

MMAT – Quality Appraisal Table

Citation (Evaluation sample size)			MMAT Q1	MMAT Q2	MMAT Q3	MMAT Q4	MMAT Q5	Summary of evaluation quality (kurz)	Overall (H/L)
Randomized Controlled Trials									
Citation (Evaluation sample size)	S1 Are there clear research questions?	S2: Do the collected data allow to address the research question?	2.1 Is randomization appropriately performed?	2.2 Are the groups comparable at baseline?	2.3 Are there complete outcome data?	2.4 Are outcome assessors blinded to the intervention provided?	2.5 Did the participants adhere to the assigned intervention?	Summary of evaluation quality based on MMAT	Overall (H/L)
Makransky & Klingenberg, 2022 (28)	Yes	Yes	Can't tell	Yes	Yes	No	Yes	RCT (main study n=28; VR n=15, personal trainer n=13; plus a pilot study with n=86). Random allocation and baseline checks are reported, and outcomes are clearly analysed. Outcome assessment was not blinded to the assigned intervention.	H
Macesker et al., 2023 (39)	Yes	Yes	Can't tell	No	No	Can't tell (probably a minor problem in terms of objective outcomes)	Yes	Randomized experiments (n=39 completers). Procedures and outcomes are described, but baseline details of groups are limited and incomplete outcome data (attrition) is a key limitation.	H
Rauffet et al., 2016 (20)	Yes	Yes	Can't tell	Yes	No	Can't tell	Yes	Randomized simulator study on manual submarine steering (n=20; analyzed n=16) comparing two control–display configurations. Measures (performance, NASA-TLX, pupil diameter) are appropriate, but eye-tracking data loss reduces outcome completeness, blinding is unclear, and the student sample/short training limits representativeness.	H
Cui et al., 2021 (108)	Yes	Yes	Can't tell	Can't tell	Can't tell	No	Yes	Large lab experiment with randomized parallel time-pressure groups (54 teams/108 students) and within-subject scenario complexity; measures and procedure are strong, but randomization/baseline checks and the impact of physiological-data exclusions are insufficiently reported, and blinding is absent.	H
Non-Randomized Studies									
Citation (Evaluation sample size)	S1 Are there clear research questions?	S2: Do the collected data allow to address the research question?	3.1 Are the participants representative of the target population?	3.2 Are measurements appropriate regarding both the outcome and intervention (or exposure)?	3.3 Are there complete outcome data?	3.4 Are the confounders accounted for in the design and analysis?	3.5 During the study period, is the intervention administered (or exposure occurred) as intended?	Summary of evaluation quality based on MMAT	Overall (H/L)
Takaseki & Okazaki, 2015 (1)	Yes	No	No	No	Can't tell	No	No	Single-participant evaluation (n=1). Multiple workload indicators collected, but measurement validity is limited (EEG device constraints) and no control of confounding; intervention delivery was disrupted (AR shutdown).	H
Wu et al., 2015 (11)	Yes	Yes	Can't tell	Yes	Can't tell	No	Yes	Small simulator experiment (n=11). Appropriate measurement instruments and clear implementation, but unclear confounder control.	H
Wu et al., 2017 (10)	Yes	Yes	Yes	Yes	Yes	No	Yes	Simulator experiment (n=11 recruited; 10 usable datasets) with experienced trainees. Measures are well described, but some outcome data are missing or excluded, and the control of confounders is unclear.	H
Boulhic et al., 2018 (39)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Controlled HMI experiment (novices n=20; “experts” n=19) comparing two color codes. The measures and statistics are clear; experts are students with experience. Measures were also included to account for bias.	L
Orlandi & Brooks, 2018 (10)	Yes	Yes	Yes	Yes	Yes	Yes	No	Within-subject simulator experiment in n=10 professional marine pilots (4 berthings, randomized order). Measures are strong, but some missing outcome data from crashes (mainly at higher difficulty).	H
Sugimoto & Murai, 2018 (6)	Yes	Yes	No	Yes	Can't tell	No	Yes	Very small simulator training study (5 trainees + 1 professional) using HRV/LF-HF monitoring before/after instruction; intervention is clear, but representativeness, confounding control and outcome completeness are not given or unclear.	H
Van Der Kleij et al., 2018 (22)	Yes	Yes	No	Yes	Yes	Yes	Yes	Simulator within-subject experiment (n=22) on change-detection support and interruptions; measures and control are strong, but the student sample limits representativeness.	H
Deng et al., 2019 (14)	Yes	Yes	No	Can't tell	No	Can't tell	Yes	Lab experiment (n=20; analyzed n=14). Mental workload is inferred from performance (error rate/response time). Task measures are clear, but representativeness is low and missing data/order effects are not controlled.	H
Okazaki et al., 2019 (20)	Yes	Yes	Yes	Yes	Can't tell	Yes	Yes	Simulator experiment (n=20 students). Delay conditions were randomly varied with appropriate analyses. However, the sample comprises experienced students and is not representative of active ship operators.	L
Nishizaki et al., 2020 (16)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Within-subject simulator study (n=16). Clear fatigue manipulation with randomized condition order and complete data and participants are representative but mainly cadets/students. Covariates are controlled through the within-subject design.	L
Žagar et al., 2020 (5)	Yes	Yes	Can't tell	Yes	No	No	Yes	Pilot simulator study (n=5). Multi-sensor MW measures, but very small sample and some measures were excluded and limited control for learning/order effects	H
Kari et al., 2022 (3)	Yes	Yes	No	Yes	Yes	No	Yes	Very small non-randomized within-subject simulator study (n=3) using EEG and NASA-TLX to contrast low vs high workload; measures are well described, but representativeness is low and fixed-order effects/confounding are likely.	H
Yan et al., 2022a (22)	Yes	Yes	No	Yes	Yes	No	Yes	Lab experiment with students (n=22). Measures and protocol are described, but the sample is non-representative and confounding (training/learning) is not well addressed.	H

Özsever & Tavacıoğlu, 2022 (12)	Yes	Yes	Yes	Yes	Yes	No	Yes	Small bridge-simulator study with licensed officers (n=12). Appropriate measures, but sequential scenarios may introduce order effects.	H
Lyu et al., 2023 (6)	Yes	Yes	No	Yes	Yes	No	Yes	Small simulator study (n=6 students) measuring ECG/SC + NASA-TLX across day/night emergency scenarios; measures are clear, but representativeness is low and fixed-order/learning effects are not analytically controlled.	H
Ronca et al., 2023 (26)	Yes	Yes	Yes	Yes	No	Can't tell	Yes	Full-mission bridge simulator quasi-experiment (13 vs 13 mariners) comparing operations without vs with a Lookout (HURID); measures and analyses are detailed, but exclusions from signal issues and lack of randomization/adjustment raise confounding and missing-outcome bias concerns.	H
Xia et al., 2023 (30)	Yes	Yes	No	Yes	Can't tell	Can't tell	Yes	VR case study (n=30) comparing joystick vs body-motion/gesture teleoperation with NASA-TLX and trajectory-performance metrics; outcomes are appropriate, but the novice convenience sample and unclear counterbalancing/learning control increase confounding and limit generalizability.	H
Laera et al., 2024 (44)	Yes	Yes	Can't tell	Yes	Yes	Yes	Yes	User study (n=44 boat-show visitors) comparing boat- vs head-stabilized AR with objective performance and RTLX/SUS/UEQ; outcomes are suitable, but representativeness remains unclear.	L
Laera et al., 2023 (45)	Yes	Yes	Can't tell	Yes	Yes	Yes	Yes	VR sailing user study (n=45) comparing AR interface vs traditional display. Outcomes include reaction time and validated questionnaires and order is controlled, but generalizability is uncertain.	L
Van De Merwe et al., 2024 (34)	Yes	Yes	Yes	Yes	No	Yes	Yes	Controlled experiment with nautical navigators (n=34) varying transparency and complexity in traffic situations. Measures, sample and analyses are appropriate, but missing data due to outlier and technical problems.	H
Quantitative Descriptive Studies									
Citation (Evaluation sample size)	S1 Are there clear research questions?	S2: Do the collected data allow to address the research question?	4.1 Is the sampling strategy relevant to address the research question?	4.2 Is the sample representative of the target population?	4.3 Are the measurements appropriate?	4.4 Is the risk of nonresponse bias low?	4.5 Is statistical analysis appropriate to answer the research question?	Summary of evaluation quality based on MMAT	Overall (H/L)
Akrimah et al., 2023 (30)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Small cross-sectional survey (n=30) of workers in a single terminal office. Measures and analyses are clearly described, and the sample represents the full workforce of this specific terminal office.	L
Bandono, 2019 (82)	Yes	Yes	Yes	Yes	Yes	Can't tell	Yes	Cross-sectional survey of warship personnel (n=82) using NASA-TLX (plus fuzzy aggregation) to compare mental workload across job types. Measures are appropriate, but the sampling representativeness and response rate are insufficiently reported, limiting inference beyond the surveyed respondents.	L
Kartoglu et al., 2023 (38)	Yes	Yes	Yes	Yes	Yes	Can't tell	Yes	Cross-sectional survey of high-speed craft navigators (n=38) using RNASA-TLX and ranking tasks. Aims and analyses are clearly described; however, nonresponse handling are insufficiently documented.	L
Ma et al., 2023 (987)	Yes	Yes	Yes	Yes	Can't tell	No	Yes	Large cross-sectional questionnaire study (n=987) using logistic regression. Adequate statistical power, but limited to inland river seafarers in the Yangtze Basin and affected by a high exclusion/dropout rate. In addition, the cognitive load classification is based on self-developed "cognition and stress adaptability" questions with a three-level outcome, and its construct validity is not clearly established.	H
Murai, Hayashi, et al., 2015 (8)	Yes	Can't tell	Yes	No	Yes	Can't tell	Can't tell	Very small convenience sample (n=8). As indicated at screening (S2 = Can't tell), the suitability of the data to fully address the research question is uncertain and findings should be interpreted with caution. The protocol and objective measures are clearly described, but results are mainly descriptive and illustrative, without statistical inference. The study represents an initial validation effort focused solely on the patch device, without comparison to an external reference or control condition.	interpret with caution
Murai, Kitamura, et al., 2015 (7)	Yes	Can't tell	Yes	No	Can't tell	Can't tell	Can't tell	Small field study (7 coordinators; 8 sessions). As indicated at screening (S2 = Can't tell), the suitability of the data to fully address the research question is uncertain and findings should be interpreted with caution. The study demonstrates feasible real-world measurement, but representativeness is limited. Analyses are mainly visual/descriptive (time series inspection and event assignment) without clear inferential statistical testing. Moreover, the evaluation approach does not fully align with the stated validation objective: the focus is on validating facial temperature as a measure, rather than deriving substantive conclusions.	interpret with caution
Kitamura, Murai, & Wakida, 2016 (5)	Yes	Can't tell	Yes	No	Can't tell (The robustness of NO3- measurement remains limited)	No	Can't tell	Very small field study (n=5). As indicated at screening (S2 = Can't tell), the suitability of the data to fully address the research question is uncertain. Measurement problems occurred for participant B, who was described as having a central role in steering the vessel, resulting in missing saliva samples and exclusion of relevant data. Objective measures are described, but key indicators such as salivary NO3- are not clearly validated as workload markers, and analyses are predominantly descriptive.	interpret with caution
Kitamura, Murai, Furusho, et al., 2016 (5)	Yes	Can't tell	Yes	No	Can't tell (The robustness of NO3- measurement remains limited)	No	No	Very small field feasibility study (n=5; pilots/captain and port coordinators). As indicated at screening (S2 = Can't tell), the suitability of the data to address the stated validation objective is uncertain, as the analysis does not adequately support the intended validation of salivary measures against HRV and remains largely descriptive. Measures are described, but the sample is highly limited and saliva sampling shows substantial missing data. Analyses are primarily descriptive and event-based, with minimal statistical support.	interpret with caution

Murai et al., 2016 (7)	Yes	Can't tell	Yes	No	Yes	Can't tell	No	Very small field feasibility study (n=7; analysis primarily focuses on one senior coordinator). As indicated at screening (S2 = Can't tell), the suitability of the data to address the stated validation objective is uncertain, as the analysis does not adequately align with the intended validation approach. Measures are described, but sampling, representativeness, and nonresponse are insufficiently documented, and the analysis is largely descriptive with minimal statistical control.	interpret with caution
Okazaki et al., 2016 (4)	Yes	Can't tell	Yes	No	Can't tell	Can't tell	No	Small simulator experiment (n=4). As indicated at screening (S2 = Can't tell), the suitability of the data to address the intended comparison is uncertain, as the analyses do not adequately support the stated comparative objective. The modified SAGAT and nasal temperature measures are described, but the validation of nasal temperature as a workload indicator is insufficiently established. The sample is highly limited, representativeness and nonresponse are unclear, and conclusions rely predominantly on descriptive tables, plots, and single-case interpretations rather than robust inferential testing.	interpret with caution
Sugimoto et al., 2016 (10)	Yes	Can't tell	Yes	No	Yes	Can't tell	Can't tell	Small simulator study (n=10) comparing operator and bridge roles using HRV. As indicated at screening (S2 = Can't tell), the suitability of the data to adequately address the broadly formulated research question is uncertain, particularly as only one operator was included, limiting meaningful role comparisons. Measures are described, but representativeness is restricted and analyses are largely descriptive. The interpretation of peaks, including whether the instructor shows similar patterns at the same events as the group, is based primarily on visual/event-based inspection rather than systematic statistical testing.	interpret with caution
Kamioka et al., 2017 (6)	Yes	Yes	Yes	No	Can't tell	No	Yes (, but limited)	Small field method study (n=6) building a duties list from salivary NO3 ⁻ peaks; sampling is limited and some events are missing/unclassifiable; descriptive classification ist limited but fits the aim.	H
Kitamura et al., 2017 (6)	Yes	Can't tell	Yes	No	Can't tell	Can't tell	No	Small simulator feasibility study (n=6) using salivary NO ₃ ⁻ and HRV. As indicated at screening (S2 = Can't tell), the suitability of the data to address the stated validation objective is uncertain, as the available data and analytical approach do not fully support a rigorous validation of the method. Measures are described, but the sample is limited and the evidence is predominantly descriptive and event-based.	interpret with caution
Liu, Sourina, et al., 2017 (12)	Yes	Yes	Yes	No	Can't tell	Can't tell	Yes, but limited	Small simulator feasibility study (n=12 cadets; results mainly illustrative for one trainee). Measures are described, but sampling is limited and analysis is largely descriptive.	H
Liu, Subramaniam, et al., 2017 (4)	Yes	Yes	Yes	No	Can't tell	Can't tell	Yes, but limited	Small simulator case study (n=4 trainees) using EEG-based workload and stress recognition. Data collection and statistical procedures are adequately described, but the sample is highly limited and the evidence remains largely descriptive, consistent with the exploratory case study design. It also remains unclear to what extent the EEG measures validly reflect mental workload.	H
Murai et al., 2017 (8)	Yes	Can't tell	Yes	No	Yes	Can't tell	No	Small simulator feasibility study (n=8) using HRV monitoring (LF/HF) during bridge exercises. As indicated at screening (S2 = Can't tell), the suitability of the data to adequately address the research question is uncertain, as the results are presented only briefly and lack sufficient analytical depth. Measures are described, but the sample is limited, nonresponse is unclear, and no formal validation of the HRV approach is provided.	interpret with caution
Nisizaki et al., 2017 (7)	Yes	Can't tell	Yes	No	Yes	Can't tell	No	Small simulator study (two experiments; very small total n=7 across scenarios). As indicated at screening (S2 = Can't tell), the suitability of the data to address the research question is uncertain, as multiple procedural changes across experiments result in no clear or consistent manipulation. Measures are described (SAGAT, nasal temperature, R-R/LF-HF), but representativeness and nonresponse are unclear, and the evidence linking evaluation outcomes to mental workload is largely descriptive.	interpret with caution
Lim et al., 2018 (7)	Yes	Yes	Yes	No	Can't tell	Can't tell	Yes, but limited	Pilot bridge-simulator study (n=7) using EEG-based workload/stress monitoring; methods are described, but the sample is very small and measurement validity/missing data reporting are limited.	H
Liu et al., 2018 (18)	Yes	Yes	Yes	Can't tell	Can't tell	Can't tell	Yes, but limited	Pilot simulator assessment study (n=18 trainees; 4 exercises each) using EEG-based workload/stress recognition; methods are described, but representativeness, missing data and criterion validity of the pass/retrain/fail decision are unclear.	H
Murai et al., 2018 (4)	Yes	Can't tell	Yes	No	Yes	Can't tell	No	Very small simulator feasibility/case study (n=4; repeated trials) using HRV (LF/HF) for real-time workload support. As indicated at screening (S2 = Can't tell), it remains unclear whether the collected data adequately address the intended research objective, as the study appears to demonstrate feasibility rather than provide a rigorous evaluation or validation of the training support approach. Measurements are described and collected simultaneously with the training evaluation, but the sample is very small, representativeness is low, and the evidence is largely descriptive.	interpret with caution
Kitamura et al., 2020 (15)	Yes	Yes	Yes	No	Yes	Can't tell	No	Small simulator method study (n=15 students) comparing two NO ₃ ⁻ -ISFET versions. The measurement setup is described in detail, but the student sample limits generalizability, and evidence for device agreement or differences is largely descriptive, with an exploratory analysis.	H
Toba & Murai, 2020 (3)	Yes	Can't tell	Yes	No	Can't tell	Can't tell	No	Very small simulator feasibility study (2 captains and 1 operator) using HRV (LF/HF). As indicated at screening (S2 = Can't tell), the suitability of the data to adequately address the research objective is uncertain, as not all collected data are systematically analysed and comparisons are limited to selected individual experiments. Measures are described, but representativeness is very limited.	interpret with caution
Wilson et al., 2021 (42)	Yes	Yes	Yes	Can't tell	Yes	No	Yes	Large real-world intensive longitudinal study in submariners (initial n=77; analyzed n=42) with validated fatigue/workload measures and appropriate multilevel modelling; however, substantial exclusions and dropout raise nonresponse/representativeness concerns.	H
Murai et al., 2021 (7)	Yes	Can't tell	Yes	No	Can't tell	Can't tell	No	Small simulator study (n=7; 3 masters and 4 cadets) demonstrating real-time HRV (LF/HF) monitoring for training support. As indicated at screening (S2 = Can't tell), the suitability of the data to adequately evaluate the system is uncertain, as the collected and analysed data do not allow a proper validation of the measurement approach. The concept and measurement are described, but the evidence is mainly descriptive, with low generalizability and limited validation or statistical evaluation.	interpret with caution

Yoshida et al., 2021 (10)	Yes	Yes	Yes	No	Yes	Can't tell	No	Small study with deck officers (n=10) using NASA-TLX within a clearly described protocol. However, representativeness is limited and the sample size is very small for the statistical analyses conducted.	H
Crestelo Moreno et al., 2023 (23)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Field study in two MRCCs with high local participation (23/28 VTSOs) using validated fatigue and workload scales plus an objective reaction-time task. Nonparticipation was mainly due to staff not being on duty. Analyses were appropriate and aligned with the research questions.	L
Mixed Methods Studies									
Citation (Evaluation sample size)	S1 Are there clear research questions?	S2: Do the collected data allow to address the research question?	5.1 Is there an adequate rationale for using a mixed methods design to address the research question?	5.2 Are the different components of the study effectively integrated to answer the research question?	5.3 Are the outputs of the integration of qualitative and quantitative components adequately interpreted?	5.4 Are divergences and inconsistencies between quantitative and qualitative results adequately addressed?	5.5 Do the different components of the study adhere to the quality criteria of each tradition of the methods involved?	Summary of evaluation quality based on MMAT	Overall (H/L)
Jensen & Oldenburg, 2021 (198)	Yes	Yes	Yes	Yes	Yes	Can't tell	Can't tell	Mixed-methods shipboard study (SMW-S workplace screening on 11 ships; interviews n=198). Rationale is clear, divergences are not explicitly handled, and quality criteria (incl. representativeness/non-response) are unclear.	L

Summary:

L

H

interpret with caution

10

28

13