Zertifikate als Fähigkeitsnachweis

Wo Zertifizierung nachweislich wirkt, wo sie gemessen und wirkungslos ist, wo nie gemessen wurde

Thomas Arends

3. November 2026

Stand: 03.11.2026. Grundlage sind fünf Recherchenotizen (Agile/PM, Fahr-/Flug-/Lkw-Lizenzen, IT-/Organisationszertifikate, Lizenzökonomie/Psychometrie, Medizin/Boards). Jede Zahl trägt in Klammern ihre Evidenzstufe (A amtlich/Register · B unabhängige Forschung mit Methodik · C Verbands-/Behördenmessreihe mit offener Methodik · D Umfrage/Selbstauskunft · E Anbieter-/Verbands-/Beraterangabe im eigenen Interesse) und den Erhebungszeitraum; „alt" markiert Daten, die älter als etwa drei Jahre sind. Kernaussagen stehen nur auf A–C. D/E erscheinen ausschließlich als markiertes Indiz mit benanntem Interesse.

Kurzfassung. Über alle untersuchten Felder hinweg gibt es genau drei Konstellationen, in denen eine Personen-Zertifizierung nachweislich eine Ergebnisgröße verbessert hat: erstens die Einführung einer Mindestschwelle in einem gefahrgeneigten Feld ohne vorherigen Standard (Hebammen-Lizenz USA 1900–1940: Müttersterblichkeit −6 bis −7 %, B, alt); zweitens Lizenzsysteme, deren wirksamer Bestandteil beaufsichtigte Praxis und Zugangsbeschränkung ist, nicht Unterricht (Graduated Driver Licensing, Begleitetes Fahren ab 17, schwedisches Übungsalter 16: −15 bis −26 % Unfälle im ersten Jahr, A/B, alt); drittens Prüfungen, deren Score als kontinuierliches Maß ein konstruktnahes Praxiskriterium vorhersagt (kanadische und US-Lizenzprüfungen für Ärzte: RR 1,38–1,51 je 2 SD für aufrechterhaltene Beschwerden, OR ≈ 0,75 je SD für Disziplinarmaßnahmen, B, Daten 1993–2017). Für die gesamte Klasse der Management-, Agile-, IT- und Security-Zertifikate (CSM, PSM, SAFe, PMP, PRINCE2, ITIL, ISTQB, IREB CPRE, CISSP, CCNA, Cloud-Zertifikate) existiert nach den Notizen

keine einzige unabhängige Studie, die das Zertifikat gegen eine objektive Ergebnisgröße misst; belegt sind nur Lohnprämien, Recruiter-Präferenzen und Selbstauskünfte. Wo Rezertifizierung oder Fortbildungspflicht mit starkem Design gemessen wurde (ABIM MOC: Quasi-Experiment und RCT; UK Revalidation; EU-Berufskraftfahrer-Weiterbildung), ist das Ergebnis null. Organisationszertifikate (ISO 9001, ISO 27001, SOC 2) zeigen Selektion statt Wirkung, und im einzigen Feld mit harter Ergebnisgröße und Vergleichssystem (Medizinprodukte EU vs. FDA) schneidet die kommerzielle Drittzertifizierung schlechter ab (HR 2,9 für Rückrufe/Warnungen, B, Daten 2005–2016). Länderunterschiede betreffen fast nur Marktstruktur und Lizenzdichte (Deutschland 33 % lizenzierte Beschäftigte, höchster EU-Wert, B/D, 2015), nicht gemessene Ergebnisse: Kein DACH-Zertifikat ist auf Outcome evaluiert.

Drei Mechanismen tragen Wirkung — und keiner davon ist „der Kurs"

Die positiven Befunde teilen eine Eigenschaft: Nie ist der Unterricht oder die Urkunde der wirksame Bestandteil, sondern eine Schwelle, eine Exposition oder ein Messinstrument.

Mindestschwelle in gefahrgeneigtem Feld. Die Hebammen-Lizenzgesetze der US-Bundesstaaten 1900–1940 senkten die Müttersterblichkeit um **6–7 %** (B, Working Paper, Primärquellen 1900–1940, alt) (NBER WP 22456). Die Lizenzökonomie-Notiz ordnet dies als einen der wenigen positiven Qualitätsbefunde der gesamten US-Literatur ein und benennt die Bedingung: hohes unmittelbares Gefahrenpotenzial *und* kein vorheriger Mindeststandard. Dieselbe Logik stützt der Lehrer-Befund von Larsen et al.: Strengere Zertifizierungsanforderungen reduzieren die Zahl der Sekundarlehrer aus wenig selektiven Hochschulen signifikant (linker Rand der Verteilung), ohne die Zahl hochqualifizierter Lehrer messbar zu verändern (B, Working Paper, Daten 1991–2007, alt) (NBER WP 28158). Zertifizierung wirkt, wenn überhaupt, als Floor — nicht als Prädiktor im Mittelfeld.

Beaufsichtigte Praxis und Zugangsbeschränkung. Der Cochrane-Review zu Graduated Driver Licensing (34 Studien, 21 Programme, Suche bis 2009, alt) findet bei 16-Jährigen im ersten Jahr nach Einführung **Gesamtunfälle Median −15,5 %** (Spanne −27 bis −8 %) und **Verletzungsunfälle Median −21 %** (B) (Cochrane CD003300). Die FARS-Analyse aller 50 US-Staaten 1986–2007 (alt) zeigt für starkes GDL **−26 % tödliche Unfälle bei 16-Jährigen, aber +12 % bei 18-Jährigen** (B) (UNC/JAMA 2011) — ein Teil des Effekts ist Verlagerung. Die IIHS-Auswertung 1993–2003 (C, Versicherer-finanziert, alt) zeigt den Mechanismus: pro 16-jähriger Bevölkerung −26 %, **pro lizenziertem 16-Jährigen praktisch unverändert (73 vs. 74 je 100.000)** (Insurance Journal 2005). Der Effekt läuft über weniger und spätere Lizenzierung, nicht über bessere Fahrer. Für Deutschland belegt der BASt-Bericht M 218 zum Begleiteten Fahren ab 17 (ca. 19.000 Befragte plus 75.000 stille Registerfälle, Projekt 2007–2011, alt) im ersten Jahr Alleinfahren **−17 %**

Unfallbeteiligungen konfunder-adjustiert, −22 % fahrleistungsbereinigt (A/B; Selbstselektion eingeräumt, Kontrolle nur über Konfunder) (BASt M218). Schweden senkte 1993 das Übungsalter auf 16 bei unverändertem Führerscheinalter 18: Fahranfänger gesamt **ca. −15 % Unfälle je 10 Mio. km, Nutzer der frühen Übung ca. −40 %, Nichtnutzer kein Effekt** (B, alt) (PubMed 10576673); nach Konfunder-Adjustierung bleiben 24–39 %, Frühstarter übten 118 statt 41–47 Stunden (B/C, alt) (DfT-Papier Gregersen). Norwegen senkte 1994 ebenfalls auf 16, der Übungsumfang stieg kaum, und das Unfallrisiko änderte sich nicht signifikant (B/D, Selbstbericht, alt) (TØI 498/2000). Die Kontrastgruppe ist eindeutig: Wo nur Unterricht variiert, verschwindet der Effekt. Die drei RCTs zur formalen Fahrausbildung (n = 17.965, 1982–1984, alt) liefern **RR 1,01 / 1,03 / 1,10** für Unfallbeteiligung und RR 1,04 für frühere Lizenzierung (B) (Cochrane CD003201). Schleudertraining erhöhte in allen sechs meta-analysierten Programmen die Unfallraten, **Pkw ca. +12 %, Rettungswagen ca. +45 %** (B-Meta-Analyse, hier über Sekundärquelle, daher C, alt) (International Fleet World).

Kriteriumsvalide Prüfungsscores. Die kanadische Kohorte von Tamblyn et al. (3.424 Ärzte, Clinical Skills Examination 1993–1996, Follow-up bis 2005, alt) zeigt je 2 SD niedrigerem Kommunikationsscore **RR 1,38 (95 %-KI 1,18–1,61)** für von der Aufsicht aufrechterhaltene Patientenbeschwerden und je 2 SD niedrigerem Clinical-Decision-Making-Score **RR 1,51 (1,25–1,84)**; das unterste Kommunikationsquartil verursacht 10,0 % aller aufrechterhaltenen Beschwerden (B, Registerdaten der Aufsichtsbehörden; Finanzierung durch den Prüfungsanbieter MCC, COI vermerkt) (JAMA 2007 via McGill). Die Peer-Review-Substichprobe (208 Ärzte Ontario) ergibt für das unterste Wissensquartil OR 3,41 und für das unterste Fertigkeitsquartil OR 4,24 für unakzeptable Praxisbewertung (B, COI MCC, breite KIs, alt) (Wenghofer 2009). In den USA zeigen 164.725 Absolventen 1994–2006 (alt) **OR 0,75 je SD Step 2 CK** für Disziplinarmaßnahmen (E, Autoren bei NBME/FSMB, aber methodisch transparent) (Cuddy 2017) und 275.392 Ärzte mit Step 3 2000–2017 **−23 % Odds je SD** (E, Prüfungsanbieter-Umfeld) (Academic Medicine 2022). Die Alberta-Kohorte 2014–2017 bestätigt +60 % Odds je ≈ 1 SD niedrigerem Part-I-Score (B, Autoren teils MCC) (JMR 2024). Entscheidend ist die Lesart der Medizin-Notiz: Das stützt die **Validität des Messinstruments**, nicht die Wirksamkeit der Zertifizierung als Maßnahme. Verhaltensmerkmale sind dabei

stärkere Prädiktoren als kognitive Scores — unprofessionelles Verhalten im Studium OR 3,0, Verantwortungslosigkeit OR 8,5, gegen MCAT OR 0,6 und Step 1 OR 0,9 (B, Fall-Kontroll, Disziplin 1990–2003, alt) (Papadakis 2005).

Dieselbe Logik gilt in der Luftfahrt: Die NTSB-Fall-Kontroll-Studie zu Wetterunfällen (72 Unfälle vs. 135 Kontrollflüge, 2003–2004, alt) findet das **Instrument Rating als signifikanten Schutzfaktor (p = 0,002)**, hohe Durchfallquoten in FAA-Tests mit Unfallbeteiligung verknüpft, und **Gesamtflugstunden nicht signifikant (p = 0,788)** (A/B) (NTSB 2005). Der Inhalt der Qualifikation prädiziert, die Stundenzahl nicht.

Management-, Agile- und IT-Zertifikate: Null Outcome-Studien, viele Gehaltszahlen

Das Fehlen von Evidenz ist hier selbst der Befund. Der IEEE-Software-Review von 2023 (24 Primärstudien 2004–2021) stellt fest: 68 % der Methoden sind Surveys, 4 % Feldforschung, eine Studie nutzte quantitative Projektmetriken; fünf Studien (5.206 Fälle) berichten positive Effekte, sechs (3.304 Fälle) keinen Zusammenhang; agile Zertifikate sind „kaum untersucht" (B) (Luja Beraza et al. 2023). Die einzige Feldstudie mit Projektakten (Strang 2021, **2.692 öffentliche Beschaffungsprojekte**, Nordamerika, Erhebungszeitraum nicht belegt) findet zertifizierte PM erfolgreicher, aber rein korrelational mit 67,3 % Klassifikationsgüte und einer Effektstärke von 12 % (B) (Strang 2021); Selektion (bessere Organisationen stellen Zertifizierte ein) ist nicht ausgeschlossen. Dagegen stehen Farashah et al. (n = 452, Daten ca. 2014, alt): **kein direkter Zusammenhang Zertifikat → Leistung**, Effekt läuft vollständig über Selbstwirksamkeit (B, Selbstauskunft) (Athabasca), Catanio et al. (n = 93 USA, alt): zertifizierte PM managen Scope, Zeit, Kosten nicht besser (B/D) (IGI Global), und Joseph & Marnewick (Südafrika, alt): IT-Projektleistung „not influenced" (B) (RePEc).

Für Scrum und SAFe gibt es nicht einmal die Variable. Die größten Datensätze — Verwijs & Russo mit 15.078 Befragten und 4.013 Teams (2021–2023) — erheben das Zertifikat nicht als Kontrollvariable; der Skalierungsansatz erklärt **1–2 % der Varianz** der Teameffektivität, und nach Kontrolle für Agile-Erfahrung

verschwinden fast alle Effekte (SAFe β = −0,042) (B/D, Preprint, COI Verwijs/The Liberators) (arXiv 2310.06599). Die dänische SAFe-Fallstudie (PFA, Interviews 2018, alt) nennt für > 300 SAFe-geschulte Personen keine Vorher/Nachher-Messung, bewertet die Kurse als „zu theoretisch" und dokumentiert den Eigenbau eines Leadership-Kurses als Ersatz (B) (Putta et al. 2024). Die schwedische PI-Planning-Befragung (293 Antworten, drei Organisationen, alt) findet wahrgenommene Produktivitätsgewinne bei 13 %, „lack of productivity" bei 29 % (B/D) (Gustavsson & Bergkvist 2019) — gegen Anbieterclaims von 20–50 % Produktivität und 30–75 % Time-to-Market, die Laanti & Kettunen als „lack empirical validation" einstufen (D, Erstautorin Beratung) (Helda). Die Scaled-Agile-Benefit-Seite selbst ist E ohne Methodik (Scaled Agile).

Prüfungsformat und Psychometrie: Alle Foundation-Level-Zertifikate prüfen Multiple Choice gegen ein Dokument — CSM 50 Fragen/74 %/60 Minuten nach 16 Stunden Kurs (E, Faktenangabe) (Scrum Alliance), PSM I 80 Fragen/85 % ohne Kurspflicht (E) (Scrum.org), CPRE FL ca. 45 Fragen/70 % (E, Prüfungsordnung 2024) (IREB). **Kein Anbieter veröffentlicht Pass-Raten, keine prädiktive Validität, keine Inter-Rater-Reliabilität** — weder Anbieter noch unabhängig. Mit **775.260 PSM-I-Zertifikaten** (E, Stand 10/2026) (Scrum.org Count) ist das Zertifikat als Knappheitssignal wertlos. Die Agile Alliance selbst stufte 2010 (alt) CSM als „knowledge-based" und „easy to earn" ein und riet Arbeitgebern, keine Zertifikate zu verlangen (E, aber gegen das eigene Marktinteresse) (Agile Alliance). Für ISTQB bestätigt ein Review mit ISTQB-nahen Autoren 2026 „no controlled study" zu Kompetenz oder Defektfindung bei > 1,2 Mio. Zertifikaten (B/D, COI) (arXiv 2603.14572); die ISTQB-Effectiveness-Survey 2013 (alt) misst Zufriedenheit und Glauben (E, konzipiert von der ISTQB Marketing Working Group) (ISTQB). Für IREB CPRE wurde keine Wirkungsstudie gefunden; Six Sigma hat Firmenergebnis-Evidenz nur für die Programm-Adoption (B, alt) (Jacobs et al. 2015), und „Black Belt" ist kein einheitliches Konstrukt (Training zwischen einer Woche und 1.000 Stunden, E) (iSixSigma).

Bei IT-/Security-Zertifikaten ist der Befund identisch: Keine Studie testet CISSP, CEH, CCNA, Cloud- oder Kubernetes-Zertifikate gegen Vorfallsraten, Fehlerraten oder Vorgesetztenbewertung. Die einzige Wirkungsuntersuchung einer Zertifikatspflicht (DoD 8570, N = 30, ein Standort, ca. 2013, alt) fällt qualitativ

negativ aus: Schulung liefert Vokabular, keine qualifizierte Belegschaft, Dozenten „teach to the test" (B-Methodik, D-Daten) (Starland 2014). Belegt sind Lohnprämien — CISSP 14–46 % je Jobfunktion, CCIE 73 % im Consulting, PMP nicht signifikant (B-Methodik, D-Daten Certification Magazine 2004, alt) (Quan et al. 2007) — und Nachfrage: > 48 % von 935 US-Cybersecurity-Anzeigen verlangen ein Branchenzertifikat (B, 2023) (Ramezan 2023). Das Konjoint-Experiment mit 1.725 Recruitern (UK/US, Aug.–Sep. 2025) zeigt, dass Zertifikate gegenüber einer glaubhaften Selbstangabe nur „modest gains" bei der Interview-Einladung bringen (B, Preprint; Geldgeber Microsoft AI Economy Institute, E-Vorbehalt) (Stephany et al. 2026). Die Nachfrage kommt messbar vom Screening, nicht von der Fachlinie: 52 % der Recruiter, aber nur 15 % der Manager sahen das PMP als wichtig (B, USA, alt, via Review) (IEEE-Software-Review).

Anbieterzahlen gehören in die Kategorie „Behauptung von X, Interesse: Y": PMI meldet eine PMP-Gehaltsprämie von **33 % (2023) und 17 % (2025)** (E, Interesse: eigenes Zertifikat; Methodenwechsel nicht erklärt) (PMI 2023, PMI 2025); Pearson VUE berichtet 79 % selbstberichtete Qualitätsverbesserung bei ~24.000 Zertifizierten (E, Interesse: Prüfungsdienstleister, 2025) (Pearson VUE); ISC2 nennt Zertifikate als häufigstes „kritisches" Einstellungskriterium bei 929 Hiring Managern (E, Interesse: verkauft CISSP, 12/2024) (ISC2 2025). Eine Zwanzig-Jahres-Analyse der PMI-Pulse-Berichte nennt sie „Advocacy-Instrument" mit „shifting metrics, selective transparency" (B, Preprint) (Zhangguan 2025).

Gemessen und null: Rezertifizierung, Fortbildungspflicht, Berufskraftfahrer, Lehrer, Zahnärzte, Rettungsdienst

Die zweite Gruppe unterscheidet sich von der ersten: Hier wurde gemessen, und das Ergebnis ist null oder nur Selektion.

Rezertifizierung (ABIM MOC). Das Grandfathering-Quasi-Experiment von Gray et al. (956 vs. 974 Internisten, 84.215 vs. 69.830 Patienten, Baseline 1999–2000, Post 2002–2005, alt) findet **keinen Unterschied bei ambulant-sensitiven Hospitalisierungen**, nur 2,5 % niedrigere Kosten nach starker Adjustierung (E, ABIM-Autoren — also selbst der Zertifizierer findet beim klinischen Outcome

nichts) (NBPAS Journal Club Gray 2014). Die unabhängige VA-Studie (105 Hausärzte, 68.213 Patienten, 2012–2013, alt) findet nach Adjustierung **keinen Unterschied in 10 Qualitätsmaßen** (B) (NBPAS Journal Club Hayes 2014). Der einzige RCT (Simpkins 2007, 16 Praxen, alt) verfehlte den primären Endpunkt (aOR 1,0) (E, ABIM-finanziert, negativ) (NBPAS 2018). Die Kosten des Programms: **5,7 Mrd. USD über zehn Jahre**, 9 von 10 Dollar Zeitkosten, 23.607 USD je Arzt (B, Modellstudie 2015, alt) (ScienceDaily 2015). Dagegen die ABIM-eigene Serie: 15–19 % relative Mortalitätsreduktion bei AMI/CHF durch erstzertifizierte Internisten (E, Pennsylvania 1993 und 2012–2017; Autoren ABIM-Mitarbeiter; „cannot establish causality") (BMJ Open 2022). Die stärkste unabhängige Mortalitätsstudie zur Erstzertifizierung — Silber 2002, 144.883 Medicare-Patienten Pennsylvania 1991–1994, alt — ergibt **adj. OR 1,13 (1,00–1,26)** für Tod bei nicht-zertifizierten Anästhesisten in der Karrieremitte, mit dem Vorbehalt der Autoren, dass die Krankenhäuser statt der Ärzte die Ursache sein könnten (B) (Silber 2002). Das Mengen-Outcome-Verhältnis ist dagegen um eine Größenordnung stärker: Low-Volume-Chirurgen fast **vierfache Mortalität** bei Pankreasresektion (B, Medicare ~474.000 Patienten, 2003, alt) (Dartmouth 2003); in Deutschland signifikante Mengen-Outcome-Beziehung bei 20 von 25 Prozeduren, offenes Bauchaortenaneurysma 4,7 % vs. 7,8 % risikoadjustierte Mortalität (A/B, DRG-Vollerhebung 2009–2014, alt) (Nimptsch & Mansky 2017). Der G-BA steuert deshalb über Mindestmengen (Struktur), nicht über Personenzertifikate.

Fortbildungspflicht und Revalidation. Für § 95d SGB V (250 Punkte je 5 Jahre) wurde keine Outcome-Evaluation gefunden; Ende 2008 hatten in Bayern nur 23 % der Vertragsärzte die Punkte erreicht (A, KVB-Verwaltungsdaten, alt) (Bayerisches Ärzteblatt 2009). Die obere Plausibilitätsgrenze liefert Cochrane: Fortbildungsveranstaltungen verbessern professionelle Praxis um **RD 6,79 %** und Patienten-Outcomes um **RD 3,30 %** (B, 215 Studien, Suche bis 2016, alt) (Cochrane CD003030) — und das für gemessene Interventionen, nicht für Punktesammeln. Die GMC-finanzierte UMbRELLA-Evaluation der UK-Revalidation (Daten zu ~281.000 Ärzten, > 85.000 Survey-Teilnehmer, 2015–2017, alt) beansprucht keine Verbesserung der Patientensicherheit und findet keinen kausalen Link zum Rückgang der Fitness-to-Practise-Referrals (B/D) (GMC UMbRELLA).

Berufskraftfahrer (RL 2003/59/EG, BKrFQG, Driver CPC). Keine gefundene Evaluation mit Unfallkennzahlen. Die britische Regierung räumte 2013 schriftlich ein, ein kausaler Nachweis sei „very difficult" (E, Regulator, alt) (GOV.UK); im Call for Evidence sahen von 748 Antwortenden **71 „significantly increased safety", 328 „positive but marginal", 349 „no road safety benefit"** (D, 2013, alt) (GOV.UK PDF). Der DCPC-Review 2021–22 nennt Lkw-Unfälle 2007–2020 **−68 % vs. Pkw −54 %**, ein unbereinigter Vorher-Nachher-Vergleich ohne Kontrolle für AEBS, Fahrleistung oder Konjunktur (C-Zeitreihe, E-Schluss) (GOV.UK Final Report). Das Impact Assessment der Kommission selbst: „the attributed expected effect are quite limited" (E, 2017, alt) (COM(2017)47). Deutsche Verbände (DVR, Fahrlehrerverbände) stellen die Wirkungsfrage nicht; die Behauptung, das System verbessere die Sicherheit „europaweit erfolgreich", steht ohne Beleg (E, Interesse: Fahrlehrerverbände verkaufen die Schulung) (Positionspapier). Die US-ELDT-Regel (2022) hat keine Behind-the-Wheel-Stunden und keine Wirkungsdaten; ATRI startete erst 2025 eine Studie (C/E) (Truckinginfo 2025).

Piloten (1.500-Stunden-Regel). Keine Studie mit Unfall-/Zwischenfallraten als Zielgröße. Rupp & Tan verwerfen Incidents explizit (480 in 2010–2018, zu wenig Varianz) und messen Verspätungen (+14,8 % kurzfristig) (B, Working Paper 2024) (Rupp/Tan). Die Pilot Source Study (n = 6.734, Einstellungen ab 08/2013, alt) zeigt **weniger Trainingsabbrüche bei ≤ 1.500 h als bei > 4.500 h** (B, Autoren an Hochschul-Flugprogrammen mit Interesse am R-ATP-Weg) (Smith et al. 2016). Die ALPA-Zahl „99,8 % weniger Tote" (E, Interesse: Angebotsverknappung) und die AAF-Zahl „alle sechs Regionalcrashs 2003–2009 mit > 1.500-h-Kapitänen" (E, Autor Ex-Board Mesa Air) sind beide nicht falsifizierbar (ALPA, AAF). CRM-Training: große Effekte auf Einstellung und Verhalten in 16 Studien, Unfallwirkung nicht berichtet (B, alt) (O'Connor 2008).

Lizenzökonomie USA. Strengere Zahnarzt-Lizenzierung: **kein Effekt auf Zahngesundheit, Preise +14–16 %** (B, Air-Force-Rekruten, 2000, Erhebungsjahr nicht belegt, alt) (NBER Digest 1997). Lehrer-Zertifizierung erklärt „at most small" Unterschiede (≈ ±0,02–0,03 SD) bei großer persistenter Varianz innerhalb der Zertifizierungsgruppen (B, NYC 1998–2005, alt) (NBER WP 12155); Zertifizierungstests heben Löhne um 2,4–3,3 %, aber nicht die Qualität (B, SASS 1987–2000, alt) (NBER WP 9545). Das Strukturmodell über 624.697 Beschäftigte

(CPS 2015–2018, alt) schätzt bei Vollumstellung auf Lizenz **Lohn +15 %, Beschäftigung −29 %, Wohlfahrtsverlust 12 % des Berufs-Surplus** (B) (Minneapolis Fed SR 590). Umgekehrt verbessert Deregulierung Outcomes: autonomerer Scope für Dentalhygieniker → weniger Zahnextraktionen (B, 2001/2014, alt) (CHWS), eigenständige Praxis für Nurse Practitioners → weniger Notaufnahmen (B, alt) (JHE 2018).

Rettungsdienst. Die Gegenrichtung: ALS (Paramedic) vs. BLS mit Defibrillation — Überleben bis Entlassung **5,0 % vs. 5,1 %** (B, OPALS, 5.638 Patienten, alt) (OPALS via Advocate); Medicare 2009–2011 (alt): **13,1 % (BLS) vs. 9,2 % (ALS)**, Propensity-Score (B) (Sanghavi 2015). Höhere Qualifikationsstufe heißt nicht besseres Outcome, wenn die zusätzliche Intervention Zeit kostet.

Organisationszertifikate zeigen Selektion; Medizinprodukte zeigen das Gegenteil der Erwartung

Für ISO 9001 ist die stärkste Evidenz gespalten, und zwar entlang der Kontrollgruppenwahl, nicht entlang der Daten. Corbett et al. (7.238 US-Hersteller, 1987–1997, alt, Matching auf Vor-ROA) finden ROA ~3,9 Prozentpunkte höher drei Jahre nach Zertifizierung (B) (Corbett 2005); Dick, Heras & Casadesús (800 Firmen Baskenland, 1994–1998, alt) zeigen mit den *noch nicht* zertifizierten Firmen als Kontrollgruppe, dass die Vorteile **vor** der Zertifizierung bestanden und danach kein Zugewinn eintrat — gleiches Muster bei Häversjö (Dänemark, 800 Firmen) und Wayhan (USA, 96 Firmen) (B) (Dick et al. 2006). Der Verlust des Zertifikats hat keinen messbaren finanziellen Effekt (B, 143 portugiesische Firmen, alt) (Cândido 2016). **Keine gefundene Studie misst ISO 9001 gegen Defekte, Rückrufe oder Reklamationen.** Qualitative Forschung belegt zeremonielle Adoption (B, 8 Organisationen, alt) (Heras-Saizarbitoria 2011) und Auditoren mit vorgefasstem Ergebnis: „a certificate … does not guarantee conformance" (B, alt) (Kuceja 2017). Die ASPICE-Assessments (34 Assessments, 2012–2019, alt) zeigen einen Prüfbarkeits-Bias: Dokumentations- und Tracing-Praktiken schneiden gut ab, Verifikationskriterien und Wirkungsanalyse am schlechtesten (B, Prozess- statt Produktmaß) (Falcini & Lami 2020). Was leicht auditierbar ist, wird gemacht.

Für ISO 27001 existiert keine Breach-Evidenz, nur eine positive Börsenreaktion auf 111 Ankündigungen (B) (Deane 2019). Die methodisch stärkste Arbeit des Felds testet **Kontrollen, nicht Zertifikate**: Die IV-Regression auf amtlichen israelischen Erhebungsdaten (993 Firmen, Erhebung 2020–2021) zeigt, dass die naive OLS-Korrelation positiv ist (Angegriffene rüsten nach) und erst die Instrumentierung den kausalen Rückgang der Vorfallswahrscheinlichkeit von 80 % auf 42 % bei allen sechs Basiskontrollen freilegt (B, NSF-finanziert) (Gandal et al.). Dieses Endogenitätsproblem gilt für jede künftige „zertifiziert vs. Breach"-Korrelation. Die britische Cyber-Essentials-Evaluation (DSIT 2024) ist das einzige Beispiel einer staatlich evaluierten Zertifizierung mit Ergebnisversuch — und sie räumt ein, Vorfälle nicht messen zu können: **nur 8 % der Nutzer bemerken weniger Vorfälle, 57 % „zu schwer zu beurteilen"** (D, Selbstselektion); die Versicherungszahl „80 % weniger Ansprüche" stammt von ungenanntem Versicherer ohne Risikobereinigung (C/E) (gov.uk 2024).

Medizinprodukte sind das einzige Feld mit allen vier Elementen: harter Ergebnisgröße, kontrafaktischem System, belegtem Anreizmechanismus und regulatorischer Konsequenz. Hwang et al. (309 Hochrisiko-Geräte, CE 2005–2010, Sicherheitsdaten bis 2016, alt): EU-zuerst zugelassene Geräte **27 % Rückruf/Warnung vs. 14 % bei US-zuerst, adjustierte HR 2,9 (1,4–6,2)**, nur Rückrufe HR 4,6 (B) (BMJ 2016). Die BMJ/Telegraph-Undercover-Recherche mit fiktiver Hüftprothese bei 14 benannten Stellen (2012, alt) dokumentiert Wettbewerb um Preis (2.750–50.000 €), Tempo und zugesagte Bestehensquote (C, investigativ) (EurekAlert/BMJ 2012). Nach den Joint Audits verloren mindestens vier von zwölf geprüften benannten Stellen die Benennung, eine wurde suspendiert — darunter EVPU, dieselbe Stelle, die 2012 „cross the line" angeboten hatte (C, 2013, alt) (Clinica via Team-NB). Der Mechanismus — der Geprüfte wählt und bezahlt den Prüfer, die Aufsicht misst die Delegation nicht — kehrt bei Wirecard (APAS-Sanktion EY 2023: zwei Jahre Mandatsverbot, 500.000 €, 7 von 12 Prüfern entzogen sich durch Lizenzrückgabe; A via Fachpresse) (Accounting Today) und Boeing (Kongressbericht 2020: FAA-Management überstimmte wiederholt technische Experten; A via Jurist) (Jurist 2020) wieder. Die deutsche Akkreditierungsaufsicht berichtet Verfahrenskennzahlen (104 Widersprüche 2018, ~0,97 % 2024), aber **keine Entzugs- oder Überwachungsstatistik und keine Wirkungskennzahlen**

(A, BT-Drs. 20/13351, 2024) (Bundestag). CRA, NIS2 und EUCC setzen damit auf einen Mechanismus, dessen Wirksamkeit in keinem Nachbarfeld belegt und in einem (Medizinprodukte vor 2013) widerlegt ist; die EU-Evaluation des Cybersecurity Act lag Stand Februar 2024 ohne Befunde vor (A, kein Ergebnis) (EC Evaluation).

DACH reguliert am dichtesten und misst am wenigsten

Die Länderunterschiede betreffen fast ausschließlich Struktur, nicht Ergebnis.

Lizenzdichte. EU-SOR 2015 (26.640 Interviews, Selbstauskunft, Auswertung B, alt): **Deutschland 32,9 % lizenzierte Beschäftigte — höchster EU-Wert**, Österreich 22,2 %, UK 19,5 %, Finnland 16,7 %, Schweden 15,3 %, Dänemark 14,0 %, EU-Ø 21,9 % (Koumenta & Pagliero). USA ≈ 29 % (2008, alt) bzw. 12–33 % je Bundesstaat (B/D) (Minneapolis Fed). Im OECD-OER-Indikator 2018–19 ist Deutschland das am stärksten regulierte Land des Regressionssamples, Schweden das schwächste; eine Senkung um eine Indikatoreinheit (≈ Deutschland → Schweden) entspricht **+1,5 bis 1,6 Prozentpunkten Produktivitätswachstum** der Durchschnittsfirma (B; Qualität nicht gemessen) (OECD WP 1605). Die Lizenzprämie ist in der EU mit ≈ 4 % deutlich niedriger als in den USA mit 14–18 % — bei höherer EU-Lizenzdichte; ≈ 2/3 der EU-Prämie werden als Signaling, ≈ 1/3 als Rente geschätzt, mit großem Standardfehler (B) (BJIR 2019). Die Hypothese „stark lizenzierte Länder haben bessere Verbraucherergebnisse" wurde von keiner Quelle (OECD, Kommission, Koumenta/Pagliero, BIS) mit Outcome-Daten getestet — sie ist unbelegt, nicht widerlegt.

Deutschland: Handwerk. Die HwO-Novelle 2004 (53 von 94 Handwerken zulassungsfrei) ist das größte natürliche Experiment der Region. Robust belegt sind Gründungsboom (B1-Betriebe 74.940 → 244.273 bis 2017; C, ZDH-Register) (DHZ 2019), **Austritte B1 +2,0–2,5 Prozentpunkte** (B, DiD, Common-Trends verletzt, ifh-nah) (Runst et al. 2019) und der Rückgang der 5-Jahres-Überlebensrate **69 % (Jahrgang 2001) → 46 % (Jahrgang 2007)** (B mit Interessenvermerk ifh, HWK-Daten, alt) (Stellungnahme Bizer 2019). **Kein Gutachten, auch kein handwerksnahes, fand kausale Belege für einen Qualitätsverlust**; die Monopolkommission stellte 2019 fest, es gebe „keine empirischen Studien" dazu (B

mit Vermerk) (DHZ 2019). Ausbildungseffekte sind umstritten: Koch & Nielen (DiD) finden „keine signifikanten Effekte" (B, FES-publiziert) (IfM Bonn), das ifh negative Effekte ab 2009, konfundiert mit der AEVO-Aussetzung 2003–2009. Der freiwillige Meistertitel wirkt als Signal: auf MyHammer (5.144 Firmen, 2013–2015, alt) **+2 Prozentpunkte beste Gesamtbewertung (93 % vs. 91 %)** bei 92 % Deckeneffekt (B mit Vermerk ifh) (Fredriksen et al.). Die Rückvermeisterung 2020 (12 Gewerke) begründet sich im Gesetzentwurf 19/14335 mit Gefahrgeneigtheit, Kulturgüterschutz und ZDH-Ausbildungszahlen (Azubis −32 %, Meisterprüfungen −72 % in den 12 Gewerken 2003–2018; A-Dokument auf C-Daten) und enthält **keine Statistik zu Mängeln, Reklamationen oder Qualität**; die Evaluationsklausel 2025 misst erneut nur Berufsbildungs-, Betriebs- und Beschäftigtendaten (BT-Drs. 19/14335). In der Anhörung mit neun Sachverständigen nannte keiner Qualitäts- oder Mängelzahlen (A, Protokoll) (Bundestag kw26 2019).

Deutschland: Projektmanagement, Fahrschule, Medizin. Im PM-Markt dominiert IPMA/GPM (> 30.000 Zertifikate 2012 vs. ~10.300 PMI, D/E, alt) mit einem 8-Stunden-Assessment plus Prüfungsgespräch und Erfahrungsnachweis auf Level A–C — formal näher an einem Kompetenznachweis als PMP oder CSM, aber ohne jede Outcome-Messung; die GPM-Studien fragen Gehalt und Karriere ab (E, Interesse: eigenes Zertifikat) (oose-Vergleich 2013, GPM 2024). Für den deutschen Fahrschulzwang wurde **keine Studie** gefunden, die ihn gegen Länder ohne Pflicht isoliert; Assailly verneint einen Systemeffekt GB/D/F auf das Jungfahrerrisiko explizit (B, 1999, alt) (TRID). Auf Aggregatebene hat Deutschland mit 32 % den höchsten Anteil junger Fahrer an Fahrer-Toten in der EU (A/C, ERSO, WHO-Daten ohne Jahr) (ERSO) und 3,35 Verkehrstote je 100.000 vs. UK 2,61, Schweden 2,17, Norwegen 2,14, USA 12,84 (A, Datenjahr 2022) (BITRE 2022) — kein Kausalbeleg, aber auf Aggregatebene unvereinbar mit „teure Pflichtausbildung → messbar sicherere Anfänger". In der Medizin sind Facharzt, Zusatzbezeichnung und § 95d ausschließlich input- und zufriedenheitsevaluiert (BÄK-Weiterbildungsevaluation: Gesamtnote 2,54 → 2,44, D, alt) (DÄ 2015); die Schlichtungsstelle Norddeutscher Ärztekammern analysiert in 41.138 Entscheidungen 2000–2015 den Facharztstatus nicht als Variable (A ohne Zertifikatsbezug, alt) (A&I 2016). Keine deutsche oder

europäische Studie zu Prüfungsscores (Staatsexamen, Facharztprüfung) und späteren Beschwerden wurde gefunden.

Österreich und Schweiz. Die Mehrphasenausbildung (seit 2003) zeigt −10,3 % Unfälle bei 18–19-Jährigen 2004–2005 (C/E, Ministerium, einfacher Vorher-Nachher, alt) (Parlament AT) bzw. −28 % bei 18-Jährigen 2000–2002 vs. 2004–2006 (C, Evaluator im Auftrag, alt) (WKO/Bartl); die BASt-finanzierte Methodenkritik (Saarland 2013) hat ihre Ergebnisse nicht veröffentlicht, und der ÖAMTC-Experte selbst schreibt den Rückgang „großteils" der Fahrzeugsicherheit zu (E, Interesse: ÖAMTC verkauft Mehrphasen-Kurse) (ÖAMTC 2023). Die WKO-Prüfungsstatistik 2024 nennt eine Erfolgsquote von 66,5 % bei Meister-/Befähigungsprüfungsmodulen (C) (WKO 2024), keine Reliabilität. Die Schweiz kennt ≈ 160 reglementierte Berufe, nur 13 mit zwingendem Berufsbildungsabschluss, Rest kantonal uneinheitlich; keine Outcome-Messung (A/B, Ecoplan 2025) (edudoc.ch). Das relativiert die deutsche Begründung „Meisterpflicht = Ausbildungsgarant": Ein starkes duales System funktioniert ohne flächige Zulassungspflicht.

USA. Das Land mit der meisten Outcome-Forschung (Medicare, FARS, NTSB, NCDR-Register) und zugleich mit der stärksten Lizenz-Rentenökonomie. Die Boards produzieren die Positiv-Evidenz überwiegend selbst (ABIM, ABS, NBME/FSMB), unabhängige Replikationen fehlen oder sind null. DoD 8140 lockert die reine 8570-Zertifikatspflicht zugunsten von Erfahrung und Ausbildung — eine Verwaltungsentscheidung ohne gefundene Ergebnisevaluation (A, GAO-Berichte thematisieren Wirkung nicht) (Defense Systems 2011). Das Deming-Feldexperiment (AER 2016, alt) liefert den direktesten Beleg für die Screening-Funktion: In Gesundheitsjobs **mit** Zertifikats- und Lizenzpflicht verschwindet der Herkunftsnachteil eines For-Profit-Abschlusses (−57 % Rückrufe ohne Pflicht, 0 % mit Pflicht) (B) (Harvard/Katz) — ein Zertifikat ist für den Arbeitgeber nützlich, wenn es Anbietervarianz neutralisiert.

UK. Der britische Staat behandelt das Herkunftslandzertifikat PRINCE2 seit 2012 nicht mehr als Kompetenznachweis: eigene Project Delivery Profession, Functional Standard GovS 002 (2018), > 2.000 MPLA/PLP-Absolventen und eine erfahrungs- und delivery-record-basierte Akkreditierung; PRINCE2 wird in keiner amtlichen

Quelle genannt (A) (IPA 2018, Smallwood 2020). Outcome-Daten zu dieser Akkreditierung fehlen ebenfalls. Der UK-Freiwilligenregister für Fitnesstrainer hatte „little impact" auf Qualifikation und Lohn, Einstiegsqualifikationen sind in ≈ 10 Tagen erwerbbar, während die französische Lizenz mit 350–500 h Praxispflicht höhere Skills stützt (B, qualitativ, 2009–2011, alt) (Cardiff ORCA).

Nordics. Die Länder mit der niedrigsten Lizenzdichte (14–15 %) und den niedrigsten Verkehrstotenraten (2,14–2,17) liefern die saubersten Fahrausbildungs-Experimente: Schwedens Übungsalter 16 wirkt (B), Norwegens nicht (B/D), norwegisches und finnisches Pflicht-Schleudertraining erhöhte Unfälle (B/C). ETSC schreibt Schwedens niedrige Jungfahrer-Todesrate (53/Mio. vs. EU 103/Mio., 2008–2010, alt) der niedrigen Lizenzquote 18–22 zu, nicht der Ausbildung (C) (ETSC PIN 21). SAFe-Adoptionsdaten aus Finnland und Dänemark sind D/B ohne Outcome.

Widersprüche zwischen und innerhalb der Notizen

Die Notizen enthalten Widersprüche, die hier offen stehen bleiben statt geglättet zu werden.

ISO 9001 — Treatment oder Selektion. Corbett et al. (Matching auf Vor-ROA, Zugewinn ~3,9 Pp) und Dick/Häversjö/Wayhan (Noch-nicht-Zertifizierte als Kontrolle, kein Zugewinn) nutzen überlappende Zeiträume und Länder und kommen zu entgegengesetzten Schlüssen. Die IT/Org-Notiz ordnet dies als Methodenstreit um die Kontrollgruppenwahl ein. Cândido (kein Effekt bei Dezertifizierung) stützt die Selektionsseite. Der Bericht übernimmt keine Seite als gesichert.

Board-Zertifizierung Ärzte — ABIM-Serie vs. unabhängige Nullbefunde. Norcini 2000/2002/2022 (E) berichten 15–19 % relative Mortalitätsreduktion; Hayes 2014 (B) und Gray 2014 (E) finden beim klinischen Outcome null, Simpkins 2007 (RCT, E) negativ. Sharp 2002 liest die Literatur als „more than half support an association" (C/E), Grosch 2006 reanalysiert dieselbe Sammlung und findet „no credible link" (B). Diese Dissonanz ist der Kern der MOC-Kontroverse und wird im Bericht als solche geführt.

Cyber Essentials — 80 % vs. 92 %. Die DSIT-Evaluation nennt 80 % weniger Versicherungsansprüche (NCSC Annual Review 2023, Daten 2022), das NCSC-Blog 92 % (Quelle: gebündelter Versicherer, Jahr ungenannt). Die Abweichung ist in keiner Quelle erklärt (gov.uk 2024, NCSC).

PMI-Gehaltsprämie — 33 % vs. 17 %. Zwischen Earning Power 2023 und 2025 fiel die PMP-Prämie um die Hälfte, ohne dass PMI den Methoden- oder Stichprobenwechsel erklärt (E).

BF17 — BASt vs. DVR. BASt M 218 nennt −17 % (konfunder-adjustiert) bzw. −22 % (fahrleistungsbereinigt) Unfallbeteiligungen; der DVR-Beschluss 2024 zitiert „−23 % Unfälle, −22 % Verkehrsauffälligkeiten" als BASt-Zahlen. Die Fahr-Notiz führt die DVR-Angabe als E (Interesse: Fahrlehrer-/Versicherer-Mitglieder).

Österreich Mehrphase — −10,3 % vs. −28 %. BMVIT (2004–2005 vs. zwei Vorjahre, 18–19-Jährige) und Gatscha/Brandstätter (2004–2006 vs. 2000–2002, 18-Jährige) nutzen verschiedene Fenster und Altersgruppen; beide sind unbereinigte Vorher-Nachher-Vergleiche. Die BASt-finanzierte Methodenkritik ist nicht veröffentlicht.

Cochrane GDL — Datierung. Die cochrane.org-Zusammenfassung nennt 01.04.2022, die Cochrane-Library-Versionsseite datiert pub3 auf 05.10.2011; Datensuche in jedem Fall bis 2009.

Tamblyn 2007 — Stufe und Zahlen. Die Psychometrie-Notiz führt die Studie als B (A-Registerdaten) mit RR 1,38 (Kommunikation) und 1,51 (Clinical Decision Making); die Medizin-Notiz führt dieselbe Studie über die MCC-Kurzfassung als „B mit COI-Vermerk" (MCC-finanziert) mit RR 1,39 für den MCCQE1-Gesamtscore und RR 1,38 für den Kommunikationssubscore. Die Zahlen sind kompatibel (verschiedene Subscores), die Stufenvergabe unterscheidet sich im COI-Vermerk; der Bericht übernimmt B mit COI-Vermerk.

Norcini 2022 — N. Abstract 184.115 vs. Ergebnisteil 184.155 Hospitalisierungen (interne Inkonsistenz der Primärquelle).

Fahrschule — RCTs vs. Nebraska. Drei RCTs (RR ≈ 1,0) gegen die Nebraska-Registerstudie (n = 151.880, 2003–2010, alt), in der Fahrschulabsolventen OR 1,21

weniger Unfälle und OR 1,79 weniger Verstöße haben als Eltern-Fahrtenbuch (B, Selbstselektion explizit nicht ausgeschlossen: Fahrtenbuch häufiger ländlich, männlich, einkommensschwach) (Nebraska DOT). Randomisierte Evidenz schlägt hier die Beobachtungsevidenz.

Fredriksen et al. — Meisteranteil. Text ¼ vs. Tabelle 0,41 (interne Inkonsistenz der Primärquelle, von der Notiz vermerkt).

Foote Partners 2006. Artikel intern widersprüchlich („dreimal schneller" vs. 4,4 % vs. 2,6 %) (E).

Reglementierte Berufe vs. Lizenzanteil. Register 2012: DE 86, UK 131, AT 151 reglementierte Berufe — aber Lizenzanteil DE 33 % vs. UK 19,5 %. Deutschland reguliert wenige, beschäftigungsstarke Berufe; die Berufszählung ist als Strenge-Maß untauglich.

Übersichtstabelle

Zertifikat / Lizenz	Gemessene Ergebnisgröße	Befund	Beste Belegstufe (Zeitraum)	Land
Hebammen-Lizenz	Müttersterblichkeit	nachweisend (−6–7 %, Floor-Effekt)	B (1900–1940, alt)	USA
Graduated Driver Licensing	Unfälle 16-Jährige	nachweisend (−15 bis −26 %), teils Verlagerung auf 18 J.; pro Lizenzinhaber ≈ 0	B/A (1986–2009, alt)	USA/CA/AU/NZ
Begleitetes Fahren ab 17	Unfallbeteiligung 1. Jahr	nachweisend (−17/−22 %), Selbstselektion offen	A/B (2007–2011, alt)	DE

Zertifikat / Lizenz	Gemessene Ergebnisgröße	Befund	Beste Belegstufe (Zeitraum)	Land
Übungsalter 16 (Schweden)	Unfälle je km	nachweisend für Nutzer (−24 bis −40 % adj.)	B (1993–1996, alt)	SE
Übungsalter 16 (Norwegen)	Unfallrisiko	Null	B/D (1998–1999, alt)	NO
Formale Fahrausbildung	Unfallbeteiligung	Null (RR 1,01–1,10), frühere Lizenz	B, 3 RCTs (1982–84, alt)	AU/US/NZ
Schleudertraining (Pflicht)	Unfallrate	negativ (+12 % Pkw, +45 % RTW)	B via C (alt)	NO/FI
Praktische Fahrprüfung (DPE)	Unfälle/ Verstöße 2 J.	Null (n = 394.101)	A/B (1993–1997, alt)	USA
Mehrphasenausbildung	Unfälle 18–19 J.	unklar (−10 bis −28 %, unbereinigt)	C/E (2004–2006, alt)	AT
Fahrschulpflicht (D) vs. UK/USA	Jungfahrer-Fatalität	keine Messung; Aggregat spricht nicht dafür	A/B (1999; 2012; 2022)	DE/UK/US
Berufskraftfahrer RL 2003/59 / CPC / BKrFQG	Unfälle Lkw/Bus	keine Messung; Regulator räumt fehlenden	D/E (2013–2022)	EU/UK/DE

Zertifikat / Lizenz	Gemessene Ergebnisgröße	Befund	Beste Belegstufe (Zeitraum)	Land
		Nachweis ein		
US CDL/ELDT	Unfälle	keine Messung	E (2016–2025)	USA
1.500-h-Regel Piloten	Unfall-/ Zwischenfallrate	keine Messung (zu seltene Ereignisse); Trainingserfolg ohne Vorteil hoher Stunden	B (2013–2018)	USA
Instrument Rating (GA)	Wetterunfälle	nachweisend (p = 0,002); Flugstunden n.s.	A/B (2003–2004, alt)	USA
CRM-Training	Einstellung/ Verhalten; Unfälle	Einstellung/ Verhalten ja; Unfälle nicht berichtet	B (alt)	intl.
Staplerschein (OSHA)	Verletzungen	keine Messung	E/D	USA
CSM / PSM / SAFe SPC	Projektergebnis, Defekte	keine Messung (Variable nicht erhoben)	—	intl.
PMP / PRINCE2 / IPMA	Projekterfolg	gemischt; 1 Feldstudie positiv korrelational,	B (2004–2021)	US/ZA/intl.

Zertifikat / Lizenz	Gemessene Ergebnisgröße	Befund	Beste Belegstufe (Zeitraum)	Land
		Mehrheit null (Selbstauskunft)		
ISTQB / IREB CPRE	Defektfindung, RE-Qualität	keine Messung („no controlled study“)	B/D (2026)	intl.
Six Sigma Black Belt	Projektergebnis	Programm-Effekt B; Person: keine Messung	B (alt)	USA
CISSP / CCNA / Cloud-Zertifikate	Vorfallsrate, Leistung	keine Messung; nur Lohn/Nachfrage	B-Methodik auf D-Daten (2004, alt; 2023)	USA/NZ
DoD 8570-Zertifikatspflicht	Arbeitsqualität (wahrgenommen)	negativ (qualitativ, N = 30)	B/D (ca. 2013, alt)	USA
Facharzt-/Board-Erstzertifizierung	Mortalität	klein positiv (OR 1,13 Nicht-Zertifizierte) mit Umfeld-Vorbehalt; ABIM-Serie 15–19 % (E)	B (1991–1994, alt)	USA
ABIM MOC (Rezertifizieru	ACSH, Qualitätsmaß	Null (Quasi-Experiment,	B/E (1999–2013, alt)	USA

Zertifikat / Lizenz	Gemessene Ergebnisgröße	Befund	Beste Belegstufe (Zeitraum)	Land
ng)	e, Asthma-Outcome	unabhängig, RCT)		
Lizenzprüfungsscores Ärzte (MCCQE, USMLE, ABIM)	Beschwerden, Disziplin, Peer-Review	nachweisend als Messinstrument (RR 1,38–1,51/2 SD; OR 0,75/SD)	B (1993–2017)	CA/USA
UK Revalidation	Patientensicherheit, FTP	Null	B/D (2015–2017, alt)	UK
§ 95d Fortbildungspflicht	Patienten-Outcome	keine Messung; CME generell RD 3,3 %	A/B (alt)	DE
Pflege-Fachzertifizierung	30-Tage-Mortalität	klein positiv (−2 %/+10 % zertifizierte BSN), nur mit BSN	B (2005–2006, alt)	USA
Chirurgenvolumen (Vergleich)	Mortalität	nachweisend (bis 4-fach; 20/25 Prozeduren)	A/B (2003; 2009–2014, alt)	USA/DE
ALS vs. BLS (Rettungsdienst)	Überleben	Null bis negativ	B (2004; 2009–2011, alt)	CA/USA
Zahnarzt-Lizenzstrenge	Zahngesundheit, Preise	Null Qualität; Preise +14–16 %	B (alt)	USA

Zertifikat / Lizenz	Gemessene Ergebnisgröße	Befund	Beste Belegstufe (Zeitraum)	Land
Lehrer-Zertifizierung	Schülerleistung	Null im Mittelfeld (±0,03 SD); Floor-Effekt am linken Rand	B (1987–2007, alt)	USA
Meisterpflicht (HwO 2004/2020)	Qualität/ Mängel	keine Messung; Überlebensrate 69 → 46 %; freiwilliger Meister +2–3 Pp Bewertung	B/C (2001–2022)	DE
ISO 9001	ROA, Überleben, Verletzungen	gespalten: Zugewinn vs. nur Selektion; Dezertifizierung ohne Effekt; keine Defektmessung	B (1987–2009, alt)	US/ES/DK/PT
ISO 27001 / SOC 2	Breach-Rate	keine Messung; nur Börsenreaktion	B	intl.
Cyber Essentials	Vorfälle, Versicherungsansprüche	nur Selbstauskunft; Versicherungs	D, C/E (2022–2023)	UK

Zertifikat / Lizenz	Gemessene Ergebnisgröße	Befund	Beste Belegstufe (Zeitraum)	Land
		zahl ohne Risikobereinigung (80 %/92 %)		
CE/benannte Stellen (Medizinprodukte)	Rückrufe/ Warnungen	negativ vs. FDA (HR 2,9)	B (2005–2016, alt)	EU/USA
CMMI / ASPICE	Programmleistung, Prozessrating	Selbstauskunft bzw. Prozessmaß; kein Produktmaß	D/E, B (2005; 2012–2019, alt)	USA/intl.
Akkreditierung (DAkkS)	Wirksamkeit Zertifizierer	keine Messung	A (2024)	DE

Was ein nachweisendes Zertifikat ausmacht

Die Kriterien folgen aus dem Kontrast zwischen den Fällen mit Wirkung und den Fällen ohne, nicht aus Meinung.

Erstens: Konstruktnähe von Prüfinhalt und Praxiskriterium. Wo das Prüfungskonstrukt dem Praxiskriterium entspricht, prädiziert der Score — Kommunikationsscore → Kommunikationsbeschwerden (RR 1,43), klinische Entscheidung → aufrechterhaltene Beschwerden (RR 1,51), Instrument Rating → Wetterunfälle. Wo es nicht entspricht, prädiziert er nicht — Fahrtest-Score → Unfälle (n = 394.101, p = 0,62), Theorietest → Bußgelder ja, Praxistest → Bußgelder nein (Litauen, B, alt) (TOTS 2018), Gesamtflugstunden → Unfälle (p = 0,788). Multiple Choice über ein Dokument (Scrum Guide, PMBOK, Syllabus) hat per

Konstruktion keine Nähe zu Lieferzeit, Defektrate oder Projekterfolg; folgerichtig existiert dort keine Validitätsstudie.

Zweitens: Berichtete prädiktive Validität gegen ein Registerkriterium. Die Sackett-et-al.-Korrektur (B, 2022) setzt den Maßstab: strukturierte Interviews ρ ≈ 0,42, Fachwissenstests 0,40, Arbeitsproben 0,33, Berufserfahrung in Jahren 0,07 (SIOP). Ein Zertifikat ist so valide wie sein Prüfungsverfahren, nicht wie seine Dauer oder sein Erfahrungsnachweis in Jahren. Die medizinischen Lizenzprüfungen sind die einzigen im gesamten Material, für die eine Validität gegen A-Daten (Aufsichtsregister) publiziert ist. Kein Management-, Agile-, IT- oder Handwerkszertifikat berichtet eine solche Zahl.

Drittens: Kontinuierlicher Score statt binärer Urkunde. Das stabilste Signal in der Medizin liegt im Score, nicht im Bestehen: Nicht-Teilnehmer der MOC-Prüfung zeigten keinen Qualitätsunterschied, die Assoziation hängt an der Prüfungsleistung (E, Holmboe 2008) (NBPAS 2018). Ein Zertifikat, das nur „bestanden" ausweist, wirft die prädiktive Information weg. Keine Regression-Discontinuity-Studie am Pass/Fail-Cutpoint wurde gefunden — der Nachweis, dass die Schwelle selbst etwas bewirkt, fehlt überall.

Viertens: Inter-Rater-Reliabilität ist erreichbar, wird aber nicht berichtet. Im Keele-OSCE (116 Studierende, 67 Examinatoren) erklärt die Examinatorenkohorte 5,7 % der Varianz und verschiebt ≈ 6 % der Bestehensentscheidungen (B, 2019) (Yeates 2019); Produktbewertung durch mehrere Fachjuroren erreicht ICC > 0,90 (E-Quelle, Design transparent) (IJ 2010). Ein IRR ≥ 0,7 ist für praktische Prüfungen also machbar — gängige Meister-, Gesellen-, IPMA- oder Facharztprüfungen berichten keine. Kammern veröffentlichen Bestehensquoten (WKO 66,5 %), keine Reliabilität.

Fünftens: Exposition unter Aufsicht ist Bestandteil, nicht Unterricht. Alle positiven Fahrbefunde (GDL, BF17, Schweden 16) enthalten beaufsichtigte Praxis in Stunden oder Monaten (118 h statt 41–47 h); alle Null- und Negativbefunde (RCTs, Schleudertraining, DPE-Fahrtest) variieren nur Unterricht oder Prüfung. In der Medizin folgt Outcome dem Volumen, nicht dem Titel; supervidierte Assistenten erreichen Facharzt-Komplikationsraten (Revisionen 2,0 % vs. 2,5 %, p = 1,0; B, 1.480 Hüft-TEP 2013–2016, alt) (Orthopäde 2022).

Sechstens: Floor-Funktion in gefahrgeneigtem Feld ohne vorherigen Standard. Hebammen 1900–1940 und der linke Rand der Lehrerverteilung sind die einzigen Fälle, in denen die Schwelle selbst wirkt. Das Kriterium der Gefahrgeneigtheit, das der Gesetzgeber 2019 für die Rückvermeisterung anführt, ist damit das einzige empirisch tragfähige Lizenzkriterium — mit der Einschränkung, dass es für die zwölf betroffenen Gewerke nicht gemessen wurde.

Siebtens: Rezertifizierung muss auf Leistungsdaten stehen, nicht auf Zeit oder Punkten. Zeit- und punktbasierte Rezertifizierung ist dreifach null (MOC, Revalidation, CPC), und selbst gemessene Fortbildung bewegt Patienten-Outcomes um RD 3,3 %. Zertifikatsverlust bei ISO 9001 hat keine finanzielle Folge. Eine Rezertifizierung auf Leistungsdaten (Registerergebnisse, Volumen, Beschwerdequoten) wurde in keiner Notiz als implementiert und evaluiert gefunden — sie ist aus der Evidenz die logische Forderung, aber selbst ungetestet.

Achtens: Unabhängigkeit des Prüfers vom Geprüften und Messung der Delegation. Der einzige Systemvergleich mit harter Ergebnisgröße (Hwang HR 2,9) zeigt die kommerzielle Drittzertifizierung unterlegen; die Undercover-Recherche zeigt den Mechanismus; Wirecard, Boeing und der SOC-2-Fall 2026 (E) zeigen ihn außerhalb der Medizinprodukte. Wo die Aufsicht die Zertifizierer tatsächlich prüfte (Joint Audits 2013), fiel ein Drittel bis die Hälfte durch. Ein Zertifikat, dessen Aussteller vom Zertifizierten bezahlt wird und dessen Akkreditierer keine Entzugsstatistik führt (DAkkS), hat keinen belegten Nachweiswert.

Neuntens: Veröffentlichte Pass-Raten und Trennschärfe. Kein Anbieter von CSM, PSM, PMP, CPRE oder ISTQB veröffentlicht Pass-Raten. Das ist mit einem Geschäftsmodell vereinbar, in dem der Kurs das Produkt ist und die Prüfung die Rechtfertigung. Eine Prüfung, die fast jeder besteht, trennt nicht — und ohne Trennschärfe gibt es keine Validität, die man messen könnte.

Ein Zertifikat, das keines dieser neun Merkmale erfüllt, ist nach der vorliegenden Evidenz ein Label: ein Screening-Filter für Personalabteilungen (Deming, Starkweather), eine Lohnprämie (Quan, Kleiner) und ein Knappheitssignal, das mit der Massenverbreitung verschwindet (775.000 PSM I).

Offene Lücken

Keine Outcome-Messung existiert nach den Notizen für folgende Zertifikate und Fragen. Die Lücken sind sortiert nach dem Umfang des betroffenen Marktes.

Scrum-Master- und SAFe-Zertifikate (CSM, PSM, SPC) — keine einzige Studie, weder positiv noch negativ, misst sie gegen Defektrate, Lieferzeit oder Projekterfolg; die größten Datensätze (4.000 Teams) erheben die Variable nicht. ISTQB, IREB CPRE, ITIL — keine kontrollierte Studie zu Defektfindung, Anforderungsqualität oder Betriebsstabilität; ITIL wurde in der Recherche gar nicht untersucht. IT-/Security-Personenzertifikate (CISSP, CEH, CCNA, CCIE-Lab, OSCP, CKA, RHCE, Cloud) — keine Vorfalls-, Fehler- oder Vorgesetztenbewertungsstudie; die These „Lab-Prüfung ist valider als Multiple Choice" ist nur Anbieterbehauptung (E). ISO 27001, SOC 2, FedRAMP — keine Breach-Inzidenz-Studie zertifiziert vs. nicht zertifiziert. ISO 9001 / IATF 16949 — keine Studie gegen Defekte, Rückrufe oder Reklamationen; VDA-/OEM-Feldausfalldaten nach Zertifikatsstatus nicht öffentlich. CMMI, Automotive SPICE, ISO 26262-Prozesszertifizierung, TÜV Functional Safety Engineer, intacs-Assessor — kein Produkt- oder Feldqualitätsmaß.

Berufskraftfahrer-Qualifikation (RL 2003/59, BKrFQG, Driver CPC, US ELDT), ADR-Gefahrgutschulung, Staplerschein, Kranführer — keine Evaluation mit Unfallkennzahlen; FMCSA-Vergleich Illinois/Indiana ohne gefundenes Ergebnis. Deutscher Pflicht-Fahrschulunterricht — keine Studie, die ihn kausal gegen Länder ohne Pflicht isoliert; Unfallrate je km im ersten Fahrjahr D vs. UK nicht gefunden. 1.500-Stunden-Regel, MPL vs. ATPL, Flight Review (BFR) — keine Unfall-/Zwischenfallraten je Flugstunde vor/nach Regeländerung.

Deutsches Medizinsystem — Facharztanerkennung, Zusatzbezeichnungen, Fachkunden (Strahlenschutz, Rettungsmedizin), § 95d: keine Patienten-Outcome-Evaluation; Behandlungsfehlerstatistiken (Schlichtungsstellen, MD, Kassen) schlüsseln nicht nach Facharztstatus auf; keine europäische Studie zu Staatsexamens- oder Facharztprüfungsscores und späteren Beschwerden. EU-Richtlinie 2005/36/EG — Ausbildungsdauer als Proxy ohne Outcome-Validierung. Schweißer (EN ISO 9606), Elektro-Sachkunde (VDE), Strahlenschutz — keine Felddaten; Zertifizierung beruht auf Prüfstück. NREMT — nur Prüfung-zu-Prüfung-Prädiktion.

Handwerk — keine deutsche Studie mit Mängel-, Gewährleistungs- oder Schadensdaten (Bauschadenberichte, Versicherer, Schlichtungsstellen) für B1- vs. A-Gewerke; die Evaluation 2025 zur Rückvermeisterung misst erneut nur Ausbildungs-, Betriebs- und Beschäftigtenzahlen; Inter-Rater-Reliabilität von Meister- und Gesellenprüfungen nicht berichtet. Länderübergreifend — keine Studie testet Lizenzanteil (DE 33 % vs. SE/DK 14–15 %) gegen Verbraucher-Outcomes (Schäden, Beschwerden, Bauschadenraten); keine Studie regressiert Ausbildungskosten je Land gegen Anfängerrisiko. Akkreditierung (DAkkS, UKAS) — keine Entzugs-, Überwachungs- oder Wirkungskennzahlen; ob akkreditierte Zertifizierer bessere Ergebnisse liefern als nicht akkreditierte, ist nicht prüfbar. EUCC/CRA/NIS2 — Ergebnis der EU-Evaluation (angekündigt H2 2024) nicht abgerufen; Zahl ausgestellter EUCC-Zertifikate unbekannt.

Methodische Lücken quer durch alle Felder: Keine Regression-Discontinuity-Studie am Pass/Fail-Cutpoint irgendeines Zertifikats; kein RCT oder Quasi-Experiment zu irgendeinem Management-, Agile- oder IT-Zertifikat; keine Meta-Analyse zur prädiktiven Validität beruflicher Zertifikate gegen Arbeitsleistung; keine Studie, die eine leistungsdatenbasierte Rezertifizierung implementiert und evaluiert.

Nicht abgerufene Primärquellen, die das Bild ändern könnten: Chen/Krumholz 2006 (unabhängige Medicare-AMI-Kohorte zu Board-Zertifizierung), Boiral 2012 und Heras-Saizarbitoria & Boiral 2013 (ISO-9000-Reviews), Naveh & Marcus 2005, Sackett et al. 2022 Volltext, Kleiner & Kudrle Volltext, Koch & Nielen Volltext, Malone & Brünken 2013 (Mehrphasen-Methodenkritik), SWOV-Faktenblatt Fahrausbildung, Starkweather & Stevenson 2011 Primärtext.

Schluss

Die Evidenz zeichnet ein konsistentes Bild, das quer zu den üblichen Debatten liegt. Die Frage „wirkt das Zertifikat?" ist in den meisten Feldern falsch gestellt, weil das Zertifikat drei verschiedene Dinge bündelt, die getrennt zu bewerten sind: eine Prüfung (die als Messinstrument valide sein kann, wie in der Medizin, oder nicht, wie beim Fahrtest), eine Zugangsregel (die als Floor in gefahrgeneigten Feldern wirkt und sonst Renten erzeugt) und einen Kurs (für den in keinem Feld eine eigenständige Wirkung belegt ist). Der Markt für Management-, Agile- und IT-

Zertifikate verkauft fast ausschließlich den dritten Bestandteil und verzichtet auf die Veröffentlichung der Daten, die eine Bewertung der ersten beiden erlauben würden. Das ist kein Zufall der Forschungslage, sondern ein Merkmal des Geschäftsmodells.

Für die Praxis folgt daraus eine Umkehr der Beweislast. Ein Zertifikat, dessen Anbieter keine Pass-Rate, keine Validität gegen ein Praxiskriterium und keine Reliabilität veröffentlicht, ist bis zum Gegenbeweis als Label zu behandeln — unabhängig davon, wie viele Inhaber es hat und wie hoch die Gehaltsprämie ist. Umgekehrt zeigt die Medizin, dass Prüfungsscores Praxis prädizieren können, wenn Konstrukt und Kriterium zusammenpassen und das Kriterium aus einem Register stammt; die Konsequenz wäre, den Score statt der Urkunde zu nutzen und Rezertifizierung an Leistungsdaten statt an Punkte zu binden. Genau dieser Schritt ist nirgends implementiert und evaluiert. Die Regulierung (CRA, NIS2, Rückvermeisterung, § 95d) baut unterdessen weiter auf Zertifizierung, ohne die Ergebnisgröße zu definieren, an der sie scheitern könnte.