

Manuel de nettoyage des données du PASEC



pasec
Programme d'analyse des systèmes
éducatifs de la confemen

Équipe du PASEC

Le présent document a été préparé par l'équipe du Programme d'analyse des systèmes éducatifs de la CONFEMEN (PASEC).

- Hilaire Hounkpodoté, Coordonnateur du PASEC
- Bassile Tankeu, Chef de la division « Statistiques, analyse des données et études » du PASEC
- Labass Lamine Diallo, Chef de la division « Conception des instruments, collecte des données et renforcement des capacités » du PASEC
- Ousmane Birba, Conseiller technique du PASEC
- Seydou Garba Hamidou, Conseiller technique du PASEC
- Tchindebé Berkeunbé, Conseiller technique du PASEC
- Valérie Vinuesa, Conseillère technique du PASEC
- Fatimé Zara Boukar, Conseillère technique du PASEC
- Adama Bologo, Chargé de projet du PASEC
- Khady Mbaye Camara, Assistante administrative principale du PASEC
- Penda Seck Fall Diouf, Assistante administrative adjointe du PASEC
- Ulrich Tadajeu Kenfack, Attaché de communication du PASEC

Relecture

Alexis LOYE, Président du comité scientifique du PASEC

Publié en 2026 par le Programme d'analyse des systèmes éducatifs de la CONFEMEN (PASEC)

Complexe Sicap Point E, Immeuble C, 3^e étage, Avenue Cheikh Anta Diop, BP : 3220 Dakar, Sénégal

© PASEC 2026

Ce document est disponible en accès ouvert sous licence Creative Commons Attribution – Partage dans les mêmes conditions 4.0 International – CC BY-SA 4.0, sauf indication contraire.

Toute utilisation, reproduction, adaptation, traduction ou diffusion, totale ou partielle, est autorisée à condition de mentionner la source, de citer correctement la publication et de diffuser toute œuvre dérivée sous une licence équivalente.

Pour citer ce document :

PASEC. 2026. *Manuel de nettoyage des données du PASEC*. Dakar : PASEC.

La CONFEMEN encourage l'utilisation et la diffusion des informations contenues dans cette publication à des fins de recherche, de formation, d'élaboration de politiques publiques et de plaidoyer en faveur d'une éducation de qualité.

Conception graphique et mise en page : Suzanne Grossmann

ISBN : 92-9133-222-4

Sommaire

Liste des tableaux et figures	4
Introduction	5
1. Qualité des données et comparabilité internationale	7
2. Concordance entre le fichier Suivi des écoles et le fichier Suivi des élèves	9
2.1. Données manquantes	9
2.2. Données qui ne respectent pas les règles prescrites pour les valeurs minimale et maximale	9
2.3. Doublons	10
2.4. Cohérence des informations entre plusieurs enregistrements dans un fichier de données	10
2.5. Cohérence des informations d'un enregistrement au sein d'un fichier de données	10
2.6. Cohérence d'information entre les deux fichiers de suivi	11
3. Concordance entre le fichier Suivi des élèves et les fichiers de Livret élèves (Tests et Questionnaire)	13
3.1. Étape 1 : détection des doublons dans chacun des fichiers	13
3.2. Étape 2 : cohérence des codes Élève	15
3.3. Étape 3 : fusion des 4 fichiers Élève en sixième ou Enseignants du primaire ou Élève en neuvième	15
3.4. Étape 4 : vérification de la concordance entre le fichier Suivi des Élèves et le fichier Élèves	16
3.5. Étape 5 : définition des élèves participants et recodage des fichiers Suivi des Élèves, Suivi des Écoles et Élèves	18
4. Concordance entre le fichier Suivi des écoles et les fichiers Questionnaires Enseignants et Questionnaires écoles	21
4.1. Vérification de l'unicité des identifiants	21
4.2. Vérification de la correspondance avec le fichier de suivi des écoles	21
4.3. Vérification des questionnaires entièrement vides	22
4.4. Contrôles complémentaires de cohérence	22
5. Concordance entre le fichier suivi des élèves et questionnaires parents	22
6. Les recodages	23
6.1. La catégorisation des données manquantes	23
6.2. Création de variable composite	25
6.3. Recodage de réponses courtes	25
6.4. Autres recodages	25
7. Vérification des variables	28
7.1. Les statistiques univariées	28
7.2. Les statistiques multivariées par pays	31
8. Imputation des données manquantes	33

Liste des tableaux et figures

Tableau

TABLEAU 1 Statut de participation des élèves 19

Figures

FIGURE 1 Distribution normale et exclusion 7

FIGURE 2 Exemple de question emboîtée 24

FIGURE 3 Question matricielle à réponses dichotomiques 24

FIGURE 4 Exemple de question emboîtée dans le questionnaire enseignant 30

FIGURE 5 Exemple de QCM comportant deux choix de réponse 30

Introduction

Avant toute opération de pondération, de calcul des scores ou de mise à l'échelle des compétences cognitives, il est indispensable de procéder à une vérification approfondie de la qualité des données collectées et saisies. Cette étape vise notamment à :

- vérifier l'intégrité de chaque fichier ;
- contrôler la cohérence interne des données ;
- s'assurer de la concordance entre les différents fichiers ;
- identifier les erreurs de saisie, de codification ou de traitement ;
- détecter les valeurs aberrantes ou incohérentes ;
- documenter et corriger les données manquantes.

Pour répondre à ces exigences de qualité, un processus structuré de nettoyage des données est mis en œuvre. Ce processus constitue une étape essentielle pour garantir la qualité scientifique de l'évaluation et assurer la comparabilité des résultats entre les pays participants. Il est réalisé sous la coordination du PASEC en étroite collaboration avec les centres nationaux. À chaque étape du processus, le PASEC transmet au centre national un rapport détaillant les incohérences, erreurs ou données manquantes détectées dans les fichiers. Le centre national est alors invité à vérifier les informations à la source et à fournir, dans les meilleurs délais, les clarifications ou corrections nécessaires. Dans la mesure du possible, les données erronées doivent être corrigées plutôt que supprimées.

Afin de garantir la transparence et la reproductibilité des traitements, toutes les modifications apportées aux données doivent être documentées et conservées. À cet effet, les corrections doivent être réalisées au moyen de programmes informatiques reproductibles (par exemple sous Stata). En outre, aucun fichier de données ne doit être écrasé : une nouvelle version du fichier doit être créée après chaque étape de traitement.

Les fichiers disponibles pour ce travail de nettoyage des données sont les suivants :

1. Niveau primaire

• Fichiers de suivi

- *Fichiers de suivi des écoles (un pour le début du primaire et un pour la fin du primaire) :* Chaque fichier contient les informations issues des fiches de suivi des écoles sélectionnées pour l'évaluation. Il recense les écoles participantes ainsi que celles échantillonnées n'ayant pas participé à l'enquête (refus, exclusion, fermeture temporaire, etc.), qu'elles aient été remplacées ou non pour chacun de ces deux niveaux ;
- *Fichiers de suivi des élèves (un pour le début du primaire et un pour la fin du primaire) :* Chaque fichier regroupe les informations contenues dans les fiches de suivi des élèves des écoles participantes pour chacun de ces deux niveaux.

• Fichiers des tests et questionnaires

- *Fichier des livrets de test des élèves du début du primaire ;*
- *Fichier des questionnaires élèves du début du primaire ;*
- *Fichiers des livrets de test élèves A, B, C et D de fin du primaire ;*
- *Fichier des questionnaires élèves de fin du primaire ;*
- *Fichiers des livrets de test enseignants A, B, C et D ;*
- *Fichier des questionnaires enseignants ;*
- *Fichier des questionnaires directeurs d'école ;*
- *Fichier des questionnaires parents d'élèves du début du primaire ;*
- *Fichier des questionnaires parents d'élèves de fin du primaire.*

2. Niveau secondaire (collège)

- **Fichiers de suivi**

- *Fichier de suivi des établissements* : contient les informations issues des fiches de suivi des établissements secondaires. Il recense les établissements participants ainsi que ceux sélectionnés mais n'ayant pas participé à l'enquête, qu'ils aient été remplacés ou non.
- *Fichier de suivi des élèves* : regroupe les informations issues des fiches de suivi des élèves des établissements participants.

- **Fichiers des tests et questionnaires**

- *Fichiers des livrets de test élèves A, B, C et D* ;
- *Fichier des questionnaires élèves* ;
- *Fichier des questionnaires enseignants* ;
- *Fichier des questionnaires directeurs d'établissement* ;
- *Fichier des questionnaires parents d'élèves*.

1. Qualité des données et comparabilité internationale

La qualité des données constitue une condition essentielle à la comparabilité internationale des résultats. Des données erronées ou incomplètes peuvent entraîner des biais dans les estimations et compromettre la validité des comparaisons entre pays. Lorsqu'une anomalie est détectée, deux approches sont possibles :

1. la suppression de la donnée concernée lorsqu'aucune correction fiable n'est possible ;
2. la consultation du centre national afin de confirmer, corriger ou compléter l'information.

Dans la mesure du possible, la suppression de données doit être évitée. En effet, toute suppression réduit la taille effective de l'échantillon et peut entraîner une perte de précision statistique ainsi qu'un risque accru de biais. Cette précaution est particulièrement importante lorsque les données manquantes ou les exclusions ne sont pas réparties aléatoirement au sein de la population étudiée.

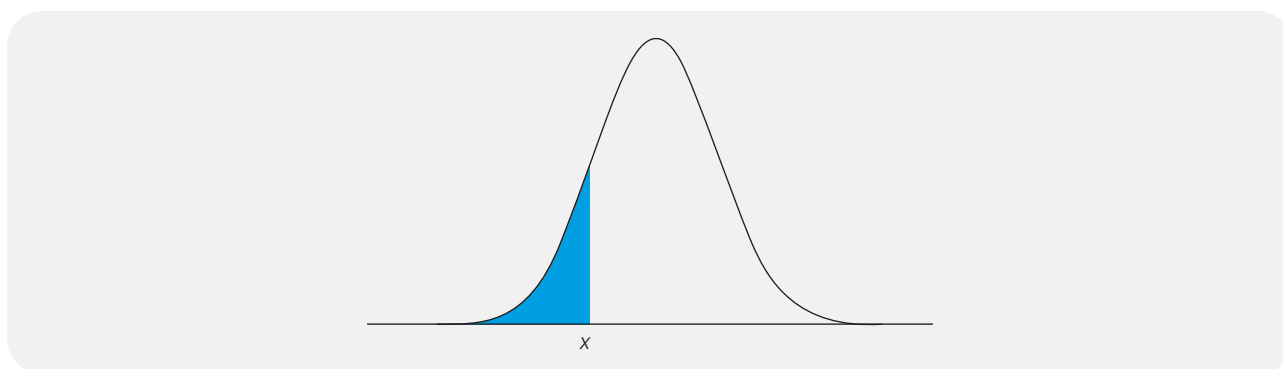
La non-réponse et les exclusions peuvent introduire un biais lorsque les unités qui ne participent pas à l'enquête diffèrent systématiquement des unités participantes relativement à la caractéristique étudiée. L'ampleur de ce biais dépend principalement du taux de participation et de la relation entre la probabilité de participer et la mesure d'intérêt.

Dans le cas d'un **échantillon aléatoire et simple**, le biais engendré par la non-réponse ou par l'exclusion est égale à :

$$\begin{aligned} \text{Biais} &= \frac{\phi(x)}{[1 - \Phi(x)]} \rho = \frac{\phi(x)}{RR} \rho \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \left(\frac{\rho}{RR}\right) \end{aligned}$$

Avec $\phi(x)$ la fonction de densité de la distribution normale, $\Phi(x)$ sa fonction de répartition, RR le taux de réponse et ρ la corrélation entre la mesure pour laquelle on souhaite déterminer le biais potentiel et la probabilité d'être exclu ou de ne pas participer.

FIGURE 1 : Distribution normale et exclusion



Supposons que le taux de non-réponse soit de 5 %. Dans un premier temps, il convient de déterminer la valeur standardisée de la distribution normale qui correspond à 5 % (soit la valeur de x dans la Figure 1). Cette valeur peut être déterminée en recourant à la fonction de répartition de la distribution normale. Pour un taux de 5 %, cette valeur est égale à $\Phi(0,05) = -1,645$.

Ensuite, il importe de calculer la fonction de densité de cette valeur, soit $\phi(-1,645) = 0,103$.

En faisant l'hypothèse d'une corrélation entre la mesure et la probabilité de réponse de 0,50, le biais, exprimé en fraction d'écart-type, est égal à :

$$\text{Biais} = \frac{\phi(x)}{RR} \rho = \frac{0,103}{(1 - 0,05)} \times 0,50 = 0,054$$

Le biais maximum, illustré à la *Figure 1*, est atteint lorsque la corrélation entre la probabilité de participer et la mesure est égale à 1.00 et correspond à la situation où les 5 % des élèves les plus faibles ont été exclus ou n'ont pas participé. Ce biais est égal à :

$$\text{Biais} = \frac{\phi(x)}{RR} \rho = \frac{0,103}{(1 - 0,05)} \times 1,00 = 0,1084$$

Cette formule montre que le biais est (i) inversement proportionnel au taux de participation (ii) proportionnel à la corrélation entre probabilité de participer et (iii) proportionnel à la mesure pour laquelle on veut déterminer le biais engendré par la non-réponse.

Dans le cadre d'un **échantillonnage à deux degrés** comme dans le cadre de l'enquête PASEC, on peut démontrer que le biais est égal à :

$$\left[\frac{1}{\sqrt{2\pi}} \right] \left\{ \left[\exp\left(-\frac{x_1^2}{2}\right) \left(\frac{\rho_1 \sqrt{p}}{R_1} \right) \right] + \left[\exp\left(-\frac{x_2^2}{2}\right) \left(\frac{\rho_2 \sqrt{1-p}}{R_2} \right) \right] \right\}$$

avec R_1 et R_2 respectivement les taux de réponses au niveau écoles et au niveau élèves, ρ_1 et ρ_2 les corrélations entre probabilité de participer et mesure respectivement au niveau 1 (écoles) et au niveau 2 (élèves) et, enfin p , la proportion de variance qui se situe au niveau 1, c'est-à-dire la proportion de variance de la variable qui se situe entre écoles. Ces formules indiquent l'importance de maximaliser les taux de réponses.

Dans la plupart des études internationales, une étape de nettoyage préliminaire est prise en charge par les centres nationaux eux-mêmes. Le nettoyage des données est donc effectué par le centre national du PASEC et l'équipe des conseillers techniques du PASEC.

Les étapes de nettoyage décrites ci-dessous doivent être effectuées pays par pays, avant la fusion des fichiers prévue à une étape ultérieure. À chaque étape de nettoyage, le PASEC enverra au centre national un rapport listant les éventuelles incohérences, manques ou erreurs dans les fichiers. Le centre national devra s'enquérir des problèmes soulevés et y apporter une réponse rapide.

Le nettoyage des données comporte un ensemble d'étapes successives. À chaque étape, des modifications pourront être apportées aux données. Il est essentiel que :

1. toutes les modifications apportées aux données soient introduites sous la forme de syntaxe Stata (ou tout autre logiciel). Afin de garantir la traçabilité, la reproductibilité et la transparence des opérations de nettoyage des données, il est recommandé de développer des programmes Stata distincts pour chacune des étapes du processus. Ces programmes peuvent être nommés selon une convention standardisée, par exemple : **cleaning_step1.do**, **cleaning_step2.do**, **cleaning_step3.do**, etc. ;
2. la même logique de dénomination doit être appliquée aux différents fichiers de données produits à chaque étape du traitement. Cette pratique facilite le suivi des modifications apportées aux données, permet de revenir facilement à une version antérieure en cas d'erreur et contribue à une meilleure documentation du processus de nettoyage. Ainsi, si le fichier initial de suivi des écoles est nommé FICHE_SUIVI_ECOLE.dta, le fichier obtenu après la deuxième étape de nettoyage pourra être nommé FICHE_SUIVI_ECOLE_2.dta, celui issu de la troisième étape FICHE_SUIVI_ECOLE_3.dta, et ainsi de suite. Cette convention de dénomination permet de conserver l'historique des traitements effectués, de faciliter les contrôles de qualité et de revenir facilement à une version antérieure du fichier en cas d'erreur ou de besoin de vérification. À cette fin, il suffit de renommer les fichiers avec une extension différente.

2. Concordance entre le fichier Suivi des écoles et le fichier Suivi des élèves

Avant de procéder au nettoyage de ces deux fichiers, une variable d'identification du pays doit être ajoutée dans les deux fichiers. Cet identifiant à deux chiffres aura été préalablement défini par le PASEC. Dans le fichier *Suivi des élèves*, un identifiant unique (TOT_ID à dix chiffres) doit également être ajouté : il s'agit d'un identifiant élève unique, résultant de la concaténation de l'identifiant pays (deux chiffres), de l'identifiant PASEC de l'école (six chiffres) et de l'identifiant élève (deux chiffres).

Le fichier *Suivi des Écoles* recense toutes les écoles de l'échantillon initial (écoles S) et comprend 1) les écoles ayant effectivement participé, 2) les écoles n'ayant pas participé quelle que soit la raison d'exclusion, 3) les écoles de remplacement (écoles R1 ou R2) qui ont effectivement participé. Ce fichier contient entre autres le niveau d'études envisagés (9^e, 6^e ou 2^e), le nombre actuel d'élèves, le nombre de classes, le nombre d'élèves dans la classe sélectionnée, le nombre d'élèves échantillonnés (en principe 30 au collège, 28 en sixième et 16 en deuxième du primaire) et parmi ceux-ci, le nombre d'élèves exclus (voir raisons possibles dans la partie du document consacrée à l'échantillonnage).

Le fichier *Suivi des Élèves* liste pour toutes les écoles participantes, des informations relatives à chaque élève échantillonné, participant, absent ou exclu. Dans ce fichier, une ligne (observation) représente donc un élève dans une école. Pour chaque élève, sont répertoriées des données administratives telles que la date de naissance et le genre et des informations relatives à la passation du test, à savoir l'éventuelle exclusion et le motif de celle-ci le cas échéant, la participation ou la non-participation à chacune des sessions des tests cognitifs et du questionnaire.

Voici, à titre d'exemples et de manière non-exhaustive, une série de vérifications à implémenter pour s'assurer de l'intégrité des données du fichier *Suivi des Écoles* et du fichier *Suivi des Élèves*.

2.1. Données manquantes

- Le fichier de *Suivi des Écoles* comporte, par école, trois séries d'informations : (i) informations pour la passation du test dans l'école, (ii) informations à compléter par l'administrateur pour la classe du niveau d'étude concerné et enfin (iii) le bilan de la collecte des données dans l'école. En principe, le fichier *Suivi des Écoles* n'admet aucune donnée manquante. Toutefois, certaines informations relatives au nom du directeur, son numéro de téléphone, la localisation de l'école ne sont pas fondamentales. Cependant, pour toute école participante, les deux derniers points d'informations doivent être entièrement complétés ;
- Le fichier *Suivi des Élèves* peut admettre des valeurs manquantes pour les variables descriptives comme les nom et prénom de l'élève, le genre, la date de naissance ou encore les variables d'heure de fin de test. Les variables d'identification et de participation sont par contre essentielles pour chaque élève.

2.2. Données qui ne respectent pas les règles prescrites pour les valeurs minimale et maximale

Pour chacun des fichiers de données, une application de saisie a été développée par le PASEC. Cette application de saisies définit, par variable, un ensemble de valeurs autorisées. En principe, toute valeur qui ne respecte pas ces règles ne peut être saisie. Malheureusement, ces règles pourraient être contournées par les utilisateurs. Il importe donc de vérifier, variable par variable, aussi bien pour le fichier *Suivi des Écoles* que pour le fichier *Suivi des Élèves*, si ces règles ont été strictement respectées. Le centre national sera consulté pour comprendre et corriger les incohérences observées dans les données saisies.

2.3. Doublons

À chaque élève et à chaque école ayant participé à l'enquête, doit correspondre un identifiant unique. Un double identifiant peut provenir d'une double saisie des données d'un élève (données identiques) ou d'une erreur de saisie dans l'un des deux identifiants. L'erreur doit être investiguée et corrigée.

2.4. Cohérence des informations entre plusieurs enregistrements dans un fichier de données

Dans le fichier de suivi des écoles, une école principale de l'échantillon ne peut être enregistrée comme participante si l'une ou l'autre de ses écoles de remplacement est également déclarée participante. En effet, le recours à une école de remplacement implique nécessairement que l'école initialement sélectionnée n'a pas participé à l'enquête. Pour chaque unité d'échantillonnage, une seule école — l'école de l'échantillon principal ou, le cas échéant, l'une de ses écoles de remplacement — peut être considérée comme participante. La vigilance est donc de mise dans l'implémentation et la vérification des données à cette étape déterminante dans la pondération des données Écoles. Le repérage d'une école de l'échantillon initial et de ses écoles de remplacement se fait sur base des identifiants école puisque ceux-ci sont liés entre eux. En effet, pour rappel, l'identifiant école comporte six chiffres : (i) deux chiffres pour identifier la strate - les strates explicites sont numérotées séquentiellement 01, 02, 03 -, (ii) un chiffre pour renseigner le ou les grade(s) évalué(s) : 1 si l'école est sélectionnée pour la 6^e uniquement, 2 si elle fait partie des deux échantillons au primaire ; '1' si l'école est sélectionnée pour la 9^e année (fin de collège) uniquement dans le cas de l'échantillon des écoles secondaires - et enfin (iii) trois chiffres d'identification de l'école. Pour ces trois derniers chiffres, le code d'une école initialement échantillonnée peut varier de 001 à 300, le code des premières écoles de remplacement de 301 à 600 (code de l'école initiale + 300) et le code des deuxièmes écoles de remplacement de 601 à 900 (code de l'école initiale + 600). Il est donc aisé de connaître l'identifiant d'une école de remplacement au départ de l'identifiant de l'école de l'échantillon initial. Ainsi, si le code de l'école échantillonnée est 031019, les codes des écoles de remplacements seront respectivement de 031319 et de 031619.

Afin de repérer ces éventuelles erreurs dans le fichier *Suivi des Écoles*, la procédure suivante sera implémentée :

- Un *identifiant temporaire* doit être créé afin de regrouper, sous une même référence, chaque école principale de l'échantillon et les éventuelles écoles de remplacement qui lui sont associées. Cet identifiant correspond au code de l'école initialement sélectionnée dans l'échantillon. Ainsi :
 - pour une école de l'échantillon principal, l'identifiant temporaire est identique à l'identifiant original ;
 - pour une première école de remplacement (R1), l'identifiant temporaire est obtenu en additionnant **300** à l'identifiant original ;
 - pour une seconde école de remplacement (R2), l'identifiant temporaire est obtenu en additionnant **600** à l'identifiant original.
- Une variable *statut* sera également créée. Cette variable aura trois modalités : S (pour Sample), R1 et R2 respectivement pour la première et la deuxième école de remplacement ;
- Les données seront ensuite transposées par identifiant temporaire, en fonction de la variable statut. La transposition des données portera sur la variable participation. Cette transformation permettra de recoder sur une seule ligne, le code de participation des écoles échantillonnées et de leurs éventuelles écoles de remplacement.

Cette procédure permet d'associer automatiquement chaque école de remplacement à l'école de l'échantillon principal qu'elle est susceptible de remplacer et facilite les contrôles de cohérence lors du nettoyage et de la validation des données.

2.5. Cohérence des informations d'un enregistrement au sein d'un fichier de données

- Le fichier *Suivi des Écoles* comporte un certain nombre de variables qui peuvent être comparées afin de détecter des données aberrantes. Par école,
 1. Diviser le nombre total d'élèves pour l'année d'étude concernée par le nombre de classes et vérifier si le nombre moyen d'élèves n'est pas trop élevé ou trop faible (pour autant qu'il y ait plus d'une classe) ;
 2. Comparer ce nombre moyen d'élèves par classe avec le nombre d'élèves de la classe sélectionnée. Si la taille de la classe sélectionnée diffère de plus de 25 % de la taille moyenne, et pour autant que cette différence représente en valeur absolue au moins 5 élèves, il y a lieu de considérer la taille de la classe échantillonnée comme suspecte ou la taille de la classe moyenne comme suspecte. Ces règles permettent ainsi de considérer comme suspectes les données d'une école avec une taille moyenne de classe de 40 élèves et de seulement 30 élèves pour la classe sélectionnée. Ce cas de figure pourrait notamment laisser penser que certains élèves de la classe sélectionnée n'ont pas été répertoriés (les directeurs et/ou les enseignants ont plutôt tendance à oublier de mentionner les élèves qui sont particulièrement faibles !) ;

3. Vérifier que le nombre d'élèves exclus ne représente pas plus de 10 % du nombre total d'élèves de la classe sélectionnée ;
 4. Vérifier que le nombre d'élèves échantillonnés au sein de la classe sélectionnée ne dépasse pas :
 - la taille de la classe,
 - la taille de la classe moins le nombre d'exclus,
 - le nombre d'élèves à échantillonner par école et par niveau d'études (30, 28 et 16 respectivement en 9^e, 6^e et en 2^e) ;
- Le fichier *Suivi des Élèves* peut également faire l'objet de certaines vérifications de cohérence. Ainsi,
 1. pour tout élève exclu, devrait figurer la raison de l'exclusion ;
 2. pour tout élève exclu, les données relatives à l'heure de fin des tests et questionnaires ne doivent que comporter le symbole A pour *Absent*.

2.6. Cohérence d'information entre les deux fichiers de suivi

Pour toute école mentionnée comme participante dans le fichier *Suivi des Écoles*, le nombre d'enregistrements dans le fichier de *Suivi des Élèves* doit impérativement correspondre au nombre d'élèves échantillonnés mentionné dans le fichier de *Suivi des Écoles*. Cette vérification peut être faite en réalisant un tableau de fréquences sur la variable *Identifiant Écoles* en utilisant le fichier de *Suivi des Élèves*. Ce tableau de fréquence est alors fusionné avec le fichier de *Suivi des Écoles*. Il reste à comparer la fréquence absolue au nombre d'élèves échantillonnés. Toute incohérence doit immédiatement être discutée avec le responsable national. Ces incohérences peuvent être de différentes natures :

1. Le nombre d'enregistrements dans le fichier *Suivi des Élèves* d'une école donnée est différent du nombre d'élèves échantillonnés et repris dans le fichier *Suivi des Écoles* ;
2. Une école, mentionnée comme non participante (qu'elle ait été exclue ou qu'elle ait refusé de participer) comporte des enregistrements dans le fichier de *Suivi des Élèves* ;
3. Une école participante ne comporte aucun enregistrement dans le fichier de *Suivi des Élèves*. Dans le cas contraire, il y a lieu de suspecter une erreur de saisie pour la variable participation dans le fichier *Suivi des Écoles*.

Les centres nationaux, en charge de la saisie des *Fiches de suivi de l'école* et des *Fiches de suivi des élèves* du pays, devront dans la mesure du possible effectuer certaines de ces premières étapes de nettoyage et corriger toutes les erreurs et incohérences avant de transmettre les fichiers au PASEC. Ainsi, les centres nationaux peuvent vérifier l'intégrité du fichier *Suivi des Écoles* simplement sur base d'une impression papier ou visuellement à l'écran. Il apparaît moins probable que les centres nationaux puissent vérifier la concordance entre les deux fichiers de suivi sans que les conseillers techniques du PASEC développent des procédures spécifiques et forment les centres nationaux à l'utilisation de ces procédures.

Il est aussi important de vérifier la concordance des noms des écoles/élèves à travers les bases de données. Cette étape est utile pour assurer que les mêmes écoles/élèves portent les mêmes identifiants à travers toutes les bases de données.

Par ailleurs, si les centres nationaux implémentent trop de procédures de vérification, les conseillers techniques ne seront plus en mesure d'estimer la qualité des données. Pour ces diverses raisons, les procédures de vérification des données au niveau national seront plutôt limitées dans leur nombre et judicieusement choisies par les conseillers techniques, en fonction des développements informatiques requis. Les capacités des superviseurs de la saisie des données seront ainsi renforcées afin de leur permettre de procéder aux premières vérifications suivantes :

- Saisie correcte des identifiants écoles, élèves, livrets par tabulations simples sur les variables concernées ;
- Saisie complète des fiches et des livrets, questionnaires enseignants et directeurs par décompte des observations dans les différentes bases de données et des comparaisons avec les nombres d'instruments revenus du terrain. Cette opération est à réaliser après suppression des doublons éventuels dans les données.

3. Concordance entre le fichier Suivi des élèves et les fichiers de Livret élèves (Tests et Questionnaire)

En rappel, pour les élèves de **fin du primaire (6^e année)** et de **fin du premier cycle du secondaire (9^e année ou niveau équivalent)**, le plan d'évaluation repose sur un **schéma de livrets tournants selon un plan incomplet équilibré** (*Balanced Incomplete Block Design – BIBD*).

Les instruments cognitifs sont répartis en **quatre livrets distincts (A, B, C et D)**. Chaque livret couvre les deux disciplines évaluées et contient, pour chacune d'elles, deux blocs d'items sélectionnés parmi un ensemble de quatre blocs disponibles. Cette organisation permet de couvrir un large éventail de compétences tout en limitant la charge de passation pour chaque élève.

En complément des épreuves cognitives, chaque livret comprend également un **questionnaire contextuel élève**.

Du point de vue du traitement des données, ce dispositif conduit à la production de **quatre fichiers distincts de réponses élèves**, correspondant respectivement aux livrets **A, B, C et D**. Ces fichiers devront être vérifiés individuellement avant d'être harmonisés et intégrés dans les étapes ultérieures du processus de nettoyage et de mise en base.

En **deuxième année du primaire**, le plan d'évaluation repose sur un livret unique. Chaque élève répond à l'ensemble des items de lecture, des items de mathématiques ainsi qu'au questionnaire contextuel. Les données correspondantes sont donc regroupées dans un seul fichier contenant l'ensemble des réponses de l'élève.

À l'inverse, pour les évaluations de **fin du primaire (6^e année)**, de **fin du collège (9^e année ou niveau équivalent)** ainsi que pour les **tests enseignants du primaire**, les données sont initialement réparties dans plusieurs fichiers correspondant aux différents livrets administrés. Le processus de nettoyage comporte donc une étape supplémentaire consistant à vérifier, harmoniser et fusionner les fichiers issus des différents livrets avant la poursuite des traitements.

Les étapes à mettre en œuvre pour le nettoyage des données élèves ou enseignants sont successivement décrites dans cette section.

3.1. Étape 1 : détection des doublons dans chacun des fichiers

Pour chacun des quatre fichiers de sixième année ou neuvième année ou enseignant du primaire et le fichier de deuxième, il importe de vérifier l'existence de doublons. Pour rappel, chaque identifiant élève total (ID_TOT) résultant de la concaténation de l'identifiant pays (2 chiffres), l'identifiant école (ID_PASEC) (6 chiffres) et l'identifiant élève au sein de l'école (2 chiffres) est unique.

Pour chaque fichier, la recherche de doublons doit être effectuée à partir de la variable **ID_TOT**, selon l'une des procédures décrites précédemment :

- produire une table de fréquence de la variable **ID_TOT** afin d'identifier les identifiants apparaissant plus d'une fois ;
- ou trier le fichier selon **ID_TOT** puis créer un numéro d'ordre temporaire à l'intérieur de chaque identifiant afin de repérer les occurrences multiples.

Toute valeur de **ID_TOT** apparaissant plus d'une fois dans un même fichier doit être examinée et corrigée avant de poursuivre les opérations de nettoyage. Le recours aux fiches de suivi élèves sera parfois nécessaire pour corriger les doublons dans les livrets.

Lorsqu'un doublon est détecté, deux cas de figures peuvent se présenter : soit un livret a été saisi à deux reprises, soit l'opérateur de saisie a commis une erreur au niveau de l'identifiant ID_TOT. Dans le premier cas, les enregistrements devraient être identiques ou quasiment identiques et un des deux enregistrements doit être supprimé. Si les deux enregistrements sont parfaitement identiques, il n'y a pas lieu de consulter le centre national PASEC. En revanche, si les deux enregistrements se ressemblent fortement mais qu'il existe des différences, il y a lieu de consulter le centre national pour vérifier lequel de ces deux enregistrements correspond au livret papier.

Si les deux enregistrements diffèrent nettement, il y a lieu de suspecter une erreur de saisie au niveau de la variable ID_TOT. On pourra dès lors implémenter successivement les vérifications suivantes pour corriger l'erreur. Par exemple, dans le cas de la sixième année du primaire, où 28 élèves de la classe ont été sélectionnés :

1. **Un des deux enregistrements correspond à un autre livret.** La fiche de *Suivi des Élèves* a été conçue de sorte que les 7 premiers élèves sélectionnés, dont le code Élève varie de 01 à 07, reçoivent le livret A, les 7 élèves suivants, dont le code varie de 08 à 14, reçoivent un livret B, et ainsi de suite. Nous disposons donc de deux informations pour déterminer si un des deux enregistrements a été saisi dans le mauvais fichier : les deux derniers chiffres de ID_TOT et la variable ID_LIVRET. S'il en est ainsi, l'enregistrement sera déplacé dans le bon fichier après vérification qu'il n'y figurait pas déjà ;
2. **Une erreur d'encodage au niveau de la variable ID_TOT est à l'origine de ce doublon.** Dans ce cas, l'erreur peut concerner le code Élève (deux derniers chiffres de ID_TOT), le code École (les 5^e à 8^e chiffres de ID_TOT) ou le code de la strate (les 3^e et 4^e chiffres de ID_TOT). Il est peu probable qu'un doublon apparaisse en lien avec le code du pays :

- **Code Élève :** pour une école donnée et pour un livret donné (c'est-à-dire au sein d'un même fichier), seuls certains codes élèves sont autorisés. Ainsi :
 - pour le livret A, les codes valides vont de **01 à 07** ;
 - pour le livret B, les codes valides vont de **08 à 14** ;
 - pour le livret C, les codes valides vont de **15 à 21** ;
 - pour le livret D, les codes valides vont de **22 à 28**.

La présence d'un identifiant dupliqué associée à l'absence d'un identifiant attendu peut révéler une erreur de saisie ou de codification du code élève.

Par exemple, si l'identifiant **XX01100101** (où XX représente le code pays) apparaît deux fois dans un même fichier alors que l'identifiant **XX01100102** est absent, il est raisonnable de supposer qu'un des deux enregistrements codés **XX01100101** correspond en réalité à l'élève **XX01100102**.

Dans une telle situation, il convient d'abord de consulter le fichier de suivi des élèves afin de vérifier si l'élève **XX01100102** est effectivement enregistré comme participant à l'évaluation. Si tel est le cas, les conseillers techniques du PASEC signaleront cette incohérence au centre national en précisant l'erreur probable observée. Le centre national sera alors invité à vérifier les documents sources et à déterminer lequel des deux enregistrements portant l'identifiant **XX01100101** doit être recodé sous l'identifiant **XX01100102**.

Cette procédure permet de corriger les erreurs d'identification tout en préservant l'intégrité et l'exhaustivité des données collectées.

- **Code École :** Lorsque l'analyse des doublons ne permet pas d'identifier une erreur au niveau du code élève, il convient d'examiner la possibilité d'une erreur dans le code de l'établissement. Par exemple, si l'identifiant **XX01100101** apparaît deux fois dans le fichier alors que les identifiants **XX01100102**, **XX01100103**, **XX01100104**, **XX01100105**, **XX01100106** et **XX01100107** sont également présents, l'ensemble des codes élèves attendus pour cette école semble complet. Dans ce cas, il est peu probable que le doublon résulte d'une erreur dans le code élève. Une erreur dans le code de l'établissement doit alors être envisagée.

La vérification doit être réalisée à l'intérieur de la même strate explicite que celle de l'établissement concerné. Dans l'exemple présenté, il s'agit de la **strate 01**. Il convient alors d'examiner l'ensemble des écoles de cette strate afin d'identifier un établissement pour lequel l'élève portant le même code élève (ici le code **01**) serait absent du fichier. Si un tel cas est identifié, il est possible qu'un élève ait été associé par erreur à un mauvais code établissement lors de la saisie ou de la codification. Cette hypothèse devra être vérifiée à l'aide du fichier de suivi des élèves afin de confirmer la participation effective de l'élève concerné.

Comme dans le cas précédent, toute correction potentielle devra être validée auprès du centre national. Les conseillers techniques du PASEC communiqueront l'anomalie détectée ainsi que l'hypothèse de correction, puis demanderont au centre national de consulter les documents sources afin de confirmer l'identifiant correct de l'élève concerné.

- **Code Strate** : Dans ce cas, et en supposant qu'une seule erreur s'est glissée dans ID_TOT, il suffira, à titre d'exemple pour le doublon 01100101, de vérifier au sein de chacune des strates l'absence de l'élève 01 pour la seule école 1001 participant à l'évaluation de fin du primaire ou de fin de collège (code 1).

S'il n'apparaît qu'aucun de ces cas de figure n'explique la présence du doublon, le centre national sera consulté pour déterminer lequel des deux enregistrements doit être conservé dans le fichier de données. Il convient toutefois de garder dans un fichier séparé l'enregistrement supprimé car la source de la double identification pourrait être identifiée lors d'une étape ultérieure.

3.2. Étape 2 : cohérence des codes Élève

Cette étape consiste à vérifier que les codes élèves correspondant aux deux derniers chiffres de la variable ID_TOT sont cohérents avec le livret auquel les élèves ont été affectés, conformément aux règles d'allocation décrites à la section 2.a..

Par conséquent, seuls les codes correspondant à l'intervalle attribué au livret considéré peuvent apparaître dans le fichier de données associé à ce livret.

Lorsqu'un code élève ne respecte pas cette règle, plusieurs vérifications doivent être effectuées.

Dans un premier temps, il convient d'examiner, pour l'établissement concerné, si l'un des élèves qui aurait dû figurer dans le fichier est absent. Cette vérification doit être réalisée à partir du fichier de suivi des élèves. Si un élève déclaré participant dans la fiche de suivi n'apparaît pas dans le fichier du livret concerné, il est possible qu'une erreur se soit produite lors de la saisie ou de la codification du code élève. Dans ce cas, le centre national sera consulté afin de vérifier cette hypothèse et d'identifier l'enregistrement devant être corrigé.

Si aucun élève attendu n'est manquant dans le fichier du livret concerné, une seconde hypothèse doit être examinée. Il est alors possible qu'un enregistrement ait été saisi dans un fichier de livret différent de celui auquel il appartient. Par exemple, si l'identifiant **XX01100111** apparaît dans le fichier du livret A alors que le code élève **11** correspond normalement au livret B, il convient de vérifier si l'élève 11 est absent du fichier du livret B pour l'établissement **XX011001**, après avoir confirmé sa participation dans le fichier de suivi des élèves.

Cette situation demeure toutefois peu probable. Les applications de saisie sont généralement conçues de manière à limiter ce type d'erreur, les variables étant spécifiques à chaque livret. Dans la plupart des cas, un opérateur de saisie détectera rapidement une incohérence de cette nature.

Si aucune des vérifications précédentes ne permet d'expliquer la présence de l'enregistrement incohérent, celui-ci devra être retiré du fichier de travail. Toutefois, afin de préserver une trace complète des opérations de nettoyage, l'enregistrement supprimé sera conservé dans un fichier distinct. Cette précaution permettra de réexaminer ultérieurement le cas si de nouvelles informations deviennent disponibles.

Au terme de cette étape, aucun identifiant dupliqué ne devrait subsister dans les fichiers concernés. Par ailleurs, l'ensemble des codes élèves présents dans les fichiers des livrets devra respecter strictement les règles d'allocation définies par le plan d'évaluation.

3.3. Étape 3 : fusion des 4 fichiers Élève en sixième ou Enseignants du primaire ou Élève en neuvième

La fusion consiste à mettre ensemble les données des quatre fichiers relatifs aux livrets A, B, C et D. Lors de l'étape 2, les fichiers de données ont été nettoyés pour garantir l'intégrité et l'unicité des identifiants des élèves. La fusion des fichiers peut alors se faire en utilisant les identifiants apurés.

Pour éviter que des enregistrements n'écrasent d'autres enregistrements, la procédure de fusion devra intégrer la variable ID_TOT et la variable ID_LIVRET. Après fusion, la présence de doublon sera implémentée selon une des deux méthodes décrites précédemment. Il est vivement conseillé de créer un identifiant ID_ECOLE, composé des 8 premiers chiffres de la variable ID_TOT et de calculer un tableau de fréquence sur cette variable ID_ECOLE pour déterminer si aucune école n'a plus de 28 enregistrements pour le cas de la sixième année ou 30 enregistrements pour le cas de la neuvième année.

Au terme de cette étape, nous disposons donc, pour la sixième année, la neuvième année ou des enseignants du primaire d'un seul fichier de données.

3.4. Étape 4 : vérification de la concordance entre le fichier Suivi des Élèves et le fichier Élèves

Cette étape vise à vérifier la cohérence entre les informations de participation et d'exclusion consignées dans le fichier de suivi des élèves et les données effectivement enregistrées dans les fichiers de réponses des élèves.

Cette vérification doit être réalisée séparément pour :

- les tests de lecture ;
- les tests de mathématiques ;
- le questionnaire contextuel.

En effet, ces différents instruments sont administrés à des moments distincts de la collecte. Il est donc possible qu'un élève participe à certaines sessions et soit absent à d'autres.

Pour chaque élève déclaré présent à une session de test ou au questionnaire contextuel dans le fichier de suivi, un enregistrement correspondant doit être retrouvé dans le fichier de données élèves. De plus, cet enregistrement doit contenir au moins une réponse valide pour l'instrument concerné. À l'inverse, un élève déclaré absent ou exclu ne devrait normalement présenter aucune donnée valide pour cet instrument.

Afin de faciliter cette vérification, il est recommandé de regrouper les variables en ensembles homogènes correspondant aux différents instruments administrés. Dans le cas des évaluations de fin du primaire ou de fin du collège, cela conduit à la création de :

- seize vecteurs de variables correspondant aux épreuves cognitives (2 disciplines × 4 livrets × 2 sessions) ;
- un dix-septième vecteur correspondant au questionnaire contextuel.

Pour chacun de ces vecteurs, il convient de calculer le nombre de valeurs manquantes ou non applicables. La comparaison entre ce nombre et le nombre total de variables composant le vecteur permet de déterminer si l'élève possède au moins une réponse valide pour l'instrument considéré.

À partir de cette information, cinq variables dichotomiques de participation peuvent être créées pour chaque élève :

- **PART_S1** : participation à la première session du test de lecture ;
- **PART_S2** : participation à la seconde session du test de lecture ;
- **PART_S3** : participation à la première session du test de mathématiques ;
- **PART_S4** : participation à la seconde session du test de mathématiques ;
- **PART_S5** : participation au questionnaire contextuel.

Chacune de ces variables prend la valeur :

- **0** si toutes les variables de l'instrument correspondant contiennent uniquement des codes « données manquantes » ou « non applicables » ;
- **1** si au moins une réponse valide est enregistrée pour cet instrument.

Les informations issues de ces variables peuvent ensuite être comparées aux statuts de participation renseignés dans le fichier de suivi des élèves. Toute divergence entre les deux sources constitue une incohérence qui doit être analysée.

Exemple

- Un élève déclaré présent à une session dans le fichier de suivi mais ne présentant aucune donnée valide dans le fichier de réponses ;
- un élève déclaré absent à une session alors que des réponses valides sont enregistrées ;
- un élève déclaré exclu alors que des données sont présentes dans le fichier.

Dans chacun de ces cas, le centre national devra être consulté afin de vérifier les documents sources et de déterminer la correction appropriée. Toutes les incohérences identifiées devront être résolues avant la poursuite des opérations de nettoyage et de mise en base.

Le fichier de suivi des élèves et le fichier de données élèves seront considérés comme cohérents si l'ensemble des conditions suivantes est respecté :

1. Présence des élèves participants

Tout élève déclaré présent à au moins une des cinq sessions de collecte (deux sessions de lecture, deux sessions de mathématiques et une session de questionnaire contextuel) doit être présent dans le fichier de données élèves. Lorsqu'un élève est déclaré présent à une seule des deux sessions d'un test cognitif, les variables correspondant à la session à laquelle il était absent doivent être codées « not applicable ».

2. Gestion des élèves exclus

Tout élève exclu de l'évaluation doit être clairement identifié comme tel dans le fichier de suivi des élèves. Aucun enregistrement relatif à cet élève ne doit être présent dans le fichier de données élèves.

3. Gestion des élèves totalement absents

Tout élève déclaré absent aux cinq sessions de collecte ne doit pas figurer dans le fichier de données élèves.

4. Codification des absences par session

Pour toute session à laquelle un élève est déclaré absent, l'ensemble des variables associées à cette session doit être codé « not applicable ».

Cette règle est essentielle puisque, dans les tests cognitifs, les codes « missing » sont généralement interprétés comme des réponses incorrectes lors du traitement des données, alors que les codes « not applicable » sont traités comme des données manquantes et n'affectent pas l'estimation du score de l'élève.

5. Concordance des informations d'identification

Les informations figurant dans le fichier de suivi des élèves et dans les fichiers de données doivent être concordantes. Cette vérification porte notamment sur :

- les identifiants des élèves ;
- les noms et prénoms des élèves, lorsque ces informations sont disponibles ;
- les identifiants des établissements ;
- les identifiants des livrets ;
- le nombre d'enregistrements attendus et observés.

6. Correspondance entre les deux fichiers

Tout élève présent dans le fichier de données doit être déclaré participant dans le fichier de suivi des élèves.

Réciproquement, tout élève déclaré participant dans le fichier de suivi des élèves doit être retrouvé dans le fichier de données correspondant.

7. Présence de données pour les élèves participants

Aucun élève déclaré participant ne doit présenter un enregistrement entièrement composé de données manquantes ou non applicables pour l'ensemble des instruments auxquels il était censé répondre.

La présence d'un tel enregistrement constitue une anomalie qui doit faire l'objet d'une vérification auprès du centre national.

Lorsque l'une de ces conditions n'est pas respectée, une incohérence est signalée dans le rapport de nettoyage et doit être examinée avant la poursuite des traitements.

3.5. Étape 5 : définition des élèves participants et recodage des fichiers Suivi des Élèves, Suivi des Écoles et Élèves

Une fois la cohérence établie entre le fichier de suivi des élèves et le fichier de données élèves, l'intégrité des données peut être considérée comme assurée. Cela signifie notamment qu'aucun enregistrement n'est manquant ou surnuméraire, que tous les identifiants sont valides et que les informations de participation sont cohérentes entre les différentes sources de données.

L'étape suivante consiste à attribuer à chaque élève un statut de participation. Quatre statuts sont théoriquement possibles :

- participant ;
- absent ;
- exclu ;
- inéligible.

Un élève **inéligible**, contrairement à un élève **exclu**, ne fait pas partie de la population cible de l'étude. Par exemple, dans une enquête telle que PISA, dont la population cible est définie par l'âge, un élève âgé de 14 ans ou de 16 ans est considéré comme inéligible puisqu'il ne répond pas au critère d'âge retenu. À l'inverse, un élève exclu appartient à la population cible, mais ne peut participer à l'évaluation en raison des critères d'exclusion définis par l'enquête.

Dans les évaluations du PASEC, la population cible étant définie à partir du niveau scolaire, un élève inéligible correspondrait par exemple à un élève inscrit dans un niveau différent de celui visé par l'évaluation. Une telle situation demeure exceptionnelle lorsque les procédures d'échantillonnage et de sélection des élèves ont été correctement appliquées.

3.5.1. Définition opérationnelle d'un élève participant

La détermination du statut de participation nécessite de définir précisément ce que l'on entend par un élève participant ou absent.

Les cas extrêmes sont simples à traiter :

- un élève ayant participé à toutes les sessions est considéré comme participant ;
- un élève n'ayant participé à aucune session est considéré comme absent.

La situation est plus complexe pour les élèves ayant participé à une partie seulement de l'évaluation. La définition retenue a des conséquences directes sur les taux de participation, les ajustements de pondération pour la non-réponse et, plus généralement, sur la qualité des estimations produites.

Les grandes enquêtes internationales, telles que PISA, TIMSS ou PIRLS, utilisent des modèles de réponse à l'item et la méthodologie des valeurs plausibles pour estimer les compétences des élèves. Ces approches permettent d'exploiter simultanément les réponses cognitives et les informations contextuelles afin d'estimer les compétences même lorsque certaines données sont manquantes. Ainsi, dans certaines éditions de PISA, un élève pouvait être considéré comme participant dès lors qu'il avait fourni suffisamment d'informations contextuelles pour permettre l'estimation de ses compétences à l'aide des modèles statistiques utilisés.

Le PASEC recourt également à la méthodologie des valeurs plausibles pour l'estimation des compétences des élèves. Toutefois, afin de maintenir un lien direct entre le statut de participation et la passation effective des épreuves cognitives, il a été décidé d'adopter une définition plus restrictive.

Dans le cadre du PASEC, un élève est considéré comme **participant** s'il a pris part à au moins une session d'un test cognitif, qu'il s'agisse de la compréhension de l'écrit ou des mathématiques.

Cette définition apparaît particulièrement adaptée au contexte du PASEC, où les épreuves sont administrées sur plusieurs sessions réparties sur deux jours consécutifs (ou davantage pour certaines populations évaluées). Dans certains contextes nationaux, l'absentéisme scolaire peut être relativement élevé et il n'est pas rare qu'un élève soit présent à une partie seulement des sessions prévues. La prise en compte de toute participation à une épreuve cognitive permet ainsi de limiter les pertes d'information tout en conservant un critère de participation directement lié aux objectifs principaux de l'évaluation.

En revanche, la seule participation au questionnaire contextuel ne suffit pas pour qu'un élève soit considéré comme participant à l'évaluation.

3.5.2. Création de la variable PART_ELEVE

L'attribution du statut de participation se traduit par la création d'une variable **PART_ELEVE** dans le fichier de données élèves.

Cette variable prend les valeurs suivantes :

- **PART_ELEVE = 1 (participant)** : l'élève a participé à au moins une session d'un test cognitif de compréhension de l'écrit ou de mathématiques ;
- **PART_ELEVE = 0 (absent)** : l'élève n'a participé à aucune session des tests cognitifs.

La détermination de cette variable repose directement sur les variables dichotomiques de participation créées à l'étape précédente (PART_S1 à PART_S5). Un élève sera considéré comme participant dès lors qu'au moins une des variables correspondant aux sessions cognitives (PART_S1 à PART_S4) prend la valeur 1.

Le *tableau 1* illustre l'attribution du statut de participation selon différentes configurations de présence ou d'absence aux sessions de l'évaluation.

TABLEAU 1 : Statut de participation des élèves

ID_PAYS	ID_PASEC	ID_ELEVE	TEST_S1	TEST_S2	TEST_S3	TEST_S4	PART_S1	PART_S2	PART_S3	PART_S4	PART_S5	PART_ELEVE
10	061125	01	Français	Français	Maths	Maths	1	1	1	1	1	Participant
10	061125	02	Français	Français	Maths	Maths	1	1	1	1	0	Participant
10	061125	06	Français	Français	Maths	Maths	1	0	0	0	0	Participant
10	061125	08	Français	Français	Maths	Maths	0	0	0	0	1	Absent
10	061125	09	Maths	Maths	Français	Français	0	1	0	1	1	Participant
10	061125	10	Maths	Maths	Français	Français	0	1	0	1	1	Participant
10	061125	11	Maths	Maths	Français	Français	0	0	0	0	0	Absent
-	-	-	-	-	-	-	-	-	-	-	-	-

Note : PART_S1 et PART_S2 font référence à la première discipline passée par l'élève (respectivement PART_S3 et PART_S4 pour la deuxième discipline), et PART_S5 fait référence au questionnaire contextuel.

3.5.3. Mise à jour des fichiers après l'attribution du statut de participation

Une fois le statut de participation attribué à chaque élève, les enregistrements correspondant aux élèves considérés comme **absents (PART_ELEVE = 0 ou absent)** doivent être retirés du fichier de données élèves. Ces enregistrements ne seront toutefois pas définitivement supprimés : ils seront conservés dans un fichier distinct afin de préserver la traçabilité des opérations de nettoyage et de permettre d'éventuelles vérifications ultérieures.

Cette étape implique également une mise à jour des informations de participation dans le fichier de suivi des élèves. Les statuts des élèves retirés du fichier de données devront être harmonisés avec la définition retenue du participant et de l'absent.

Les contrôles suivants doivent ensuite être réalisés :

1. Calcul des effectifs par école

Pour chaque établissement, calculer :

- le nombre d'enregistrements supprimés du fichier élèves ;
- le nombre d'enregistrements conservés dans le fichier élèves après nettoyage.

Ces deux informations doivent être intégrées au fichier de suivi des écoles afin de documenter précisément les effectifs retenus pour l'analyse.

2. Vérification de la cohérence des effectifs

Pour chaque école, vérifier que :

Nombre initial d'élèves participants – Nombre d'élèves supprimés = Nombre d'élèves conservés dans le fichier élèves

Toute divergence entre ces effectifs doit être examinée et corrigée avant la poursuite des traitements.

À l'issue de cette étape, les fichiers **Suivi des Élèves**, **Suivi des Écoles** et **Élèves** doivent présenter des informations parfaitement cohérentes en matière de participation. Les effectifs d'élèves participants doivent être identiques dans l'ensemble des fichiers et aucune incohérence relative aux statuts de participation ne doit subsister.

Une fois ces vérifications réalisées et les éventuelles anomalies corrigées, les bases de données peuvent être considérées comme prêtes pour les étapes ultérieures de pondération, d'ajustement pour la non-réponse et d'estimation des compétences des élèves.



1

2

3

4

5

6

7

8

4. Concordance entre le fichier Suivi des écoles et les fichiers Questionnaires Enseignants et Questionnaires écoles

Les fichiers des **Questionnaires enseignants** et des **Questionnaires établissements** doivent faire l'objet de contrôles de cohérence similaires à ceux réalisés pour les fichiers élèves.

4.1. Vérification de l'unicité des identifiants

Dans chacun de ces fichiers, chaque identifiant d'établissement (**ID_PASEC**) ne doit apparaître qu'une seule fois. La présence d'identifiants dupliqués constitue une anomalie qui doit être analysée.

Deux situations peuvent se présenter :

- **Les deux enregistrements sont strictement identiques.** Dans ce cas, il s'agit très probablement d'une double saisie. L'un des deux enregistrements doit être supprimé.
- **Les deux enregistrements présentent des informations différentes.** Cette situation traduit vraisemblablement une erreur d'identification ou une erreur de saisie. Les conseillers techniques du PASEC devront signaler cette incohérence au centre national afin qu'il vérifie les questionnaires originaux et précise quel enregistrement est correct.

Si, malgré ces vérifications, il n'est pas possible d'identifier avec certitude l'enregistrement valide, les deux enregistrements devront être retirés du fichier de travail. Comme pour les fichiers des livrets élèves, cette situation doit demeurer exceptionnelle. Tous les efforts devront être déployés afin d'identifier l'origine de l'erreur et de la corriger avant d'envisager la suppression des données.

4.2. Vérification de la correspondance avec le fichier de suivi des écoles

Tout établissement (école) présent dans les fichiers **Questionnaires enseignants** ou **Questionnaires établissements** doit également être présent dans le fichier **Suivi des Écoles**.

L'absence d'un identifiant d'établissement dans le fichier de suivi résulte généralement :

- d'une erreur de saisie de l'identifiant PASEC dans le questionnaire ;
- d'une erreur de transcription de l'identifiant lors de la collecte des données ;

Ces anomalies devront être signalées au centre national afin que les documents sources soient consultés et que les corrections nécessaires soient apportées ;

Inversement, tout établissement pour lequel le fichier de suivi des écoles indique qu'un questionnaire enseignant ou un questionnaire établissement a été complété doit obligatoirement être retrouvé dans le fichier correspondant ;

Si un questionnaire annoncé comme complété est absent du fichier de données, il est probable qu'il n'ait pas été saisi. Le centre national devra alors rechercher le questionnaire papier (ou sa version électronique) afin de procéder à sa saisie et de compléter la base de données.

4.3. Vérification des questionnaires entièrement vides

Un questionnaire dont toutes les variables sont codées comme données manquantes correspond, dans la pratique, à un questionnaire non complété.

Dans ce cas :

- l'enregistrement doit être retiré du fichier de données concerné ;
- le statut correspondant dans le fichier **Suivi des Écoles** doit être mis à jour afin d'indiquer que le questionnaire n'a pas été complété.

4.4. Contrôles complémentaires de cohérence

Les vérifications suivantes doivent également être réalisées :

- vérifier la concordance des noms des établissements entre les différents fichiers de données ;
- vérifier la cohérence des effectifs déclarés par les enseignants et les directeurs d'établissement avec les informations consignées dans les fiches de suivi ;
- comparer la taille totale de l'établissement avec la somme des effectifs des classes évaluées (2^e année, 6^e année et/ou 9^e année, selon les niveaux concernés). La taille totale de l'établissement doit être supérieure ou, au minimum, égale à la somme des effectifs des classes évaluées. Toute situation contraire constitue une incohérence qui devra être examinée par le centre national.

5. Concordance entre le fichier suivi des élèves et questionnaires parents

Dans les fichiers questionnaires parents, chaque identifiant parent doit correspondre à un identifiant élève dans la fiche de suivi élève. Le questionnaire parent est lié au questionnaire élève. Toutefois, il convient de vérifier la concordance des noms des élèves dans les deux fichiers.

6. Les recodages

6.1. La catégorisation des données manquantes

Il importe de distinguer les catégories de données manquantes dans les tests cognitifs. Le recodage de celles-ci permet de distinguer :

- les questions isolées non répondues. Ces non-réponses sont codées 9 ;
- les questions non atteintes. Il s'agit des questions auxquelles l'élève n'a pas répondu faute de temps ou par renoncement. Lorsqu'un minimum de 2 réponses en fin de série (une série étant une session dans un livret) sont manquantes, la première réponse manquante de la suite conserve le code 9, les dernières d'entre-elles sont codées 8 pour questions non atteintes. Ce codage doit être considéré séparément, pour chacune des deux sessions des deux tests cognitifs ;

Ce recodage est essentiel pour le traitement ultérieur des données. Ainsi, les questions non atteintes (code 8) ne seront pas prises en considération pour le calcul des paramètres des items alors que les questions non répondues (code 9) interviendront dans ces estimations. En revanche, l'estimation de la performance des élèves considérera les codes 9 et les codes 8 comme des réponses incorrectes.

Dans les questionnaires contextuels, les réponses manquantes peuvent également être de plusieurs types.

- Les questions emboîtées. Il s'agit d'un premier item du type OUI/NON suivi d'un second item (à choix multiples ou réponse chiffrée ouverte) applicable uniquement si la réponse au premier item est OUI (*Si oui,....*). Une non-réponse au second item aura un sens différent selon la réponse au premier :
 - Premier item, réponse OUI (code 1) : l'absence de réponse au second item est une réelle réponse manquante et conserve le code 9 ;
 - Premier item, réponse manquante (code 9) : l'absence de réponse au second item est aussi une réelle réponse manquante et conserve le code 9 ;
 - Premier item, réponse NON (code 2) : le second item est non applicable et tout code 9 initialement attribué au second item doit être recodé en code 7 – non applicable ;
- Les questions à choix multiple autorisant de cocher plusieurs modalités de réponse (voir exemple à la *Figure 2*). Chaque modalité de la question constitue en fait un item de type OUI/NON. Dès qu'une modalité au moins a été cochée, les autres peuvent être considérées comme des réponses négatives et non des réponses manquantes.

FIGURE 2 : Exemple de question emboîtée

7. Avez-vous bénéficié de formations complémentaires en cours d'emploi (stage pédagogique, séminaire de formation, cellule d'animation pédagogique...) durant les deux dernières années ?

☐ OUI ☐ NON QM6_7A

Si oui, précisez la durée totale :
(Indiquez-la en chiffres) _____ jours QM6_7B

Si oui, dans quel(s) domaine(s) avez-vous reçu ces formations complémentaires ?

Didactique du français	☐ OUI	☐ NON	QM6_7C
Didactique des mathématiques	☐ OUI	☐ NON	QM6_7D
Didactiques dans d'autres matières que le français et les mathématiques (histoires, sciences, TIC, etc.)	☐ OUI	☐ NON	QM6_7E
Approche par compétences	☐ OUI	☐ NON	QM6_7F
Approche par situation	☐ OUI	☐ NON	QM6_7G
Gestion des classes multigrades	☐ OUI	☐ NON	QM6_7H
Pédagogie des grands groupes	☐ OUI	☐ NON	QM6_7I
Pédagogie différenciée	☐ OUI	☐ NON	QM6_7J
Pédagogie par objectifs	☐ OUI	☐ NON	QM6_7K
Ecole amie des enfants	☐ OUI	☐ NON	QM6_7L
Education inclusive (handicap ou autres)	☐ OUI	☐ NON	QM6_7M

Notons que nous ne recommandons pas ce format de questions. Nous conseillons d'opter pour une question matricielle à réponses dichotomiques (OUI/NON) qui permet de distinguer plus facilement les réponses négatives des réponses manquantes.

FIGURE 3 : Question matricielle à réponses dichotomiques

QEP_07. Indiquez si vous avez bénéficié de formation continue en cours d'emploi durant les deux dernières années (stage pédagogique, séminaire de formation, cellule d'animation pédagogique) ainsi que l'objet et la durée de ces formations :	Non	Oui	Durée en jours	
A. Didactique des mathématiques	☐	☐		
B. Didactique du français ou langue d'instruction	☐	☐		
C. Didactique des disciplines autres que le français et les mathématiques (histoire, sciences, etc.)	☐	☐		
D. Thématique spécifique en lien avec les enjeux liés au genre	☐	☐		
E. Thématique spécifique en lien avec le développement durable ou la santé	☐	☐		
F. Thématique spécifique en lien avec l'éducation inclusive	☐	☐		
G. Thématique spécifique en lien avec le climat scolaire	☐	☐		
H. Pédagogie	☐	☐		
I. Formation en lien avec l'approche par compétences ou l'approche par situation	☐	☐		
J. Formation en lien avec l'intégration des technologies à l'école	☐	☐		
K. Évaluation des apprentissages	☐	☐		
L. Autre	☐	☐		

Si toutefois, des questions de ce type sont maintenues dans les questionnaires, elles devront faire l'objet du recodage suivant :

- une ou plusieurs items de la question ont été cochés (code 1 - OUI) : tous les autres items doivent recevoir le code 0 (NON) et non le code 9 *missing* ;
- aucun item de la question n'a été coché : tous les items de la question doivent recevoir le code 9.

6.2. Création de variable composite

Dans certains cas, plusieurs items portant sur un même contenu ou mesurant un même construit peuvent être combinés afin de créer une **variable composite**. Cette opération consiste à synthétiser l'information contenue dans plusieurs items en une seule variable dérivée.

Par exemple, à partir de quatre items de type **Vrai/Faux** associés à un même texte, il serait possible de créer une variable dichotomique prenant :

- la valeur **1** si l'élève répond correctement à au moins x des quatre items ;
- la valeur **0** dans le cas contraire ;
- la valeur **9** si la nouvelle variable ne peut être calculée en raison de données manquantes.

Lorsqu'une telle synthèse est réalisée, une nouvelle variable doit être créée dans la base de données. Son nom doit permettre de l'identifier clairement comme une **variable composite** ou **variable dérivée**, afin d'éviter toute confusion avec les variables issues directement des questionnaires.

Les items ayant servi à construire cette variable doivent néanmoins être conservés dans la base de données. Ils constituent les données originales de l'enquête et pourront être nécessaires pour des analyses complémentaires, des contrôles de qualité ou des traitements statistiques ultérieurs.

Dans le cadre de la présente enquête PASEC, la création de variables composites à partir des items des tests cognitifs n'est pas envisagée. Les données seront analysées à partir des réponses individuelles aux items, conformément aux méthodes d'estimation fondées sur la théorie des réponses à l'item.

6.3. Recodage de réponses courtes

Dans certaines enquêtes, les réponses aux questions ouvertes courtes sont saisies directement telles qu'elles sont fournies par le répondant, sans qu'un codage préalable ne soit effectué. C'est notamment le cas de certaines questions appelant une réponse numérique ou textuelle très courte.

Par exemple, pour un item de mathématiques demandant de calculer $12 - 5$, la réponse de l'élève pourrait être saisie directement sous la forme de la valeur **7**, plutôt que d'être codée immédiatement comme réponse correcte ou incorrecte.

Lorsque cette approche est retenue, une étape supplémentaire de traitement est nécessaire. Les réponses saisies doivent être recodées afin d'obtenir les catégories utilisées dans les analyses, par exemple :

- **1** : réponse correcte ;
- **0** : réponse incorrecte ;
- **9** : réponse manquante.

Cette procédure présente toutefois plusieurs inconvénients. Elle nécessite un traitement supplémentaire, augmente les risques d'erreurs de saisie et complique les opérations de contrôle de qualité et de nettoyage des données.

C'est pourquoi les grandes évaluations internationales, telles que **TIMSS**, **PIRLS** et **PISA**, privilégient un codage préalable des réponses avant leur saisie dans la base de données. Les opérateurs de saisie enregistrent directement les codes attribués aux réponses plutôt que les réponses elles-mêmes.

Le PASEC adopte cette même approche. Les réponses aux questions ouvertes courtes doivent être codées avant la saisie, ce qui simplifie le processus de traitement des données, réduit les risques d'erreur et garantit une meilleure harmonisation des procédures entre les pays participants.

6.4. Autres recodages

D'une manière générale, les éventuelles incohérences relevées dans les bases de données et qui n'auraient pu être corrigées seront recodées 9, 99, 999 (non-réponse) selon la longueur retenue pour la variable, ou en code 7, 97, 997 pour les cas non applicables. Pour rappel, le code *missing* est utilisé lorsque le sujet devait répondre mais n'a pas répondu alors que le code « not applicable » est utilisé lorsque l'élève ne devait ou ne pouvait pas répondre.

À titre d'exemple, considérons à nouveau la question 7 du questionnaire enseignant présentée dans la *Figure 2*. Cette question comporte un premier item dichotomique (QM6_7A) à réponses « Oui/Non ». Si la réponse à l'item QM6_7A est « Non » (code 2), aucun autre item (QM6_7C à QM6_7M) ne devrait avoir été choisi. Ces items seront donc recodés comme *not applicable* (Non applicable) pour différencier la situation où un ne coche pas le premier item et aucun des suivants.

Si pour cette question (QM6_7A), la réponse est « Non » (code 2 a été attribué à la fois au premier question) et un ou plusieurs autres questions (QM6_7C à QM6_7M) ayant des réponses « Non », il ne serait pas correct de traiter cette incohérence en donnant prérogative à l'une ou l'autre logique. Il sera préalablement demandé au centre national de vérifier cette incohérence. Si l'incohérence est bien dans la version papier du questionnaire enseignant, alors l'ensemble des items seront recodés comme *missing*.

Le code 7 « non applicable » sera également utilisé pour les éventuels items d'un test ou d'un questionnaire qui seraient affectés par une erreur systématique non remédiable, dans un pays donné. Ce serait le cas par exemple, d'un item mal traduit dans un pays testant dans une autre langue que le français. Cette solution doit cependant rester extrêmement marginale.

Enfin, si un pays a délibérément supprimé un ou plusieurs items d'un questionnaire contextuel, c'est encore le code 7 qui sera attribué pour ces items dans la base de données du pays.

Recodage des incohérences résiduelles

D'une manière générale, toute incohérence détectée lors du processus de nettoyage des données et n'ayant pu être résolue après vérification auprès du centre national devra être recodée à l'aide des codes prévus pour les données manquantes ou les données non applicables.

Selon le cas, les codes suivants seront utilisés :

- Code « *missing* » : 9, 99 ou 999, selon le format de la variable, lorsqu'un répondant devait répondre à une question mais qu'aucune réponse valide n'a pu être obtenue ou confirmée.
- Code « *not applicable* » : 7, 97 ou 997, selon le format de la variable, lorsqu'une question ne s'applique pas au répondant ou lorsque celui-ci ne devait pas y répondre en raison du cheminement du questionnaire ou des règles de passation.

Il est important de distinguer ces deux situations. Le code « *missing* » traduit une absence de réponse à une question qui aurait dû être renseignée, tandis que le code « *not applicable* » indique qu'aucune réponse n'était attendue pour le répondant concerné.

Exemple de recodage

- Considérons la question QM6_7 du questionnaire enseignant. Le premier item (QM6_7A) est une question dichotomique « Oui/Non » qui conditionne les réponses aux items suivants (QM6_7C à QM6_7M).
- Lorsque la réponse à QM6_7A est « Non », les items QM6_7C à QM6_7M ne doivent pas être renseignés. Si ces variables contiennent néanmoins des données, celles-ci ne sont pas interprétables dans le cadre de la logique du questionnaire. Après vérification auprès du centre national, ces items seront donc recodés « *not applicable* », afin de refléter le fait que ces questions ne devaient pas être complétées.
- En revanche, si une incohérence est observée, par exemple lorsque la réponse « Non » est enregistrée à QM6_7A alors qu'une ou plusieurs réponses sont également enregistrées aux items QM6_7C à QM6_7M, il ne convient pas d'accorder automatiquement la priorité à l'une des réponses. Cette situation doit d'abord être signalée au centre national afin qu'il consulte le questionnaire original et détermine l'origine de l'erreur.
- Si la vérification confirme que l'incohérence est bien présente sur le questionnaire source et qu'il est impossible de déterminer l'intention réelle du répondant, l'ensemble des variables concernées par cette question sera recodé « *missing* », la réponse ne pouvant être interprétée de manière fiable.

Utilisation du code « not applicable » dans des situations particulières

Le code « *not applicable* » est également utilisé lorsque certaines données ne peuvent être exploitées pour des raisons indépendantes du répondant.

Il peut notamment s'agir :

- d'un item affecté par une erreur systématique non remédiable dans un pays donné, par exemple une erreur de traduction ou une erreur de mise en page rendant l'item inexploitable ;
- d'un ou plusieurs items d'un questionnaire contextuel qui auraient été volontairement supprimés par un pays, avec l'accord préalable du PASEC.

Ces situations doivent toutefois demeurer exceptionnelles et être soigneusement documentées. Le recours au code « *not applicable* » ne doit intervenir qu'après validation par les conseillers techniques du PASEC, afin de garantir l'harmonisation des traitements entre les pays participants.

7. Vérification des variables

Les étapes précédentes du nettoyage des données étaient principalement consacrées à la vérification des **observations**, c'est-à-dire des enregistrements (lignes) des différentes bases de données. Une fois ces contrôles réalisés, il convient de procéder à une seconde série de vérifications portant sur les **variables** (colonnes).

Ces vérifications ont pour objectif de détecter d'éventuelles anomalies dans la distribution des variables, d'identifier des erreurs de codification ou de saisie qui n'auraient pas été mises en évidence lors des contrôles effectués au niveau des enregistrements, et de s'assurer de la cohérence globale des données.

Elles se déclinent en deux étapes complémentaires :

1. **l'analyse des statistiques univariées**, qui consiste à examiner chaque variable individuellement (fréquences, valeurs minimales et maximales, taux de données manquantes, etc.) ;
2. **l'analyse des statistiques multivariées**, qui vise à vérifier la cohérence des relations entre plusieurs variables (tableaux croisés, corrélations, contrôles logiques, etc.).

Ces analyses constituent une étape essentielle du processus de nettoyage des données. Elles doivent être réalisées **séparément pour chaque pays participant**, avant toute opération de pondération, de mise à l'échelle des compétences ou de production des indicateurs internationaux.

7.1. Les statistiques univariées

7.1.1. Variables catégorielles

L'ensemble des variables catégorielles fait l'objet d'une analyse des fréquences non pondérées. Cette étape vise à vérifier que seules les modalités de réponse autorisées sont présentes dans les bases de données et que les distributions observées sont plausibles au regard du plan d'enquête et de la structure des questionnaires. Toute modalité inattendue ou fréquence anormale devra faire l'objet d'une investigation avant la poursuite du traitement des données.

Variables issues des tests cognitifs

Pour les items à choix multiple (QCM), les codes **1 à 4** correspondent aux différentes modalités de réponse, tandis que les codes **7** (non applicable) et **9** (réponse manquante) peuvent également être observés selon les circonstances. Le code **8** (« item non atteint ») ne devrait apparaître que pour les items situés en fin de bloc de test, lorsque l'élève n'a pas eu le temps d'atteindre ces questions. Sa présence sur des items placés au début d'un bloc devra être considérée comme une anomalie.

Pour les items à réponse construite, seuls les codes **0, 1, 7 et 9** sont attendus. Comme pour les QCM, le code **8** ne peut apparaître que pour les items situés en fin de bloc.

Variables issues des questionnaires contextuels

Les questionnaires contextuels du PASEC comportent des variables de nature diverse : questions dichotomiques (Oui/Non, Vrai/Faux), questions à choix multiples comportant entre deux et quinze modalités de réponse, variables ordinales, réponses numériques ainsi que de nombreuses questions conditionnelles. Les contrôles réalisés devront tenir compte de cette diversité.

Les principales vérifications sont présentées ci-après.

• Vérification des modalités de réponse

Pour chaque variable, il convient de vérifier que les codes observés correspondent exclusivement aux modalités prévues dans le questionnaire ainsi qu'aux codes réservés (non applicable, réponse manquante, etc.). Toute modalité non autorisée devra être identifiée et corrigée conformément aux règles de codification établies par le PASEC.

• Vérification de la cohérence entre plusieurs sources

Certaines informations sont collectées dans plusieurs fichiers et doivent être parfaitement cohérentes.

À titre d'exemple, le sexe de l'élève renseigné dans le questionnaire élève devra être comparé à celui enregistré dans le fichier de suivi des élèves. En cas de divergence, l'information figurant dans le fichier de suivi des élèves fera foi et sera conservée comme valeur de référence.

• Vérification des questions emboîtées

Les questionnaires comportent de nombreuses questions conditionnelles, généralement introduites par des instructions telles que « **Si oui...** » ou « **Si non, passez à la question...** ». Ces questions doivent respecter une cohérence logique entre la question principale et les questions dépendantes.

Le principe général est le suivant :

- lorsqu'une réponse à la question principale rend les questions suivantes non applicables, celles-ci ne doivent comporter aucune réponse ;
- inversement, lorsqu'une réponse rend une question dépendante applicable, celle-ci doit obligatoirement être renseignée.

Les deux principales situations d'incohérence sont les suivantes :

Premier cas : réponse négative suivie d'une réponse à une question non applicable

L'enquête répond « **Non** » à la question principale tout en renseignant une question qui ne devait pas être administrée.

Deux règles de correction sont envisageables selon les décisions méthodologiques du PASEC :

- **préséance accordée à la question dépendante** : la réponse à la question principale est recodée en « **Oui** » ;
- **préséance accordée à la question principale** : la réponse à la question dépendante est supprimée et recodée en « Non applicable » (code 7).

Deuxième cas : réponse manquante suivie d'une réponse à la question dépendante

Lorsque la question principale est laissée vide alors qu'une réponse est fournie à la question dépendante, cette dernière constitue généralement une preuve implicite que la condition était satisfaite. La question principale pourra alors être recodée en « **Oui** », sous réserve de validation par les règles de nettoyage retenues par le PASEC.

Exemples d'application

Dans le **Questionnaire élève**, si un élève répond « **Non** » à la question demandant s'il dispose d'un manuel de français, la question suivante portant sur la possibilité d'emporter ce manuel à la maison ne doit pas être renseignée. Une réponse enregistrée à cette seconde question constitue une incohérence. À l'inverse, lorsqu'un élève déclare disposer d'un manuel, l'absence de réponse à la question suivante est également incohérente.

Le même principe s'applique au **Questionnaire enseignant**. Par exemple, lorsqu'un enseignant indique ne pas avoir bénéficié d'une formation continue, les questions demandant de préciser le type de formation ou sa durée ne doivent comporter aucune réponse. En revanche, si une formation est déclarée, les informations complémentaires correspondantes doivent obligatoirement être renseignées.

Des contrôles similaires devront être réalisés dans les **questionnaires de direction d'école primaire et de collège**. Ainsi, lorsqu'un directeur indique que son établissement ne possède pas de classes multigrades, aucune valeur ne doit être renseignée pour le nombre de classes multigrades. De même, si l'établissement est déclaré accessible toute l'année scolaire, la question portant sur la principale cause d'inaccessibilité ne doit comporter aucune réponse. À l'inverse, lorsqu'un établissement est déclaré inaccessible pendant une partie de l'année, cette cause devra obligatoirement être précisée.

De façon générale, l'ensemble des embranchements présents dans les questionnaires devra être contrôlé systématiquement. Toute réponse fournie à une question non applicable, ou toute absence de réponse à une question rendue obligatoire par un embranchement, devra être considérée comme une incohérence et traitée conformément aux règles de nettoyage définies par le PASEC.

FIGURE 4 : Exemple de question emboîtée dans le questionnaire enseignant

24. Indiquez le nombre de manuels par élève de votre classe :
(Veuillez ne cocher qu'une seule case par ligne)

	1 manuel par élève	1 manuel pour 2 élèves	1 manuel pour 3 élèves	1 manuel pour 4 élèves	1 manuel pour plus de 4 élèves	Aucun manuel	
Français							QM6_24A
Mathématiques							QM6_24B

• Si les élèves disposent de manuels de français, sont-ils pourvus par l'école, par les familles ou bien ces manuels sont-ils achetés la fois par l'école et les familles
(Veuillez ne cocher qu'une seule case)

☐ Les manuels sont pourvus par l'école
☐ Les familles achètent les manuels
☐ Les manuels sont achetés conjointement par l'école et les familles
QM6_24C

• Si les élèves disposent de manuels de mathématiques, sont-ils pourvus par l'école, par les familles ou bien ces manuels sont-ils achetés la fois par l'école et les familles
(Veuillez ne cocher qu'une seule case)

☐ Les manuels sont pourvus par l'école
☐ Les familles achètent les manuels
☐ Les manuels sont achetés conjointement par l'école et les familles
QM6_24D

• Si les élèves disposent chacun d'un manuel peuvent-ils l'amener à la maison ?
(Veuillez ne cocher qu'une seule case par ligne)

• Le manuel de français ☐ OUI ☐ NON QM6_24E
• Le manuel de mathématiques ☐ OUI ☐ NON QM6_24F

Les questions à choix multiple comportant un nombre fixe de réponses attendues (supérieur à une seule modalité).

Certaines questions demandent explicitement au répondant de sélectionner un nombre déterminé de modalités (par exemple, « Cochez les deux principales raisons » ou « Choisissez au maximum trois réponses »). Pour ces questions, il convient de vérifier que le nombre de modalités cochées correspond exactement au nombre attendu. Lorsqu'un nombre de réponses supérieur au maximum autorisé est observé, l'enregistrement est considéré comme incohérent et doit être corrigé conformément aux règles de nettoyage définies par le PASEC. Plusieurs stratégies peuvent être retenues, notamment : (i) invalider l'ensemble des réponses à cette question en les recodant en **non applicable** (code 7) ou en **réponse manquante** (code 9), selon les conventions adoptées ; (ii) appliquer une règle de priorisation préalablement définie lorsque celle-ci est justifiée par la conception du questionnaire ou les objectifs analytiques. En revanche, la sélection aléatoire d'un sous-ensemble de réponses est généralement déconseillée, car elle introduit une information qui n'a pas été fournie par le répondant et peut générer un biais dans les analyses.

FIGURE 5 : Exemple de QCM comportant deux choix de réponse

35. En français, quels sont les deux objectifs auxquels vous accordez le plus de temps ?
(Cochez les deux objectifs les plus importants)

☐ La lecture	QM6_35A
☐ La compréhension écrite	QM6_35B
☐ L'expression écrite	QM6_35C
☐ L'orthographe, la grammaire et la syntaxe	QM6_35D
☐ La compréhension orale	QM6_35E
☐ L'expression orale	QM6_35F

7.1.2. Variables continues

Avant d'entreprendre les analyses statistiques, il est recommandé de vérifier la présence de valeurs aberrantes (outliers) pour les variables susceptibles d'en contenir. L'objectif est de détecter les erreurs de saisie, de codification ou de déclaration qui pourraient affecter la qualité des analyses. Les principales vérifications à effectuer sont les suivantes :

- **Respect des limites théoriques des variables.** Certaines variables sont soumises à des bornes naturelles ou définies par le questionnaire. Par exemple, une variable exprimée en pourcentage ne peut prendre une valeur inférieure à 0 ni supérieure à 100. Toute valeur située en dehors de cet intervalle constitue une anomalie qui doit être corrigée.

- **Détection des valeurs statistiquement atypiques.** Les observations s'écartant de plus de **trois écarts-types** de la moyenne peuvent être considérées comme potentiellement aberrantes et doivent faire l'objet d'un examen approfondi. Toutefois, ce critère ne doit pas être appliqué de manière automatique. En pratique, il est préférable de définir, lorsque cela est pertinent, des seuils de contrôle adaptés au contexte national et à la nature de chaque variable. Ces seuils permettent d'identifier plus efficacement les valeurs susceptibles de résulter d'une erreur.
- **Vérification des plafonds propres à certaines variables.** Pour certaines questions, des limites maximales (ou minimales) peuvent être fixées en fonction de la logique du questionnaire ou des caractéristiques du système éducatif. Par exemple, le nombre d'élèves dans une classe, le nombre d'années d'expérience ou le nombre d'heures d'enseignement hebdomadaires ne peuvent raisonnablement dépasser certains seuils. Ces seuils de contrôle sont définis par le PASEC, variable par variable, en tenant compte des spécificités des pays participants.
- Toute valeur identifiée comme potentiellement aberrante doit être signalée au centre national afin qu'il procède aux vérifications nécessaires à partir des documents sources. Si la valeur est confirmée, elle est conservée dans la base de données. En revanche, si elle résulte d'une erreur et qu'aucune correction fiable ne peut être apportée, elle est recodée en **donnée manquante (missing)** conformément aux règles de codification du PASEC.

7.1.3. Les rapports de nettoyage

Pendant le processus de nettoyage des données, il est recommandé de transmettre régulièrement aux centres nationaux des **rapports de nettoyage** présentant les résultats des différentes procédures de vérification. Ces rapports doivent recenser les incohérences, anomalies et données manquantes détectées dans les bases de données, et inviter les centres nationaux à en vérifier l'origine et à apporter les clarifications ou corrections nécessaires à partir des documents sources.

Cette collaboration étroite entre le PASEC et les centres nationaux est essentielle pour garantir la qualité des données, assurer la traçabilité des corrections apportées et préserver la comparabilité internationale des résultats.

7.2. Les statistiques multivariées par pays

Les analyses multivariées constituent la seconde étape de vérification des variables après les analyses univariées. Elles consistent à confronter les informations issues de plusieurs variables afin de vérifier leur cohérence logique, leur plausibilité et leur conformité avec les règles de collecte définies par le PASEC.

Contrairement aux analyses univariées, qui examinent chaque variable individuellement, les analyses multivariées permettent de détecter des incohérences qui ne peuvent être mises en évidence qu'en croisant plusieurs variables d'un même questionnaire ou de questionnaires différents.

Ces contrôles doivent être réalisés séparément pour chaque pays participant. Toute anomalie détectée doit être documentée dans le rapport de nettoyage puis soumise au centre national afin qu'il procède aux vérifications nécessaires à partir des questionnaires originaux ou des documents de collecte. Selon les cas, les anomalies pourront être corrigées lorsqu'elles résultent d'une erreur de saisie ou de codification, ou être maintenues lorsque le centre national confirme qu'elles reflètent une situation réelle, bien qu'atypique.

Les exemples présentés ci-dessous sont fournis à titre indicatif.

7.2.1. Questionnaire établissement

Les principaux contrôles recommandés sont les suivants.

- **Âge du directeur et expérience professionnelle**

Comparer l'âge du directeur (QDE_01) avec son expérience professionnelle totale, obtenue en additionnant le nombre d'années d'enseignement avant sa nomination comme directeur (QDE_04) et le nombre d'années exercées en tant que directeur (QD_03).

La différence entre l'âge du directeur et son expérience professionnelle totale ne peut être inférieure à **20 ans**, âge minimal raisonnable d'entrée dans la profession.

7.2.2. Questionnaire enseignant

Les principaux contrôles de cohérence portent notamment sur les aspects suivants.

- **Âge et expérience professionnelle**

Comparer l'âge de l'enseignant (QEP_02) au nombre d'années d'enseignement (QEP_10). La différence entre ces deux valeurs ne peut être inférieure à **20 ans**.

- **Effectif de la classe**

Comparer les effectifs déclarés par l'enseignant (nombre de filles et de garçons : QEP_23) avec ceux enregistrés dans le fichier de suivi des écoles. Les écarts éventuels doivent être examinés.

- **Temps hebdomadaire d'enseignement du français**

Examiner le volume horaire hebdomadaire consacré au français (QEP_13), puis la distribution de cette variable au niveau national afin d'identifier les valeurs atypiques. Les valeurs situées à plus de trois écarts-types de la moyenne ou dépassant des seuils définis par le PASEC devront être vérifiées.

- **Temps hebdomadaire d'enseignement des mathématiques**

Procéder de manière identique en utilisant le nombre moyen de leçons hebdomadaires de mathématiques et leur durée moyenne.

- **Temps hebdomadaire total d'enseignement**

Calculer, pour chaque enseignant, le nombre total d'heures consacrées à l'enseignement du français et des mathématiques. Examiner ensuite la distribution de cette variable afin d'identifier les valeurs extrêmes ou manifestement incompatibles avec l'organisation scolaire du pays.

7.2.3. Questionnaire Élève

Le questionnaire élève ne comporte pas de questions ouvertes ni de variables nécessitant des calculs de cohérence complexes. Les contrôles portent essentiellement sur la logique de cheminement des questions, la cohérence entre variables et le respect des filtres définis dans le questionnaire.

8. Imputation des données manquantes

Dans le cadre du traitement international des données du PASEC, les données manquantes ne font pas l'objet d'une imputation systématique. Les bases de données finales conservent les codes de données manquantes conformément aux conventions de codification définies dans le présent manuel.

Toutefois, dans le cadre d'analyses statistiques spécifiques, notamment lors de modélisations économétriques réalisées à des fins de recherche, des méthodes d'imputation peuvent être envisagées pour certaines variables contextuelles lorsque le taux de données manquantes demeure limité et que les hypothèses d'imputation sont jugées acceptables.

En règle générale, il est recommandé de ne recourir à l'imputation que lorsque le taux de données manquantes n'excède pas **30 %**, et d'utiliser des méthodes adaptées à la nature des variables et au mécanisme de non-réponse.

Les évaluations internationales PASEC : de 2014 à 2024

Évolution du dispositif, spécificités des cycles et principales innovations

Depuis la réforme méthodologique engagée au début des années 2010, le **Programme d'analyse des systèmes éducatifs de la CONFEMEN (PASEC) conduit des évaluations internationales comparatives à grande échelle** afin de produire des données fiables et comparables sur les apprentissages, leur environnement et l'équité des systèmes éducatifs. Les cycles PASEC2014, PASEC2019 et PASEC2024 s'inscrivent dans une même architecture méthodologique, tout en intégrant progressivement de nouvelles populations, de nouveaux champs d'analyse et de nouvelles modalités de collecte. Cette évolution répond à un double objectif : préserver la comparabilité dans le temps et accroître la capacité du dispositif à éclairer le pilotage des politiques éducatives.

1. Un socle méthodologique commun

Les trois cycles reposent sur une évaluation externe et standardisée des apprentissages, conduite selon des procédures harmonisées dans les pays participants. Les populations scolaires sont étudiées à partir d'échantillons probabilistes représentatifs, permettant d'estimer les performances au niveau national sans procéder à une enquête exhaustive.

Le dispositif combine des tests de compétences en langue d'enseignement/lecture et en mathématiques avec des questionnaires contextuels. Les scores sont positionnés sur des échelles internationales et interprétés à travers des niveaux de compétences. Les informations contextuelles recueillies auprès des différents acteurs permettent de relier les performances aux caractéristiques des élèves, des enseignants, des établissements et de leur environnement.

La comparabilité repose également sur des procédures d'assurance qualité couvrant la conception et la mise à l'essai des instruments, leur adaptation linguistique, l'échantillonnage, l'administration, le contrôle de la collecte, le traitement des données et les analyses psychométriques. Les équipes nationales PASEC sont associées à ces différentes étapes, dans une logique de renforcement durable des capacités nationales en évaluation.

2. PASEC2014 : établir la référence internationale

PASEC2014 constitue le premier cycle international standardisé du Programme. Mis en œuvre dans dix pays d'Afrique subsaharienne — Bénin, Burkina Faso, Burundi, Cameroun, Congo, Côte d'Ivoire, Niger, Sénégal, Tchad et Togo — il évalue les élèves en début de scolarité primaire (2^e année) et en fin de scolarité primaire (6^e année) en langue d'enseignement/lecture et en mathématiques. Des données contextuelles sont parallèlement recueillies sur les élèves, les enseignants, les directeurs, les classes et les écoles.

En début de primaire, les épreuves sont administrées individuellement afin d'apprécier précocement les compétences fondamentales. En fin de primaire, l'évaluation repose principalement sur des épreuves papier-crayon et des cahiers tournants. Les résultats sont analysés en termes de performance, de niveaux de compétences, d'efficacité et d'équité des systèmes éducatifs.

L'innovation structurante de PASEC2014 est la création d'une métrique internationale commune. Le cycle transforme les expériences nationales antérieures du PASEC en un dispositif comparatif harmonisé et fournit une première référence commune pour suivre les systèmes éducatifs participants dans l'espace et dans le temps. L'évaluation simultanée du début et de la fin du primaire permet en outre de documenter à la fois les premiers apprentissages et les compétences indispensables à la poursuite de la scolarité.

3. PASEC2019 : consolider les tendances et intégrer les enseignants

PASEC2019 constitue le deuxième cycle international et élargit la participation à quatorze pays. Aux dix pays de 2014 s'ajoutent le Gabon, la Guinée, Madagascar et la République démocratique du Congo. L'architecture

d'évaluation des élèves en début et en fin de primaire est maintenue, ce qui permet, pour les dix pays communs aux deux cycles, d'analyser les évolutions des performances et des disparités entre 2014 et 2019.

Les questionnaires contextuels sont également révisés afin de mieux décrire l'environnement d'apprentissage : questionnaire Élève, questionnaire Enseignant/Classe et questionnaire Directeur/École. Les procédures d'adaptation linguistique, de mise à l'essai et d'assurance qualité sont renforcées afin de sécuriser la comparabilité nationale et internationale des instruments.

La principale innovation de PASEC2019 est l'introduction d'une enquête internationale sur les enseignants du primaire. Celle-ci mesure leurs connaissances disciplinaires en compréhension de l'écrit et en mathématiques ainsi que leurs connaissances didactiques, notamment leur capacité à planifier une situation d'apprentissage et à identifier ou analyser les erreurs des élèves. Elle recueille également des informations sur la formation, l'expérience professionnelle et les conditions de travail. Le diagnostic ne porte donc plus seulement sur les acquis des élèves et leur contexte : il intègre directement les connaissances et compétences professionnelles des enseignants, avec des implications pour les politiques de formation initiale et continue.

4. PASEC2024 : élargir le champ de l'évaluation et moderniser la collecte

PASEC2024 constitue le troisième cycle international et le plus large conduit par le Programme à ce jour, avec seize pays. Il conserve les composantes du primaire nécessaires au suivi des tendances, tout en élargissant le périmètre géographique (16 pays), linguistique et méthodologique de l'évaluation.

La première innovation majeure est l'extension de l'évaluation au collège / à la fin du premier cycle du secondaire. PASEC2024 permet ainsi d'appréhender plus largement le continuum de l'éducation de base, en complétant des mesures réalisées en début et en fin de primaire. Cette extension renforce notamment la contribution du dispositif au suivi de la cible 4.1.1 de l'ODD 4, qui porte sur les niveaux minimaux de compétence à différents moments de la scolarité.

PASEC2024 introduit aussi une innovation linguistique importante. Le cycle marque à la fois une évolution des langues internationales mobilisées, avec l'intégration du portugais et l'élargissement de l'anglais notamment au Nigéria, et une prise en compte accrue des langues nationales utilisées comme langues d'enseignement, particulièrement dans les premières années de scolarité. Cette double évolution permet de mieux refléter la réalité des contextes éducatifs multilingues, de rapprocher les instruments des conditions effectives d'apprentissage et de renforcer l'équité de la mesure, tout en exigeant un travail plus poussé d'harmonisation, d'adaptation et de vérification linguistique afin de préserver l'équivalence des items entre versions.

Le cycle 2024 enrichit également l'analyse de l'environnement des apprentissages. La composante consacrée aux enseignants, introduite en 2019, est maintenue et renforcée ; les chefs d'établissement demeurent une source centrale d'informations sur les établissements ; les parents ou familles sont intégrés parmi les répondants afin de mieux documenter l'environnement familial, le soutien aux apprentissages et certains déterminants extrascolaires. Une attention plus explicite est par ailleurs accordée à l'inclusion et à la prise en compte de la diversité des élèves dans les procédures d'administration.

Une autre innovation structurante concerne la transformation numérique de la collecte. Pour la première fois dans une évaluation internationale PASEC, les élèves de début de primaire sont évalués au moyen de tablettes et d'une application CAPI (Computer-Assisted Personal Interviewing). Cette évolution modernise la chaîne de collecte, facilite certains contrôles de cohérence et contribue à la sécurisation et à la standardisation des données, tout en maintenant des dispositifs adaptés aux différentes composantes de l'enquête.

5. Lecture synthétique de l'évolution des trois cycles

Dimension	PASEC2014	PASEC2019	PASEC2024
Positionnement	1 ^{er} cycle international standardisé	2 ^e cycle ; consolidation et tendances	3 ^e cycle ; changement d'échelle et modernisation
Pays participants	10	14	16
Niveaux élèves	Début et fin du primaire	Début et fin du primaire	Début primaire, fin primaire et collège / fin de l'éducation de base
Domaines élèves	Langue/lecture et mathématiques	Langue/lecture et mathématiques	Lecture/langue et mathématiques, avec extension au collège
Enquête enseignants	Données contextuelles sur les enseignants	Tests disciplinaires et didactiques + questionnaire enseignant	Composante enseignants renforcée ; volume nettement accru
Familles / parents	Informations familiales principalement via les élèves	Informations familiales principalement via les élèves	Parents/familles intégrés comme répondants
Chefs d'établissement	Questionnaires directeurs	Questionnaire Directeur/École	Chefs d'établissement intégrés à la collecte
Collecte	Principalement papier-crayon ; administration individuelle en début de primaire	Papier-crayon ; cahiers tournants ; procédures QA renforcées	Introduction du CAPI/tablettes en début de primaire ; dispositifs papier et numériques selon les composantes
Nombre et types de langues d'évaluation	3 langues : français, anglais et kirundi	7 langues : français, anglais, arabe et langues nationales africaines	14 langues d'enseignement, combinant langues internationales/officielles, dont l'anglais et le portugais, et langues nationales
Langues nationales prises en compte	Kirundi	Kirundi, malagasy, haoussa et zarma	Élargissement à plusieurs langues nationales, dont kirundi, malagasy, haoussa, zarma, bamanankan, wolof, soninké, sérère, pulaar et mandingue
Innovation et adaptation linguistique	Première prise en compte de contextes linguistiques différenciés dans un cadre international commun	Diversification des versions et formalisation des procédures de traduction, conciliation, vérification et validation	Changement d'échelle multilingue : extension des langues internationales, meilleure prise en compte des langues nationales d'enseignement et renforcement de l'assurance qualité linguistique
Comparabilité temporelle	Création de la référence	Mesure des tendances 2014–2019 pour les pays communs	Continuité des séries du primaire + nouvelles composantes
Innovation dominante	Métrique internationale commune	Évaluation des compétences des enseignants	Collège, familles, numérique, inclusion, élargissement géographique et linguistique
Lien avec l'ODD 4	Contribution à la mesure des apprentissages	Suivi renforcé des apprentissages et de l'équité	Couverture élargie de la cible 4.1.1 jusqu'à la fin du premier cycle du secondaire

En synthèse, PASEC2014 a standardisé et créé la référence internationale ; PASEC2019 a consolidé cette référence, rendu possible l'analyse des tendances et intégré les compétences des enseignants ; PASEC2024 conserve les séries comparables tout en élargissant l'évaluation à de nouvelles populations et à la fin de l'éducation de base, et en modernisant

les modalités de collecte. Cette progression traduit une évolution d'un dispositif centré sur la mesure des acquis des élèves vers un système plus complet sur les conditions de production des apprentissages et les leviers d'amélioration de la qualité de l'éducation.

PASEC

Évaluer les apprentissages pour transformer les politiques éducatives

Créé en 1991, le Programme d'analyse des systèmes éducatifs de la CONFEMEN (PASEC) s'est imposé comme l'instrument de référence au service du pilotage des systèmes éducatifs des États et gouvernements membres de la CONFEMEN. Il a pour vocation d'éclairer la décision publique en mesurant l'évolution des performances des systèmes éducatifs, pour orienter l'élaboration, le suivi et l'évaluation des politiques éducatives et, à travers elles, l'amélioration de la qualité de l'éducation.

Son action ne se limite pas à la mesure. Le PASEC renforce les capacités nationales d'évaluation, favorise l'appropriation et l'usage effectif des résultats par les acteurs éducatifs, et soutient l'essor de la recherche en éducation. Par ses évaluations et ses travaux d'analyse, il met à la disposition des décideurs des données fiables et comparables ainsi que des analyses rigoureuses, leviers d'une éducation plus performante, plus équitable et plus efficace.

Le PASEC concourt enfin à la production de connaissances en accompagnant les chercheurs et en mettant à leur disposition ses données et ses cadres d'évaluation. Il nourrit des travaux utiles tant aux systèmes éducatifs qu'à la communauté universitaire.

CONFEMEN – PASEC

Complexe SICAP Point E, Immeuble C
3^e étage, avenue Cheikh Anta Diop
BP 3220 Dakar, Sénégal
Tél. : +221 33 859 29 79 / +221 33 859 29 91
Fax : +221 33 825 17 70
Courriel : pasec@confemen.org
Site web : www.pasec.confemen.org

@CONFEMEN



Avec le soutien de

