.

Engeland, A., Roysamb, E., Smedslund, G., & al., e. (2002). Effects of first aid training in junior high schools . *International Journal of Injury Control and Safety Promotion* , 99-106.

Ennis, R., & Millman, K. (1985). *Cornell Critical Thinking Tests, Levels X and Z .* Pacific Grove, Calif: Midwest Publications .

Epstein, R. (2007). Assessment in Medical Education. *New England Journal of Medicine* , 387-396.

Everett-Thomas, R., Yero-Aguayo, M., Valdes, B., Valdes, G., Shekhter, I., Rosen, L., et al. (2016). An assessment of CPR skills using simulation: Are first responders prepared to save lives. *Nurse Education in Practice*, 58-62 .

Eze, C., Ebuehi, O., Brig, F., Otte, W., & Igwe, S. (2015). Effect of health education on trainee teachers' knowledge, attitudes and first aid management of epilepsy: An interventional study . *Seizure*, 46-53.

Fishbein, M. (1967). *Readings in attitude theory and measurement .* New York: Wiley.

Fishbein, M. (2000). The role of theory in HIV prevention. *AIDS care* , 273-278.

Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behaviour: the reasoned action approach.* New York: Pscyhology Press .

Frederick, K., Bixby, E., Orzel, M., Stewart-Brown, S., & Willet, K. (2000). An evaluation of the effectiveness of the Injury Minimization Programme for Schools (IMPS). *Injury Prevention*, 92-95.

Gelerstein, D., del Rio, R., Nussbaum, M., Chiuminato, P., & Lopez, X. (2016). Designing and implementing a test for measuring critical thinking in primary school. *Thinking Skills and Creativity*, 40-49.

Gibbons, F., & Gerrard, M. (1997). Health images and their effects on health behavior. In B. Buunk, & F. Gibbons, *Health, coping and well-being: Perspectives from social comparison theory* (pp. 63-94). Mahwah, NJ: Lawrence Erlbaum Associates Publishers .

Gibbons, F., Gerrard, M., Blanton, H., & Russell, D. (1998). Reasoned action and social reaction: willingness and intention as independent predictors of health risk. *Journal of Personality and Social Psychology*, 1164-1180.

Grumm, M., & Collani, G. (2007). Measuring Big-Five personality dimensions with the implicit association test - Implicit personality traits or self-esteem . *Personality and Individual Differences* , 2205-2217.

Hardeman, W., Johnston, M., Johnston, D., Bonneti, D., Wareham, N., & Kinmonth, A. (2002). Application of the theory of planned behaviour in behaviour change interventions: A systematic review . *Psychology and Health*, 123-158.

Hawks, S., & Egan, M. (1998). The impact of three different first aid curricula on emergency helping among college students . *Journal of Health Education*, 289-293.

He, Z., Wynn, P., & Kendrick, D. (2013). Non-resuscitative first-aid training for children and laypeople: a systematic review . *Emergency Medicine Journal*, 763-768.

Helping behaviour. (n.d.). Retrieved 09 2017, from Psychology: https://psychology.iresearchnet.com/social-psychology/prosocial-behavior/helping-behavior/

Hemmerdinger, J., Stoddart, S., & Lilford, R. (2007). A systematic review of tests of empathy in medicine . *BMC Medical Education*.

Hernandez-Padilla, J., Suthers, F., Fernandez-Sola, C., & Granero-Molina, J. (2016). Development and psychometric assessment of the Basic Resuscitation Skills Self-Efficacy Scale. *European Journal of Cardiovascular Nursing* , 10-18.

Hibbard, J., & Gilburt, H. (2014). *Supporting people to manage their health: An introduction to patient activation .* The Kings Fund .

Hughes, R., & Huby, M. (2002). The application of vignettes in social and nursing research. *Journal of Advanced Nursing*, 382-386.

IFRC. (2016). *International first aid and resuscitation guidelines.* Geneva.

Jansen, J., Tan, L., Van der Vleuten, C., Luijk, S., Rethans, J., & Grol, R. (1995). Assessment of competence in technical clinical skills of general practitioners . *Medical Education* , 247-253.

Jenkins, H. (1985). Improving clinical decisio making in nursing. *Journal of Nursing Education* , 242-243.

Kalebic-Maglica, B., & Martinac-Dorcic, T. (2016). Components of the Prototype/Willingness Model as Concurrent and Prospective Predictors of Adolescents' Alcohol Consumption. *Journal of Child Adolescent Substance Abuse* , 159-168.

Kelly, N., Huffman, L., Mendoza, F., & Robinson, T. (2003). Effects of a Videotape to Increase Use of Poison Control Centres by Low-Income and Spanish-Speaking Families: A Randomized Control Trial . *Pediatrics*, 21-26.

Kendrick, D., Young, B., Mason-Jones, A., Ilyas, N., Achana, F., Cooper, N., et al. (2012). Home safety education and provision of safety equipment for injury prevention. *Cochrane Database of Systematic Reviews* .

Kirkpatrick, D., & Kirkpatrick, J. (2006). *Evaluating Training Programs: The Four Levels (3rd Edition).* San Francisco, CA: Berrett-Koehler Publishers .

Klassen, R. (2004). Optimism and realism: A review of self-efficacy from a cross-cultural perspective. *International Journal of Psychology*, 205-230.

Kroenung, E., & Eckhardt, A. (2015). The attitude cube- A three-dimensional model of situational factors in IS adoption and their impact on the attitude-behavior relationship. *Journal of Information and Management* , 611-627.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 1121-1134.

Lai, E. (2011). *Critical Thinking: A Literature Review.* Pearson .

Latane, B., & Darley, J. (1969). Bystander "Apathy". *American Scientist*, 244-268.

Lavrakas, P. (2008). Attitude Measurement . In P. Lavrakas, *Encyclopedia of Survey Research Methods .*

Levine, R. (2003). *Measuring Helping Behavior Across Cultures .* Online Readings in Psychology and Culture .

Lippmann, J., Livingston, P., & Craike, M. (2011). Comparison of two modes of delivery of first aid training including basic life support. *Health Education Journal*, 131-140.

Liu, O., Frankel, L., & Roohr, K. (2014). *Assessing Critical Thinking in Higher Education: Current State and Directions for Next-Generation Assessment .* ETS Research Report Series .

Macpherson, K., & Owen, C. (2010). Assessment of critical thinking ability in medical students . *Assessment and Evaluation in Higher Education*, 41-54.

Mahony, P., Griffiths, R., Larsen, P., & Powell, D. (2008). Retention of knowledge and skills in first aid and resuscitation by airline cabin crew. *Resuscitation*, 413-418.

Major, D. (2005). OSCEs - seven years on the bandwagon: the progress of an objective structured clinical evaluation programme. *Nurse Education Today*, 442-454.

Malek, C. (1986). A model for teaching critical thinking. *Nurse Educator*, 20-23.

Matterne, U., Diepgen, T., & Weisshaar, E. (2011). A longitudinal application of three behaviour models in the context of skin protection behaviour in individuals with occupational skin disease. *Psychology and Health*, 1118-1207.

McCrae, R., & John, O. (1992). An introduction to the Five-Factor Model and its applications . *Journal of Personality* , 175-215.

McDaniel, M., Morgeson, F., Finnegan, E., Campion, M., & Braverman, E. (2001). Predicting job performance using situational judgment tests: A clarification of the literature . *Journal of Applied Psychology*, 730-740 .

McEachan, R., Conner, M., Taylor, N., & Lawton, R. (2011). Prospective prediction of health-related behaviours with the theory of planned behaviour: A meta-analysis. *Health Psychology Review* , 97-144.

McLeod, S. (2009). *Attitude Measurement*. Retrieved 09 2017, from Simply Psychology: https://www.simplypsychology.org/attitude-measurement.html

McWhirter, J., & Wetton, N. (1994). Children's perceptions of risk. *Journal of Health and Safety*, 21-29.

Merrell, K. (2003). *Behavioral, social and emotional assessment of children and adolescents .* Psychology press .

Middlestadt, S. E. (2012). Beliefs underlying eating better and moving more: lessons learned from comparative salient belief elicitations with adults and youths . *Annals of the American Academy of Political and Social Science* , 81-100.

Miller, G. (1990). The Assessment of Clinical Skills / Competence / Performance . *Academic Medicine* , 563-567.

Miller, M. (1992). Outcomes evaluation: measuring critical thinking. *Journal of Advanced Nursing*, 1401-1407.

Mitchell, M., Henderson, A., Groves, M., Dalton, M., & Nulty, D. (2009). The objective structured clinical examination (SCE): Optimising its value in the undergraduate nursing curriculum. *Nurse Education Today*.

Moghavvemi, S., Salleh, N., Sulaiman, A., & Abessi, M. (2015). Effect of external factors on intention-behaviour gap. *Behaviour and Information Technology*, 1171-1185.

Montano, D., & Kasprzyk, D. (2015). Theory of Reasoned Action, Theory of Planned Behaviour, and the Integrated Behavioural Model . In K. Glanz, B. Rimer, & K. Viswanath, *Health Behavior: Theory, Research and Practice* (pp. 95-119).

Moore, J. (1987). Effects of assertion training and first aid instruction on children's autonomy and self care agency. *Research in Nursing and Health* , 101-109.

Morris, J., Maizano, M., Dandy, N., & O'Brien, L. (2012). *Theories and models of behaviour and behaviour change .* Retrieved 09 2017, from Forest Research .

Munshi, F., Lababidi, H., & Alyousef, S. (2015). Low versus high fidelity simulations in teaching and assessing clinical skills . *Journal of Taibah University Medical Sciences* , 12-15 .

Nehring, W., Ellis, W., & Lashley, F. (2001). Human Patient Simulators in Nursing Education: An overview . *Simulation and Gaming* , 194-204.

Nesdale, D., Griffith, J., Durkin, K., & Maass, A. (2005). Empathy, group norms and children's ethnic attitudes . *Journal of Applied Developmental Psychology* , 623-637.

NHS England . (n.d.). *Patient activation and PAM FAQS*. Retrieved 09 2017, from NHS England : https://www.england.nhs.uk/ourwork/patient-participation/self-care/patient-activation/pa-faqs/

Nhu Lam, N., Li, F., Anh Tuan, C., & Huong, H. (2017). To evaluate first aid knowledge on burns management amongst high risk groups . *Burns Open*, 29-32 .

Nicholas, M., & Labig, C. (2013). Faculty approaches to assessing critical thinking in the humanities and the natural and social sciences: Implications for general education. *The Journal of General Education* , 297-319.

OECD. (2012). *Education at a glance 2012: OECD indicators .* Paris, France: OECD Publishing.

Oxford Dictionary. (n.d.). Retrieved 09 2017, from https://en.oxforddictionaries.com/definition/knowledge

Oxford Dictionary. (n.d.). Retrieved 09 2017, from https://en.oxforddictionaries.com/definition/skill

Papathanasiou, I., Kleisiaris, C., Fradelos, E., Kakou, K., & Kourkouta, L. (2014). Critical Thinking: The Development of an Essential Skill for Nursing Students. *Acta Informatica Medica*, 283-286.

Patterson, F., Kerrin, M., Edwards, H., Ashworth, V., & Baron, H. (2015). *Validation of the F1 Selection Tools .* Work Psychology Group .

Peterson, L. (1984). The "safe at home" game. Training comprehensive prevention skills in latchkey children. *Behavior Modification*.

Pomerance, M., & Converse, P. (2014). Investigating context specificity, self-schema characteristics and personality test validity . *Personality and Individual Differences* , 54-59.

Rahman, M. (2007, May 19). Tackling situational judgment tests . *BMJ*.

Rammstedt, B., & John, O. (2007). Measuring personality in one minute or less: A 10 item short version of the Big Five Inventory in English and German . *Journal of Research in Personality* , 203-212.

Reynolds, D., Sammons, P., De Fraine, B., Van Damme, J., Townsend, T., Teddlie, C., et al. (2014). Educational effectiveness research (EER): a state-of-the-art review. *School Effectiveness and School Improvement*, 197-230.

Reynolds, W., Scott, B., & Jessiman, W. (1999). Empathy has not been measured in clients' terms or effectively taught: a review of the literature. *Journal of Advanced Nursing* , 1177-1185.

Rossetto, A., Jorm, A., & Reavley, N. (2016). Predictors of adults' helping intentions and behaviours towards a person with a mental illness: A six-month follow-up study. *Psychiatry Research* , 170-176.

Rothschild, D., & Wolfers, J. (2011). *Forecasting elections: Voter intentions versus expectations.* Retrieved from http://assets.wharton.upenn.edu/~rothscdm/RothschildExpectations.pdf

Rudinow, J., & Barry, V. (1999). *Invitation to critical thinking (4th ed).* Fort Worth TX: Harcourt Brace .

Sangowawa, A., & Owoaje, E. (2012). Building capacity of drivers in Nigeria to provide first aid for road crash victims . *Injury Prevention* , 62-65.

Schoenberg, N., & Ravdal, H. (2000). Using vignettes in awareness and attitudinal research. *International Journal of Social Research Methodology*, 63-74.

Schumann, S., Schimelpfenig, T., Sibthorp, J., & Collins, R. (2012). An Examination of Wilderness First Aid Knowledge, Self-Efficacy and Skill Retention. *Wilderness and Environmental Medicine*, 218-287.

Schwarzer, R., & Jerusalem, M. (1995). Generalised Self-Efficacy Scale . In J. Weinman, S. Wright, & M. Johnston, *Measures in health psychology: A user's portfolio. Causal and control beliefs* (pp. 35-37). Windsor, UK: NFER, NELSON .

Sheeran, P. (2002). Intention-Behavior Relations: A conceptual and empirical review . *European Review of Social Psychology*, 1-36.

Sheppard, B., Hartwick, J., & Warshaw, P. (1988). The theory of reasoned action: A meta-analysis of past research with recommendations for modifications and future research. *Journal of Consumer Research*, 325-343.

Shotland, R., & Heinold, W. (1985). Bystander response to arterial bleeding: helping skills, the decision-making process and differentiating the helping response . *Journal of Personality and Social Psychology*, 347-356.

Sitzmann, T., Ely, K., Brown, K., & Bauer, K. (2010). Self-assessment of knowledge: a cognitive learning or affective measure . *Academy of Management Learning and Education* , 169- 191.

Sniehotta, F., Presseau, J., & Araujo-Soares, V. (2014). Time to retire the theory of planned behaviour. *Health Psychology Review*, 1-7.

Speechley, M., Weston, W., Dickie, G., & Orr, V. (1994). Self assessed competence: before and after residency. *Canadian Family Physician*, 459-464.

Spreng, N., McKinnon, M., Mar, R., & Levine, B. (2009). The Toronto Empathy Questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy. *Journal of Personality Assessment* , 62-71.

Stewart, J., O'Halloran, C., Barton, R., Singleton, S., Harrigan, P., & Spencer, J. (2000). Clarifying the concepts of confidence and competence to produce appropriate self-evaluation measurement scales . *Medical Education*, 903-909.

Stocker, B. (2016). Validity and reliability of personality tests. *Comparative European Research* , 43-46.

Swor, R., Khan, I., Domeier, R., Honeycutt, L., Chu, K., & Compton, S. (2006). CPR Training and CPR Performance: Do CPR trained Bystanders Perform CPR. *Society for Academic Emergency Medicine*, 596-601.

Tai, W. (2006). Effects of training framing, general self-efficacy and training motivation on trainees' training effectiveness . *Personnel Review*, 51-65.

Taylor, N., Mehra, S., Elley, K., Patterson, F., & Cousands, F. (2016). The value of situational judgement tests for assessing non-academic attributes in dental selection. *British Dental Journal*.

Timko, T., & Sainato, D. (1999). Effects of First Aid Training Using Small Group Instruction With Young Children With Disabilities. *Journal of Early Intervention*.

Todd, J., Kothe, E., Mullan, B., & Monds, L. (2014). Reasoned versus reactive prediction of behaviour: a meta analysis of the prototype willingness model. *Health Psychology Review*, 1-24.

Toure-Tillery, M., & Fishbach, A. (2014). How to measure motivation: A guide for the experimental social psychologist . *Social and Personality Psychology Compass* , 328-341.

Tracey, J., Arroll, B., Barham, P., & Richmond, D. (1997). The validity of general practitioners' self assessment of knowledge: cross sectional study. *General Practice*.

UNESCO. (2000, April 26-28). Dakar Framework for Action. Education for All: Meeting Our Collective Commitments. Adopted by the world education forum, Dakar, Senegal. Paris.

Ungar, M., Liebenberg, L., Boothroyd, R., Kwong, W., Lee, T., Leblanc, J., et al. (2008). The study of young resilience across cultures: Lessons from a pilot study of measurement development. *Research in Human Development*, 166-180.

Van de Velde, S., Heselmans, A., Roex, A., Vandekerckhove, P., Ramaekers, D., & Aertgeets, B. (2009). Effectiveness of Nonresuscitative First Aid Training in Laypersons: A Systematic Review. *Annals of Emrgency Medicine* , 447-457.

Vuma, S., & Sa, B. (2017). A comparison of clinical-scenario versus stand-alone multiple choice questions in a problem-based learning environment in undergraduate medicine . *Journal of Taibah University Medical Sciences* , 14-26.

Ward, H., & Barratt, J. (2005). Assessment of nurse practitioner advanced clinical practice skills: using the objective structured clinical examination (OSCE) Helen Ward and Julian Barrat examine how OSCEs can be developed to ensure a robust assessment of clinical competence . *Primary Health Care*, 37-41.

Ward, H., & Willis, A. (2006). Assessing advanced clinical practice skills: Helen Ward and Annaliese Willis show how the development of an objective structured clinical assessment (OSCA) has enabled assessment of nurse practitioner advanced practice clinical skills at Masters level. *Primary Health Care*, 22-24 .

Warshaw, P., & Davis, F. (1984). *Disentangling behavioral intention (BI) and behavioral expectation (BE): The latter predicts better.*

Weekley, J., & Ployhart, R. (2006). An introduction to situational judgment testing. In J. Weekley, & R. Ployhart, *Situational Judgment tests .* Mahwah, NJ : Lawrence Erlbaum Associates .

Weinstein, N. (2007). Misleading tests of health behavior theories. *Annals of Behavioral Medicine*, 1-10.

Whetzel, D., & McDaniel, M. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review*, 188-202 .

White, J., & McNulty, A. (2011). *Assessing the links between first aid training and community resilience .*

Willet, L., Heudebert, G., Palonen, K., Massie, S., Kiefe, C., & Allison, J. (2009). The importance of measuring competency based outcomes: standard evaluation measures are not surrogates for clinical performance of internal medicine residents . *Teaching and Learning in Medicine* , 87-93.

Windle, G., Bennett, K., & Noyes, J. (2011). A methodological review of resilience measurement scales. *Health and Quality of Life Outcomes* .

Witt, P., & Wheeles, L. (2001). An experimental study of teachers' verbal and nonverbal immediacy and students affective and cognitive learning . *Communication Education*, 327-342.

Yap, M., & Jorm, A. (2012). Young people's mental health first aid intentions and beliefs prospectively predict their actions: Findings from an Australian National Survey of Youth . *Psychiatry Research*, 315-319.

Yu, J., & Kirk, M. (2008). Measurement of empathy in nursing research: systematic review. *Journal of Advanced Nursing* , 440-454.

Yzer, M. (2012). The integrated model of behavioral prediction as a tool for designing health meassages . In H. Cho, *Health Communication Message Design* (pp. 21-40). Sage: Ohio, USA .

Appendix 1: Table 5 Outline of measures used in studies

Knowledge	Skills	Helping behaviours	Confidence	Attitudes	Other
Interview administered survey, asked about their knowledge of first aid kits in the home and the correct steps to respond to a first aid emergency (Campbell, 2001)	Observation of children's performance in a situation against set criteria (Bollig, 2009)	Assessed helping behaviours through a deception experiment of a simulated scenario (Shotland, 1985)	Confidence in first aid skills, five items on a four point Likert scale, e.g. how confident they felt accessing the emergency services (Campbell, 2001)	Measured changes in attitudes towards seat belt use (Azeredo, 2003)	Measured attendance to A&E and whether it decreased as a result of the first aid intervention (Kendrick, 1999)
First aid knowledge assessed through number of questions answered correctly (Esposito, 1985)	Asked to respond to two audio recorded scenarios, their responses were coded in terms of the steps chosen and order completed in. Assessed at baseline, after intervention and one year follow up (Campbell, 2001)	Used a deception experiment to measure helping behaviour, measured the time taken to help and actions taken (Van de Velde, 2012)	Postal questionnaires asking about confidence in dealing with first aid (Kendrick, 1999)	Attitudes towards first aid and intended behaviours measured on a 1 – 100 scale (Engeland, 2002)	Assessors also asked to generally grade the participants overall performance from 0-10 (Lippmann, 2010)
Test for response on burns and poisoning (Zhao, 2006)	Demonstration of different skills on a manikin pre and post intervention (Esposito, 1985)	Observation of helping behaviours in a deception test of a simulated emergency (Hawks, 1998)	Measured self-efficacy related to barriers (Van de Velde, 2012)	Draw and write method used to measure attitudes and hypothetical	

				behaviour towards safety (Frederick, 2000)	
Assessment of first aid knowledge on a written test (Moore, 1987)	Assessment of first aid skills on a written test (Moore, 1987)		Self-efficacy measured using a self-efficacy scale where participants rated their perceived ability to perform specific skills on an 11 point scale (Schumann et al, 2012)	Nine questions exploring how positive teachers attitudes were towards epilepsy (Eze et al, 2015)	
Written test and follow up interviews assessing knowledge of poisoning (Kelly, 2003)	Interviews by a nurse where participants were asked to demonstrate on a manikin their reaction to a stated scenario, response graded by the nurse. Comparison to a control group at 1 week, 1 month and 13 months (Capone, 2000)		Participants self-confidence of knowledge and skills assessed through a five point Likert scale (Mahony et al, 2008)		
Assessed first aid knowledge through a written postal questionnaire (Kendrick,	Used an OSCE assessment of different selected practical skills				

1999)	against a predefined assessment criteria (Ertl, 2007)				
Written test assessing first aid knowledge through 37 multiple choice questions (Li, 2011)	Practical assessment of skills where participants were asked to respond to different situations and graded from not competent to very good (Lippmann, 2010).				
Written test assessing first aid knowledge through 24 multiple choice questions (Lippmann, 2010)	Observation of number of steps completed successfully (Timko, 1999)				
Knowledge measured through 25 questions (Sangowawa, 2012)	Skills measured through 24 questions and observations of skills mentioned and demonstrated in simulated scenarios (Sangowawa, 2012)				
Multiple choice test to assess knowledge about steps to take in a first aid situation (Van de Velde, 2012)	Observed first aid skills (Shotland, 1985)				
Knowledge measured through true and false / multiple choice questions	Skills observed in scenarios to measure skill retention and behaviours				

(age dependent) (Azeredo, 2003)	(Frederick, 2000)				
43 questions on first aid knowledge and safety/ prevention (Carruth, 2009)	Scenarios for participants to role play their first aid skills, correct responses measured (Peterson, 1984)				
Knowledge measured on 53 items (Engeland, 2002)	Practical skills performance test (Breivik, 1980)				
Knowledge measured through a quiz (Frederick, 2000)	Practical skills assessed using observations (Hawks, 1998)				
Knowledge measured through 13 multiple choice questions on the topics covered in the intervention (Lubrano, 2005).	Nine questions assessing teachers actions in a given scenario (Eze et al, 2015)				
Written knowledge test (Breivik, 1980)	Simulated test scenario where participants actions were measured using a checklist by an observer (Mahony et al, 2008)				
Written knowledge test using multiple choice questions (Hawks, 1998)	Observations of participants actions in a choking scenario (Anderson et al, 2011)				
Self-administered	Simulated scenarios				

questionnaires with 28 closed questions on knowledge of epilepsy (Eze et al, 2015)	where participants actions were evaluated on a predefined observation form (Everett-Thomas et al, 2016)				
Knowledge measured on 25 multiple choice questions (Schumann et al, 2012)					
12 multiple choice tick box questions on knowledge about burns (Nhu Lam et al, 2017)					
Self-assessment of knowledge in three situations, open questions where participants were asked to describe in detail their first aid actions (Mahony et al, 2008)					
Multiple choice written exam (Anderson et al, 2011)					

Appendix 2: Links to more information on existing tools or websites from which tools can be purchased

- Cornell Critical Thinking Test http://faculty.education.illinois.edu/rhennis/cornellclassreas.pdf
- The California Critical Thinking Test https://www.insightassessment.com/Products/Products-Summary/Critical-Thinking-Skills-Tests/California-Critical-Thinking-Skills-Test-CCTST#sthash.0cnhglYK.dpbs
- The Watson Glaser Critical Thinking Appraisal https://www.assessmentday.co.uk/watson-glaser-critical-thinking.htm
- Californian Critical Thinking Disposition Inventory https://www.insightassessment.com/Products/Products-Summary/Critical-Thinking-Attributes-Tests/California-Critical-Thinking-Disposition-Inventory-CCTDI
- The Connor-Davidson Resilience scale http://www.connordavidson-resiliencescale.com/
- The Resilience Scale for Adults
- The Brief Resilience Coping Scale file:///H:/Downloads/Brief%20Resilience%20Scale.pdf
- The Resilience Scale http://www.resiliencecenter.com/resilience-scale/
- The Baruth Protective Factors Inventory http://connection.ebscohost.com/c/articles/9054662/formal-assessment-resilience-baruth-protective-factors-inventory
- PISA tests http://www.oecd.org/pisa/
- Emergency response confidence tool (Arnold et al, 2009) http://www.sciencedirect.com/science/article/pii/S1876139908000595
- General self efficacy scale (Schwarzer and Jerusalem, 1995) http://userpage.fu-berlin.de/health/selfscal.htm
- Ajzen's TPB questionnaire design guidance http://people.umass.edu/aizen/tpb.html
- Thematic apperception task http://www.utpsyc.org/TATintro/
- Rorschach Inkblot task http://theinkblot.com/
- Draw a person task https://janaenahirney.wordpress.com/2015/10/28/the-draw-a-person-personality-test/
- Attitude semantic differential scale
- Firefighter NFQ https://www.gov.uk/government/publications/national-firefighter-selection-process-national-firefighter-questionnaire-technical-manual
- PAM https://www.england.nhs.uk/ourwork/patient-participation/self-care/patient-activation/pa-faqs/
- Colour works personality test http://www.thecolourworks.com/
- Myers Briggs Type Indicator http://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/home.htm?bhcp=1
- Big 5 aspects scale http://ipip.ori.org/BFASKeys.htm
- BFI-10 https://www.ocf.berkeley.edu/~johnlab/pdfs/BFI-10.doc
  The Kirkpatrick Model https://www.kirkpatrickpartners.com/Our-Philosophy/The-Kirkpatrick-Model