Deux précautions accompagnent ce recueil. La première est éthique : pendant l'étude 2, les comptes étaient ouverts par un code de classe et un code personnel, sans aucun nom, si bien que les traces relèvent d'un pseudonymat, l'enseignant seul pouvant relier un code à un élève ; leur recueil est annoncé à l'enseignant dans la fiche G2, et aucun relevé individuel n'est communiqué à un tiers. La seconde est interprétative : une trace atteste d'un accès, non d'un usage pédagogique. Un chapitre consulté pendant douze minutes peut avoir été travaillé ou laissé ouvert. Les traces établissent donc des bornes et non des faits d'usage, et elles ne prennent leur sens que croisées avec le journal et l'auto-observation.

7.8. Le calendrier du terrain

Trois points de contrôle ont jalonné cette séquence. Les instruments ont été figés avant le T0, sans pré-test : aucune modification n'a été admise une fois le T0 engagé. Après le T0, le taux de présence a été vérifié dans les deux classes, 34 élèves sur 36 et 36 sur 38, soit plus de neuf élèves sur dix dans chaque classe. En cours d'usage, un journal non rempli le dimanche soir déclenchait une relance le lundi ; une seule relance a été nécessaire, en semaine 2.

VN-Ch07-6 : La grille d'évaluation experte F1 et la question du fractionnement

Complément du chapitre 7, section 7.5.1

7.5. L'instrument de l'étude 4

7.5.1. La grille d'évaluation experte

L'étude 4 confie l'évaluation du dispositif à des experts extérieurs à sa conception, au moyen d'une grille de soixante-seize critères organisés en quatre dimensions et seize sous-dimensions.

La première dimension, pédagogique, couvre le contenu, la progression, l'interactivité et l'engagement, la différenciation et l'évaluation. La deuxième, technique, couvre l'ergonomie et l'interface, l'accessibilité, la performance et la fiabilité, et la compatibilité avec les équipements disponibles. La troisième porte sur l'organisation de l'information, la navigation et la recherche, et la médiation documentaire. La quatrième porte sur la personnalisation, la flexibilité d'usage, la collaboration et le partage, et l'autonomie de l'utilisateur.

Cette structure, qui n'a pas été construite pour la circonstance, recoupe pourtant le cadre de la partie II. La troisième dimension correspond à la dimension documentaire du chapitre 3, et la quatrième décline les conditions de l'appropriation au sens du chapitre 4 : maîtrise, intégration

aux pratiques, possibilité de création. La grille permet donc de faire juger par des tiers, sur des critères observables, ce que le cadre théorique avait posé comme constitutif du dispositif.

Chaque critère est coté sur une échelle en quatre niveaux, de l'insuffisant à l'excellent, assortie d'une modalité « non applicable ». Le choix d'une échelle paire est délibéré : il interdit le refuge dans une position médiane et oblige l'évaluateur à se prononcer dans un sens ou dans l'autre. Chaque dimension se clôt par un espace de commentaire libre, dont les réponses sont analysées au chapitre 14 à côté des cotations.

Le panel visait douze à quinze experts répartis sur quatre profils : des enseignants expérimentés du niveau concerné, des inspecteurs pédagogiques, des chercheurs en didactique des sciences ou en sciences de l'information et de la communication, et des concepteurs de ressources numériques éducatives. Seize experts ont été sollicités en octobre 2026, treize ont accepté et dix ont achevé la grille au cours du mois : quatre enseignants, deux chercheurs en didactique, deux en sciences de l'information et de la communication, et deux concepteurs. Le formulaire F1 n'a recueilli, pour chaque expert, que son profil et son ancienneté dans ce profil. Aucun des trois inspecteurs sollicités n'a achevé le formulaire, absence que le chapitre 14 signale comme une limite du jugement de conformité au cadre prescrit. Les experts ont examiné le prototype figé avant l'usage, celui que la classe utilisatrice de l'étude 2 a employé ; l'écart entre cet état et la version du dispositif décrite au chapitre 14 explique certains de leurs commentaires. L'effectif minimal opérationnel fixé à l'avance par le document F3 est de dix experts, avec au moins deux experts par profil. Le premier seuil est atteint, le second ne l'est pas pour les inspecteurs, dont aucun ne figure dans le panel. La diversité des profils, qui donne son sens à l'exercice, est donc incomplète.

7.5.2. La question du fractionnement

Soixante-seize critères représentent une durée de passation que l'expérience des enquêtes en ligne situe au-delà du seuil d'abandon habituel. Deux options se présentaient : réduire le nombre de critères, ou fractionner la passation.

La réduction a été écartée parce qu'elle supposerait d'arbitrer entre des critères dont la pertinence n'a pas encore été éprouvée, arbitrage qu'aucune donnée ne permet aujourd'hui de fonder. Le dispositif retenu est donc le maintien de la grille complète en un envoi unique, assorti d'une version fractionnée en quatre envois espacés, préparée à l'avance pour le cas où le taux d'abandon du premier envoi dépasserait le tiers. Cette solution conservait l'information et reportait le risque sur la logistique plutôt que sur le contenu. La version fractionnée, préparée

dans le document F2, n’a pas été déclenchée : la grille complète a été administrée en un seul envoi.

VN-Ch07-7 : Paramètres des analyses de sensibilité, cohérence interne et indices des items

Complément du chapitre 7, section 7.7.1

7.6.2. Absence de pré-test et cohérence interne

La cohérence interne des échelles est estimée par le coefficient alpha de Cronbach (1951), calculé par sous-échelle et non pour l’instrument entier, un alpha global sur un questionnaire multidimensionnel n’ayant guère de sens. La valeur de 0,70 sert de repère usuel de lecture. Faute de pré-test, les coefficients ont été calculés après le recueil, sur les effectifs complets, et ils sont rapportés au chapitre 9.

7.6.4. Qualité des items du test de connaissances

La quatrième question est propre au test d’acquisition. Après la passation en T0, chaque item fait l’objet de deux indices classiques. L’indice de difficulté rapporte le nombre de réponses correctes à l’effectif : un item réussi par la quasi-totalité ou échoué par la quasi-totalité n’apporte aucune information discriminante. L’indice de discrimination, estimé par la corrélation point-bisériale corrigée, mesure la capacité de l’item à séparer les élèves qui réussissent l’ensemble du test de ceux qui y échouent : un item que les meilleurs réussissent moins souvent que les plus faibles est un item défectueux.

7.7.1. Traitement des données quantitatives

Des analyses de sensibilité accompagnent ces traitements. Elles ont été définies après le recueil, au moment des analyses, et elles ne modifient aucun critère : elles disent la solidité des résultats. Le rééchantillonnage avec remise (Efron et Tibshirani, 1993), 5 000 tirages au même effectif et à graine fixée, éprouve la stabilité des classements et des moyennes ; pour l’accord entre experts, ce sont les critères et non les experts qui sont rééchantillonnés, la duplication d’un expert gonflant artificiellement l’accord. Un test de permutation, sur 10 000 permutations des étiquettes de classe, éprouve l’écart entre les deux classes de l’étude 2 sans hypothèse de distribution. Une analyse multivers (Steegen et al., 2016) rapporte ensemble tous les choix défendables de score et de modèle. La sensibilité à l’effet de grappe recalcule l’erreur type de