The Assessment and Generation of Evidence for XR Approaches in Training

Mayowa Olonilua
DSTL
UNITED KINGDOM

moolonilua@dstl.gov.uk

Samantha Black
NSC
UNITED KINGDOM

Samantha.black@nsc.co.uk

Eleanor Forrest
Bright HF Limited
UNITED KINGDOM

eleanor@brighthf.co.uk

Petra Saskia Bayerl
CENTRIC
UNITED KINGDOM

p.s.bayerl@shu.ac.uk

ABSTRACT

Virtual, Augmented and Mixed Reality (collectively known as XR) are rapidly maturing technologies that have attracted a great deal of attention from the training and education research community. How can military personnel best assess the breadth of literature on this topic as part of their evidence-based decision-making process? In this paper we present two new frameworks for assessing the quality of XR evidence and for running trials that will produce quality evidence.

The Quality Assessment Framework delivers a bespoke research output review protocol combining best practice methodologies and quality criteria from existing industry frameworks with technology factors specific to the use of XR. The Trials Planning Framework delivers a modular and adaptive framework designed to support the evaluation of disparate training and assessment goals relating to new XR technologies.

Together these frameworks support an XR for Training and Education (XR4TE) Portfolio of evidence (PoE), a key enabler of the exploitation of XR technologies in defence. The Quality Assessment protocol will be applied when entering research papers into the XR4TE PoE ensuring PoE output is usable and meaningful. The protocol will also feed into the Trials Planning Framework which will provide a gold-standard for the evaluation of AR/VR/XR capabilities within DSTL[1].

1.0 INTRODUCTION

There is a need to integrate novel training interventions to enhance current and future training solutions. This is to ensure that the UK armed forces are prepared to operate in rapidly changing, complex and demanding operational environments. These anticipated complexities are not easily replicated in the live environment, and the use of live assets has other implications. Therefore, the UK MoD has set out a challenge (Defence Innovation Priorities, ref [1]) to better exploit novel technologies to develop more effective training environments. Rapidly maturing immersive technologies such as Augmented Reality (AR), Mixed Reality (MR) and Virtual Reality (VR) (collectively known as XR) have been identified as a potential tool to enable more effective training.

Recent developments in commercial off the shelf (COTS) XR technologies have made them viable and accessible to the defence training and education community. It is anticipated that these technologies could provide a step change in capability, however science and technology (S&T) is required to demonstrate the effectiveness of XR, so that interested parties within the United Kingdom (UK) Ministry of Defence (MoD) can make evidence based decisions about adopting XR within current and future training solutions.

The MoD have tasked the UK Defence Science and Technology Laboratory (Dstl) to evaluate the benefits of XR technology for Defence training and education as part of a 3-4 year research study. These benefits may include increased training throughput, reduced training time or a more compelling representation of the operational environment. However, these benefits will not be realised until XR technology has been proven to be an effective training tool.

To understand the current effectiveness of XR technology Dstl, in partnership with NSC, BrightHF and CENTRIC (Sheffield Hallam University), have been developing an approach to systematically gather and assess evidence relating to the use of XR for defence Training & Education (T&E).

The work uses a two-pronged approach. Firstly, it creates a Quality Assessment Framework that defines criteria by which to evaluate the quality of existing trials (i.e. studies investigating the effectiveness of XR systems). The purpose of this work is to enable interested stakeholders such as those who design training solutions or work in a capability branch to understand the effectiveness of XR for their particular use case. Secondly, it develops a Trials Planning Framework, which provides practical guidance on how to conduct studies that produce high-quality evidence as defined in the former's quality criteria. These frameworks will serve to strengthen the evidence base as well as develop the ability of end users to generate their own evidence which will be used to support their business case as well as develop an intelligent customer capability within MoD.

The anticipated outputs and recommendations from this work will be used to advise on the effectiveness of XR solutions, which will allow MoD to maximise the impact of immersive technology on training and education.

1.1 FRAMEWORKS

The Quality Assessment Framework delivers a bespoke research asset review protocol combining best practice methodologies and quality criteria from existing industry frameworks with technology factors specific to the use of XR. An asset in this context is a report or published paper documenting the use of XR in a training and education setting and draws conclusions on the suitability of the technology. The Trials Planning Framework delivers a modular and adaptive framework designed to support the evaluation of disparate training and assessment goals relating to new XR technologies.

Together these frameworks support an XR for Training and Education (XR4TE) Portfolio of evidence (PoE),

a key enabler of the exploitation of XR technologies in defence.

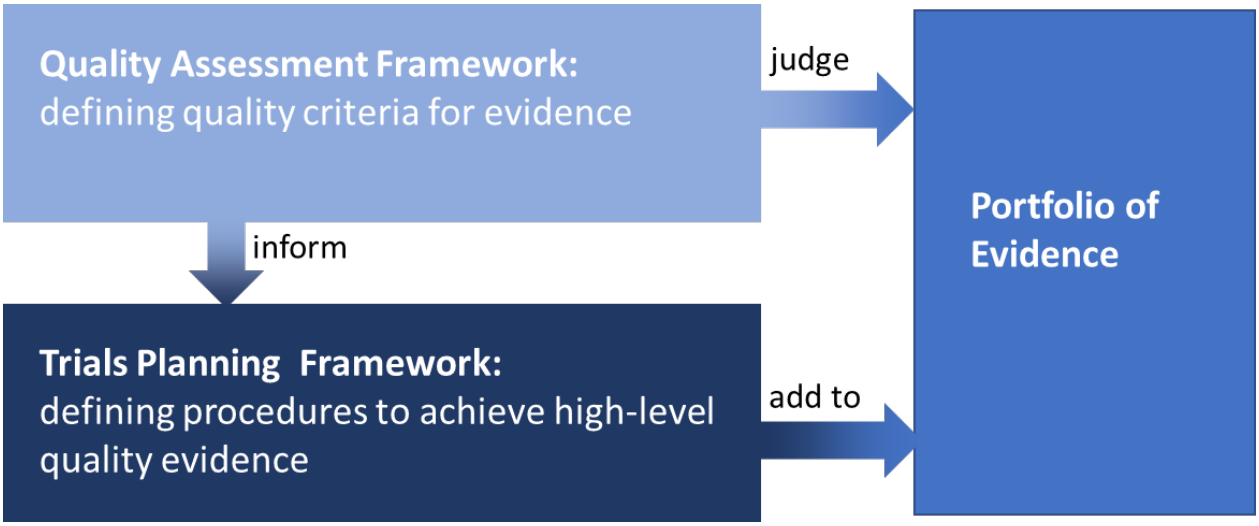


Figure 1-1: Frameworks as they support expansion and use of the XR4TE POE

The Quality Assessment protocol will be applied when entering assets into the XR4TE PoE ensuring the PoE output is usable and meaningful to stakeholders. The protocol will also feed into the Trials Planning Framework which will provide a gold-standard for the evaluation of AR/VR/XR capabilities within UK MoD.

2.0 THE QUALITY ASSESSMENT FRAMEWORK

When identifying and providing evidence of the effectiveness of XR4TE one key question to address is '*What counts as good evidence?*'.

2.1 Why a bespoke framework?

In early 2020 a review was conducted to identify existing quality assessment frameworks from across a range of industries. It was found that no one quality assessment framework or standard nomenclature has been universally endorsed by all industries to assess study quality.

The number of quality criteria used ranged from 4-28 and the type and the origin of quality criteria was highly variable. Variations were found in terms of: what frameworks aim to assess, with some appraising the quality of studies in terms of specific study designs; applicability to varying contexts, with fewer applicable across multiple disciplines; how they measure quality, with a variety of checklists, hierarchies, ready reckoners, toolkits available; the number, type and origin of quality criteria used and, the type of assessment and scoring types used. It was decided that adapting one of these existing frameworks to evaluate XR4TE evidence would lack validity or incompletely cover important study quality elements.

Therefore, a bespoke Quality Assessment Framework is being developed suitable for determining the intrinsic quality of evidence for XR4TE. This aims to determine the extent to which evidence entered is accurate, relevant, and reliable to promote greater precision, transparency and evaluation of research evidence. The bespoke Quality Assessment Framework will: draw on best practice in terms of criteria, scoring methods and overall quality classification methods; be guided by the context of XR4TE research standards, and be co-created with stakeholders and subject matter experts. The aim is to ensure that it is

quick and simple to apply, easy to understand and meaningful and useful to the full range of end users.

2.2 Overview of Quality Assessment Framework elements

The key elements of the Quality Assessment Framework are the criteria, scoring and presentation of results to the end user.

2.2.1 Criteria

The quality criteria are grouped into three subjects: data collection, analysis and reporting each with supporting quality considerations (see Table 2-1).

Table 2-1: Proposed Quality Criteria

Quality	Quality Considerations
Data collection	Robust: Does the study use a research method well suited to the research question? Suitable measures: Does what is being measured reflect what is being studied?
Analysis	Rigour: Is the analysis process transparent and well described?
Reporting	Context and purpose: Are the research aims clearly stated? Does the study acknowledge existing research or theory? Coherent: Is there a clear statement of the findings? Are the findings and conclusions supported by evidence? Does the author consider the study's limitations or offer alternative interpretations of the analysis?

2.2.2 Scoring and quality classification

The research paper reviewer will rate each aspect using a 7-point Likert Scale, Likert, R. (1932) ref [2]. The output is then translated into the overall quality classification of:

- **Strong:** comprehensively addresses multiple quality criteria.
- **Moderate**: some deficiencies in attention to quality criteria.
- **Weak**: major deficiencies in attention to quality criteria.

Ideally, each research paper would be reviewed by two researchers and any discrepancies resolved through a consensus discussion to align the quality scores and classification.

2.2.3 Presentation of the quality assessment

The quality classification would then be presented to the end user along with other complementary assessment criteria to provide a multi-dimensional picture (see Figure 2-1). This could include a score of journal prestige using an online source such as Scimago, ref [3], an indicator of the effectiveness of the XR4TE (i.e. lacking, approaching, equivalent to, exceeding) compared to traditional methods of training and education, and an Evidence Readiness Level (ERL).

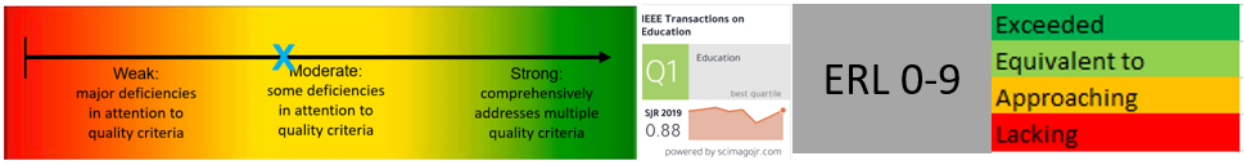


Figure 2-1: Proposed Summary Presentation of Assessments

ERLs (see Table 2-2) developed and used on previous training research projects, ref [4], can provide some additional background context as to the nature of the type and scale of the study covered in the research paper. Note that ERLs 0-4 may not be rated for quality due to the nature of the type of research paper they typically involve (e.g. blog articles, literature reviews etc.).

Table 2-2: Evidence Readiness Levels (ERL)

ERL	Definition
0	Opinion expressed in an opinion document, 'think piece' or during team or customer meeting/symposium.
1	Basic principles observed and reported. Typically, a Secondary Evidence review (e.g. literature review).
2	Concept and/or method formulated. Typically, extended secondary evidence or State of the Art Review (SOAR).
3	Analytical and experimental critical function and/or characteristic proof-of-concept. Typically, secondary evidence further theorised or Lit Review/SOAR with critical analysis.
4	Basic validation in a controlled environment. Typically, primary evidence methodology based on secondary or Lit review/SOAR with Critical Analysis and "So what" thinking.
5	Experimental validation in a relevant environment. Typically, primary evidence – small base.
6	Experimental prototype demonstration in a relevant environment. Typically, primary evidence – medium base. Current practice in a working environment and demonstrated widely.
7	Experimental prototype demonstration in an operational environment. Typically, primary evidence – large base.
8	Principles Concept and Method justified and qualified through Peer review or Short-Term Study. Short term justified primary evidence.
9	Principles Concept and Method qualified through Longitudinal Study or successful exploitation. Long term justified primary evidence.

2.3 Next steps

To complete the initial version of the framework the scoring method will be confirmed to ensure that it enables sufficient discrimination between scoring classifications. This will be done by testing the criteria, scoring and overall classification methods on sample research papers. Finally, the proposed method of

presenting the outputs will be reviewed by end users and subject matter experts.

It is expected that this framework will continue to be adapted as understanding of the technology and stakeholder requirements matures.

3.0 TRIAL PLANNING FRAMEWORK

The Trial Planning Framework (TPF) is created as a modular and thus flexible and adaptive framework to accommodate the evaluation of disparate training and assessment goals in new XR technologies. TPF provides two elements:

- Templates for the most frequent trial designs in the form of stand-alone protocols
- Guidance on scientific and pragmatic considerations that support the execution of the trial protocols to ensure the creation of high-quality evidence, including advice on ethics and mitigation of common risks

Together these two elements aim to ensure that XR evaluations result in evidence that is academically sound as well as practically useful. In effect, TPF is a guide on how to conduct trials that fulfil the quality requirements developed in the Quality Assessment Framework (see section 2.0).

Intended users are practitioners that plan XR evaluation trials as well as decision-makers that need to evaluate trial proposals on whether they are likely to deliver evidence of sufficient quality.

3.1 Outline of TPF

TPF offers templates for three general evaluation streams which can be employed individually or in combination:

- (1) functionality and usability of the technology and training scenarios
- (2) trainee reactions such as perceived usefulness and satisfaction
- (3) impact assessment to determine whether the training achieves its intended effect(s)

It further addresses XR as a means for training as well as assessment and the comparative assessments of technologies and training approaches. The approach is comprehensive in terms of outcome levels (attitudes, emotions, skills, knowledge, results, behaviours, etc.), focus (individual and groups), time (short- and long-term, accommodating single and repeated measures) as well as disparate methodologies depending on evaluation needs. With this TPF ensures the systematic evidence collection, analysis and presentation under consideration of best-practice research practices and ethics. Figure 3-1 presents the overall structure of the TPF document.

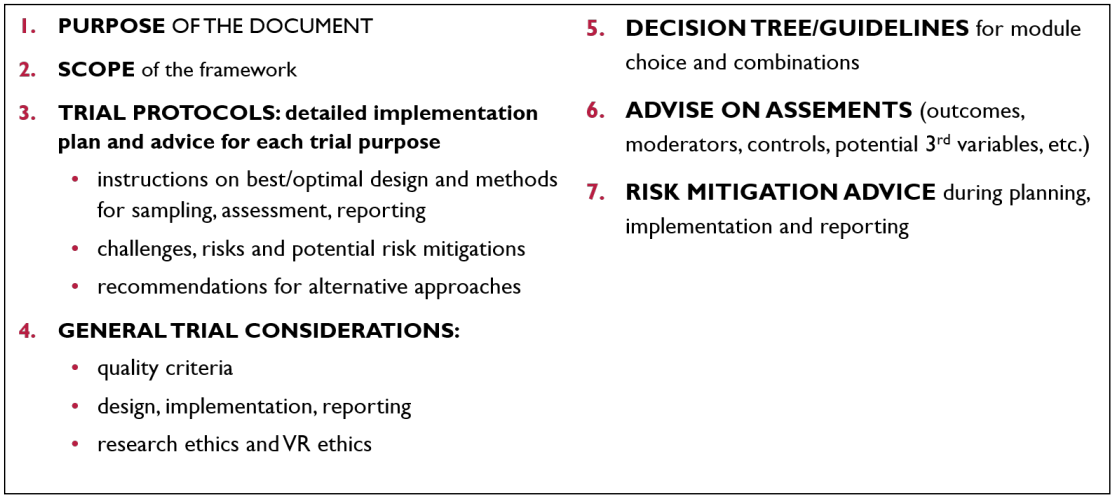


Figure 3-1: Outline of the TPF structure and content.

Below we provide details on some of the Framework elements.

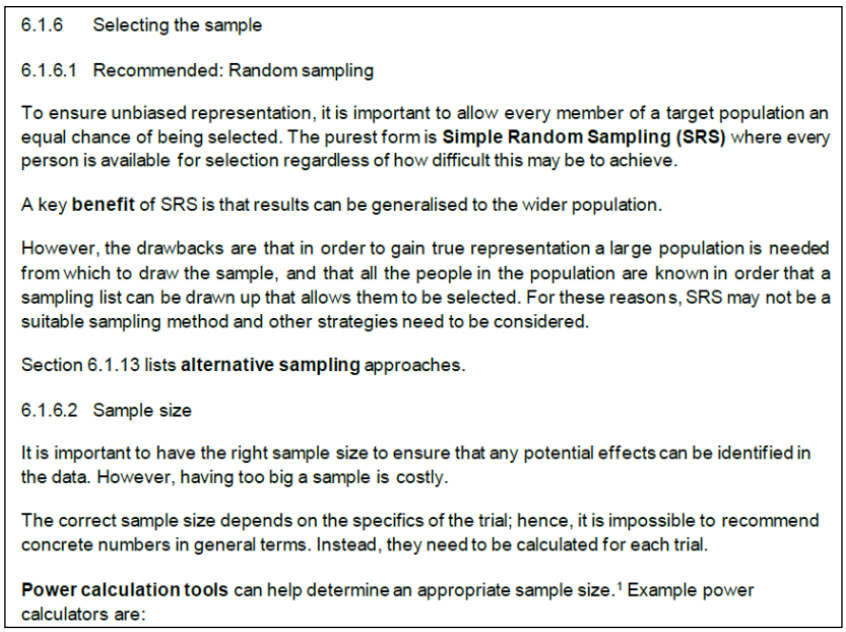
3.1.1 Evaluation purposes covered in the framework

The general aim of XR evaluation trials is to establish the impact(s) of interventions on specific and measurable outcomes (e.g. reduction in errors or improvements in knowledge or skills). More specifically they aim to assess whether the technology is fit-for-purpose and efficient in reaching its intended effects. The first design choice depends on the exact question(s) the trial aims to answer. TPF's approach is comprehensive in that is covers eight disparate evaluation purposes (adopted from Bayerl et al, ref [5]):

- Establish whether the **XR system itself** (including technology, setup and XR scenarios) is functional, usable and efficient given the purpose and target group(s) it was developed for
- Understand whether the **context** in which the XR system is used is fit for purpose
- Capture **immediate user reactions** to understand, e.g., whether users feel comfortable using the XR system, perceive it as useful and are satisfied with the experience
- Conduct an **impact assessment** to understand whether the XR system delivers the intended outcomes short- and/or long-term
- **Comparative evaluation:** comparing different XR systems, setups or usage contexts to understand which one is more effective/efficient
- **Transferability:** test whether an existing (usage of an) XR system can be transferred to other settings (e.g. establishing whether it can be used equally effectively/efficiently in a different task, cultural context or professional discipline)
- **Long-term viability:** test whether an XR system and/or its usage still delivers the expected outcomes or may need adaptation
- **Influencing factors:** investigate whether factors other than the XR system influence its usability, its reception by users or its impact; such factors may either be expected and systematically explored (e.g. differences in the speed of learning a task between people with little or a lot of VR experience) or appear as unintended consequences (e.g. people can react unexpectedly negative to scenarios if they triggers fears or memories of bad experiences).

3.1.2 Trial protocols

The trial protocols outline which steps and procedures need to be undertaken to achieve good quality results. These steps and procedures differ depending on the purpose of each evaluation. Hence, the Framework offers disparate trial protocols for each of the different evaluation purposes. These protocols are kept concise on purpose with the aim to allow readers a quick and efficient understanding of what is involved in conducting a specific type of trial. The protocols follow a standardised structure: description of the evaluation purpose, proposed trial design, key features of the design, advice on sampling and data collection procedures, advice on reporting, ethical considerations, potential limitations and risks and (where applicable) alternative design choices. Part of a trial protocol is shown as an example in Figure 3-2. Following the steps in a protocol should allow a trial to achieve the highest-possible quality evidence in terms of data collection, analysis and reporting, as defined in the Quality Assessment Framework.



6.1.6 Selecting the sample

6.1.6.1 Recommended: Random sampling

To ensure unbiased representation, it is important to allow every member of a target population an equal chance of being selected. The purest form is **Simple Random Sampling (SRS)** where every person is available for selection regardless of how difficult this may be to achieve.

A key **benefit** of SRS is that results can be generalised to the wider population.

However, the drawbacks are that in order to gain true representation a large population is needed from which to draw the sample, and that all the people in the population are known in order that a sampling list can be drawn up that allows them to be selected. For these reasons, SRS may not be a suitable sampling method and other strategies need to be considered.

Section 6.1.13 lists **alternative sampling** approaches.

6.1.6.2 Sample size

It is important to have the right sample size to ensure that any potential effects can be identified in the data. However, having too big a sample is costly.

The correct sample size depends on the specifics of the trial; hence, it is impossible to recommend concrete numbers in general terms. Instead, they need to be calculated for each trial.

Power calculation tools can help determine an appropriate sample size.[1] Example power calculators are:

Figure 3-2: Example of a trial protocol (beginning of Randomized Control Trials).

Trial protocols are provided for five overarching evaluation needs:

- (1) investigating effects of the setting (technology, delivery and/or content) on expected training outcomes
- (2) assessing different types of training outcomes (e.g. attitudes, emotions, skills, knowledge, behaviour)
- (3) assessing influences to training effectiveness (e.g. moderators, mediators, controls/3rd variables)
- (4) location of training effects (e.g. individual, group)
- (5) conducting longitudinal designs

3.1.3 Background of academic and practical guidance

The trial protocols constitute the core of TPF. This is accompanied by a compendium of academic and practical guidance as background information to support the decision-making along the evaluation process. This second part of the TPF details scientific best and common practices for the designing, conducting and reporting of XR evaluation trials, as well as information on trial features that facilitate better training

outcomes (see Figure 3-3 as example). It further offers readers practical advice and recommendations on handling common risks of evaluation trials and research and VR ethics. Pointers to example studies and further reading guide end users to additional material either as illustrations of good practice or additional background information.

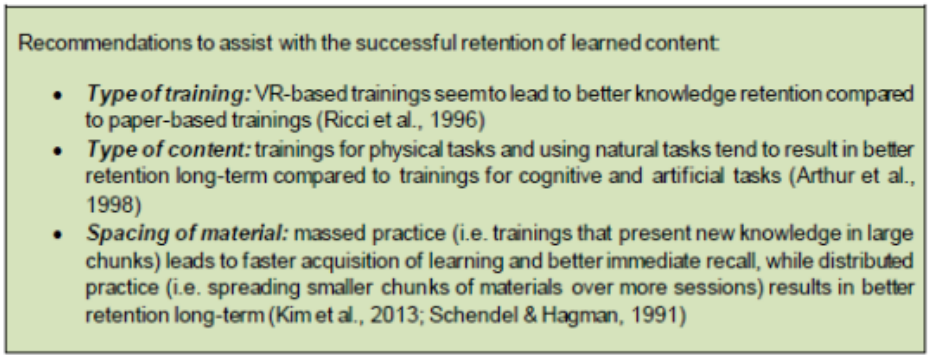


Figure 3-3: Example of practical recommendations in TPF (here: retention of learned content).

3.2 Next steps

In a next step, the TPF is intended to be validated in actual trials. These will confirm whether the TPF acts as intended (i.e. supports the creation of high-quality evidence) and links well into the Quality Assessment Framework (i.e. leads to studies that can feed effectively into the body of works within the PoE).

As with the trial framework, this framework will continue to be adapted as understanding of the technology and stakeholder requirements matures.

4.0 OUTCOMES

The research outlined above aims to develop a common and useable methodology for assessing XR technology for defence T&E. This will help the UK MoD build a body of evidence that supports the acquisition of XR technology at the right time and in the right place, with the ultimate objective of transitioning XR technology from the innovation space to business as usual.

The Quality Assessment Framework and the Trials Planning Framework together form a dual approach that address how MoD generate and assess high quality evidence. The work will develop MoD's understanding of the benefits of XR for T&E and enable decision makers to make informed decisions using technically robust evidence.

The following are recommendations for the validation of the chosen approach and for ensuring its full potential:

- Continue to develop the trials and assessment frameworks together to ensure that the Trials Planning Framework allows users to create high quality evidence,
- Test and refine the criteria, scoring and classification methods on sample research methods,
- Test and refine the Trials Planning Framework using an example trial,
- Consider proposing that other NATO member nations adopt this common approach to ensure that the assessment of XR technology is consistent.

REFERENCES

[1] UK MoD. (2019). Defence Innovation Priorities, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/831427/20190906-InnovationPrioritiesPub_Final_.pdf

[2] Likert, R. (1932). A technique for the measurement of attitudes. Archives of Psychology, 140, 1-55.

[3] SCImago, Scimago Journal & Country Rank, https://www.scimagojr.com/

[4] Cullingford, E. (2019). Augmented, Virtual and Mixed Reality Technologies for Training and Education (XR4TE) – Final Report, QINETIQ/19/00657. QinetiQ Ltd.

[5] Bayerl, P.S., Davey, S., Lohrmann, P., & Saunders, J. (2019). Evaluating serious game trainings. In B. Akhgar (ed.), Serious Games for Enhancing Law Enforcement Agencies. From Virtual Reality to Augmented Reality (p. 149-169). Cham: Springer.