RESSOURCES

Ce qui fait échouer un assessment center

Cinq défaillances silencieuses, leurs contre-feux, et l'objection la plus sérieuse faite à la méthode

---

L'assessment center occupe une place particulière parmi les méthodes d'évaluation : c'est la seule dont la qualité ne se voie pas dans son produit. Un test mal étalonné donne des scores aberrants. Un entretien bâclé se sent. Un assessment center mal conduit, lui, produit exactement ce qu'on attend de lui — des niveaux par compétence, une synthèse, une recommandation — et rien dans ce résultat n'indique qu'il ne mesure plus rien.

Cette propriété est ce qui rend la méthode à la fois puissante et dangereuse. Puissante, parce qu'elle observe des comportements en situation plutôt que de recueillir des déclarations sur soi, et que c'est ce qui lui vaut son rang parmi les prédicteurs de la performance au travail. Dangereuse, parce que son mode de défaillance est **silencieux** : rien ne signale à l'évaluateur que sa journée d'observation n'a produit qu'une impression habillée de chiffres.

La question posée ici n'est donc pas celle de la validité de la méthode, sur laquelle la littérature est abondante. Elle est de savoir ce qui, dans la conduite ordinaire d'une session, détruit cette validité sans que personne ne s'en aperçoive — et ce qu'on peut y opposer qui ne repose pas sur la bonne volonté des évaluateurs.

> **Précision de méthode**
>
> Les chiffres de validité cités ici proviennent de méta-analyses internationales, majoritairement anglo-saxonnes, dont le critère est la performance au poste évaluée par la hiérarchie — un critère lui-même imparfait. Ces estimations ont été révisées à la baisse en 2022 par Sackett et ses collègues, qui montrent que les corrections pour restriction de variance appliquées depuis trente ans surestimaient systématiquement la validité de l'ensemble des prédicteurs. Les ordres de grandeur donnés plus bas se lisent comme des repères de comparaison entre méthodes, non comme des mesures absolues. La section 7 expose l'objection la plus sérieuse faite à la méthode, telle que ses auteurs l'articulent — elle vise directement ce que les sections précédentes tiennent pour acquis.

1. Le mode de défaillance est silencieux

Un dispositif d'évaluation peut échouer de deux façons. Il peut produire un résultat manifestement faux — et l'erreur se corrige. Ou il peut produire un résultat plausible, cohérent, bien présenté, et sans rapport avec ce qu'il prétend mesurer. Le second cas est le propre de l'assessment center, pour une raison structurelle : le dispositif garantit la *forme* de la mesure — des compétences nommées, une échelle, des observateurs — indépendamment de sa substance.

Retirez d'une corbeille de courrier ses contradictions délibérées et ses échéances non négociables : elle reste administrable, elle se cote normalement, elle produit des notes. Elle ne discrimine plus. Aucun indicateur interne ne le signale. C'est ce qui distingue

l'assessment center d'un instrument psychométrique, où la fidélité et l'étalonnage sont mesurables sur l'instrument lui-même.

La conséquence pratique est que la qualité d'un assessment center ne se contrôle pas *a posteriori*, sur ses résultats. Elle se garantit *a priori*, par la procédure. Les cinq défaillances qui suivent sont les cinq points où cette procédure cède le plus souvent — et aucune ne relève de l'incompétence : ce sont des comportements normaux d'évaluateurs expérimentés.

2. L'évaluateur seul

C'est la défaillance la plus fréquente, et celle qu'on s'autorise pour de bonnes raisons — la disponibilité des managers, le coût d'une journée à deux, la confiance qu'on accorde à un observateur aguerri.

Un évaluateur seul ne produit pas une mesure. Il produit un jugement, qui peut être excellent, et dont rien ne dit s'il l'est. La fiabilité d'une observation comportementale ne se démontre que par la convergence : deux personnes qui ont vu la même séquence et qui aboutissent au même niveau, pour les mêmes raisons, produisent une information d'une autre nature que celle d'un observateur isolé, si compétent soit-il. Les recommandations professionnelles internationales font de la pluralité des évaluateurs et de l'intégration des données par confrontation deux éléments constitutifs de la méthode, pas des raffinements optionnels (International Taskforce on Assessment Center Guidelines, 2015).

Le contre-feu

Deux évaluateurs par candidat au minimum, avec des rôles distincts. L'acteur d'un jeu de rôle ne cote pas : il ne peut pas tenir un rôle et observer simultanément, et son ressenti de partenaire est une donnée d'un autre ordre — précieuse, mais qui se transmet séparément, après la cotation, sans participer à la décision de niveau.

Un second point, souvent négligé dans les dispositifs internes : un évaluateur ne doit pas coter une personne dont il est, ou sera, le responsable hiérarchique direct. Il ne peut pas ignorer ce qu'il sait d'elle, et sa décision est en général déjà prise. Quand c'est inévitable, la seule conduite tenable est de le dire au candidat et de faire coter par un tiers.

3. Coter de mémoire

La deuxième défaillance tient à un décalage de quelques dizaines de minutes. L'évaluateur observe, prend quelques notes, puis renseigne sa grille à la fin de la séquence — parfois après le déjeuner, parfois le soir. Il coterait mieux, pense-t-il, avec le recul.

C'est l'inverse. La mémoire ne conserve pas, elle reconstruit : elle lisse les contradictions, retient ce qui confirme l'impression d'ensemble et efface le reste. Un évaluateur qui cote une heure après ne cote pas le candidat, il cote le souvenir qu'il en a — et ce souvenir a déjà été mis en forme par son jugement. Le phénomène est d'autant plus insidieux qu'il

produit une *plus grande* cohérence apparente : le profil obtenu paraît net, articulé, convaincant.

Le contre-feu

Séparer strictement deux gestes que l'habitude confond. Pendant l'épreuve, on consigne *ce que le candidat a dit ou fait*, horodaté, aussi littéralement que possible, sans attribuer aucun niveau. Après la séquence, on cote à partir de ces notes.

Le test de qualité d'une note d'observation est simple : permettrait-elle à un collègue absent de se faire une idée ? « Manque de recul » n'est pas une observation, c'est une conclusion. L'observation serait : *« interrogé sur les causes du décrochage commercial, il a cité trois éléments du dossier sans les hiérarchiser ni les relier, et n'a pas repris la question sur ce qui relevait de son périmètre »*. Une note qui ne contient ni verbe d'action, ni citation, ni moment n'est pas une note d'observation.

> **L'effet secondaire est en réalité l'essentiel.**
>
> Cette exigence rend la cotation discutable. Deux évaluateurs peuvent confronter des faits. Ils ne peuvent rien faire de deux impressions.

4. L'effet de halo, et sa variante la plus coûteuse

Une impression dominante colore l'ensemble des compétences cotées. Le candidat brillant à l'oral se retrouve noté au plus haut niveau en analyse, alors que sa production écrite était moyenne. C'est le risque numéro un du dispositif, documenté depuis un siècle, et le plus difficile à repérer de l'intérieur : l'évaluateur qui en est victime a le sentiment d'une évaluation particulièrement cohérente.

La variante la plus coûteuse en assessment center mérite son propre nom : **l'aisance verbale prise pour de la pensée**. Le candidat qui parle avec fluidité, vocabulaire riche et assurance paraît analyser mieux qu'il ne le fait. L'inverse est vrai aussi, et plus injuste : un candidat qui cherche ses mots devant deux observateurs silencieux paraît confus alors que son écrit était structuré.

Les contre-feux

— Coter compétence par compétence à travers les candidats, et non candidat par candidat. Traiter les neuf compétences d'une même personne d'affilée fabrique de la cohérence là où il n'y en a pas nécessairement : un bon analyste peut être un piètre communicant, et le profil n'a d'intérêt que si les compétences ont été cotées indépendamment.

— Attribuer l'aisance orale à la compétence qui lui revient — la communication — et revenir à l'écrit pour juger de l'analyse, l'écrit ne bénéficiant d'aucun charme.

— Vérifier la distribution de ses propres niveaux après trois ou quatre candidats. Un évaluateur qui n'a jamais posé le niveau le plus bas ni le plus haut n'observe pas des candidats moyens : il a une dérive personnelle, et elle est corrigible une fois nommée.

5. La moyenne, qui n'est pas un jugement

Vient le moment de la confrontation des cotations, en fin de journée, quand chacun voudrait rentrer. Deux évaluateurs annoncent leurs niveaux, constatent un écart, et retiennent la moyenne. L'opération paraît équitable. Elle est le contraire d'une mesure.

Une moyenne entre un 2 et un 4 donne 3, c'est-à-dire le seul niveau qu'*aucun des deux observateurs n'a constaté*. Elle escamote précisément ce que l'écart contenait d'information : soit l'un des deux a vu quelque chose que l'autre n'a pas vu, soit les deux n'entendent pas la compétence de la même façon. Dans le premier cas, la discussion enrichit le jugement ; dans le second, elle révèle un défaut de définition qui affectera tous les candidats suivants.

Une seconde erreur se loge au même endroit, plus discrète : **l'effet d'ancrage**. Le premier qui annonce un chiffre fixe la discussion. Une fois « moi je l'ai mis à 4 » prononcé, on ne parle plus du candidat, on parle de ce 4. Et si celui qui a parlé le premier est le plus gradé, la confrontation devient une validation.

Les contre-feux

— Chacun écrit ses niveaux avant d'entrer dans la salle et ne les modifie plus avant de les avoir tous annoncés.

— On commence par les observations, pas par les chiffres : chacun expose ce qu'il a vu sur la compétence traitée, et les niveaux ne se révèlent qu'ensuite, simultanément. Quand ils doivent l'être l'un après l'autre, le moins gradé parle en premier.

— On traite une compétence à la fois, à travers toutes les épreuves — jamais « alors, ce candidat, qu'en pensez-vous ? », question qui appelle une impression globale, c'est-à-dire exactement ce que le dispositif cherche à éviter.

— Quand un désaccord de deux niveaux persiste après retour aux faits, on retient le niveau le plus bas en consignant l'écart. Un désaccord tracé vaut mieux qu'un consensus de façade : il documente l'incertitude, et il servira à la session suivante.

Un écart marqué entre deux épreuves pour une même compétence n'est pas un problème à résoudre : c'est une information. Un candidat qui analyse finement par écrit et superficiellement à l'oral ne se moyenne pas, il se décrit.

6. La synthèse sans faits

La dernière défaillance est la plus visible de l'extérieur, et c'est celle qui expose juridiquement. Une synthèse qui énonce « bonne capacité d'analyse », « manque de recul », « profil encore immature » ne dit rien d'observable. Le candidat ne peut rien y contester : on n'argumente pas contre une étiquette. Et l'évaluateur qui l'a rédigée ne peut rien y défendre, si la décision vient à être contestée.

La règle tient en une phrase : **chaque appréciation s'adosse à un comportement observé, cité**. Non pas « bonne capacité d'analyse », mais « a relevé la contradiction entre deux documents du dossier sans y avoir été invité, et en a tiré une conséquence sur son plan d'action ». Cette exigence protège deux personnes à la fois : le candidat, qui peut discuter d'un fait ; l'évaluateur, dont le jugement adossé à des comportements datés se défend, alors qu'un jugement adossé à des qualités abstraites ne se défend pas.

Elle a un corollaire que les rédacteurs franchissent souvent sans y penser : les qualificatifs de personnalité — *rigide, immature, fragile, dominateur* — ne relèvent pas de ce qui a été observé. Ils relèvent d'une inférence sur ce que la personne *est*, à partir d'une journée de situations simulées. C'est sortir du domaine de validité de l'outil, et c'est le point sur lequel une contestation prospère.

Il faut y ajouter la question du potentiel, qui sera posée. Un assessment center observe des comportements dans des situations proches d'*un* poste. Il ne dit pas ce que la personne deviendra dans deux ans. Quand aucun comportement au-delà de l'exigence du poste n'a été observé, la formulation honnête est celle-là même — et non la conclusion à une absence de potentiel, que l'outil ne mesure pas.

> **Une synthèse sans point faible n'est pas bienveillante, elle est inutile.**
>
> Les faits la démentiront dans les six mois. Dire clairement ne veut pas dire écraser : on décrit le comportement, sa conséquence observée, et ce qui le rendrait différent. Aucun avis défavorable ne devrait sortir sans un engagement sur les moyens et une date de réexamen.

7. L'objection la plus sérieuse, et ce qu'elle change

Tout ce qui précède suppose que l'on cote des *compétences*. Cette supposition est contestée depuis quarante ans, et par les travaux les plus solides du domaine.

En 1982, Sackett et Dreher examinent les cotations de trois assessment centers et constatent un résultat déconcertant : les cotations d'une même compétence à travers plusieurs exercices corrèlent faiblement entre elles, tandis que les cotations de compétences différentes au sein d'un même exercice corrèlent fortement. Les analyses factorielles font apparaître des facteurs d'*exercice*, non des facteurs de *compétence*. Autrement dit, ce que l'évaluateur cote ressemble davantage à « comment ce candidat s'est comporté dans cet exercice » qu'à « quel est son niveau sur cette dimension ». Le résultat a été répliqué à de nombreuses reprises, en sélection comme en développement. On l'appelle l'*effet d'exercice*.

Cette objection est sérieuse et elle n'est pas résolue. Elle a une conséquence directe sur ce qui précède : l'écart entre deux épreuves sur une même compétence, que la section 5 présente comme une information sur le candidat, pourrait n'être qu'un artefact de la méthode.

Trois éléments viennent cependant nuancer sa portée.

Le premier est que la validité prédictive de la méthode, elle, ne s'effondre pas. La méta-analyse de Gaugler et ses collègues (1987) établissait la validité de l'évaluation globale autour de .37, soit environ 14 % de la variance de la performance expliquée. Arthur et ses collègues (2003) sont allés plus loin : en regroupant 168 intitulés de dimensions en six dimensions générales, ils obtiennent des validités estimées comprises entre .25 et .39, et surtout une combinaison de quatre dimensions qui explique davantage de variance (environ 20 %) que l'évaluation globale. Des cotations dimensionnelles qui

prédisent mieux que la note d'ensemble ne peuvent pas être entièrement dépourvues de contenu.

Le deuxième est que l'effet d'exercice est peut-être moins un défaut de la méthode qu'un fait sur les personnes. Qu'un manager analyse mieux par écrit qu'en réunion n'est pas une anomalie de mesure : c'est une information de premier ordre pour qui doit décider de lui confier un poste. La difficulté n'est pas que le comportement varie selon la situation ; elle est que le dispositif force ensuite cette variation dans une note unique par compétence.

Le troisième est qu'aucune des estimations citées ne doit être prise pour une mesure absolue. Les révisions de Sackett et ses collègues (2022) montrent que les corrections pour restriction de variance appliquées depuis trente ans ont surestimé la validité de l'ensemble des méthodes de sélection — l'entretien structuré est ainsi ramené à .42 et les tests d'aptitude cognitive à .31. Ce qui reste solide est l'ordre relatif des méthodes, pas la valeur de chacune.

Ce que cette objection impose en pratique tient en deux points. D'abord, ne jamais présenter une cotation par compétence comme la mesure d'un trait stable : c'est un comportement observé dans des situations données, et la synthèse doit le dire ainsi. Ensuite, multiplier les situations plutôt que les dimensions. Un dispositif qui observe six compétences dans cinq épreuves variées vaut mieux qu'un dispositif qui en observe douze dans deux exercices — lesquels ne mesureront, au mieux, que deux choses.

C'est du reste la ligne générale à tenir sur l'évaluation : préférer les **référentiels de situations** aux référentiels d'attributs, et les ancrages comportementaux aux adjectifs. La compétence n'est pas observable ; on n'observe que des performances dont on l'infère.

8. Le contrôle qui remplace la certification

Reste une question que la plupart des dispositifs internes éludent : comment savoir que ses évaluateurs évaluent bien ? La réponse habituelle est la certification — on forme, on certifie, on considère le problème réglé. Elle a deux défauts : elle est coûteuse, et elle ne mesure rien après coup.

Il existe un contrôle plus simple, que tout dispositif produit sans effort supplémentaire : **l'accord inter-juges**. À partir des cotations individuelles relevées avant discussion, trois indicateurs suffisent — la proportion de compétences où les deux évaluateurs ont posé le même niveau, la proportion où l'écart ne dépasse pas un niveau, et l'écart moyen signé de chaque évaluateur sur l'ensemble de ses cotations. Le troisième est le plus instructif : s'il est systématiquement positif chez quelqu'un, cette personne est indulgente ; négatif, elle est sévère. C'est une dérive personnelle, stable, et corrigible une fois nommée.

Sur une dizaine de candidats, ces relevés disent lequel des évaluateurs dérive, dans quel sens, et quelles compétences sont mal définies dans l'équipe — trois informations qu'aucune certification ne fournirait. Contrôle en aval plutôt que diplôme en amont.

Ce qui distingue un dispositif sérieux

Les cinq défaillances décrites ici ont un trait commun : aucune ne relève de la mauvaise foi ni de l'incompétence. Elles sont le fonctionnement normal du jugement humain sous contrainte de temps. On ne les supprime pas par la vigilance ou l'expérience — on les contrarie par une procédure qui ne dépend pas de la vertu de celui qui l'applique.

C'est le critère auquel se reconnaît un dispositif sérieux, et la question à poser à tout fournisseur : non pas « comment vos évaluateurs sont-ils formés ? », mais « qu'est-ce qui, dans votre procédure, rend une erreur d'évaluation *visible* ? ». Un assessment center dont le mode de défaillance reste silencieux n'est pas un instrument de mesure. C'est une mise en scène de la mesure, et elle coûte une journée à tout le monde.

Une dernière distinction, qui commande toutes les autres : séparer l'évaluation qui *décide* de l'évaluation qui *développe*. Les deux n'appellent ni la même rigueur procédurale, ni la même restitution, ni le même degré de prudence dans les conclusions. Les confondre est la manière la plus sûre de rater les deux.

Références

Arthur, W. Jr., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. Personnel Psychology, 56(1), 125-153.

Code du travail, articles L. 1221-6, L. 1221-8 et L. 1221-9 (recrutement : lien direct et nécessaire, information préalable du candidat sur les méthodes, confidentialité des résultats, pertinence au regard de la finalité) et L. 1222-3 (évaluation d'un salarié en poste).

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C. III, & Bentson, C. (1987). Meta-analysis of assessment center validity. Journal of Applied Psychology, 72(3), 493-511.

International Taskforce on Assessment Center Guidelines (2015). Guidelines and ethical considerations for assessment center operations (6e éd.). Journal of Management, 41(4), 1244-1273.

ISO 10667-1 et 10667-2. Livraison d'un service d'évaluation : modes opératoires et méthodes d'évaluation des personnes au travail et des paramètres organisationnels. Partie 1 : exigences pour le client. Partie 2 : exigences pour le prestataire.

Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions : some troubling empirical findings. Journal of Applied Psychology, 67(4), 401-410.

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2022). Revisiting meta-analytic estimates of validity in personnel selection : addressing systematic overcorrection for restriction of range. Journal of Applied Psychology, 107(11), 2040-2068.

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2023). Revisiting the design of selection systems in light of new findings regarding the validity of widely used predictors. Industrial and Organizational Psychology, 16(3), 283-300.

Réserve de comparabilité : les méta-analyses citées portent sur des dispositifs anglo-saxons, avec pour critère la performance au poste évaluée par la hiérarchie. Les coefficients ne sont pas transposables tels quels à un assessment center français de promotion interne, et les estimations d'avant 2022 sont surévaluées pour une raison méthodologique désormais documentée. Leur intérêt est comparatif.

Document préparé avec l'assistance d'une IA (recherche et rédaction), puis révisé et validé par François Labbas, qui en assume la responsabilité éditoriale. Chiffres et références à vérifier à leur source.