


# 12 Evaluasi pembelajaran dan dampak

Prof Ir Rudy C Tarumingkeng, PhD

www.rudyct.com/cv.pdf

## Abstrak

*Artikel ini membahas evaluasi pembelajaran sebagai proses sistematis untuk menilai mutu pelaksanaan, penguasaan, transfer ke pekerjaan, dan kontribusi terhadap hasil organisasi. Evaluasi dibedakan dari pelaporan aktivitas, survei kepuasan, dan klaim dampak yang tidak didukung desain pembanding. Artikel menguraikan teori perubahan, pemilihan pertanyaan dan indikator, validitas pengukuran, pengukuran transfer, kerangka evaluasi, desain kausal dan kuasi-eksperimental, analisis kontribusi, ROI, evaluasi digital dan AI, etika data, metode campuran, tata kelola, serta penerapan di Indonesia. Prinsip utamanya ialah menyelaraskan klaim dengan kekuatan bukti: aktivitas menunjukkan jangkauan, asesmen menunjukkan kemampuan pada kondisi tertentu, observasi dan data kerja membantu menilai transfer, sementara klaim dampak memerlukan pembanding dan pengujian penjelasan alternatif. Evaluasi yang baik mendukung keputusan untuk memperluas, memperbaiki, atau menghentikan program, serta melibatkan pekerja dan menjaga privasi.*

### Kata Kunci

Evaluasi pembelajaran; evaluasi pelatihan; transfer belajar; dampak organisasi; validitas pengukuran; desain kuasi-eksperimental; ROI; evaluasi SDM

## Pendahuluan: dari kegiatan belajar menuju bukti perubahan

Organisasi mengalokasikan waktu, tenaga, dan anggaran untuk mengembangkan kompetensi. Namun, jumlah sesi yang diselenggarakan atau peserta yang menyelesaikan kursus belum menunjukkan apakah kemampuan benar-benar bertambah, apakah keterampilan dipakai dalam pekerjaan, atau apakah perubahan tersebut berkontribusi pada hasil organisasi. Evaluasi pembelajaran diperlukan untuk memahami rangkaian itu. Ia membantu pengelola memperbaiki desain, manajer mengatasi hambatan transfer, pimpinan menilai pilihan investasi, serta pekerja memahami kemajuan kompetensinya. Evaluasi yang baik bukan pemeriksaan administratif di akhir program, melainkan bagian dari keputusan sejak perencanaan.

Istilah dampak sering digunakan secara longgar. Peserta mungkin merasa puas, memperoleh nilai tes yang lebih tinggi, atau melaporkan bahwa pelatihan berguna; semua itu merupakan bukti yang bernilai, tetapi belum sama dengan dampak organisasi. Dampak dapat merujuk pada perubahan perilaku, mutu layanan, keselamatan, produktivitas, mobilitas karier, atau indikator sosial yang dituju. Klaim bahwa pelatihan menyebabkan perubahan membutuhkan pembanding yang masuk akal dan pertimbangan terhadap faktor lain. Karena itu, laporan perlu membedakan hasil yang diamati, kontribusi yang masuk akal, serta efek kausal yang telah didukung desain evaluasi.

Evaluasi berfungsi untuk belajar dan mempertanggungjawabkan. Fungsi formatif memberi informasi selama program berjalan agar instruksi, materi, dukungan, atau jadwal dapat diperbaiki. Fungsi sumatif menilai hasil setelah pelaksanaan untuk menentukan apakah program diteruskan, diperluas, diubah, atau dihentikan. Fungsi akuntabilitas menunjukkan bagaimana sumber daya dipakai dan apa yang diperoleh. Fungsi pengembangan pengetahuan membantu organisasi mengetahui apa yang berhasil untuk siapa, dalam kondisi apa, dan melalui mekanisme apa. Keempat fungsi dapat berjalan bersama, tetapi kebutuhan datanya berbeda.

Bab ini menyajikan pendekatan evaluasi yang berangkat dari teori perubahan dan pertanyaan keputusan, bukan dari indikator yang mudah tersedia. Pembahasan mencakup definisi hasil belajar, validitas pengukuran, transfer, desain kausal, kerangka evaluasi, nilai investasi, dampak yang tidak diinginkan, pembelajaran digital dan AI, privasi, tata kelola, serta penerapan di Indonesia. Rujukan klasik tentang evaluasi pelatihan menekankan perlunya menilai beragam hasil kognitif, keterampilan, dan afektif secara sesuai tujuan (Kraiger et al., 1993). Meta-analisis menunjukkan bahwa hasil pelatihan dapat positif pada reaksi, belajar, perilaku, dan hasil, tetapi besar dan kepastian temuan bergantung pada jenis intervensi dan cara mengukur (Arthur et al., 2003).

Gagasan utama bab ini ialah evaluasi harus menjaga rantai inferensi tetap jelas: kegiatan apa yang dilakukan, perubahan apa yang diharapkan, bukti apa yang diamati, penjelasan alternatif apa yang mungkin, dan keputusan apa yang akan diambil. Jumlah klik atau tingkat penyelesaian dapat menjawab pertanyaan jangkauan, tetapi tidak menjawab penguasaan. Nilai pascates dapat menunjukkan performa pada situasi tes, tetapi tidak selalu menggambarkan transfer. Indikator bisnis bisa bergerak setelah pelatihan, tetapi pergerakan itu mungkin dipengaruhi perubahan proses, pasar, teknologi, kepemimpinan, atau komposisi tenaga kerja.

Contoh dalam artikel ini bersifat ilustratif, bukan laporan tentang organisasi tertentu. Tujuannya memberi kerangka yang dapat disesuaikan dengan program pengembangan kepemimpinan, orientasi, keselamatan, layanan pelanggan, literasi digital, maupun reskilling. Evaluasi sebaiknya proporsional: semakin tinggi risiko keputusan dan besar klaim dampak, semakin kuat bukti yang diperlukan. Untuk program kecil berisiko rendah, umpan balik pengguna dan tes penguasaan sederhana mungkin cukup. Untuk keputusan yang menyangkut keselamatan, promosi, atau investasi besar, diperlukan desain, validasi, tata kelola, dan dokumentasi yang lebih cermat.

## 1. Apa yang dievaluasi: proses, hasil, dan dampak

Evaluasi dimulai dengan mendefinisikan objeknya. Program pembelajaran dapat mencakup analisis kebutuhan, rancangan kurikulum, kualitas fasilitator, materi, teknologi, dukungan atasan,

praktik, dan perubahan proses kerja. Jika evaluator hanya menilai kursus, ia mungkin melewatkan kondisi organisasi yang menentukan keberhasilan. Pekerja dapat belajar prosedur baru, tetapi jika sistem operasional belum diperbarui atau alat tidak tersedia, perilaku yang diharapkan tidak dapat dilakukan. Objek evaluasi perlu mencakup intervensi yang benar-benar dialami peserta, bukan hanya silabus yang direncanakan.

Hasil belajar mencakup pengetahuan, keterampilan, sikap, kepercayaan diri yang terkalibrasi, strategi pemecahan masalah, atau kemampuan mengambil keputusan. Hasil perlu dipisahkan menurut sifatnya. Tes ingatan mengukur pengetahuan deklaratif; simulasi dapat menguji keterampilan; penilaian kasus menilai penalaran; refleksi dapat menunjukkan kesadaran atas strategi; dan observasi di tempat kerja melihat aplikasi. Tidak ada satu ukuran yang mewakili seluruh kompetensi. Instrumen dipilih berdasarkan tujuan, peserta, konsekuensi keputusan, dan kondisi pemakaian.

Proses pembelajaran mencakup apakah peserta menerima dosis dan urutan yang diharapkan, bagaimana fasilitasi berlangsung, apakah konten relevan dan aksesibel, serta apakah dukungan teknis tersedia. Data proses membantu menjelaskan mengapa hasil berbeda. Jika program gagal menjangkau sebagian shift, hasil rata-rata bisa terlihat baik meski akses timpang. Jika fasilitator memodifikasi materi, perubahan itu perlu dicatat agar temuan dapat ditafsirkan. Fidelity tidak selalu berarti setiap sesi identik; evaluasi mencatat unsur inti yang harus dipertahankan serta adaptasi yang dapat diterima.

Hasil proksimal muncul dekat dengan pengalaman belajar, misalnya pemahaman, kemampuan simulasi, niat menerapkan, atau rencana tindakan. Hasil menengah tampak dalam perilaku kerja, mutu keputusan, kolaborasi, atau penggunaan alat. Hasil distal dapat berupa produktivitas, retensi pegawai, kepuasan pelanggan, keselamatan, kualitas, atau mobilitas tenaga kerja. Horizon waktu dan tingkat pengaruh berbeda. Perubahan perilaku mungkin terlihat dalam beberapa minggu, sedangkan hasil organisasi memerlukan waktu lebih panjang dan lebih dipengaruhi konteks.

Dampak mengacu pada perbedaan yang relevan antara hasil yang terjadi dengan intervensi dan hasil yang mungkin terjadi tanpa intervensi atau di bawah pilihan lain. Dalam praktik, kondisi tanpa program tidak selalu dapat diamati secara langsung. Desain evaluasi berupaya membangun pembanding dan mengurangi penjelasan alternatif. Jika hal itu tidak mungkin, evaluasi dapat menilai kontribusi melalui bukti mekanisme, urutan waktu, triangulasi, dan penjelasan konteks, dengan menyebut tingkat kepastian secara jujur.

Pembedaan antara output, outcome, dan impact mencegah laporan mencampur jumlah kegiatan dengan perubahan. Output seperti peserta, jam kelas, dan modul terbit menjelaskan apa yang dilakukan. Outcome menguraikan perubahan pengetahuan, perilaku, atau proses. Impact

menilai konsekuensi yang lebih luas dan kontribusi intervensi. Organisasi boleh melaporkan semua lapisan, tetapi tidak menyebut output sebagai dampak. Kamus indikator membantu menjaga konsistensi definisi antartim dan periode.

## 2. Teori perubahan dan model logika evaluasi

Teori perubahan menjelaskan bagaimana aktivitas pembelajaran diharapkan menghasilkan perubahan. Rangkaian sederhana dapat memuat kebutuhan, input, kegiatan, keluaran, hasil awal, hasil kerja, dan dampak. Akan tetapi, diagram tidak cukup jika panah antarunsurnya tidak memiliki asumsi. Misalnya, kursus komunikasi diharapkan meningkatkan keterampilan mendengar; keterampilan tersebut diperkirakan memperbaiki penanganan keluhan; dan perbaikan layanan diperkirakan meningkatkan kepuasan. Agar rantai ini bekerja, peserta perlu berlatih, manajer perlu memberi kesempatan, prosedur layanan harus memungkinkan respons, dan ukuran kepuasan harus sensitif pada perubahan.

Setiap hubungan kausal memiliki asumsi dan risiko. Peserta mungkin sudah memiliki kompetensi awal tinggi, materi mungkin tidak cocok dengan tugas, atau perubahan target kinerja mungkin datang bersamaan dengan pelatihan. Teori perubahan membuat faktor tersebut terlihat sebelum data dikumpulkan. Tim menandai asumsi yang paling rapuh dan menentukan bukti untuk mengujinya. Jika program menargetkan perilaku tertentu, evaluator perlu mengamati apakah perilaku itu memang terjadi sebelum berharap dampak bisnis muncul.

Model logika memetakan sumber daya dan tahapan program ke hasil yang diharapkan. Input dapat berupa waktu ahli, anggaran, platform, dan dukungan manajer. Aktivitas meliputi latihan, coaching, simulasi, dan tindak lanjut. Output mencatat jangkauan dan paparan. Outcome menjelaskan perubahan kemampuan dan praktik. Dampak menggambarkan hasil yang lebih luas. Model logika berguna untuk menyepakati definisi dan tanggung jawab, tetapi bukan bukti bahwa hubungan itu benar. Setiap panah adalah hipotesis yang perlu didukung bukti.

Analisis kontribusi mempertimbangkan beberapa pihak dan penyebab. Pelatihan dapat menjadi salah satu faktor dalam perubahan kinerja bersama perubahan teknologi, kebijakan, sistem penghargaan, atau permintaan pasar. Evaluator mengumpulkan bukti yang menguji hipotesis kontribusi: apakah perubahan dimulai setelah intervensi, apakah kelompok yang terpapar menunjukkan pola berbeda, apakah mekanisme belajar tampak, serta penjelasan alternatif mana yang masih mungkin. Pendekatan ini lebih realistis daripada mengklaim bahwa seluruh hasil organisasi berasal dari satu kegiatan.

Rantai perubahan perlu memperhitungkan waktu. Pengetahuan dapat meningkat segera setelah pelatihan lalu berkurang jika tidak dipraktikkan. Perilaku dapat memerlukan kesempatan berulang sebelum menjadi kebiasaan. Hasil organisasi mungkin baru tampak setelah proses lain

berubah. Jadwal evaluasi harus mengikuti teori perubahan, bukan hanya kalender anggaran. Pengukuran terlalu dini dapat melewatkan transfer; terlalu lambat dapat menyulitkan atribusi karena perubahan lain telah terjadi.

Model teori juga membedakan target yang diinginkan dan dampak samping. Program kepemimpinan dapat meningkatkan kepercayaan diri, tetapi juga mendorong keputusan terlalu cepat jika tidak mengembangkan refleksi. Program produktivitas dapat mempercepat proses sekaligus menambah beban atau mengurangi kualitas. Teori perubahan yang matang mencatat mekanisme positif dan risiko negatif, serta memasukkan ukuran penyeimbang untuk mendeteksinya.

## 3. Pertanyaan evaluasi dan pemilihan indikator

Pertanyaan evaluasi menjadi kompas pengumpulan dan analisis data. Pertanyaan perlu mengarah pada keputusan yang jelas: apakah program layak diperluas, bagian mana yang perlu diubah, kelompok mana yang membutuhkan dukungan, atau apakah investasi perlu dialihkan. Pertanyaan “apakah program berhasil?” terlalu luas. Pertanyaan yang lebih operasional menanyakan apakah simulasi meningkatkan penguasaan dibanding pembelajaran mandiri, apakah coaching meningkatkan penerapan setelah dua bulan, atau apakah perubahan perilaku berkontribusi pada penurunan kesalahan dalam konteks tertentu.

Pertanyaan dapat dikelompokkan ke dalam relevansi, desain, implementasi, efektivitas, pemerataan, efisiensi, keberlanjutan, dan dampak. Evaluasi relevansi menilai kecocokan program dengan kebutuhan. Evaluasi implementasi memeriksa apa yang benar-benar terjadi. Evaluasi efektivitas mengukur hasil. Pemerataan menilai distribusi akses dan manfaat. Efisiensi membandingkan sumber daya dengan hasil yang sesuai. Keberlanjutan menanyakan apakah kemampuan tetap digunakan dan dipelihara. Dampak menguji perubahan yang lebih luas serta kontribusi intervensi.

Indikator menerjemahkan pertanyaan menjadi bukti. Setiap indikator memiliki definisi, sumber data, periode, unit analisis, pemilik, dan batas interpretasi. Contoh indikator penguasaan adalah ketepatan respons pada kasus baru dengan rubrik yang disepakati. “Keaktifan” harus didefinisikan, apakah kehadiran, kontribusi diskusi, atau penyelesaian tugas. Indikator yang dinamai sama tetapi dirumuskan berbeda tidak dapat dibandingkan. Kamus data mencegah perubahan definisi diam-diam.

Pemilihan indikator memperhatikan validitas, reliabilitas, sensitivitas perubahan, beban responden, biaya, aksesibilitas, dan kemungkinan manipulasi. Ukuran yang mudah dikumpulkan belum tentu relevan. Jumlah klik dapat meningkat karena navigasi membingungkan. Skor

kepuasan dapat tinggi karena peserta menyukai suasana walau tidak menguasai materi. Indikator harus menjawab konstruk yang dimaksud dan diuji terhadap konteks penggunaannya.

Ukuran perlu seimbang. Kecepatan penyelesaian dilengkapi akurasi; produktivitas dilengkapi mutu dan keselamatan; tingkat partisipasi dilengkapi penguasaan; penghematan biaya dilengkapi akses dan kesejahteraan. Ketika sebuah indikator menjadi target bonus, perilaku dapat diarahkan pada angka daripada tujuan. Audit sampel dan diskusi kualitatif membantu mendeteksi efek permainan metrik.

Pertanyaan evaluasi juga menentukan analisis kelompok. Perbedaan menurut lokasi, jabatan, shift, masa kerja, atau aksesibilitas dapat memperlihatkan variasi. Segmentasi hanya dilakukan bila ukuran cukup dan penggunaan data sah. Evaluator perlu menghindari pembuatan banyak subkelompok kecil yang membuka identitas atau menghasilkan kesimpulan tidak stabil. Tujuan segmentasi ialah memperbaiki dukungan dan desain, bukan memberi cap pada peserta.

## 4. Pengukuran pembelajaran: konstruk, instrumen, dan validitas

Pengukuran pembelajaran dimulai dengan mendefinisikan konstruk. “Kepemimpinan”, “literasi data”, atau “pelayanan unggul” terlalu luas tanpa rincian perilaku atau pengetahuan. Tim memecah kompetensi menjadi komponen yang relevan dengan tugas: mengenali informasi penting, memeriksa sumber, memilih tindakan, menjelaskan alasan, dan mengetahui kapan harus eskalasi. Definisi ini memandu pembuatan tes, simulasi, rubrik, dan contoh bukti.

Validitas menyangkut apakah interpretasi skor cukup didukung untuk tujuan tertentu. Tes yang dikembangkan untuk pembelajaran formatif belum tentu layak untuk keputusan promosi. Isi instrumen diperiksa agar mewakili kompetensi, proses respons ditelaah agar peserta memahami tugas sebagaimana dimaksud, dan hubungan skor dengan ukuran lain dianalisis jika diperlukan. Validitas bukan cap permanen pada instrumen; ia bergantung pada populasi, konteks, dan penggunaan.

Reliabilitas mengacu pada konsistensi pengukuran. Skor dapat bervariasi karena kondisi tes, penilai, soal, atau suasana peserta. Untuk observasi kinerja, rubrik perlu contoh perilaku dan pelatihan penilai; kesepakatan antarpemberi nilai dapat diperiksa. Untuk asesmen pilihan ganda, kesulitan dan diskriminasi soal ditinjau. Reliabilitas tinggi tidak menjamin validitas: sebuah alat dapat secara konsisten mengukur hal yang tidak relevan.

Tes sebelum dan sesudah program membantu memperkirakan perubahan, tetapi perlu mempertimbangkan efek latihan, perubahan tingkat kesulitan, dan regresi menuju rata-rata. Bentuk alternatif dapat mengurangi pengulangan jawaban. Jika peserta mengingat item, kenaikan skor mungkin bukan karena penguasaan yang dapat ditransfer. Ukuran retensi setelah jeda, soal baru, dan variasi kasus membantu menilai pemahaman yang lebih tahan lama.

Asesmen keterampilan perlu mendekati kondisi penggunaan. Simulasi dapat menilai urutan tindakan, keputusan, komunikasi, dan respons terhadap pengecualian. Penilaian portofolio dapat menunjukkan pekerjaan nyata tetapi memerlukan kriteria konsisten dan pemeriksaan kontribusi individu. Observasi kerja memberi bukti autentik, namun dipengaruhi konteks dan dapat mengubah perilaku karena pekerja tahu sedang diamati. Triangulasi beberapa metode membantu membangun gambaran yang lebih utuh.

Instrumen harus adil bagi kelompok yang dinilai. Bahasa terlalu rumit, format yang tidak aksesibel, atau prasyarat teknologi yang tidak diperlukan dapat mengubah tes kompetensi menjadi tes kemampuan lain. Adaptasi format perlu mempertahankan konstruk inti sambil menghapus hambatan yang tidak relevan. Evaluator mendokumentasikan penyesuaian, memeriksa distribusi hasil, dan menyediakan kesempatan koreksi atau asesmen ulang bila prosedur terganggu.

## 5. Evaluasi reaksi, pengalaman, dan keterlibatan

Reaksi peserta mengungkap relevansi yang dirasakan, mutu fasilitasi, aksesibilitas, beban, dan niat menggunakan materi. Data ini penting bagi perbaikan, tetapi tidak boleh disamakan dengan belajar atau dampak. Peserta dapat menyukai program yang menghibur tetapi tidak menambah kompetensi; mereka juga dapat menilai sesi menantang secara negatif padahal latihan tersebut meningkatkan kemampuan. Reaksi adalah salah satu hasil yang perlu ditafsirkan, bukan pengganti hasil lain.

Survei pengalaman perlu memiliki tujuan spesifik. Pertanyaan seperti “apakah program bermanfaat?” dapat memberi kesan positif umum, tetapi tidak menunjukkan bagian mana yang berguna atau untuk tugas apa. Pertanyaan yang lebih informatif menilai relevansi contoh, kejelasan tujuan, kesempatan berlatih, kualitas umpan balik, kecepatan materi, akses teknis, dan keyakinan peserta untuk menerapkan. Pertanyaan terbuka dapat menangkap masalah yang belum masuk ke skala, tetapi responden perlu tahu bagaimana masukan digunakan.

Bias respons muncul ketika peserta takut mengkritik, merasa perlu menyenangkan fasilitator, atau menjawab segera sebelum mencoba tugas. Survei anonim dapat mengurangi tekanan, tetapi ukuran kelompok kecil tetap berisiko identifikasi. Pengumpulan dilakukan oleh pihak yang tidak memberi nilai langsung bila memungkinkan. Waktu pengisian yang wajar dan bahasa yang mudah dipahami meningkatkan kualitas respons. Tingkat respons dilaporkan bersama hasil agar pembaca mengetahui siapa yang mungkin tidak terwakili.

Keterlibatan digital sering diukur melalui login, lama sesi, klik, atau konten yang dilihat. Indikator tersebut menjelaskan perilaku penggunaan, tetapi tidak langsung menunjukkan perhatian atau pemahaman. Seseorang dapat membuka materi sambil terputus koneksi, meninggalkan layar

terbuka, atau mengakses melalui perangkat bersama. Analitik platform digunakan sebagai sinyal untuk menyelidiki pengalaman, bukan sebagai bukti final motivasi atau komitmen.

Pengalaman peserta dapat ditriangulasi melalui wawancara, diskusi kelompok, catatan fasilitator, analitik, dan pengamatan. Kualitatif membantu memahami mengapa suatu fitur dipakai atau dihindari. Sampel peserta dipilih agar mencakup variasi pengalaman, bukan hanya peserta yang antusias. Pertanyaan tidak menggiring jawaban dan memberi ruang untuk menyebut kendala sistemik.

Reaksi ditindaklanjuti secara terlihat. Jika peserta melaporkan materi tidak relevan, tim memeriksa apakah tujuan kompetensi atau contoh kasus yang keliru. Jika kendala teknis berulang, pemilik platform bertanggung jawab memperbaiki. Jika peserta menyebut beban kerja menghambat praktik, manajer menilai ulang penjadwalan. Menutup lingkaran umpan balik menjaga kepercayaan dan mengubah survei dari ritual menjadi alat perbaikan.

## 6. Mengukur transfer ke tempat kerja

Transfer adalah penggunaan kompetensi yang diperoleh dalam konteks kerja. Transfer dekat terjadi pada tugas yang mirip dengan latihan; transfer jauh memerlukan adaptasi pada situasi baru. Generalisasi menjelaskan kemampuan menerapkan pengetahuan pada beragam kasus, sedangkan retensi menunjukkan apakah kemampuan tetap tersedia setelah waktu berlalu. Evaluasi menentukan bentuk transfer yang penting untuk program tersebut, bukan menggunakan satu pengertian umum.

Bukti transfer dapat berasal dari observasi perilaku, sampel produk kerja, audit kualitas, simulasi pascapelatihan, laporan supervisor, umpan balik pelanggan, atau catatan penggunaan prosedur. Setiap sumber memiliki keterbatasan. Penilaian atasan dapat dipengaruhi hubungan kerja dan kesempatan mengamati; data pelanggan dipengaruhi faktor lain; pengamatan langsung dapat mengubah perilaku; laporan diri rentan pada ingatan dan keinginan memberi kesan baik. Karena itu, ukuran sebaiknya dipadukan dan keterbatasannya dijelaskan.

Tindak lanjut perlu dilakukan pada jarak waktu yang cukup. Pengukuran terlalu dekat dengan sesi belajar hanya menunjukkan kesiapan awal. Pengukuran setelah beberapa minggu atau bulan dapat melihat penerapan serta kendala, sedangkan pengukuran berulang menilai keberlanjutan. Jadwal disesuaikan dengan frekuensi tugas. Jika kompetensi jarang dipakai, simulasi berkala dapat menilai retensi tanpa menunggu kesempatan kerja yang langka.

Transfer dipengaruhi karakteristik peserta, rancangan pelatihan, dan lingkungan kerja. Tinjauan Grossman dan Salas (2011) mengelompokkan faktor yang berkaitan dengan kesiapan individu, desain latihan, serta dukungan dan kesempatan di tempat kerja. Meta-analisis Blume et al. (2010) juga menunjukkan peran faktor peserta, intervensi, dan lingkungan, disertai variasi

pengukuran transfer. Evaluasi perlu melihat apakah pekerja memiliki alat, wewenang, waktu, dukungan, dan insentif untuk menerapkan keterampilan, bukan menyimpulkan kegagalan berasal dari peserta semata.

Rencana transfer sebaiknya dibangun sejak pelatihan dirancang: tujuan penerapan, tugas praktik, komitmen manajer, pasangan umpan balik, alat bantu kerja, serta waktu tindak lanjut. Setelah program, pekerja dapat memilih satu atau dua sasaran penerapan, mendiskusikannya dengan supervisor, lalu mencatat hambatan dan bukti. Rencana tersebut tidak digunakan untuk mengawasi setiap langkah; fungsinya membuat penerapan lebih nyata dan memberi data perbaikan.

Instrumen transfer perlu jelas mengenai apa yang diamati, siapa menilai, kapan, dan berdasarkan standar apa. Schoeb et al. (2021) meninjau pengukuran transfer dan menemukan ketidaksesuaian antara definisi dan instrumen serta pelaporan parameter pelatihan yang terbatas. Konsistensi istilah dan pelaporan membantu organisasi membandingkan hasil, namun perbandingan hanya masuk akal jika tugas, populasi, dan kondisi kerja sebanding.

## 7. Kerangka evaluasi: manfaat dan batas penggunaan

Model empat tingkat Kirkpatrick—reaksi, pembelajaran, perilaku, dan hasil—sering dipakai untuk mengorganisasikan pertanyaan evaluasi (Kirkpatrick & Kirkpatrick, 2006; Alsalamah & Callinan, 2021; Rouse, 2011). Ia membantu mengingatkan bahwa kepuasan peserta hanyalah satu jenis bukti, dan organisasi juga perlu memeriksa kemampuan, aplikasi, serta hasil. Namun, model tidak dengan sendirinya menentukan desain pengukuran, metode atribusi, kualitas indikator, atau cara mengendalikan penjelasan alternatif. Tingkat yang lebih tinggi tidak otomatis lebih baik jika indikatornya tidak valid.

Hubungan antarlevel tidak selalu berurutan atau linear. Peserta dapat belajar tanpa menyukai sesi; perilaku dapat berubah karena umpan balik manajer meski skor tes tidak banyak berubah; hasil bisnis dapat membaik karena proses lain tanpa perubahan perilaku peserta. Karena itu, korelasi antarlevel bukan bukti bahwa satu level menyebabkan level lain. Evaluator menguji hubungan berdasarkan teori dan data, bukan mengasumsikan bahwa setiap tingkat harus meningkat bersama.

Model CIPP (context, input, process, product) dapat membantu menilai kebutuhan, sumber daya, pelaksanaan, dan hasil. Pendekatan ini berguna ketika keputusan menyangkut perbaikan atau akuntabilitas program secara menyeluruh. Namun, daftar kategori tetap memerlukan pertanyaan spesifik. Model evaluasi tujuan atau goal-free dapat memberi sudut pandang lain saat hasil yang tidak direncanakan perlu dicari. Pilihan kerangka bergantung pada keputusan, konteks, dan sumber daya.

Kerangka pembelajaran berbasis hasil membedakan perubahan kognitif, keterampilan, dan afektif serta mendorong ukuran yang sesuai dengan masing-masing konstruk (Kraiger et al., 1993). Pendekatan ini menghindari penggunaan satu survei kepuasan sebagai proksi untuk semua hasil. Program pengetahuan dapat menggunakan tes dan kasus; program keterampilan menggunakan demonstrasi; intervensi sikap mengukur orientasi dan perilaku; pengembangan tim melihat koordinasi dan hasil tugas.

Setiap model adalah peta, bukan realitas. Evaluator dapat menggabungkan kerangka jika peran tiap model jelas, tetapi tidak perlu menumpuk istilah yang membingungkan. Model tidak menggantikan penalaran kausal atau kompetensi pengukuran. Laporan perlu menyebut kerangka yang dipakai dan alasan pemilihan, kemudian memperlihatkan bagaimana pertanyaan, indikator, dan keputusan saling berhubungan.

Kritik terhadap penggunaan mekanis model Kirkpatrick menekankan bahwa pengguna kadang menganggap empat level sebagai tangga sebab-akibat, padahal model awalnya mengategorikan kriteria evaluasi. Artikel Reio et al. (2017) membahas kegunaan dan batas model dalam praktik evaluasi; kritik serupa menyoroti risiko memperlakukan level sebagai rantai kausal dan pentingnya prinsip kemanfaatan bagi peserta (Bates, 2004). Organisasi dapat tetap menggunakan kerangka tersebut sebagai daftar lensa, dengan pengakuan bahwa bukti dampak membutuhkan pendekatan tambahan.

## 8. Desain evaluasi kausal dan pembanding

Evaluasi dampak bertanya apa yang mungkin terjadi jika program tidak diberikan atau jika alternatif lain dipilih. Masalah dasarnya ialah setiap individu hanya dapat diamati dalam satu kondisi pada waktu yang sama. Evaluasi perlu membangun pembanding yang kredibel. Randomisasi membantu menyeimbangkan faktor awal secara rata-rata, tetapi tidak selalu layak karena alasan etis, operasional, kapasitas, atau kewajiban untuk memberi pelatihan kepada semua pekerja.

Uji acak dapat dilakukan pada tingkat individu, tim, atau unit kerja. Jika kontaminasi antarpegawai mungkin terjadi, randomisasi kelompok dapat lebih sesuai. Peluncuran bertahap memungkinkan semua unit akhirnya menerima program, sementara waktu implementasi menciptakan pembanding sementara. Desain harus menetapkan protokol, hasil utama, ukuran sampel, analisis, penanganan peserta yang tidak mengikuti, serta aturan penghentian sebelum hasil dilihat.

Desain kuasi-eksperimental seperti difference-in-differences, regresi diskontinuitas, pencocokan, atau interrupted time series dapat digunakan ketika randomisasi tidak tersedia. Masing-masing bergantung pada asumsi. Difference-in-differences membutuhkan tren paralel;

regresi diskontinuitas memerlukan aturan ambang yang tidak dimanipulasi; pencocokan hanya menyeimbangkan faktor yang terukur; deret waktu memerlukan pengamatan yang cukup dan tidak terganggu perubahan serentak. Asumsi dan uji sensitivitas harus dilaporkan.

Perbandingan sebelum-sesudah tanpa kelompok pembanding berguna untuk memantau perubahan, tetapi tidak cukup untuk mengisolasi efek program. Maturasi, perubahan beban kerja, pergantian pimpinan, sistem baru, dan pola musiman dapat menjelaskan perbedaan. Pengukuran berulang memberi informasi lebih baik daripada satu pascates, tetapi tidak otomatis mengatasi faktor pengganggu. Laporan menyebut desain sebagai pre-post deskriptif jika tidak ada basis inferensi kausal.

Ukuran efek perlu disertai interval ketidakpastian, ukuran sampel, kehilangan responden, dan distribusi hasil. Signifikansi statistik tidak sama dengan makna praktis. Perubahan kecil mungkin penting pada tugas keselamatan berfrekuensi tinggi, sedangkan perubahan besar pada skor yang tidak valid tidak mendukung keputusan. Analisis subkelompok sebaiknya direncanakan sebelumnya atau dinyatakan eksploratif untuk menghindari temuan kebetulan.

Randomisasi tidak selalu menjawab semua pertanyaan. Evaluasi proses dan wawancara membantu menjelaskan bagaimana intervensi bekerja dan mengapa dampak berbeda. Metode campuran menggabungkan inferensi kuantitatif dengan konteks, pengalaman, dan mekanisme. Tujuannya bukan membuat satu metode terlihat lebih unggul, melainkan memastikan klaim sesuai kekuatan bukti dan kebutuhan keputusan.

## 9. Analisis kontribusi, konteks, dan variasi hasil

Dalam organisasi, program berjalan bersamaan dengan intervensi lain. Analisis kontribusi menyusun argumen yang didukung bukti bahwa program merupakan salah satu penyebab perubahan. Evaluator meninjau teori perubahan, urutan waktu, kesesuaian mekanisme, data hasil, serta penjelasan alternatif. Kesimpulan dapat menyatakan kontribusi kemungkinan besar terjadi, terjadi pada kondisi tertentu, atau belum dapat dibedakan dari faktor lain. Bahasa ini lebih informatif daripada klaim “program menghasilkan” tanpa kualifikasi.

Konteks meliputi beban kerja, struktur supervisi, kebijakan, teknologi, komposisi tim, budaya, kondisi pasar, serta kejadian eksternal. Program yang sama dapat memberi hasil berbeda pada cabang dengan dukungan atasan yang kuat dibanding unit yang kekurangan kapasitas. Analisis konteks tidak mencari alasan untuk membenarkan kegagalan; ia menentukan kondisi yang perlu dipenuhi agar desain dapat bekerja dan aspek mana yang dapat disesuaikan.

Variasi hasil dapat terjadi antara peserta, tim, lokasi, fasilitator, tugas, dan waktu. Rata-rata menyembunyikan siapa yang memperoleh manfaat dan siapa yang tertinggal. Analisis moderator memeriksa faktor yang berhubungan dengan besar hasil, tetapi perlu jumlah data memadai dan

hipotesis yang masuk akal. Temuan kelompok kecil harus ditafsirkan hati-hati dan tidak digunakan untuk memberi label potensi.

Evaluasi juga mencari efek tak terduga, termasuk perubahan kesejahteraan akibat beban atau pengalaman belajar (Watson et al., 2018). Program keterampilan negosiasi mungkin meningkatkan keberanian berbicara, tetapi menambah konflik bila tim tidak memiliki norma diskusi. Pelatihan efisiensi dapat mempercepat penyelesaian sambil mengurangi pemeriksaan mutu. Wawancara, pengaduan, pengukuran beban, dan audit kualitas dapat menemukan konsekuensi yang tidak tercantum dalam tujuan awal.

Triangulasi sumber tidak berarti semua data harus sama. Perbedaan antara laporan peserta, observasi atasan, dan hasil operasional dapat menunjukkan waktu pengukuran berbeda, definisi yang berlainan, atau variasi konteks. Evaluator menjelaskan bagaimana sumber-sumber ditimbang, bukan memilih hanya yang mendukung klaim. Catatan keputusan dan data mentah yang terlindungi membuat interpretasi dapat ditinjau.

Pendekatan studi kasus keberhasilan dapat melengkapi data agregat dengan menelaah contoh penggunaan dan kondisi yang memungkinkan hasil baik (Brinkerhoff, 2006). Hasil nol atau negatif tetap berguna. Program mungkin tidak cukup kuat, tidak terimplementasi sesuai rancangan, tidak cocok dengan konteks, atau sasaran tidak bergantung pada kompetensi. Evaluasi memberi petunjuk untuk membedakan kemungkinan tersebut. Keputusan menghentikan program setelah bukti memadai adalah bentuk penggunaan evaluasi, bukan kegagalan profesi pembelajaran.

## 10. Dampak organisasi dan indikator bisnis

Hasil organisasi mencakup indikator finansial dan nonfinansial. Produktivitas, biaya, waktu siklus, mutu, retensi, keselamatan, kepuasan pelanggan, kepatuhan, inovasi, atau mobilitas internal dapat relevan bila terhubung dengan tujuan program. Indikator dipilih setelah pemetaan mekanisme: bagaimana perubahan kompetensi memengaruhi proses, dan bagaimana proses memengaruhi hasil. Tanpa hubungan ini, indikator bisnis hanya ditempelkan pada laporan pelatihan.

Kualitas data operasional perlu diperiksa sebelum digunakan. Apakah definisi cacat atau keluhan berubah selama periode? Apakah sistem pencatatan konsisten antarunit? Apakah ada perubahan komposisi pelanggan atau tingkat permintaan? Data yang dikumpulkan untuk tujuan operasi mungkin tidak dirancang untuk evaluasi, sehingga missing data, pembaruan kode, dan penundaan pencatatan memengaruhi analisis.

Horizon pengamatan disesuaikan dengan seberapa cepat hasil dapat berubah. Untuk keselamatan, near-miss dan kepatuhan prosedur dapat dipantau lebih cepat daripada kecelakaan serius yang jarang terjadi. Untuk retensi, waktu pengamatan harus cukup panjang dan

memperhitungkan mobilitas internal. Untuk inovasi, jumlah ide saja tidak menggambarkan nilai implementasi atau dampak pada layanan.

Indikator bisnis sering dipengaruhi faktor di luar program. Penurunan waktu layanan dapat terjadi karena otomatisasi proses; kenaikan penjualan dapat dipengaruhi kampanye; retensi dapat berubah karena pasar tenaga kerja. Desain pembanding, deret waktu, dan pencatatan kejadian membantu membangun interpretasi. Jika inferensi kausal tidak kuat, laporan menyebut perubahan terjadi setelah program dan membahas faktor lain sebagai kemungkinan.

Hasil organisasi juga memiliki distribusi. Rata-rata produktivitas dapat meningkat sementara beban kerja terkonsentrasi pada kelompok tertentu. Layanan mungkin membaik bagi pelanggan mayoritas tetapi memburuk bagi kelompok dengan kebutuhan khusus. Evaluasi memeriksa siapa yang menikmati manfaat dan menanggung biaya, sejauh data dapat dikumpulkan secara aman. Ukuran pemerataan dan kesejahteraan menjadi pasangan bagi indikator efisiensi.

Organisasi dapat menggunakan scorecard yang menghubungkan indikator pembelajaran, perilaku, proses, dan hasil. Namun scorecard harus ringkas dan menunjukkan hubungan, bukan menampilkan banyak angka tanpa keputusan. Setiap indikator memiliki pemilik dan tindakan jika hasil bergerak ke arah yang tidak diinginkan. Pertemuan tinjauan menafsirkan bukti bersama pemilik proses, manajer, analis, dan pekerja yang memahami konteks.

## 11. Nilai investasi, biaya, dan ROI

Analisis nilai investasi membandingkan sumber daya yang digunakan dengan manfaat dan risiko yang diperoleh. Biaya langsung dapat meliputi desain, fasilitator, platform, perjalanan, materi, lisensi, dan dukungan. Biaya kesempatan mencakup jam kerja peserta, waktu manajer, gangguan operasional, dan pemeliharaan. Mengabaikan waktu peserta dapat membuat program tampak murah padahal organisasi menanggung biaya besar dalam kapasitas kerja.

Manfaat yang dapat dikuantifikasi mungkin meliputi kesalahan lebih sedikit, waktu orientasi lebih singkat, pengurangan kerja ulang, produktivitas, retensi, atau peningkatan mutu. Estimasi harus memiliki dasar data dan metode konversi yang dapat diperiksa. Manfaat nonfinansial seperti keselamatan, kepercayaan, kesiapan menghadapi perubahan, dan akses karier dapat dilaporkan dengan indikator serta narasi tanpa dipaksa menjadi nominal uang yang spekulatif.

ROI dihitung dengan membandingkan manfaat bersih terhadap biaya, tetapi angka ini sensitif pada asumsi atribusi, nilai unit, masa berlaku, dan biaya yang dimasukkan. Organisasi menyatakan bagaimana kontribusi pelatihan diisolasi, misalnya menggunakan kelompok pembanding, estimasi manajer yang dikalibrasi, atau model tren. Jika atribusi sangat tidak pasti, laporkan rentang dan analisis sensitivitas, bukan satu angka presisi semu (Phillips & Phillips, 2004, 2007).

Biaya total kepemilikan platform mencakup lisensi, integrasi, keamanan, migrasi, konten, dukungan, pembaruan, dan biaya keluar. Pembelajaran digital dapat menurunkan biaya perjalanan tetapi menambah kebutuhan kurasi dan dukungan teknis. Perbandingan alternatif menghitung biaya sepanjang siklus hidup dan menilai fleksibilitas, aksesibilitas, portabilitas data, serta risiko vendor. Harga awal saja tidak menunjukkan nilai jangka panjang.

Analisis cost-effectiveness berguna ketika hasil tidak mudah dinilai dalam uang. Biaya per peserta yang mencapai standar kompetensi atau biaya per pengurangan kesalahan dapat membantu membandingkan pendekatan. Namun perbandingan hanya valid jika populasi, definisi hasil, horizon, dan kualitas program sebanding. Efisiensi biaya juga tidak membenarkan pengurangan akses bagi kelompok yang membutuhkan dukungan tambahan.

Keputusan investasi mempertimbangkan manfaat, risiko, dan kesesuaian strategi. ROI positif tidak berarti program adil atau berkelanjutan; ROI rendah tidak berarti program tidak bernilai jika ia memenuhi kewajiban keselamatan, akses, atau hak pekerja. Laporan memisahkan temuan empiris dari penilaian normatif dan memperlihatkan asumsi sehingga pimpinan dapat membuat keputusan yang bertanggung jawab.

## 12. Evaluasi program digital dan berbasis AI

Pembelajaran digital menyediakan jejak data yang lebih rinci, seperti akses, interaksi, jawaban, waktu, rekomendasi, dan penggunaan bantuan. Data ini memperluas kemampuan pemantauan, tetapi tidak secara otomatis meningkatkan kualitas evaluasi. Klik tidak membuktikan perhatian; durasi tidak membuktikan pemahaman; rekomendasi yang diterima tidak membuktikan kecocokan. Setiap variabel diberi definisi dan dihubungkan dengan konstruk yang tepat.

AI dapat dipakai untuk membuat soal, memberi umpan balik, merekomendasikan materi, atau mengadaptasi tingkat kesulitan. Evaluasi menguji akurasi, relevansi, keselamatan konten, aksesibilitas, bias, dan pengalaman pengguna. Perubahan versi model dapat mengubah keluaran, sehingga setiap studi mencatat versi, konfigurasi, sumber konten, dan periode. Model perlu diuji oleh ahli dan pengguna sebelum hasilnya dipakai untuk keputusan penting.

Fitur adaptif dapat mengubah pengalaman antar peserta, sehingga kelompok pembanding dan dosis intervensi menjadi lebih rumit. Evaluator mencatat apa yang dilihat setiap peserta, jumlah latihan, urutan, dan bantuan yang tersedia. Bila personalisasi menjadi bagian perlakuan, evaluasi menilai manfaat tambahannya dibanding jalur standar yang relevan, bukan sekadar membandingkan skor peserta yang menerima rekomendasi berbeda tanpa mengendalikan seleksi.

Analitik prediktif mungkin mengidentifikasi peserta yang diperkirakan membutuhkan bantuan. Label risiko dapat membantu jika menghasilkan dukungan tepat waktu, tetapi berbahaya jika dilihat sebagai sifat individu atau digunakan untuk menolak peluang. Evaluator menilai false

positive dan false negative, dampak pada kelompok, transparansi, kemungkinan koreksi, dan tindak lanjut manusia. Prediksi yang akurat secara rata-rata tetap dapat merugikan kelompok tertentu.

Data pembelajaran dan data kinerja dapat dihubungkan untuk menilai transfer, tetapi pemadanan harus memiliki tujuan yang sah, pembatasan akses, retensi, dan perlindungan. Jika data dikumpulkan untuk mendukung pembelajaran, penggunaan baru untuk promosi atau disiplin memerlukan penilaian terpisah dan komunikasi. Privasi dan kepercayaan memengaruhi respons peserta, sehingga tata kelola data juga memengaruhi validitas evaluasi.

Evaluasi digital perlu menyertakan jalur non-digital atau alternatif bagi pekerja yang aksesnya terbatas. Jika konektivitas menjadi hambatan, data platform meremehkan partisipasi kelompok tersebut. Organisasi menyediakan perangkat, waktu, format aksesibel, dan dukungan teknis. Kinerja sistem dilaporkan bersama tingkat akses agar hasil tidak mewakili hanya pengguna paling terhubung.

## 13. Evaluasi yang etis, aman, dan berkeadilan

Evaluasi mengumpulkan informasi tentang pekerja, sehingga tujuan, dasar, dan batas penggunaannya harus jelas. Prinsip minimisasi mengharuskan pengumpulan data yang diperlukan untuk pertanyaan tertentu. Data sensitif tidak dikumpulkan hanya karena dapat tersedia. Pemberitahuan menjelaskan tujuan, siapa yang memiliki akses, lama penyimpanan, penggunaan sekunder, serta cara meminta koreksi sesuai ketentuan yang berlaku.

Kerahasiaan dan anonimitas bukan hal yang sama. Data yang diberi kode masih dapat dihubungkan kembali jika kunci tersedia; kelompok kecil dapat dikenali dari konteks; dan kutipan wawancara dapat mengungkap identitas. Protokol pelaporan mengatur ambang ukuran kelompok, penghapusan pengenal, penanganan kutipan, serta akses terbatas. Data disimpan aman dan dihapus sesuai jadwal.

Partisipasi sukarela penting untuk survei atau wawancara yang tidak diwajibkan, tetapi relasi kerja dapat membuat pekerja merasa sulit menolak. Peneliti internal menjelaskan bahwa keputusan kerja tidak bergantung pada jawaban pribadi, sejauh hal itu benar. Bila evaluasi terkait sertifikasi wajib, peserta mengetahui konsekuensi dan jalur banding. Penilaian formatif dan sumatif dipisahkan agar ruang latihan tidak menjadi alat pengawasan tanpa penjelasan.

Evaluasi berkeadilan memeriksa apakah instrumen dan proses memiliki dampak berbeda menurut bahasa, aksesibilitas, shift, lokasi, status kerja, dan pengalaman awal. Analisis dilakukan dalam batas privasi. Perbedaan hasil memicu pemeriksaan atas proses, bukan kesimpulan otomatis tentang kemampuan kelompok. Penguji juga memeriksa bias penilai dan kesesuaian tugas asesmen dengan kompetensi.

Risiko bahaya mencakup stres, stigma, ketakutan terhadap pembalasan, kebocoran informasi, kelelahan survei, dan pemanfaatan hasil di luar tujuan. Evaluator menyiapkan prosedur penanganan keluhan, dukungan peserta, serta respons insiden. Pihak yang tidak memiliki konflik kepentingan dapat meninjau evaluasi dengan konsekuensi besar.

Etika meningkatkan mutu bukti. Ketika pekerja memahami tujuan dan merasa aman, mereka cenderung memberi masukan lebih jujur. Jika evaluasi dipersepsikan sebagai pengawasan terselubung, data menjadi kurang akurat dan partisipasi menurun. Transparansi, proporsionalitas, dan kendali yang layak adalah prasyarat kepercayaan sekaligus kualitas analitik.

## 14. Metode campuran dan evaluasi kualitatif

Metode kuantitatif dapat menilai perubahan skor, perbedaan kelompok, distribusi, dan tren. Metode kualitatif membantu mengurai makna, mekanisme, hambatan, dan kondisi yang memengaruhi hasil. Kombinasi keduanya dapat menjawab pertanyaan yang tidak dapat dijelaskan oleh satu sumber. Contohnya, data menunjukkan transfer rendah; wawancara kemudian menemukan bahwa pekerja tidak memiliki wewenang menerapkan prosedur baru.

Wawancara semi-terstruktur dapat membahas pengalaman sebelum dan sesudah program, contoh penerapan, dukungan, hambatan, dan konsekuensi. Pertanyaan meminta kejadian konkret agar jawaban tidak berhenti pada pendapat umum. Fasilitator netral dan panduan konsisten meningkatkan keterbandingan, sementara pertanyaan terbuka memungkinkan tema yang tidak diprediksi.

Diskusi kelompok memperlihatkan norma bersama dan perbedaan pandangan, tetapi peserta dapat menahan kritik di hadapan atasan atau rekan. Kelompok dibentuk dengan mempertimbangkan relasi kuasa, ukuran, lokasi, dan keamanan. Untuk isu sensitif, wawancara individual atau kanal tertulis mungkin lebih tepat. Kerahasiaan batas-batasnya dijelaskan.

Observasi kerja menangkap praktik dalam konteks, termasuk interaksi dengan alat, kolega, dan pelanggan. Catatan lapangan menggunakan fokus dan protokol yang jelas. Pengamat membedakan deskripsi dari interpretasi dan mengakui kehadiran mereka dapat mengubah perilaku. Observasi perlu dijadwalkan dengan mempertimbangkan beban dan privasi pihak lain.

Dokumen, catatan mutu, materi pelatihan, panduan kerja, dan rapat tindak lanjut dapat membantu merekonstruksi implementasi. Analisis dokumen memeriksa tanggal versi, pihak yang membuat, tujuan, dan keterbatasan pencatatan. Materi resmi belum tentu menunjukkan pengalaman aktual, sehingga dibandingkan dengan sumber lain.

Integrasi data dilakukan pada tahap desain, analisis, atau interpretasi. Evaluator dapat menggunakan hasil kuantitatif untuk memilih peserta wawancara, atau membandingkan pola antarsumber melalui matriks. Ketidakselarasan ditelaah sebagai temuan, bukan dihapus. Laporan

menyajikan kutipan secukupnya tanpa mengorbankan identitas dan menyertakan bukti negatif atau kasus menyimpang untuk menguji kesimpulan.

## 15. Menerjemahkan temuan menjadi perbaikan program

Evaluasi hanya bermanfaat jika temuan masuk ke keputusan. Laporan sebaiknya menyajikan pertanyaan, desain, ukuran, hasil, konteks, keterbatasan, dan rekomendasi yang berkaitan langsung dengan bukti. Ringkasan eksekutif tidak boleh melebih-lebihkan kepastian. Pemangku kepentingan perlu mengetahui apa yang dapat disimpulkan dan apa yang masih terbuka.

Rekomendasi dibedakan menurut tingkat: desain materi, metode fasilitasi, akses, dukungan atasan, perbaikan proses, kebijakan, atau alokasi sumber daya. Setiap rekomendasi menetapkan pemilik, prioritas, tenggat, kebutuhan sumber daya, dan indikator tindak lanjut. Rekomendasi "tingkatkan komunikasi" terlalu kabur tanpa sasaran, kanal, pesan, dan tanggung jawab.

Sesi sensemaking mempertemukan evaluator dengan manajer, fasilitator, analis, peserta, dan pimpinan. Mereka meninjau bukti, membandingkan interpretasi, serta menguji alternatif. Diskusi tidak mengubah data agar sesuai harapan; ia menambah konteks yang dapat diuji dan mencegah kesalahan penafsiran. Perbedaan pendapat dicatat bersama dasar bukti masing-masing.

Siklus perbaikan dapat menggunakan uji kecil: mengubah satu elemen, menentukan indikator, menjalankan pada kelompok terbatas, dan menilai dampak serta efek samping. Perubahan dilakukan secara terdokumentasi agar dapat dibandingkan. Jika beberapa elemen diubah sekaligus, organisasi perlu mencatatnya karena kontribusi tiap perubahan akan lebih sulit dipisahkan.

Tindak lanjut mengukur apakah rekomendasi dilaksanakan dan apakah perubahan memperbaiki hasil. Program tidak seharusnya terus berjalan hanya karena sudah dianggarkan atau memiliki sponsor. Jika hasil lemah, organisasi menentukan apakah program diperbaiki, diintegrasikan dengan perubahan kerja, atau dihentikan. Jika hasil positif, perluasan mempertimbangkan kapasitas dan variasi konteks.

Pembelajaran evaluasi juga dibagikan secara aman kepada unit lain. Catatan metode, instrumen, data definisi, dan temuan memudahkan replikasi yang bertanggung jawab. Berbagi tidak berarti menyalin program tanpa adaptasi. Dokumen mencantumkan konteks, karakteristik peserta, pelaksanaan, dan keterbatasan agar pengguna berikutnya memahami kondisi penerapannya.

## 16. Model operasi dan tata kelola evaluasi

Organisasi memerlukan tata kelola yang menjelaskan siapa menetapkan pertanyaan, mengelola data, memvalidasi instrumen, menganalisis hasil, menyetujui rekomendasi, dan memantau tindak lanjut. Evaluator dapat berasal dari tim internal, pihak eksternal, atau kemitraan.

Independensi dan pemahaman konteks sama-sama penting; pengaturan perlu mengelola konflik kepentingan dan memastikan akses pada data yang sah.

Kompetensi evaluasi dapat dibangun di fungsi L&D dan SDM. Kemampuan inti meliputi teori perubahan, pengukuran, metode kuantitatif dan kualitatif, etika, analisis biaya, komunikasi, dan fasilitasi keputusan. Tidak setiap program memerlukan analis penuh waktu; organisasi dapat membuat standar minimum dan mengakses keahlian khusus untuk proyek berisiko tinggi.

Portofolio evaluasi menetapkan tingkat kedalaman sesuai biaya, risiko, skala, dan kepentingan strategi. Program rutin kecil dapat memakai pemantauan ringan dan tes kompetensi. Program yang memengaruhi keselamatan, banyak pekerja, atau investasi besar memerlukan desain lebih kuat dan review independen. Proporsionalitas mencegah evaluasi menjadi beban berlebihan sekaligus mencegah klaim besar tanpa bukti.

Data dictionary, repositori instrumen, protokol, dan kalender evaluasi menjaga konsistensi. Versi instrumen dicatat; perubahan indikator dijelaskan; akses dibatasi sesuai peran. Repositori menyimpan juga evaluasi yang hasilnya nol atau negatif agar organisasi tidak mengulang kesalahan dan tidak hanya mengarsipkan keberhasilan.

Komite evaluasi dapat menghubungkan HR, pembelajaran, operasional, keuangan, teknologi, hukum, privasi, dan perwakilan pekerja. Komite meninjau desain berisiko, bukan mengambil alih seluruh evaluasi. Keputusan dan alasannya tercatat. Jika data menemukan persoalan keselamatan atau perlakuan tidak adil, mekanisme eskalasi mengikuti kebijakan yang sesuai.

Pelaporan kepada pimpinan perlu menyelaraskan kedalaman teknis dengan kebutuhan keputusan. Dashboard memberi sinyal, tetapi laporan naratif menerangkan alasan, ketidakpastian, dan konsekuensi. Pimpinan yang menerima hasil bertanggung jawab memutuskan tindakan atau menjelaskan mengapa rekomendasi tidak dilaksanakan. Dengan demikian evaluasi membentuk siklus akuntabilitas dua arah.

## 17. Studi kasus ilustratif: evaluasi program layanan pelanggan

Bayangkan sebuah organisasi layanan melatih petugas untuk menggunakan prosedur baru dalam menangani keluhan. Tujuan program ialah meningkatkan diagnosis kebutuhan, komunikasi empatik, ketepatan eskalasi, dan penyelesaian kasus. Program mencakup modul digital, simulasi percakapan, coaching supervisor, serta panduan kerja. Manajemen ingin mengetahui apakah program memperbaiki mutu dan mempercepat penyelesaian tanpa menambah keluhan berulang.

Sebelum pelaksanaan, tim membangun teori perubahan. Simulasi diharapkan meningkatkan keterampilan bertanya; keterampilan itu harus digunakan pada kontak nyata; supervisor memberi umpan balik; sistem kasus harus memuat pilihan eskalasi yang benar; dan mutu percakapan

diharapkan menurunkan pengulangan kontak. Asumsi mencakup tersedianya waktu latihan, kestabilan prosedur, dan pencatatan kasus yang konsisten.

Garis dasar mengukur hasil simulasi, tingkat kontak ulang, penyelesaian kasus, kepuasan pelanggan, waktu penanganan, dan distribusi beban. Rubrik simulasi dikalibrasi oleh penilai. Unit pelatihan diluncurkan bertahap sehingga tim dapat membandingkan cabang yang sudah dan belum menjalani program dalam periode transisi. Semua cabang kemudian menerima pelatihan. Analisis mempertimbangkan volume, jenis kasus, pengalaman petugas, dan perubahan sistem.

Data proses menunjukkan peserta menyelesaikan modul, tetapi tingkat penggunaan panduan berbeda karena perangkat cabang dan akses jaringan. Simulasi meningkat pada keterampilan bertanya dan ketepatan eskalasi. Pada periode berikutnya, kontak ulang turun di beberapa cabang tetapi tidak di semua. Wawancara mengungkap cabang dengan antrean puncak tidak memiliki waktu coaching dan supervisor baru belum memahami rubrik. Unit dengan beban berat memperlihatkan waktu penanganan lebih lama meski mutu diagnosis membaik.

Evaluasi tidak langsung menyimpulkan program gagal atau berhasil sepenuhnya. Bukti menunjukkan kemampuan meningkat pada simulasi; transfer terjadi lebih konsisten ketika dukungan dan waktu tersedia; hasil bisnis bercampur dan dipengaruhi kapasitas cabang. Tim menyesuaikan jadwal coaching, melatih supervisor, dan memperbaiki akses panduan. Analisis lanjutan melihat apakah perubahan tersebut mengurangi variasi hasil.

Manfaat finansial diperkirakan dari biaya kontak ulang dan kerja koreksi, dengan asumsi atribusi konservatif. Hasil dilaporkan sebagai rentang dan tidak mengubah semua perubahan menjadi nilai uang. Efek pada kepuasan dan beban kerja disajikan bersama kecepatan. Organisasi menetapkan ukuran tindak lanjut dan memeriksa apakah program menambah intensitas kerja.

Kasus ini memperlihatkan bagaimana evaluasi dapat memperbaiki sistem. Satu angka rata-rata tidak akan menunjukkan perbedaan kondisi cabang. Pembelajaran, dukungan, proses, dan hasil perlu dinilai bersama. Perluasan program dilakukan setelah faktor yang mendukung transfer dipahami, bukan hanya karena survei peserta positif.

## 18. Evaluasi pembelajaran dalam konteks Indonesia

Organisasi Indonesia memiliki variasi besar dalam jenis pekerjaan, skala, konektivitas, bahasa, sektor, dan akses terhadap dukungan SDM. Metode evaluasi perlu disesuaikan dengan kondisi tersebut. Pekerja lapangan dan shift mungkin tidak tercatat pada platform yang sama dengan karyawan kantor. Penjadwalan wawancara dan survei harus memperhitungkan jam kerja serta perangkat yang tersedia. Jika sebagian responden tidak terjangkau, laporan menyebut siapa yang tidak terwakili.

Bahasa instrumen memengaruhi validitas. Terjemahan istilah teknis perlu diperiksa agar makna pertanyaan konsisten lintas wilayah. Uji coba kognitif menanyakan bagaimana peserta memahami soal atau skala. Takarir, panduan audio, glosarium, serta fasilitator lokal dapat memperluas partisipasi. Adaptasi bahasa dicatat dan ditinjau agar skor dapat dibandingkan dengan kehati-hatian.

UMKM sering memiliki data operasional terbatas dan kapasitas analitik kecil. Evaluasi dapat memulai dari indikator yang sudah tersedia dan dapat dipercaya, contoh kerja sebelum-sesudah, wawancara, serta observasi. Desain sederhana tetapi konsisten lebih baik daripada dashboard rumit tanpa pemilik data. Asosiasi usaha, lembaga pelatihan, atau perguruan tinggi dapat membantu membangun instrumen bersama, dengan perjanjian jelas mengenai data dan publikasi.

Organisasi besar, BUMN, pemerintah, dan sektor yang diatur memerlukan dokumentasi, audit, dan kepatuhan yang lebih formal. Program nasional dapat menghasilkan variasi implementasi antarwilayah. Evaluasi perlu membedakan standar inti yang wajib sama dari adaptasi lokal yang memperbaiki kecocokan. Pelibatan pelaksana daerah membantu menjelaskan bagaimana kebijakan diterjemahkan di lapangan.

Data personalia dan pembelajaran harus dikelola sesuai peraturan yang berlaku, termasuk UU Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi. Pertanyaan evaluasi tidak membenarkan pengumpulan data berlebihan. Organisasi memeriksa dasar pemrosesan, keamanan, peran pengendali dan prosesor, masa simpan, akses subjek data, serta penilaian risiko sesuai konteks. Nasihat hukum diperlukan untuk keputusan spesifik.

Evaluasi juga dapat mendukung agenda produktivitas dan pengembangan tenaga kerja nasional bila hasil dibagikan secara agregat dan aman. Publikasi evaluasi perlu menyertakan metode, konteks, batasan, serta konflik kepentingan. Temuan lokal memperkaya bukti internasional yang sering berasal dari sektor dan negara berbeda. Kolaborasi penelitian menjaga independensi analisis serta memberi organisasi kesempatan menerapkan temuan secara bertanggung jawab.

## 19. Risiko salah evaluasi dan cara menghindarinya

Risiko pertama ialah menjadikan kepuasan sebagai bukti efektivitas. Pencegahannya ialah melaporkan reaksi sebagai pengalaman, lalu mengukur kemampuan dan penerapan secara terpisah. Risiko kedua ialah mengambil indikator yang mudah tetapi tidak relevan; kamus konstruk dan telaah ahli membantu menjaga kesesuaian. Risiko ketiga ialah menyimpulkan kausalitas dari perubahan sebelum-sesudah; pembanding, catatan konteks, dan bahasa kesimpulan yang tepat mengurangi risiko tersebut.

Risiko lain adalah memilih hanya indikator yang menguntungkan, mengganti definisi setelah program, atau mengabaikan data yang hilang. Protokol analisis, dokumentasi perubahan, dan pelaporan hasil nol melindungi integritas. Jika indikator utama tidak tersedia, laporan menjelaskan dampaknya pada kesimpulan. Data yang kurang tidak dapat diatasi dengan lebih banyak grafik.

Efek Hawthorne atau perubahan perilaku karena sedang diamati dapat memengaruhi observasi. Pengukuran berulang, penggunaan data rutin yang sah, serta penjelasan tujuan dapat mengurangi distorsi. Namun penggunaan data operasional tetap dapat mengubah perilaku; pekerja perlu tahu apakah dan bagaimana data diakses. Observasi harus proporsional dan tidak menjadi pengawasan terus-menerus.

Bias seleksi muncul ketika peserta yang mengikuti program berbeda dari yang tidak ikut. Peserta mungkin lebih termotivasi atau dipilih karena kinerja tinggi; hasil kelompok tidak langsung mewakili seluruh tenaga kerja. Evaluator memeriksa proses nominasi, kondisi awal, dan alasan ketidakikutsertaan. Jika pelatihan wajib, ketidakhadiran juga dianalisis sebagai masalah akses atau kapasitas, bukan hanya ketidakpatuhan.

Manipulasi metrik terjadi saat angka dijadikan sasaran dengan insentif kuat. Pegawai dapat menyelesaikan modul tanpa belajar atau memilih kasus yang mudah. Ukuran penyeimbang, audit sampel, dan diskusi praktik mengungkap perilaku tersebut. Indikator tidak dipakai sebagai target kinerja individu tanpa validasi dan konsekuensi yang jelas.

Terakhir, evaluasi dapat membebani pekerja melalui survei terlalu sering, wawancara berulang, atau permintaan dokumentasi tambahan. Pengumpulan data perlu dikoordinasikan, singkat, dan memberi manfaat yang tampak. Jika tidak ada keputusan yang akan berubah berdasarkan data, pengumpulan sebaiknya ditinjau kembali. Prinsip ini membantu evaluator tetap fokus pada pertanyaan yang penting.

## 20. Kualitas data dan keterwakilan sampel

Kualitas data membatasi kekuatan setiap kesimpulan. Sebelum menghitung perubahan, evaluator perlu mengetahui unit analisis, sumber, periode, aturan pencatatan, serta tingkat kelengkapan. Data dapat berbentuk individu, tim, cabang, atau organisasi. Skor pekerja tidak dapat langsung dibandingkan dengan indikator cabang jika ukuran tim dan jumlah peserta berbeda. Definisi unit, numerator, denominator, dan cara agregasi ditetapkan sejak awal.

Sampel menentukan siapa yang terwakili. Jika evaluasi hanya mencakup peserta yang menyelesaikan program, hasil tidak menjelaskan pengalaman pekerja yang tidak mendaftar, tidak hadir, atau berhenti. Catat alur dari pekerja yang memenuhi syarat hingga mereka yang diukur pada tindak lanjut. Tahap kehilangan peserta dapat berkaitan dengan jadwal shift, akses internet,

dukungan atasan, disabilitas, atau relevansi isi. Hambatan tersebut juga merupakan temuan evaluasi, bukan gangguan administratif belaka.

Perbandingan peserta dan nonpeserta dapat mengungkap bias seleksi. Peserta mungkin lebih termotivasi, dinominasikan karena kinerja tinggi, atau dipilih karena dianggap memerlukan perbaikan. Karena itu, kelompok yang berbeda tidak langsung sebanding. Data kondisi awal, proses nominasi, alasan ketidakikutsertaan, dan karakteristik kerja membantu menilai batas inferensi. Bila informasi kelompok nonpeserta tidak dapat digunakan secara sah, evaluasi menyatakan keterbatasan dan memilih pembanding lain.

Ukuran sampel yang diperlukan bergantung pada perubahan yang berarti bagi keputusan, variasi data, desain, dan jumlah unit yang dibandingkan. Nilai-p tidak perlu menjadi satu-satunya penentu. Besaran efek, rentang ketidakpastian, serta konsekuensi praktis menjelaskan makna hasil. Program dengan peserta sedikit dapat menggunakan pengukuran berulang, dokumentasi kasus, dan bukti proses, tetapi tidak mengklaim generalisasi luas. Bila pertanyaan membutuhkan banyak unit, organisasi dapat menggabungkan beberapa angkatan atau memperpanjang periode pengamatan.

Data yang hilang perlu dianalisis, bukan sekadar dihapus. Ketidakhadiran pada asesmen lanjutan mungkin berhubungan dengan beban kerja, perpindahan tugas, pengalaman program, atau akses digital. Jika pekerja yang mengalami kesulitan lebih jarang menjawab, rata-rata responden yang tersisa dapat tampak lebih baik. Laporan mencantumkan tingkat respons menurut unit, waktu, dan kelompok relevan. Analisis sensitivitas dapat menunjukkan apakah kesimpulan berubah ketika asumsi tentang responden yang hilang dibuat berbeda.

Jumlah individu yang besar tidak selalu berarti bukti kuat. Jika kebijakan pembelajaran diberikan pada tingkat cabang, jumlah cabang yang dibandingkan menentukan banyaknya unit independen. Seribu pekerja dalam empat cabang tidak setara dengan seribu unit terpisah. Pengelompokan membuat pengalaman di cabang sama cenderung mirip. Analis perlu menyesuaikan metode, atau organisasi memilih desain yang lebih sederhana dan menghindari presisi semu. Penjelasan struktur data memudahkan pembaca memahami dasar estimasi.

Kamus data menjaga konsistensi indikator lintas waktu. Untuk setiap ukuran, dokumentasikan definisi, rumus, sumber, frekuensi pembaruan, pemilik, aturan data hilang, dan perubahan sistem. Istilah “kesalahan”, misalnya, dapat merujuk pada catatan yang dikoreksi, keluhan pelanggan, pelanggaran prosedur, atau insiden keselamatan. Angka dari definisi yang berbeda tidak dapat dipertukarkan. Jika sistem pencatatan atau definisi berubah, tren sebelum dan sesudah diberi penanda.

Kualitas data adalah tanggung jawab lintas fungsi. Analis memahami distribusi dan ketidaklengkapan; manajer operasional mengetahui bagaimana data dicatat; pekerja mengetahui

tugas yang tidak tampak dalam sistem; tim teknologi menjelaskan batas ekstraksi. Pemeriksaan bersama menemukan duplikasi, keterlambatan entri, atau perbedaan kode antar unit. Sampel catatan dapat dibandingkan dengan sumber aslinya dengan akses terbatas. Keputusan pembersihan disimpan agar analisis dapat ditinjau.

Uji coba pengumpulan data sebelum program dimulai memeriksa apakah pertanyaan dipahami, formulir dapat diakses, waktu pengisian wajar, dan data tersambung sesuai izin. Uji coba mencakup seluruh proses mulai pemberitahuan dan pengumpulan hingga penyimpanan serta pelaporan. Instrumen yang baik di atas kertas dapat gagal karena pekerja tidak memiliki perangkat atau tidak memahami istilah. Memperbaiki alur sebelum data terkumpul menghemat waktu dan mengurangi bias.

Preregistrasi atau rencana analisis terdokumentasi berguna untuk keputusan berisiko tinggi. Dokumen menyebut pertanyaan utama, hasil primer, waktu ukur, kriteria pengecualian, dan metode analisis. Perubahan tetap dapat dilakukan ketika keadaan menuntut, tetapi alasan dan waktunya dicatat. Praktik ini membedakan pengujian yang direncanakan dari penjelajahan setelah hasil terlihat. Ia juga mencegah pemilihan hanya indikator yang menguntungkan untuk masuk ke ringkasan pimpinan.

Penyajian distribusi melengkapi rata-rata. Dua unit bisa memiliki skor rata-rata sama tetapi kesenjangan internal berbeda. Variasi menunjukkan siapa yang membutuhkan dukungan dan di mana praktik baik muncul. Dalam laporan kecil, rincian jabatan, lokasi, usia, dan karakteristik lain dapat mengungkap identitas. Tim menetapkan batas pelaporan, menggabungkan kategori, atau menahan angka ketika kelompok terlalu kecil. Perlindungan ini dirancang sejak awal, bukan setelah laporan hampir diterbitkan.

Kecukupan bukti selalu bergantung pada tujuan. Evaluasi eksploratif dapat mencari mekanisme dan hambatan; evaluasi dampak memerlukan pembanding yang masuk akal dan unit yang memadai. Bila organisasi tidak dapat memenuhi kebutuhan tersebut, laporan membatasi klaim, menyarankan tahap pengumpulan tambahan, atau menggunakan pendekatan kontribusi. Ketidakpastian yang dijelaskan memberi dasar lebih aman untuk keputusan daripada angka presisi yang dibangun dari data yang tidak memadai.

## 21. Retensi, lintasan belajar, dan waktu dampak

Pembelajaran dan dampak berkembang dalam rentang waktu yang berbeda. Pengetahuan faktual mungkin berubah segera setelah modul, keterampilan memerlukan latihan, perilaku kerja bergantung pada kesempatan, sedangkan hasil organisasi dapat merespons lebih lambat. Satu pengukuran pascaprogram tidak menggambarkan seluruh urutan tersebut. Rencana longitudinal

menetapkan titik ukur sebelum pembelajaran, sesudahnya, dan setelah peserta kembali bekerja, dengan jadwal yang disesuaikan pada mekanisme dan risiko.

Pengukuran berulang membantu menilai retensi, pemudaran, dan lintasan penerapan. Peserta dapat memperoleh skor tinggi setelah sesi lalu kehilangan sebagian kemampuan jika tidak digunakan. Sebaliknya, kompetensi tertentu menguat setelah praktik beberapa kali dan menerima umpan balik. Perubahan antarwaktu tidak selalu linear. Evaluator dapat menambahkan titik ukur ketika penggunaan diperkirakan mulai terjadi, tetapi menghindari survei berulang yang membebani pekerja tanpa tujuan jelas.

Retensi tidak cukup diukur dengan mengulang soal yang sama. Soal identik dapat memunculkan efek latihan atau ingatan jawaban. Bentuk asesmen alternatif, kasus baru dengan tingkat kesulitan sepadan, serta tugas kerja autentik memberi bukti lebih kuat. Untuk prosedur teknis, demonstrasi berulang menggunakan rubrik yang sama. Penilai perlu dilatih agar perubahan skor tidak mencerminkan standar pengamatan yang berubah. Jika beberapa penilai terlibat, kesepakatan antarpenilai diperiksa pada sampel.

Perubahan individu dan hasil organisasi merupakan tingkat analisis berbeda. Pegawai dapat menguasai keterampilan, tetapi jarang menggunakannya karena kesempatan yang sesuai sedikit. Dalam kasus itu, perilaku per kesempatan lebih informatif daripada jumlah tindakan mentah. Indikator organisasi juga dipengaruhi musim, siklus permintaan, perubahan teknologi, dan kebijakan. Beberapa periode sebelum intervensi membantu melihat tren yang sudah berjalan. Bila terjadi perubahan lain bersamaan, urutannya dicatat.

Komposisi tenaga kerja bisa berubah selama tindak lanjut. Pekerja pindah, keluar, berganti atasan, atau menerima program lain. Evaluator memutuskan apakah pertanyaan menyangkut kohort peserta awal, mereka yang masih bekerja, atau seluruh tenaga kerja pada setiap waktu. Setiap pilihan memberi jawaban berbeda. Analisis kohort mengamati perjalanan peserta awal; pemantauan operasional melihat keadaan pekerja saat ini. Definisi populasi dicantumkan agar hasil tidak disalahartikan sebagai representasi kelompok lain.

Peta waktu dampak membantu menghindari evaluasi yang terlalu cepat atau terlalu lambat. Program keselamatan dapat mengharapkan kepatuhan prosedur berubah lebih awal daripada kecelakaan serius yang jarang. Program kepemimpinan dapat menilai percakapan umpan balik dalam beberapa bulan, sedangkan retensi atau perencanaan suksesi memerlukan waktu lebih lama. Peta tersebut menghubungkan indikator awal dengan hasil akhir dan menetapkan kapan keputusan sementara serta tinjauan berikutnya dilakukan.

Data longitudinal menunjukkan urutan, tetapi urutan waktu saja tidak membuktikan sebab. Skor yang meningkat sesudah program bisa dipengaruhi praktik kerja, perubahan atasan, atau pelatihan lain. Evaluator membandingkan tren dengan kelompok relevan atau menggunakan deret

waktu yang cukup panjang. Tanpa pembanding, kesimpulan menjelaskan perubahan teramati, mekanisme yang didukung sumber lain, serta faktor alternatif yang masih mungkin.

Pengukuran pada beberapa tahap bersifat diagnostik. Jika pengetahuan naik tetapi simulasi tidak berubah, peserta mungkin memahami konsep tanpa latihan yang cukup. Jika simulasi membaik tetapi perilaku kerja tidak berubah, dukungan atasan atau desain pekerjaan mungkin menjadi hambatan. Jika perilaku berubah tetapi indikator layanan tetap, hubungan perilaku dengan hasil atau kapasitas proses perlu ditinjau. Rangkaian ini mencegah respons otomatis berupa menambah jam pelatihan tanpa memeriksa letak masalah.

Retensi kemampuan perlu dibedakan dari retensi kesempatan. Pekerja dapat tetap mampu menjawab soal tetapi jarang memakai keterampilan karena prosedur baru belum tersedia di unitnya. Evaluasi menggabungkan tes, jumlah kesempatan, observasi, serta percakapan tentang lingkungan kerja. Latihan penguatan, pengingat, coaching, atau komunitas praktik dapat diuji sebagai komponen tersendiri. Dengan demikian, organisasi mengetahui apakah perubahan memerlukan desain instruksional, dukungan setelah belajar, atau keduanya.

Titik ukur dipilih berdasarkan keputusan yang akan dibuat. Bila pimpinan perlu memutuskan apakah modul dapat dipersingkat, asesmen penguasaan langsung mungkin cukup. Bila keputusan menyangkut perluasan program ke seluruh cabang, diperlukan observasi transfer dan kondisi implementasi. Bila tujuan menyangkut mobilitas karier, tindak lanjut harus cukup lama untuk memeriksa perpindahan tugas, penempatan, dan perkembangan. Pengukuran tidak dibiarkan berjalan semata karena sistem dapat mengumpulkan log.

Aturan berhenti juga bagian dari rancangan. Setelah pertanyaan terjawab, data yang dibutuhkan untuk analisis disimpan sesuai jadwal, data identitas dihapus atau dipisahkan, dan hak akses ditinjau. Log platform dapat menggambarkan pola belajar individual secara rinci. Evaluator hanya menggunakan tingkat perincian yang perlu untuk tujuan yang telah diberitahukan. Pengumpulan terus-menerus tanpa keputusan yang terkait dapat mengikis kepercayaan dan tidak otomatis menambah mutu bukti.

Hasil longitudinal dapat disajikan sebagai lintasan agregat, bukan profil yang menempel permanen pada pekerja. Visualisasi menunjukkan pola umum, variasi unit, dan rentang ketidakpastian. Jika sampel berubah, jumlah orang pada setiap gelombang disebutkan. Peserta yang pindah atau keluar tidak disembunyikan dari laporan alur. Informasi ini membuat pimpinan memahami apakah perubahan berasal dari perkembangan orang yang sama atau pergantian komposisi.

Jeda antara belajar dan dampak menjadi ruang intervensi manajerial. Atasan dapat menyediakan waktu praktik, umpan balik, pasangan kerja, atau alat bantu. Evaluasi mencatat apakah dukungan ini diberikan dan kapan. Jika program hanya menyediakan kursus tetapi

keberhasilan bergantung pada coaching, paket intervensi harus mendefinisikan coaching sebagai komponen inti. Hal ini memudahkan evaluasi berikutnya membedakan kegagalan teori dari kegagalan pelaksanaan.

## 22. Asesmen kompetensi yang valid dan berkeadilan

Penilaian kompetensi dapat memengaruhi akses pelatihan, sertifikasi, penugasan, dan promosi. Karena konsekuensinya besar, evaluator membedakan asesmen untuk belajar dari asesmen untuk keputusan administratif. Asesmen formatif memberi umpan balik dan kesempatan mencoba dengan aman. Asesmen berisiko tinggi memerlukan kriteria transparan, bukti yang relevan, pemeriksaan konsistensi, serta mekanisme peninjauan. Satu skor tidak semestinya digunakan untuk semua keputusan tanpa dukungan validitas.

Rubrik menjabarkan kinerja yang dapat diamati pada beberapa tingkat kualitas. Kata seperti “unggul” atau “belum siap” terlalu kabur jika tidak disertai perilaku dan contoh. Untuk pengambilan keputusan, rubrik dapat menilai identifikasi fakta, pertimbangan risiko, penggunaan bukti, kejelasan alasan, dan tindak lanjut. Contoh respons jangkar membantu penilai menggunakan standar yang serupa. Rubrik diuji pada tugas nyata dan direvisi ketika interpretasi berbeda.

Penilai membawa pengalaman, ekspektasi, dan bias. Kalibrasi dengan contoh kasus bersama memperlihatkan perbedaan standar. Tim memeriksa apakah skor berubah menurut unit asal, gender, masa kerja, gaya bicara, atau kedekatan dengan penilai, sejauh pengumpulan data dapat dilakukan dengan tepat. Kalibrasi bukan memaksa distribusi skor tertentu. Tujuannya memperjelas kriteria, mengenali pola bias, dan memperbaiki proses. Penilai tetap mencatat bukti yang mendasari penilaian.

Metode asesmen dipilih sesuai kompetensi. Pengetahuan diuji dengan soal penerapan, bukan hafalan istilah semata. Keterampilan teknis mungkin memerlukan demonstrasi, simulasi, atau sampel kerja. Pertimbangan profesional dapat dinilai melalui skenario, alasan yang disampaikan, dan refleksi atas konsekuensi. Pernyataan diri memberi informasi tentang persepsi, tetapi tidak otomatis menunjukkan kemampuan aktual. Semakin jauh asesmen dari tugas, semakin hati-hati interpretasi skor.

Keadilan mencakup aksesibilitas dan akomodasi yang wajar. Waktu, perangkat, bahasa, disabilitas, serta kondisi kerja dapat memengaruhi performa. Jika kompetensi yang diukur ialah diagnosis masalah, gangguan antarmuka yang tidak relevan seharusnya tidak menentukan skor. Tim membedakan tuntutan inti pekerjaan dari hambatan desain instrumen. Uji coba dengan pekerja dari konteks beragam dapat menemukan istilah, jadwal, atau teknologi yang membuat asesmen tidak setara.

Perbandingan antar kelompok mensyaratkan bahwa instrumen memiliki makna sebanding. Instrumen yang dikembangkan bagi satu wilayah atau profesi belum tentu mengukur konstruk yang sama pada kelompok lain. Wawancara kognitif, analisis butir, dan pemeriksaan pola respons membantu menemukan perbedaan tafsir. Jika bukti kesetaraan belum cukup, laporan tidak membuat peringkat kelompok seolah presisi. Keterbatasan disampaikan dan asesmen ditinjau sebelum digunakan untuk keputusan besar.

Pekerja perlu mengetahui apa yang diukur, bagaimana hasil ditafsirkan, siapa yang dapat mengaksesnya, dan keputusan apa yang mungkin dipengaruhi. Umpan balik individual memuat bukti, kekuatan, area pengembangan, serta dukungan yang tersedia. Label tetap seperti “bakat rendah” dapat membatasi peluang dan menyamarkan potensi belajar. Evaluasi sebaiknya membantu perkembangan kompetensi, bukan membekukan identitas berdasarkan satu pengukuran.

Validitas penggunaan berbeda dari konsistensi skor. Penilai mungkin memberi skor serupa, tetapi skor tetap tidak layak untuk promosi jika asesmen tidak mencakup tugas penting jabatan. Argumen validitas menghubungkan konstruk, tugas, bukti, kelompok yang diukur, dan konsekuensi penggunaan. Tim menilai apakah isi instrumen mewakili pekerjaan, apakah faktor tidak relevan memengaruhi hasil, dan apakah interpretasi yang direncanakan didukung bukti.

Uji coba perlu mengikutsertakan pekerja yang mewakili variasi pengalaman dan konteks, bukan hanya perancang. Wawancara kognitif meminta peserta menjelaskan cara memahami butir, memilih jawaban, dan menggunakan skala. Observasi menunjukkan apakah instruksi jelas dan waktu cukup. Istilah sederhana bagi pembuat instrumen dapat memiliki arti berbeda menurut wilayah atau profesi. Revisi dicatat agar versi dan dasar perubahan dapat ditelusuri.

Penilaian multi-sumber menyediakan perspektif yang berbeda, tetapi tidak menghasilkan kebenaran otomatis. Atasan melihat sebagian perilaku, rekan mengalami kolaborasi, bawahan menyaksikan kepemimpinan, dan individu mengetahui niatnya sendiri. Ketidaksepakatan dapat menunjukkan variasi konteks, kesempatan observasi, atau relasi. Skor tidak dirata-ratakan tanpa mempertimbangkan makna. Umpan balik sensitif disampaikan dengan fasilitasi dan rencana pengembangan yang jelas.

Asesmen juga dapat mengubah perilaku jika dikaitkan dengan insentif. Pekerja mungkin berlatih untuk tes, menghindari tugas sulit, atau menampilkan perilaku yang terlihat saja. Organisasi menggunakan beberapa jenis bukti yang relevan, memisahkan pembinaan dari keputusan ketika memungkinkan, dan memantau akibat pada kolaborasi. Ukuran kompetensi yang baik tidak menjamin penggunaan yang adil bila konsekuensi dan aksesnya tidak diawasi.

Keputusan penting perlu memiliki jalur peninjauan. Pekerja dapat meminta klarifikasi dan mengoreksi kesalahan faktual. Penilai lain dapat memeriksa kembali bukti jika prosedur diduga

tidak konsisten. Mekanisme keberatan memungkinkan organisasi menemukan masalah instrumen dan memperbaikinya. Hasil asesmen bukan label permanen; kesempatan belajar ulang dan menunjukkan kemajuan dengan standar yang sama mendukung prinsip pengembangan.

Audit pemerataan memeriksa siapa yang mendapat kesempatan belajar, siapa yang lulus, dan siapa yang dapat menerapkan kompetensi. Perbandingan memperhatikan denominator, konteks jabatan, masa kerja, dukungan, serta kesempatan praktik. Kelompok kecil dilaporkan secara agregat atau tidak ditampilkan bila identitas mudah diketahui. Batas publikasi ditentukan sebelum hasil terlihat agar keputusan privasi tidak dipengaruhi apakah temuan menguntungkan.

Keadilan merupakan unsur mutu metodologis. Hasil hanya berguna jika interpretasi dapat dipertanggungjawabkan bagi kelompok yang terkena dampak. Pemeriksaan akses dan instrumen sebelum program membantu mencegah hambatan, sedangkan pengukuran setelahnya menunjukkan apakah manfaat serta beban terbagi tidak merata. Respons dapat berupa perubahan materi, perangkat, jadwal, dukungan manajer, kriteria, atau kebijakan karier, bukan semata pelatihan tambahan bagi pekerja yang skornya lebih rendah.

## 23. Menyesuaikan desain evaluasi dengan program

Tidak ada satu rancangan yang sesuai untuk setiap intervensi. Program pengetahuan standar dapat memakai tes penguasaan, pemeriksaan retensi, dan perbandingan cara penyampaian. Pelatihan praktik lebih tepat dinilai melalui demonstrasi atau simulasi yang menyerupai tugas. Program kepemimpinan memerlukan pengamatan berulang, umpan balik beberapa pihak, serta penilaian kondisi tim. Pertanyaan evaluasi dimulai dari bentuk perubahan yang diharapkan, bukan dari instrumen yang kebetulan tersedia.

Pelatihan kepatuhan sering dinilai dari tingkat penyelesaian modul. Penyelesaian penting untuk dokumentasi tetapi tidak membuktikan peserta dapat mengenali risiko atau bertindak tepat. Skenario bercabang dapat meminta peserta memilih dan menjelaskan tindakan. Sampel pekerjaan kemudian memeriksa apakah perilaku diterapkan. Jika insiden serius jarang terjadi, indikator awal seperti pemeriksaan keselamatan dan mutu pelaporan dapat melengkapi data, dengan penjelasan bahwa indikator tersebut bukan pengganti langsung insiden.

Program orientasi dinilai melalui kecepatan dan kualitas adaptasi, bukan hanya kehadiran. Ukuran dapat mencakup pemahaman peran, waktu mencapai kompetensi minimum, kesalahan awal, hubungan dengan mentor, norma keselamatan, dan pengalaman pekerja baru. Jadwal pengukuran mengikuti masa adaptasi. Pembanding historis perlu ditafsirkan dengan memperhatikan perubahan rekrutmen, kondisi kerja, dukungan manajer, serta desain peran yang mungkin terjadi bersamaan.

Reskilling menghubungkan belajar dengan perpindahan tugas dan keberlanjutan karier. Sertifikat tidak menunjukkan dengan sendirinya apakah keterampilan relevan atau tersedia posisi yang menggunakannya. Evaluasi menelusuri perjalanan peserta dari pendaftaran dan penyelesaian hingga asesmen, wawancara, penempatan, serta retensi. Titik keluar dapat menunjukkan masalah pada rancangan program, kebijakan karier internal, waktu yang tersedia, atau kebutuhan tenaga kerja. Hasil dibahas bersama unit bisnis agar investasi kompetensi sesuai dengan kebutuhan riil.

Coaching dan mentoring memiliki mekanisme berbeda dari kursus. Mutu hubungan, tujuan, frekuensi, kerahasiaan, kompetensi pendamping, dan dukungan organisasi memengaruhi hasil. Lebih banyak sesi tidak otomatis berarti manfaat lebih besar. Catatan evaluasi menjaga ruang aman dalam hubungan; organisasi menilai hasil dan pola agregat tanpa meminta seluruh isi percakapan pribadi. Wawancara terpisah dapat mengeksplorasi manfaat, hambatan, dan kualitas dukungan.

Pembelajaran berbasis proyek menghasilkan kompetensi sekaligus keluaran kerja. Evaluasi dapat menilai kualitas produk, kolaborasi, keputusan, refleksi, dan penggunaan pembelajaran pada proyek berikutnya. Hasil tim tidak seluruhnya dapat diatribusikan kepada seorang anggota. Rubrik proses dan catatan kontribusi membantu, dengan kriteria yang disepakati agar penilaian tidak berubah menjadi popularitas. Tekanan target perlu diseimbangkan dengan waktu untuk refleksi dan perbaikan.

Program kesejahteraan dan pembelajaran yang berhubungan dengan kesehatan kerja mengukur manfaat sekaligus beban tambahan. Materi dapat meningkatkan rasa mampu, tetapi kursus di luar jam kerja dapat menambah tekanan. Data kesejahteraan dikumpulkan secara proporsional, dengan akses dibatasi. Evaluator tidak menyimpulkan bahwa pekerja bertanggung jawab mengatasi masalah struktural melalui kursus ketahanan. Temuan tentang beban, desain kerja, dan dukungan perlu dibahas bersama hasil individu.

Portofolio program memerlukan prioritas evaluasi. Organisasi dapat mempertimbangkan biaya, risiko, jumlah peserta, kepentingan strategis, dan ketidakpastian bukti. Program dengan paparan luas atau konsekuensi besar menerima evaluasi lebih mendalam. Inisiatif kecil dapat memakai pemantauan ringan dan siklus perbaikan cepat. Prioritas menjaga kapasitas evaluasi tetap terarah pada keputusan yang paling berpengaruh, bukan mengharuskan setiap program menjalankan desain mahal yang sama.

Perbandingan program hanya adil bila sasaran dan konteksnya serupa. Biaya per peserta yang mencapai standar kompetensi dapat membandingkan alternatif jika definisi standar, populasi, horizon, dan mutu bukti konsisten. Program yang melayani pekerjaan berisiko tinggi tidak dapat langsung dinilai dengan ukuran sama seperti kursus pengenalan yang berdampak rendah. Tabel

portofolio mencantumkan tujuan, penerima, mekanisme, hasil, biaya, risiko, dan keputusan berikutnya agar perbedaan tidak tersembunyi di balik satu peringkat.

Ketika program terdiri dari beberapa komponen, evaluasi dapat menguji paket utuh dahulu sebelum membandingkan unsur kecil. Jika hasil paket membaik, analisis pelaksanaan dan variasi unit menunjukkan komponen yang mungkin penting. Eksperimen berikutnya dapat membandingkan jadwal latihan, dukungan supervisor, atau media. Pemisahan komponen tidak dilakukan sebelum diketahui bahwa komponen itu aman dan mekanismenya masuk akal. Keputusan bertahap mengurangi risiko mengoptimalkan bagian kecil sementara program secara keseluruhan tidak relevan.

Pemilihan desain mempertimbangkan risiko tidak melakukan intervensi. Untuk keterampilan keselamatan yang diwajibkan, penugasan acak ke kondisi tanpa pelatihan mungkin tidak etis. Tim dapat menguji urutan atau format sambil memastikan semua pekerja menerima standar yang diperlukan. Untuk program pengembangan sukarela, pengundian akses bertahap mungkin dapat diterima jika kapasitas terbatas dan semua peserta pada akhirnya memperoleh kesempatan. Alasan etik, operasional, dan analitik ditulis bersama.

Program yang berubah sepanjang waktu perlu mencatat versinya. Materi, fasilitator, durasi, teknologi, atau dukungan manajer mungkin berbeda antarangkatan. Jika versi tidak terdokumentasi, perubahan hasil sulit dijelaskan. Catatan fidelity membantu membedakan apakah intervensi yang diuji terlaksana sesuai rancangan atau mengalami adaptasi. Penyesuaian lokal dapat menjadi kekuatan, selama unsur inti dan dasar modifikasi diketahui.

Pada akhirnya, evaluasi program bukan perlombaan memilih metode tercanggih. Metode dipilih karena dapat menjawab keputusan secara proporsional. Bagi sebagian kegiatan, data penyelesaian dan tes cukup untuk mengendalikan mutu dasar. Bagi keputusan perluasan besar, pembanding dan pemantauan transfer layak diinvestasikan. Pertanyaan, dampak keputusan, dan kapasitas organisasi menentukan tingkat bukti yang diperlukan.

## 24. Agenda pengembangan evaluasi dan kesimpulan

Agenda pengembangan evaluasi mencakup pengukuran transfer yang lebih konsisten, desain yang mampu membedakan kontribusi pelatihan dari perubahan konteks, dan evaluasi jangka panjang atas karier serta kapabilitas. Organisasi juga memerlukan bukti mengenai program hibrida, dukungan AI, pembelajaran sosial, dan reskilling lintas pekerjaan. Penelitian perlu menjelaskan instrumen, konteks, karakteristik peserta, pelaksanaan, serta keterbatasan agar hasil dapat ditafsirkan dan direplikasi.

Kemajuan metodologis perlu beriringan dengan pertanyaan normatif: siapa menentukan hasil yang bernilai, siapa memperoleh manfaat, dan siapa menanggung biaya evaluasi? Program dapat

meningkatkan produktivitas tanpa memperbaiki keamanan atau kualitas kerja. Evaluasi harus menyertakan pekerja dalam menentukan pertanyaan, menafsirkan konteks, dan memahami penggunaan data. Partisipasi bukan pengganti metode yang kuat, tetapi memperbaiki relevansi serta keadilan.

Evaluasi yang matang menggabungkan teori perubahan, indikator valid, bukti proses, desain dampak yang proporsional, analisis konteks, dan tata kelola etis. Ia membedakan aktivitas dari hasil, pengamatan dari atribusi, serta korelasi dari kausalitas. Ia mengakui ketidakpastian dan melaporkan temuan negatif. Model seperti Kirkpatrick, CIPP, atau kerangka lain memberi struktur, tetapi tidak menggantikan desain dan penalaran.

Bagi pimpinan, pertanyaan utama bukan hanya apakah peserta menyukai program atau apakah dashboard menunjukkan kenaikan. Pimpinan perlu bertanya kemampuan apa yang meningkat, apakah keterampilan dipakai, kondisi apa yang mendukungnya, apakah hasil relevan bagi organisasi dan pekerja, serta bagaimana biaya dan manfaat didistribusikan. Jawaban yang dapat dipertanggungjawabkan lebih berguna daripada angka ROI yang tampak presisi namun dibangun dari asumsi yang tersembunyi.

Evaluasi menjadi bagian dari pembelajaran organisasi ketika temuannya mengubah keputusan dan memperbaiki kondisi kerja. Program dapat diperluas, disesuaikan, atau dihentikan berdasarkan bukti. Manajer menyiapkan kesempatan transfer; fungsi SDM menghubungkan kompetensi dengan karier; analis menjaga mutu inferensi; dan pekerja mengetahui bagaimana data digunakan. Dengan cara ini, investasi pembelajaran menghasilkan pemahaman sekaligus kapabilitas yang dapat dipelihara.

Pada akhirnya, dampak tidak diukur dengan satu metode atau satu indikator. Dampak dinilai melalui rangkaian bukti yang masuk akal, diperiksa terhadap penjelasan alternatif, dan ditafsirkan dalam konteks. Evaluasi yang jujur tidak menjanjikan kepastian yang tidak tersedia. Ia memberi dasar terbaik untuk tindakan berikutnya dan menunjukkan apa yang masih perlu dipelajari.

## Glossarium

**Analisis kontribusi —** Pendekatan untuk menilai apakah dan bagaimana intervensi turut menyebabkan hasil bersama faktor lain.

**Baseline —** Ukuran kondisi sebelum intervensi atau sebelum perubahan yang dievaluasi.

**Dampak —** Perubahan atau konsekuensi yang relevan dan dikaitkan dengan suatu intervensi melalui argumen bukti.

**Desain kuasi-eksperimental —** Desain pembanding tanpa penugasan acak yang menggunakan strategi untuk mengurangi bias.

**Efek kausal —** Perbedaan hasil yang disebabkan oleh intervensi dibandingkan kondisi alternatif.

**Fidelity implementasi —** Derajat pelaksanaan program sesuai unsur inti rancangan dan protokol.

**Indikator —** Ukuran terdefinisi yang digunakan untuk menilai proses atau hasil.

**Kamus metrik —** Dokumentasi definisi, sumber, rumus, periode, pemilik, dan keterbatasan setiap indikator.

**Model logika —** Pemetaan hubungan antara input, aktivitas, output, hasil, dan dampak yang diharapkan.

**ROI —** Return on Investment; perbandingan manfaat bersih yang diestimasi dengan biaya investasi.

**Teori perubahan —** Penjelasan tentang mekanisme dan asumsi yang menghubungkan kegiatan dengan hasil.

**Transfer belajar —** Penerapan pengetahuan atau keterampilan yang diperoleh dalam pekerjaan.

**Validitas —** Derajat dukungan bukti dan teori bagi interpretasi serta penggunaan hasil ukur.

**Reliabilitas —** Konsistensi hasil pengukuran dalam kondisi yang sebanding.

**Triangulasi —** Penggunaan beberapa sumber atau metode untuk menguji dan memperkaya interpretasi.

**Hasil proksimal —** Perubahan yang relatif dekat dengan intervensi, seperti pengetahuan atau keterampilan.

**Kelompok pembanding —** Kelompok atau kondisi yang membantu memperkirakan hasil tanpa program atau dengan alternatif lain.

**Analisis sensitivitas —** Pemeriksaan bagaimana kesimpulan berubah ketika asumsi atau parameter diubah.

**Retensi belajar —** Kemampuan mempertahankan pengetahuan atau keterampilan setelah jeda waktu.

**Efek samping —** Konsekuensi yang tidak direncanakan, positif atau negatif, dari program.

## Daftar Pustaka

Alsalamah, A., & Callinan, C. (2021). Adaptation of Kirkpatrick's four-level model of training criteria to evaluate the effectiveness of educational programs for head teachers. *Education Sciences, 11*(3), Article 116. https://doi.org/10.3390/educsci11030116

Arthur, W., Jr., Bennett, W., Jr., Edens, P. S., & Bell, S. T. (2003). Effectiveness of training in organizations: A meta-analysis of design and evaluation features. *Journal of Applied Psychology, 88*(2), 234–245. https://doi.org/10.1037/0021-9010.88.2.234

Baldwin, T. T., & Ford, J. K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology, 41*(1), 63–105. https://doi.org/10.1111/j.1744-6570.1988.tb00632.x

Bates, R. (2004). A critical analysis of evaluation practice: The Kirkpatrick model and the principle of beneficence. *Evaluation and Program Planning, 27*(3), 341–347. https://doi.org/10.1016/j.evalprogplan.2004.04.011

Blume, B. D., Ford, J. K., Baldwin, T. T., & Huang, J. L. (2010). Transfer of training: A meta-analytic review. *Journal of Management, 36*(4), 1065–1105. https://doi.org/10.1177/0149206309352880

Brinkerhoff, R. O. (2006). *Telling training's story: Evaluation made simple, credible, and effective* (2nd ed.). Berrett-Koehler Publishers.

Gegenfurtner, A., & Ebner, C. (2019). Toward a model of learning transfer: Revalidation of the Learning Transfer System Inventory in Austria. *Zeitschrift für Erziehungswissenschaft, 22*, 137–161. https://doi.org/10.1007/s11618-018-0848-y

Grossman, R., & Salas, E. (2011). The transfer of training: What really matters. *International Journal of Training and Development, 15*(2), 103–120. https://doi.org/10.1111/j.1468-2419.2011.00373.x

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2006). *Evaluating training programs: The four levels* (3rd ed.). Berrett-Koehler Publishers.

Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology, 78*(2), 311–328. https://doi.org/10.1037/0021-9010.78.2.311

Phillips, J. J., & Phillips, P. P. (2007). Show me the money: The use of ROI in performance improvement, part 1. *Performance Improvement, 46*(3), 9–15. https://doi.org/10.1002/pfi.160

Phillips, P. P., & Phillips, J. J. (2004). ROI in the public sector: Myths and realities. *Public Personnel Management, 33*(2), 139–149. https://doi.org/10.1177/009102600403300201

Reio, T. G., Rocco, T. S., Smith, D. H., & Chang, E. (2017). A critique of Kirkpatrick's evaluation model. *New Horizons in Adult Education and Human Resource Development, 29*(2), 35–53. https://doi.org/10.1002/nha3.20178

Republik Indonesia. (2022). Undang-Undang Republik Indonesia Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi.

Rouse, D. N. (2011). Employing Kirkpatrick's evaluation framework to determine the effectiveness of health information management courses and programs. *Perspectives in Health Information Management, 8*(Spring), 1c.

Schoeb, G., Lafrenière-Carrier, B., Lauzier, M., & Courcy, F. (2021). Measuring transfer of training: Review and implications for future research. *Canadian Journal of Administrative Sciences, 38*(1), 17–28. https://doi.org/10.1002/cjas.1577

Watson, D., Tregaskis, O., Gedikli, C., Vaughn, O., & Ahmed, R. (2018). Well-being through learning: A systematic review of learning interventions in the workplace and their impact on well-being. *European Journal of Work and Organizational Psychology, 27*(2), 247–268. https://doi.org/10.1080/1359432X.2018.1435529