# Certificates as proof of competence

## Where certification demonstrably works, where it was measured and had no effect, where nobody ever measured

Thomas Arends

3 November 2026

**As of: 3 November 2026.** The basis is five research notes (Agile/PM, driving/aviation/truck licences, IT/organisational certificates, licensing economics/psychometrics, medicine/boards). Every figure carries its evidence grade in parentheses (A official/register · B independent research with methodology · C association/agency measurement series with open methodology · D survey/self-report · E provider/association/consultant statement in its own interest) and its data period; "old" marks data older than about three years. Core statements rest only on A–C. D/E appear exclusively as marked indications with the interest named.

**Summary.** Across all fields examined there are exactly three constellations in which a personal certification has demonstrably improved an outcome measure: first, the introduction of a minimum threshold in a hazard-prone field with no prior standard (US midwife licensing 1900–1940: maternal mortality −6 to −7 %, B, old); second, licensing systems whose effective component is supervised practice and access restriction, not instruction (Graduated Driver Licensing, accompanied driving from 17 (BF17), Swedish practice age 16: −15 to −26 % crashes in the first year, A/B, old); third, examinations whose score, as a continuous measure, predicts a construct-proximate practice criterion (Canadian and US licensing examinations for physicians: RR 1.38–1.51 per 2 SD for upheld complaints, OR ≈ 0.75 per SD for disciplinary action, B, data 1993–2017). For the entire class of management, Agile, IT and security certificates (CSM, PSM, SAFe, PMP, PRINCE2, ITIL, ISTQB, IREB CPRE, CISSP, CCNA, cloud certificates), according to the notes **not a single independent study** exists that measures the certificate against an objective outcome measure; what is documented are only wage premiums, recruiter preferences and self-reports. Where recertification or mandatory continuing education has been measured with a strong design (ABIM MOC: quasi-experiment and RCT; UK Revalidation; EU periodic training for professional drivers (Directive 2003/59/EC)), the result is null. Organisational certificates (ISO 9001, ISO 27001, SOC 2) show selection rather than effect, and in the only field with a hard outcome measure and a comparison system (medical devices EU vs. FDA) commercial third-party certification performs worse (HR 2.9 for recalls/warnings, B, data 2005–2016). Country differences concern almost exclusively market structure and licensing density (Germany 33 % licensed employees, highest EU value, B/D, 2015), not measured outcomes: no DACH certificate has been evaluated on outcome.

---

## Three mechanisms carry effect — and none of them is "the course"

The positive findings share one property: the effective component is never the instruction or the certificate itself, but a threshold, an exposure or a measurement instrument.

**Minimum threshold in a hazard-prone field.** The midwife licensing laws of the US states 1900–1940 reduced maternal mortality by **6–7 %** (B, working paper, primary sources 1900–1940, old) (NBER WP 22456). The licensing economics note classifies this as one of the few positive quality findings in the entire US literature and names the condition: high immediate hazard potential *and* no prior minimum standard. The same logic is supported by the teacher finding of Larsen et al.: stricter certification requirements significantly reduce the number of secondary teachers from less selective colleges (left tail of the distribution) without measurably changing the number of highly qualified teachers (B, working paper, data 1991–2007, old) (NBER WP 28158). Certification works, if at all, as a floor — not as a predictor in the middle of the distribution.

**Supervised practice and access restriction.** The Cochrane review on Graduated Driver Licensing (GDL) (34 studies, 21 programmes, search up to 2009, old) finds for 16-year-olds in the first year after introduction **total crashes median −15.5 %** (range −27 to −8 %) and **injury crashes median −21 %** (B) (Cochrane CD003300). The FARS analysis of all 50 US states 1986–2007 (old) shows for strong GDL **−26 % fatal crashes among 16-year-olds, but +12 % among 18-year-olds** (B) (UNC/JAMA 2011) — part of the effect is displacement. The IIHS evaluation 1993–2003 (C, insurer-funded, old) shows the mechanism: per 16-year-old population −26 %, **per licensed 16-year-old practically unchanged (73 vs. 74 per 100,000)** (Insurance Journal 2005). The effect runs through fewer and later licences, not through better drivers. For Germany, BASt (Federal Highway Research Institute) report M 218 on accompanied driving from 17 (BF17) (approx. 19,000 respondents plus 75,000 silent register cases, project 2007–2011, old) documents in the first year of solo driving **−17 % crash involvement confounder-adjusted, −22 % mileage-adjusted** (A/B; self-selection acknowledged, control only via confounders) (BASt M218). Sweden lowered the practice age to 16 in 1993 while keeping the licensing age at 18: novice drivers overall **approx. −15 % crashes per 10 million km, users of early practice approx. −40 %, non-users no effect** (B, old) (PubMed 10576673); after confounder adjustment 24–39 % remain, early starters practised 118 instead of 41–47 hours (B/C, old) (DfT paper Gregersen). Norway likewise lowered the age to 16 in 1994, the amount of practice barely rose, and crash risk did not change significantly (B/D, self-report, old) (TØI 498/2000). The contrast group is unambiguous: where only instruction varies, the effect disappears. The three RCTs on formal driver education (n = 17,965, 1982–1984, old) yield **RR 1.01 / 1.03 / 1.10** for crash involvement and RR 1.04 for earlier licensing (B) (Cochrane CD003201). Skid training increased crash rates in all six meta-analysed programmes, **passenger cars approx. +12 %, ambulances approx. +45 %** (B meta-analysis, here via secondary source, hence C, old) (International Fleet World).

**Criterion-valid exam scores.** The Canadian cohort of Tamblyn et al. (3,424 physicians, Clinical Skills Examination 1993–1996, follow-up to 2005, old) shows per 2 SD lower communication score **RR 1.38 (95 % CI 1.18–1.61)** for patient complaints upheld by the regulator and per 2 SD lower clinical decision-making score **RR 1.51 (1.25–1.84)**; the lowest communication quartile accounts for 10.0 % of all upheld complaints (B, register data of the

regulatory authorities; funded by the examination provider MCC, COI noted) (JAMA 2007 via McGill). The peer-review subsample (208 physicians, Ontario) yields OR 3.41 for the lowest knowledge quartile and OR 4.24 for the lowest skills quartile for unacceptable practice assessment (B, COI MCC, wide CIs, old) (Wenghofer 2009). In the US, 164,725 graduates 1994–2006 (old) show **OR 0.75 per SD Step 2 CK** for disciplinary action (E, authors at NBME/FSMB, but methodologically transparent) (Cuddy 2017) and 275,392 physicians with Step 3 2000–2017 **−23 % odds per SD** (E, examination-provider environment) (Academic Medicine 2022). The Alberta cohort 2014–2017 confirms +60 % odds per ≈ 1 SD lower Part I score (B, authors partly MCC) (JMR 2024). The decisive point is the reading given in the medicine note: this supports the **validity of the measurement instrument**, not the effectiveness of certification as an intervention. Behavioural traits are stronger predictors here than cognitive scores — unprofessional behaviour during medical school OR 3.0, irresponsibility OR 8.5, versus MCAT OR 0.6 and Step 1 OR 0.9 (B, case-control, discipline 1990–2003, old) (Papadakis 2005).

The same logic applies in aviation: the NTSB case-control study on weather accidents (72 accidents vs. 135 control flights, 2003–2004, old) finds the **instrument rating a significant protective factor (p = 0.002)**, high failure rates in FAA tests linked to accident involvement, and **total flight hours not significant (p = 0.788)** (A/B) (NTSB 2005). The content of the qualification predicts; the number of hours does not.

## Management, Agile and IT certificates: zero outcome studies, many salary figures

Here the absence of evidence is itself the finding. The IEEE Software review of 2023 (24 primary studies 2004–2021) states: 68 % of the methods are surveys, 4 % field research, one study used quantitative project metrics; five studies (5,206 cases) report positive effects, six (3,304 cases) no relationship; Agile certificates are “barely examined” (B) (Luja Beraza et al. 2023). The only field study with project files (Strang 2021, **2,692 public procurement projects**, North America, data period not documented) finds certified PMs more successful, but purely correlationally with 67.3 % classification accuracy and an effect size of 12 % (B) (Strang 2021); selection (better organisations hire certified people) is not ruled out. Against this stand Farashah et al. (n = 452, data approx. 2014, old): **no direct relationship certificate → performance**, the effect runs entirely through self-efficacy (B, self-report) (Athabasca), Catanio et al. (n = 93 USA, old): certified PMs do not manage scope, time and cost any better (B/D) (IGI Global), and Joseph & Marnewick (South Africa, old): IT project performance “not influenced” (B) (RePEc).

For Scrum and SAFe the variable does not even exist. The largest datasets — Verwijs & Russo with 15,078 respondents and 4,013 teams (2021–2023) — do not collect the certificate as a control variable; the scaling approach explains **1–2 % of the variance** in team effectiveness, and after controlling for Agile experience almost all effects vanish (SAFe $\beta = -0.042$) (B/D, preprint, COI Verwijs/The Liberators) (arXiv 2310.06599). The Danish SAFe case study (PFA, interviews 2018, old) reports no before/after measurement for > 300 SAFe-trained people, rates the courses as “too theoretical” and documents the in-house construction of a leadership course as a substitute (B) (Putta et al. 2024). The Swedish PI-planning survey (293 responses, three organisations, old) finds perceived productivity gains at 13 %, “lack of productivity” at 29 %

(B/D) (Gustavsson & Bergkvist 2019) — against provider claims of 20–50 % productivity and 30–75 % time-to-market, which Laanti & Kettunen classify as "lack empirical validation" (D, first author in consulting) (Helda). The Scaled Agile benefits page itself is E without methodology (Scaled Agile).

Exam format and psychometrics: all foundation-level certificates test multiple choice against a document — CSM 50 questions/74 %/60 minutes after a 16-hour course (E, factual statement) (Scrum Alliance), PSM I 80 questions/85 % with no course requirement (E) (Scrum.org), CPRE FL approx. 45 questions/70 % (E, examination regulations 2024) (IREB). **No provider publishes pass rates, no predictive validity, no inter-rater reliability** — neither the providers nor anyone independent. With **775,260 PSM I certificates** (E, as of 10/2026) (Scrum.org Count) the certificate is worthless as a scarcity signal. The Agile Alliance itself classified CSM in 2010 (old) as "knowledge-based" and "easy to earn" and advised employers not to require certificates (E, but against its own market interest) (Agile Alliance). For ISTQB, a 2026 review with ISTQB-affiliated authors confirms "no controlled study" on competence or defect detection for > 1.2 million certificates (B/D, COI) (arXiv 2603.14572); the ISTQB Effectiveness Survey 2013 (old) measures satisfaction and belief (E, designed by the ISTQB Marketing Working Group) (ISTQB). For IREB CPRE no effectiveness study was found; Six Sigma has firm-performance evidence only for programme adoption (B, old) (Jacobs et al. 2015), and "Black Belt" is not a uniform construct (training between one week and 1,000 hours, E) (iSixSigma).

For IT/security certificates the finding is identical: no study tests CISSP, CEH, CCNA, cloud or Kubernetes certificates against incident rates, error rates or supervisor ratings. The only effectiveness study of a certification mandate (DoD 8570, N = 30, one site, approx. 2013, old) comes out qualitatively negative: training delivers vocabulary, not a qualified workforce, instructors "teach to the test" (B methodology, D data) (Starland 2014). What is documented are wage premiums — CISSP 14–46 % depending on job function, CCIE 73 % in consulting, PMP not significant (B methodology, D data Certification Magazine 2004, old) (Quan et al. 2007) — and demand: > 48 % of 935 US cybersecurity job postings require an industry certificate (B, 2023) (Ramezan 2023). The conjoint experiment with 1,725 recruiters (UK/US, Aug.–Sep. 2025) shows that certificates bring only "modest gains" in interview invitations compared with a credible self-statement (B, preprint; funder Microsoft AI Economy Institute, E caveat) (Stephany et al. 2026). Demand measurably comes from screening, not from the technical line: 52 % of recruiters but only 15 % of managers saw the PMP as important (B, USA, old, via review) (IEEE Software review).

Provider figures belong in the category "claim by X, interest: Y": PMI reports a PMP salary premium of **33 % (2023) and 17 % (2025)** (E, interest: own certificate; change of method not explained) (PMI 2023, PMI 2025); Pearson VUE reports 79 % self-reported quality improvement among ~24,000 certificate holders (E, interest: examination service provider, 2025) (Pearson VUE); ISC2 names certificates as the most frequent "critical" hiring criterion among 929 hiring managers (E, interest: sells CISSP, 12/2024) (ISC2 2025). A twenty-year analysis of the PMI Pulse reports calls them an "advocacy instrument" with "shifting metrics, selective transparency" (B, preprint) (Zhangguan 2025).

## Measured and null: recertification, mandatory continuing education, professional drivers, teachers, dentists, emergency medical services

The second group differs from the first: here measurement took place, and the result is null or selection only.

**Recertification (ABIM Maintenance of Certification (MOC)).** The grandfathering quasi-experiment of Gray et al. (956 vs. 974 internists, 84,215 vs. 69,830 patients, baseline 1999–2000, post 2002–2005, old) finds **no difference in ambulatory-care-sensitive hospitalisations**, only 2.5 % lower costs after heavy adjustment (E, ABIM authors — so even the certifier finds nothing on the clinical outcome) (NBPAS Journal Club Gray 2014). The independent VA study (105 primary care physicians, 68,213 patients, 2012–2013, old) finds after adjustment **no difference in 10 quality measures** (B) (NBPAS Journal Club Hayes 2014). The only RCT (Simpkins 2007, 16 practices, old) missed its primary endpoint (aOR 1.0) (E, ABIM-funded, negative) (NBPAS 2018). The cost of the programme: **USD 5.7 billion over ten years**, 9 of every 10 dollars time costs, USD 23,607 per physician (B, modelling study 2015, old) (ScienceDaily 2015). Against this stands ABIM's own series: 15–19 % relative mortality reduction in AMI/CHF by initially certified internists (E, Pennsylvania 1993 and 2012–2017; authors ABIM staff; "cannot establish causality") (BMJ Open 2022). The strongest independent mortality study on initial certification — Silber 2002, 144,883 Medicare patients Pennsylvania 1991–1994, old — yields **adj. OR 1.13 (1.00–1.26)** for death with non-certified mid-career anaesthesiologists, with the authors' caveat that the hospitals rather than the physicians could be the cause (B) (Silber 2002). The volume–outcome relationship, by contrast, is an order of magnitude stronger: low-volume surgeons almost **fourfold mortality** in pancreatic resection (B, Medicare ~474,000 patients, 2003, old) (Dartmouth 2003); in Germany a significant volume–outcome relationship in 20 of 25 procedures, open abdominal aortic aneurysm 4.7 % vs. 7.8 % risk-adjusted mortality (A/B, full DRG census 2009–2014, old) (Nimptsch & Mansky 2017). The G-BA (Federal Joint Committee) therefore steers via minimum volumes (structure), not via personal certificates.

**Mandatory continuing education and revalidation.** For § 95d SGB V (250 continuing medical education points (§ 95d SGB V) per 5 years) no outcome evaluation was found; at the end of 2008 only 23 % of statutory health insurance physicians in Bavaria had reached the points (A, KVB administrative data, old) (Bayerisches Ärzteblatt 2009). Cochrane supplies the upper plausibility bound: educational meetings improve professional practice by **RD 6.79 %** and patient outcomes by **RD 3.30 %** (B, 215 studies, search up to 2016, old) (Cochrane CD003030) — and that for measured interventions, not for collecting points. The GMC-funded UMbRELLA evaluation of UK Revalidation (data on ~281,000 doctors, > 85,000 survey participants, 2015–2017, old) claims no improvement in patient safety and finds no causal link to the decline in fitness-to-practise referrals (B/D) (GMC UMbRELLA).

**Professional drivers (Directive 2003/59/EC, BKrFQG (German Professional Driver Qualification Act), Driver CPC).** No evaluation with crash metrics was found. The British government conceded in writing in 2013 that causal proof was "very difficult" (E, regulator, old) (GOV.UK); in the call for evidence, of 748 respondents **71 saw "significantly increased safety", 328 "positive but marginal", 349 "no road safety benefit"** (D, 2013, old) (GOV.UK PDF). The DCPC review 2021–22 cites truck crashes 2007–2020 **–68 % vs. cars –54 %**, an

unadjusted before/after comparison with no control for AEBS, mileage or the business cycle (C time series, E conclusion) (GOV.UK Final Report). The Commission's own impact assessment: "the attributed expected effect are quite limited" (E, 2017, old) (COM(2017)47). German associations (DVR (German Road Safety Council), driving instructor associations) do not pose the effectiveness question; the claim that the system improves safety "successfully across Europe" stands without evidence (E, interest: driving instructor associations sell the training) (position paper). The US ELDT rule (2022) has no behind-the-wheel hours and no effectiveness data; ATRI only started a study in 2025 (C/E) (Truckinginfo 2025).

**Pilots (1,500-hour rule).** No study with accident/incident rates as the target measure. Rupp & Tan explicitly discard incidents (480 in 2010–2018, too little variance) and measure delays (+14.8 % short-term) (B, working paper 2024) (Rupp/Tan). The Pilot Source Study (n = 6,734, hires from 08/2013, old) shows **fewer training dropouts at ≤ 1,500 h than at > 4,500 h** (B, authors at university flight programmes with an interest in the R-ATP pathway) (Smith et al. 2016). The ALPA figure "99.8 % fewer deaths" (E, interest: supply restriction) and the AAF figure "all six regional crashes 2003–2009 with > 1,500-h captains" (E, author ex-board Mesa Air) are both unfalsifiable (ALPA, AAF). CRM training: large effects on attitude and behaviour in 16 studies, accident effect not reported (B, old) (O'Connor 2008).

**Licensing economics USA.** Stricter dentist licensing: **no effect on dental health, prices +14–16 %** (B, Air Force recruits, 2000, survey year not documented, old) (NBER Digest 1997). Teacher certification explains "at most small" differences (≈ ±0.02–0.03 SD) against large persistent variance within the certification groups (B, NYC 1998–2005, old) (NBER WP 12155); certification tests raise wages by 2.4–3.3 % but not quality (B, SASS 1987–2000, old) (NBER WP 9545). The structural model across 624,697 employees (CPS 2015–2018, old) estimates, for full conversion to licensing, **wages +15 %, employment −29 %, welfare loss 12 % of the occupational surplus** (B) (Minneapolis Fed SR 590). Conversely, deregulation improves outcomes: more autonomous scope for dental hygienists → fewer tooth extractions (B, 2001/2014, old) (CHWS), independent practice for nurse practitioners → fewer emergency department visits (B, old) (JHE 2018).

**Emergency medical services.** The reverse direction: ALS (paramedic) vs. BLS with defibrillation — survival to discharge **5.0 % vs. 5.1 %** (B, OPALS, 5,638 patients, old) (OPALS via Advocate); Medicare 2009–2011 (old): **13.1 % (BLS) vs. 9.2 % (ALS)**, propensity score (B) (Sanghavi 2015). A higher qualification level does not mean a better outcome when the additional intervention costs time.

## Organisational certificates show selection; medical devices show the opposite of what was expected

For ISO 9001 the strongest evidence is split, and split along the choice of control group, not along the data. Corbett et al. (7,238 US manufacturers, 1987–1997, old, matched on prior ROA) find ROA ~3.9 percentage points higher three years after certification (B) (Corbett 2005); Dick, Heras & Casadesús (800 firms, Basque Country, 1994–1998, old) show, using the *not yet* certified firms as the control group, that the advantages existed **before** certification and no gain followed — the same pattern in Häversjö (Denmark, 800 firms) and Wayhan (USA, 96 firms)

(B) (Dick et al. 2006). Loss of the certificate has no measurable financial effect (B, 143 Portuguese firms, old) (Cândido 2016). **No study found measures ISO 9001 against defects, recalls or complaints.** Qualitative research documents ceremonial adoption (B, 8 organisations, old) (Heras-Saizarbitoria 2011) and auditors with a predetermined result: "a certificate … does not guarantee conformance" (B, old) (Kuceja 2017). The ASPICE assessments (34 assessments, 2012–2019, old) show an auditability bias: documentation and tracing practices score well, verification criteria and impact analysis worst (B, process rather than product measure) (Falcini & Lami 2020). What is easy to audit gets done.

For ISO 27001 no breach evidence exists, only a positive stock market reaction to 111 announcements (B) (Deane 2019). The methodologically strongest work in the field tests **controls, not certificates**: the IV regression on official Israeli survey data (993 firms, survey 2020–2021) shows that the naive OLS correlation is positive (those attacked upgrade afterwards) and only instrumentation reveals the causal drop in incident probability from 80 % to 42 % with all six basic controls (B, NSF-funded) (Gandal et al.). This endogeneity problem applies to every future "certified vs. breach" correlation. The British Cyber Essentials evaluation (DSIT 2024) is the only example of a state-evaluated certification with an attempt at outcomes — and it concedes that it cannot measure incidents: **only 8 % of users notice fewer incidents, 57 % "too difficult to judge"** (D, self-selection); the insurance figure "80 % fewer claims" comes from an unnamed insurer with no risk adjustment (C/E) (gov.uk 2024).

Medical devices are the only field with all four elements: a hard outcome measure, a counterfactual system, a documented incentive mechanism and regulatory consequence. Hwang et al. (309 high-risk devices, CE 2005–2010, safety data up to 2016, old): devices approved first in the EU **27 % recall/warning vs. 14 % for US-first, adjusted HR 2.9 (1.4–6.2)**, recalls only HR 4.6 (B) (BMJ 2016). The BMJ/Telegraph undercover investigation with a fictitious hip prosthesis at 14 notified bodies (2012, old) documents competition on price (EUR 2,750–50,000), speed and promised pass rate (C, investigative) (EurekAlert/BMJ 2012). After the joint audits at least four of twelve audited notified bodies lost their designation, one was suspended — among them EVPU, the same body that had offered to "cross the line" in 2012 (C, 2013, old) (Clinica via Team-NB). The mechanism — the audited party chooses and pays the auditor, and oversight does not measure the delegation — recurs at Wirecard (APAS sanction against EY 2023: two-year ban on mandates, EUR 500,000, 7 of 12 auditors escaped by surrendering their licences; A via trade press) (Accounting Today) and Boeing (congressional report 2020: FAA management repeatedly overruled technical experts; A via Jurist) (Jurist 2020). The German accreditation oversight reports procedural metrics (104 objections 2018, ~0.97 % 2024), but **no withdrawal or surveillance statistics and no effectiveness metrics** (A, Bundestag printed paper 20/13351, 2024) (Bundestag). CRA, NIS2 and EUCC thus rely on a mechanism whose effectiveness is documented in no neighbouring field and refuted in one (medical devices before 2013); the EU evaluation of the Cybersecurity Act was, as of February 2024, available without findings (A, no result) (EC Evaluation).

## DACH regulates most densely and measures least

The country differences concern almost exclusively structure, not outcome.

**Licensing density.** EU-SOR 2015 (26,640 interviews, self-report, analysis B, old): **Germany 32.9 % licensed employees — highest EU value**, Austria 22.2 %, UK 19.5 %, Finland 16.7 %, Sweden 15.3 %, Denmark 14.0 %, EU average 21.9 % (Koumenta & Pagliero). USA ≈ 29 % (2008, old) or 12–33 % depending on state (B/D) (Minneapolis Fed). In the OECD OER indicator 2018–19 Germany is the most heavily regulated country in the regression sample, Sweden the least; a reduction by one indicator unit (≈ Germany → Sweden) corresponds to **+1.5 to 1.6 percentage points of productivity growth** for the average firm (B; quality not measured) (OECD WP 1605). The licensing premium in the EU, at ≈ 4 %, is markedly lower than in the US at 14–18 % — with higher EU licensing density; ≈ 2/3 of the EU premium is estimated as signalling, ≈ 1/3 as rent, with a large standard error (B) (BJIR 2019). The hypothesis "heavily licensed countries have better consumer outcomes" has been tested with outcome data by no source (OECD, Commission, Koumenta/Pagliero, BIS) — it is unproven, not refuted.

**Germany: skilled crafts.** The 2004 amendment of the Handwerksordnung (Crafts Code) (53 of 94 trades made licence-free) is the region's largest natural experiment. Robustly documented are a founding boom (B1 businesses 74,940 → 244,273 by 2017; C, ZDH (German Confederation of Skilled Crafts) register) (DHZ 2019), **exits B1 +2.0–2.5 percentage points** (B, DiD, common trends violated, ifh-affiliated) (Runst et al. 2019) and the decline in the 5-year survival rate **69 % (cohort 2001) → 46 % (cohort 2007)** (B with interest note ifh, chamber of crafts data, old) (statement Bizer 2019). **No expert report, not even one close to the crafts sector, found causal evidence of a loss of quality**; the Federal Monopolies Commission (Monopolkommission) stated in 2019 that there were "no empirical studies" on it (B with note) (DHZ 2019). Training effects are disputed: Koch & Nielen (DiD) find "no significant effects" (B, FES-published) (IfM Bonn), the ifh negative effects from 2009, confounded with the AEVO suspension 2003–2009. The voluntary master craftsman title works as a signal: on MyHammer (5,144 firms, 2013–2015, old) **+2 percentage points top overall rating (93 % vs. 91 %)** with a 92 % ceiling effect (B with note ifh) (Fredriksen et al.). The 2020 reinstatement of the Master craftsman requirement (Meisterpflicht) (12 trades) is justified in draft bill 19/14335 by hazard propensity, protection of cultural assets and ZDH training figures (apprentices –32 %, master examinations –72 % in the 12 trades 2003–2018; A document on C data) and contains **no statistics on defects, complaints or quality**; the 2025 evaluation clause again measures only vocational training, business and employment data (Bundestag printed paper 19/14335). In the hearing with nine experts, none cited quality or defect figures (A, minutes) (Bundestag kw26 2019).

**Germany: project management, driving school, medicine.** The PM market is dominated by IPMA/GPM (> 30,000 certificates in 2012 vs. ~10,300 PMI, D/E, old) with an 8-hour assessment plus oral examination and proof of experience at levels A–C — formally closer to a competence assessment than PMP or CSM, but without any outcome measurement; the GPM studies ask about salary and career (E, interest: own certificate) (oose comparison 2013, GPM 2024). For the German mandatory driving school **no study** was found that isolates it against countries without the requirement; Assailly explicitly denies a system effect GB/D/F on young driver risk (B, 1999, old) (TRID). At the aggregate level Germany has, at 32 %, the highest share of young drivers among driver fatalities in the EU (A/C, ERSO, WHO data without year) (ERSO) and 3.35 road deaths per 100,000 vs. UK 2.61, Sweden 2.17, Norway 2.14, USA 12.84 (A, data year 2022) (BITRE 2022) — no causal proof, but at the aggregate level incompatible with "expensive mandatory training → measurably safer novices". In medicine, specialist title, additional

designation and § 95d are evaluated exclusively on input and satisfaction (BÄK (German Medical Association) residency evaluation: overall grade 2.54 → 2.44, D, old) (DÄ 2015); the arbitration board of the North German medical chambers does not analyse specialist status as a variable in 41,138 decisions 2000–2015 (A without certificate reference, old) (A&I 2016). No German or European study on exam scores (state examination, specialist examination) and later complaints was found.

**Austria and Switzerland.** The multi-phase driver training (since 2003) shows −10.3 % crashes among 18–19-year-olds 2004–2005 (C/E, ministry, simple before/after, old) (Parliament AT) or −28 % among 18-year-olds 2000–2002 vs. 2004–2006 (C, commissioned evaluator, old) (WKO/Bartl); the BASt-funded methodological critique (Saarland 2013) has not published its results, and the ÖAMTC expert himself attributes the decline "largely" to vehicle safety (E, interest: ÖAMTC sells multi-phase courses) (ÖAMTC 2023). The WKO (Austrian Economic Chamber) examination statistics 2024 report a pass rate of 66.5 % for master/qualification examination modules (C) (WKO 2024), no reliability. Switzerland has ≈ 160 regulated professions, only 13 with a mandatory vocational qualification, the rest cantonally inconsistent; no outcome measurement (A/B, Ecoplan 2025) (edudoc.ch). This puts the German justification "Meisterpflicht = guarantor of training" into perspective: a strong dual system works without a blanket licensing requirement.

**USA.** The country with the most outcome research (Medicare, FARS, NTSB, NCDR registers) and at the same time the strongest licensing rent economy. The boards produce the positive evidence mostly themselves (ABIM, ABS, NBME/FSMB); independent replications are absent or null. DoD 8140 relaxes the pure 8570 certificate mandate in favour of experience and education — an administrative decision with no outcome evaluation found (A, GAO reports do not address effectiveness) (Defense Systems 2011). The Deming field experiment (AER 2016, old) provides the most direct evidence for the screening function: in healthcare jobs **with** certificate and licence requirements, the origin penalty of a for-profit degree disappears (−57 % callbacks without requirement, 0 % with requirement) (B) (Harvard/Katz) — a certificate is useful to the employer when it neutralises provider variance.

**UK.** Since 2012 the British state has no longer treated the home-grown certificate PRINCE2 as proof of competence: its own Project Delivery Profession, Functional Standard GovS 002 (2018), > 2,000 MPLA/PLP graduates and an experience- and delivery-record-based accreditation; PRINCE2 is named in no official source (A) (IPA 2018, Smallwood 2020). Outcome data on this accreditation are also absent. The UK voluntary register for fitness trainers had "little impact" on qualification and pay, entry qualifications can be acquired in ≈ 10 days, while the French licence with 350–500 h of mandatory practice supports higher skills (B, qualitative, 2009–2011, old) (Cardiff ORCA).

**Nordics.** The countries with the lowest licensing density (14–15 %) and the lowest road death rates (2.14–2.17) provide the cleanest driver training experiments: Sweden's practice age 16 works (B), Norway's does not (B/D), Norwegian and Finnish mandatory skid training increased crashes (B/C). ETSC attributes Sweden's low young driver death rate (53/million vs. EU 103/million, 2008–2010, old) to the low licensing rate at 18–22, not to training (C) (ETSC PIN 21). SAFe adoption data from Finland and Denmark are D/B without outcome.

## Contradictions between and within the notes

The notes contain contradictions that are left open here rather than smoothed over.

**ISO 9001 — treatment or selection.** Corbett et al. (matched on prior ROA, gain ~3.9 pp) and Dick/Häversjö/Wayhan (not-yet-certified as control, no gain) use overlapping periods and countries and reach opposite conclusions. The IT/org note classifies this as a methodological dispute over the choice of control group. Cândido (no effect on decertification) supports the selection side. The report adopts neither side as established.

**Board certification of physicians — ABIM series vs. independent null findings.** Norcini 2000/2002/2022 (E) report 15–19 % relative mortality reduction; Hayes 2014 (B) and Gray 2014 (E) find null on the clinical outcome, Simpkins 2007 (RCT, E) negative. Sharp 2002 reads the literature as "more than half support an association" (C/E), Grosch 2006 reanalyses the same collection and finds "no credible link" (B). This dissonance is the core of the MOC controversy and is carried in the report as such.

**Cyber Essentials — 80 % vs. 92 %.** The DSIT evaluation cites 80 % fewer insurance claims (NCSC Annual Review 2023, data 2022), the NCSC blog 92 % (source: a pooled insurer, year not stated). The discrepancy is explained in no source (gov.uk 2024, NCSC).

**PMI salary premium — 33 % vs. 17 %.** Between Earning Power 2023 and 2025 the PMP premium halved without PMI explaining the change of method or sample (E).

**BF17 — BASt vs. DVR.** BASt M 218 cites −17 % (confounder-adjusted) or −22 % (mileage-adjusted) crash involvement; the DVR resolution of 2024 quotes "−23 % crashes, −22 % traffic offences" as BASt figures. The driving note lists the DVR statement as E (interest: driving instructor/insurer members).

**Austria multi-phase — −10.3 % vs. −28 %.** BMVIT (2004–2005 vs. the two preceding years, 18–19-year-olds) and Gatscha/Brandstätter (2004–2006 vs. 2000–2002, 18-year-olds) use different windows and age groups; both are unadjusted before/after comparisons. The BASt-funded methodological critique is unpublished.

**Cochrane GDL — dating.** The cochrane.org summary gives 1 April 2022, the Cochrane Library version page dates pub3 to 5 October 2011; in either case the search covers data up to 2009.

**Tamblyn 2007 — grade and figures.** The psychometrics note lists the study as B (A register data) with RR 1.38 (communication) and 1.51 (clinical decision making); the medicine note lists the same study via the MCC summary as "B with COI note" (MCC-funded) with RR 1.39 for the MCCQE1 total score and RR 1.38 for the communication subscore. The figures are compatible (different subscores); the grading differs in the COI note; the report adopts B with COI note.

**Norcini 2022 — N.** Abstract 184,115 vs. results section 184,155 hospitalisations (internal inconsistency of the primary source).

**Driving school — RCTs vs. Nebraska.** Three RCTs (RR ≈ 1.0) against the Nebraska register study (n = 151,880, 2003–2010, old), in which driving school graduates have OR 1.21 fewer crashes and OR 1.79 fewer violations than the parent logbook route (B, self-selection explicitly

not ruled out: logbook more often rural, male, low-income) (Nebraska DOT). Randomised evidence beats observational evidence here.

**Fredriksen et al. — master craftsman share.** Text ¼ vs. table 0.41 (internal inconsistency of the primary source, noted by the note).

**Foote Partners 2006.** Article internally contradictory ("three times faster" vs. 4.4 % vs. 2.6 %) (E).

**Regulated professions vs. licensing share.** Register 2012: DE 86, UK 131, AT 151 regulated professions — but licensing share DE 33 % vs. UK 19.5 %. Germany regulates few, employment-heavy professions; the count of professions is unsuitable as a measure of stringency.

## Summary table

| Certificate / licence | Outcome measure measured | Finding | Best evidence grade (period) | Country |
|---|---|---|---|---|
| Midwife licence | Maternal mortality | demonstrated (−6–7 %, floor effect) | B (1900–1940, old) | USA |
| Graduated Driver Licensing | Crashes, 16-year-olds | demonstrated (−15 to −26 %), partly displacement to age 18; per licence holder ≈ 0 | B/A (1986–2009, old) | USA/CA/AU/NZ |
| Accompanied driving from 17 (BF17) | Crash involvement, 1st year | demonstrated (−17/−22 %), self-selection open | A/B (2007–2011, old) | DE |
| Practice age 16 (Sweden) | Crashes per km | demonstrated for users (−24 to −40 % adj.) | B (1993–1996, old) | SE |
| Practice age 16 (Norway) | Crash risk | null | B/D (1998–1999, old) | NO |
| Formal driver education | Crash involvement | null (RR 1.01–1.10), earlier licence | B, 3 RCTs (1982–84, old) | AU/US/NZ |
| Skid training (mandatory) | Crash rate | negative (+12 % cars, +45 % ambulances) | B via C (old) | NO/FI |
| Practical driving test (DPE) | Crashes/ violations, 2 yrs | null (n = 394,101) | A/B (1993–1997, old) | USA |

| Certificate / licence | Outcome measure measured | Finding | Best evidence grade (period) | Country |
|---|---|---|---|---|
| Multi-phase driver training | Crashes, 18–19-year-olds | unclear (−10 to −28 %, unadjusted) | C/E (2004–2006, old) | AT |
| Mandatory driving school (D) vs. UK/USA | Young driver fatality | no measurement; aggregate does not support it | A/B (1999; 2012; 2022) | DE/UK/US |
| Professional drivers Directive 2003/59 / CPC / BKrFQG | Crashes, trucks/buses | no measurement; regulator concedes lack of proof | D/E (2013–2022) | EU/UK/DE |
| US CDL/ELDT | Crashes | no measurement | E (2016–2025) | USA |
| 1,500-h rule, pilots | Accident/ incident rate | no measurement (events too rare); training success without advantage of high hours | B (2013–2018) | USA |
| Instrument rating (GA) | Weather accidents | demonstrated ($p = 0.002$); flight hours n.s. | A/B (2003–2004, old) | USA |
| CRM training | Attitude/ behaviour; accidents | attitude/ behaviour yes; accidents not reported | B (old) | intl. |
| Forklift licence (OSHA) | Injuries | no measurement | E/D | USA |
| CSM / PSM / SAFe SPC | Project outcome, defects | no measurement (variable not collected) | — | intl. |
| PMP / PRINCE2 / IPMA | Project success | mixed; 1 field study positive correlational, majority null (self-report) | B (2004–2021) | US/ZA/intl. |
| ISTQB / IREB CPRE | Defect detection, RE quality | no measurement (“no controlled study”) | B/D (2026) | intl. |
| Six Sigma Black | Project outcome | programme | B (old) | USA |

| Certificate / licence | Outcome measure measured | Finding | Best evidence grade (period) | Country |
|---|---|---|---|---|
| Belt | | effect B; person: no measurement | | |
| CISSP / CCNA / cloud certificates | Incident rate, performance | no measurement; only wages/demand | B methodology on D data (2004, old; 2023) | USA/NZ |
| DoD 8570 certificate mandate | Work quality (perceived) | negative (qualitative, N = 30) | B/D (approx. 2013, old) | USA |
| Specialist/board initial certification | Mortality | small positive (OR 1.13 non-certified) with environment caveat; ABIM series 15–19 % (E) | B (1991–1994, old) | USA |
| ABIM MOC (recertification) | ACSH, quality measures, asthma outcome | null (quasi-experiment, independent, RCT) | B/E (1999–2013, old) | USA |
| Licensing exam scores, physicians (MCCQE, USMLE, ABIM) | Complaints, discipline, peer review | demonstrated as measurement instrument (RR 1.38–1.51/2 SD; OR 0.75/SD) | B (1993–2017) | CA/USA |
| UK Revalidation | Patient safety, FTP | null | B/D (2015–2017, old) | UK |
| § 95d mandatory continuing education | Patient outcome | no measurement; CME in general RD 3.3 % | A/B (old) | DE |
| Nursing specialty certification | 30-day mortality | small positive (−2 %/+10 % certified BSN), only with BSN | B (2005–2006, old) | USA |
| Surgeon volume (comparison) | Mortality | demonstrated (up to 4-fold; 20/25 procedures) | A/B (2003; 2009–2014, old) | USA/DE |
| ALS vs. BLS | Survival | null to negative | B (2004; 2009– | CA/USA |

| Certificate / licence | Outcome measure measured | Finding | Best evidence grade (period) | Country |
|---|---|---|---|---|
| (emergency medical services) | | | 2011, old) | |
| Dentist licensing stringency | Dental health, prices | null quality; prices +14–16 % | B (old) | USA |
| Teacher certification | Student achievement | null in the middle (±0.03 SD); floor effect at the left tail | B (1987–2007, old) | USA |
| Meisterpflicht (HwO 2004/2020) | Quality/defects | no measurement; survival rate 69 → 46 %; voluntary master +2–3 pp rating | B/C (2001–2022) | DE |
| ISO 9001 | ROA, survival, injuries | split: gain vs. selection only; decertification without effect; no defect measurement | B (1987–2009, old) | US/ES/DK/PT |
| ISO 27001 / SOC 2 | Breach rate | no measurement; only stock market reaction | B | intl. |
| Cyber Essentials | Incidents, insurance claims | self-report only; insurance figure without risk adjustment (80 %/92 %) | D, C/E (2022–2023) | UK |
| CE/notified bodies (medical devices) | Recalls/ warnings | negative vs. FDA (HR 2.9) | B (2005–2016, old) | EU/USA |
| CMMI / ASPICE | Programme performance, process rating | self-report or process measure; no product measure | D/E, B (2005; 2012–2019, old) | USA/intl. |
| Accreditation | Effectiveness of | no measurement | A (2024) | DE |

| Certificate / licence | Outcome measure measured | Finding | Best evidence grade (period) | Country |
|---|---|---|---|---|
| (DAkkS) | certifiers | | | |

## What makes a certificate demonstrative

The criteria follow from the contrast between the cases with effect and the cases without, not from opinion.

**First: construct proximity of exam content and practice criterion.** Where the exam construct matches the practice criterion, the score predicts — communication score → communication complaints (RR 1.43), clinical decision making → upheld complaints (RR 1.51), instrument rating → weather accidents. Where it does not match, it does not predict — driving test score → crashes (n = 394,101, p = 0.62), theory test → fines yes, practical test → fines no (Lithuania, B, old) (TOTS 2018), total flight hours → accidents (p = 0.788). Multiple choice over a document (Scrum Guide, PMBOK, syllabus) has, by construction, no proximity to delivery time, defect rate or project success; consequently no validity study exists there.

**Second: reported predictive validity against a register criterion.** The Sackett et al. correction (B, 2022) sets the benchmark: structured interviews ρ ≈ 0.42, job knowledge tests 0.40, work samples 0.33, years of job experience 0.07 (SIOP). A certificate is as valid as its examination procedure, not as its duration or its proof of experience in years. The medical licensing examinations are the only ones in the entire material for which validity against A data (regulatory registers) is published. No management, Agile, IT or crafts certificate reports such a figure.

**Third: continuous score instead of binary certificate.** The most stable signal in medicine lies in the score, not in passing: non-participants in the MOC examination showed no quality difference; the association depends on exam performance (E, Holmboe 2008) (NBPAS 2018). A certificate that only states “passed” throws away the predictive information. No regression discontinuity study at the pass/fail cut point was found — proof that the threshold itself achieves anything is missing everywhere.

**Fourth: inter-rater reliability is achievable but not reported.** In the Keele OSCE (116 students, 67 examiners) the examiner cohort explains 5.7 % of the variance and shifts ≈ 6 % of pass decisions (B, 2019) (Yeates 2019); product assessment by several expert jurors reaches ICC > 0.90 (E source, design transparent) (IJ 2010). An IRR ≥ 0.7 is therefore feasible for practical examinations — common master, journeyman, IPMA or specialist examinations report none. Chambers publish pass rates (WKO 66.5 %), not reliability.

**Fifth: supervised exposure is a component, not instruction.** All positive driving findings (GDL, BF17, Sweden 16) contain supervised practice in hours or months (118 h instead of 41–47 h); all null and negative findings (RCTs, skid training, DPE driving test) vary only instruction or examination. In medicine, outcome follows volume, not title; supervised residents reach specialist complication rates (revisions 2.0 % vs. 2.5 %, p = 1.0; B, 1,480 total hip arthroplasties 2013–2016, old) (Orthopäde 2022).

**Sixth: floor function in a hazard-prone field with no prior standard.** Midwives 1900–1940 and the left tail of the teacher distribution are the only cases in which the threshold itself works. The criterion of hazard propensity that the legislator cited in 2019 for reinstating the Meisterpflicht is thus the only empirically sustainable licensing criterion — with the qualification that it was not measured for the twelve trades concerned.

**Seventh: recertification must rest on performance data, not on time or points.** Time- and points-based recertification is null three times over (MOC, Revalidation, CPC), and even measured continuing education moves patient outcomes by RD 3.3 %. Loss of an ISO 9001 certificate has no financial consequence. A recertification based on performance data (register outcomes, volume, complaint rates) was found in no note as implemented and evaluated — it is the logical demand arising from the evidence, but itself untested.

**Eighth: independence of the examiner from the examined and measurement of the delegation.** The only system comparison with a hard outcome measure (Hwang HR 2.9) shows commercial third-party certification inferior; the undercover investigation shows the mechanism; Wirecard, Boeing and the SOC 2 case of 2026 (E) show it outside medical devices. Where oversight actually examined the certifiers (joint audits 2013), a third to a half failed. A certificate whose issuer is paid by the certified party and whose accreditor keeps no withdrawal statistics (German accreditation body (DAkkS)) has no documented evidential value.

**Ninth: published pass rates and discriminatory power.** No provider of CSM, PSM, PMP, CPRE or ISTQB publishes pass rates. That is consistent with a business model in which the course is the product and the exam the justification. An exam that almost everyone passes does not discriminate — and without discriminatory power there is no validity that could be measured.

A certificate that meets none of these nine criteria is, on the available evidence, a label: a screening filter for HR departments (Deming, Starkweather), a wage premium (Quan, Kleiner) and a scarcity signal that vanishes with mass distribution (775,000 PSM I).

## Open gaps

According to the notes, no outcome measurement exists for the following certificates and questions. The gaps are sorted by the size of the market concerned.

Scrum Master and SAFe certificates (CSM, PSM, SPC) — not a single study, positive or negative, measures them against defect rate, delivery time or project success; the largest datasets (4,000 teams) do not collect the variable. ISTQB, IREB CPRE, ITIL — no controlled study on defect detection, requirements quality or operational stability; ITIL was not examined in the research at all. IT/security personal certificates (CISSP, CEH, CCNA, CCIE Lab, OSCP, CKA, RHCE, cloud) — no incident, error or supervisor rating study; the thesis “a lab exam is more valid than multiple choice” is only a provider claim (E). ISO 27001, SOC 2, FedRAMP — no breach incidence study certified vs. not certified. ISO 9001 / IATF 16949 — no study against defects, recalls or complaints; VDA/OEM field failure data by certificate status not public. CMMI, Automotive SPICE, ISO 26262 process certification, TÜV Functional Safety Engineer, intacs assessor — no product or field quality measure.

Professional driver qualification (Directive 2003/59, BKrFQG, Driver CPC, US ELDT), ADR dangerous goods training, forklift licence, crane operator — no evaluation with crash metrics; FMCSA comparison Illinois/Indiana with no result found. German mandatory driving school instruction — no study that isolates it causally against countries without the requirement; crash rate per km in the first year of driving D vs. UK not found. 1,500-hour rule, MPL vs. ATPL, Flight Review (BFR) — no accident/incident rates per flight hour before/after the rule change.

German medical system — specialist recognition, additional designations, special competences (radiation protection, emergency medicine), § 95d: no patient outcome evaluation; treatment error statistics (arbitration boards, MD, insurers) are not broken down by specialist status; no European study on state examination or specialist examination scores and later complaints. EU Directive 2005/36/EC — training duration as a proxy without outcome validation. Welders (EN ISO 9606), electrical competence (VDE), radiation protection — no field data; certification rests on a test piece. NREMT — exam-to-exam prediction only.

Skilled crafts — no German study with defect, warranty or damage data (building damage reports, insurers, arbitration boards) for B1 vs. A trades; the 2025 evaluation of the Meisterpflicht reinstatement again measures only training, business and employment figures; inter-rater reliability of master and journeyman examinations not reported. Cross-country — no study tests licensing share (DE 33 % vs. SE/DK 14–15 %) against consumer outcomes (damage, complaints, building damage rates); no study regresses training costs per country against novice risk. Accreditation (DAkkS, UKAS) — no withdrawal, surveillance or effectiveness metrics; whether accredited certifiers deliver better results than non-accredited ones cannot be tested. EUCC/CRA/NIS2 — result of the EU evaluation (announced for H2 2024) not retrieved; number of EUCC certificates issued unknown.

Methodological gaps across all fields: no regression discontinuity study at the pass/fail cut point of any certificate; no RCT or quasi-experiment on any management, Agile or IT certificate; no meta-analysis on the predictive validity of professional certificates against job performance; no study that implements and evaluates a performance-data-based recertification.

Primary sources not retrieved that could change the picture: Chen/Krumholz 2006 (independent Medicare AMI cohort on board certification), Boiral 2012 and Heras-Saizarbitoria & Boiral 2013 (ISO 9000 reviews), Naveh & Marcus 2005, Sackett et al. 2022 full text, Kleiner & Kudrle full text, Koch & Nielen full text, Malone & Brünken 2013 (multi-phase methodological critique), SWOV fact sheet on driver training, Starkweather & Stevenson 2011 primary text.

## Conclusion

The evidence draws a consistent picture that cuts across the usual debates. The question “does the certificate work?” is wrongly posed in most fields, because the certificate bundles three different things that must be assessed separately: an examination (which can be valid as a measurement instrument, as in medicine, or not, as in the driving test), an access rule (which works as a floor in hazard-prone fields and otherwise generates rents) and a course (for which no independent effect is documented in any field). The market for management, Agile and IT certificates sells almost exclusively the third component and refrains from publishing the data

that would allow an assessment of the first two. That is not an accident of the state of research but a feature of the business model.

For practice this implies a reversal of the burden of proof. A certificate whose provider publishes no pass rate, no validity against a practice criterion and no reliability is to be treated as a label until proven otherwise — regardless of how many holders it has and how high the salary premium is. Conversely, medicine shows that exam scores can predict practice when construct and criterion fit together and the criterion comes from a register; the consequence would be to use the score instead of the certificate and to tie recertification to performance data instead of points. Precisely this step is implemented and evaluated nowhere. Regulation (CRA, NIS2, Meisterpflicht reinstatement, § 95d) meanwhile continues to build on certification without defining the outcome measure against which it could fail.