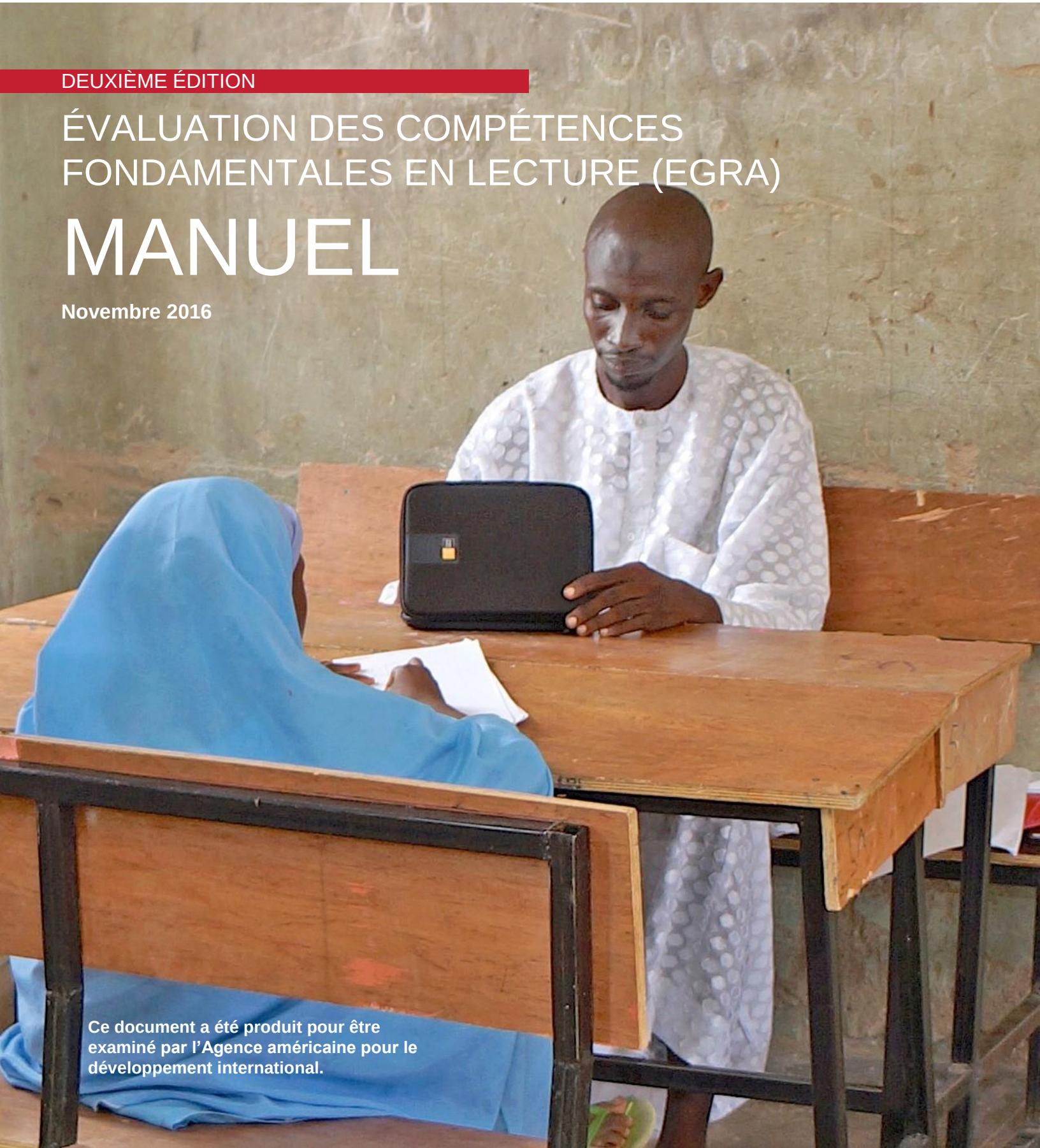
USAID
FROM THE AMERICAN PEOPLE

DEUXIÈME ÉDITION

ÉVALUATION DES COMPÉTENCES FONDAMENTALES EN LECTURE (EGRA)

MANUEL

Novembre 2016



Ce document a été produit pour être examiné par l'Agence américaine pour le développement international.

ÉVALUATION DES COMPÉTENCES FONDAMENTALES EN LECTURE, DEUXIÈME ÉDITION

PHOTO DE COUVERTURE : ÉQUIPE DE PROJET RTI, USAID/INITIATIVE POUR L'ÉDUCATION DANS LES ÉTATS DU NORD DU NIGÉRIA

CETTE PUBLICATION A ÉTÉ PRODUITE POUR L'AGENCE AMÉRICAINE POUR LE DÉVELOPPEMENT INTERNATIONAL PAR RTI INTERNATIONAL DANS LE CADRE DU PROJETS EDDATA II (EDUCATION DATA FOR DECISION MAKING), MESURE ET RECHERCHE À L'APPUI DE L'OBJECTIF STRATÉGIQUE 1 EN ÉDUCATION, NUMÉRO DE TÂCHE AID-OAA-12-BC-00003 (NUMÉRO DE TÂCHE RTI 20).

RTI INTERNATIONAL. 2016. MANUEL POUR L'ÉVALUATION DES COMPÉTENCES FONDAMENTALES EN LECTURE (EGRA), DEUXIÈME ÉDITION. WASHINGTON, DC : AGENCE AMÉRICAINE POUR LE DÉVELOPPEMENT INTERNATIONAL.

COPYRIGHT © 2016 RTI INTERNATIONAL

RTI INTERNATIONAL EST UNE MARQUE DE COMMERCE ET UN NOM COMMERCIAL DÉPOSÉS DU RESEARCH TRIANGLE INSTITUTE.

CE DOCUMENT EST PROTÉGÉ SOUS LICENCE INTERNATIONAL CREATIVE COMMONS ATTRIBUTION 4.0. UNE COPIE DE CETTE LICENCE EST DISPONIBLE SUR HTTP://CREATIVECOMMONS.ORG/LICENSES/BY/4.0/. AUX TERMES DE LA LICENCE CREATIVE COMMONS ATTRIBUTION, VOUS ÊTES LIBRE DE REPRODUIRE, DISTRIBUER, TRANSMETTRE ET ADAPTER CE DOCUMENT DANS LES CONDITIONS SUIVANTES :

ATTRIBUTION—SI VOUS REPRODUISEZ ET DISTRIBUEZ CE DOCUMENT DANS SON INTÉGRALITÉ, SANS EN MODIFIER LE CONTENU OU LES ILLUSTRATIONS, VOUS ÊTES PRIÉ DE LE CITER COMME SUIT : REPRODUCTION D'UNE ŒUVRE ORIGINALE PUBLIÉE PAR RTI INTERNATIONAL ET SOUS LICENCE INTERNATIONALE CREATIVE COMMONS ATTRIBUTION 4.0..

TRADUCTIONS—SI VOUS CRÉEZ UNE TRADUCTION DE CE DOCUMENT, VOUS ÊTES PRIÉ D'Y APPOSER L'ÉTIQUETTE SUIVANTE : TRADUCTION D'UNE ŒUVRE ORIGINALE PUBLIÉE PAR RTI INTERNATIONAL ET SOUS LICENCE INTERNATIONALE CREATIVE COMMONS ATTRIBUTION 4.0 INTERNATIONAL LICENSE.

ADAPTATIONS—SI VOUS CRÉEZ UNE ADAPTATION DE CE DOCUMENT, VOUS ÊTES PRIÉ D'Y APPOSER L'ÉTIQUETTE SUIVANTE : ADAPTATION D'UNE ŒUVRE ORIGINALE PUBLIÉE PAR RTI INTERNATIONAL ET SOUS LICENCE INTERNATIONALE CREATIVE COMMONS ATTRIBUTION 4.0 INTERNATIONAL LICENSE.

REMERCIEMENTS

Ce document est une adaptation du Manuel d'évaluation des compétences fondamentales en lecture (EGRA), deuxième édition, rédigé par RTI International en mars 2016 pour l'Agence américaine pour le développement international (USAID). La version française de ce manuel n'aurait pu être réalisée sans la contribution du Dr Liliane Sprenger-Charolles qui a adapté le contenu technique de plusieurs sections de la version anglaise pour guider et faciliter la mise en œuvre d'évaluations EGRA dans des contextes francophones. Un merci tout particulier à Foreign Language Services, Inc. (FLS) pour son excellent travail dans la traduction d'une grande partie du contenu présenté dans ce document.

Ce manuel pour l'Évaluation des compétences fondamentales en lecture (EGRA) est le produit d'une collaboration soutenue parmi les membres d'une communauté de chercheurs, spécialistes, représentants de gouvernement et professionnels du perfectionnement en éducation pour faire avancer la cause de l'acquisition et de l'évaluation de compétences fondamentales en lecture parmi les enfants scolarisés en primaire dans les pays à faible revenu.

Bien qu'il ne soit pas possible de reconnaître tous ceux qui ont contribué à la révision et à la mise à jour de ce manuel, cette tâche n'aurait pas pu être accomplie sans la participation de Melissa Chiappetta d'Abt Associates, Ray Adams et Juliette Mendelovits du Conseil australien de la recheche en éducation (ACER), Pooja Reddy Nakamura et Zarko Vukmirovic des Instituts américains de recherche (AIR), Fathi El Ashry de Creative Associates International, Elena Vinogradova de l'Education Development Center, Inc. (EDC), Matt Sloan de Mathematica Policy Research, Thomaz Alvares et Abdullah Ferdous de Management Systems International (MSI), Roger Stanton du groupe Optimal Solutions, Kellie Betts, Chris Cummiskey, Margaret Dubeck, Karon Harden, Simon King, Jessica Mejia, Erin Newton, Alison Pflepson, Lilly Piper, Sarah Pouezevara et Jonathan Stern de RTI International, Elliot Friedlander et Carol da Silva de Save the Children et Aglaia Zafeirakou de la Banque mondiale. Nous souhaitons également remercier et rendre un hommage spécial à Truphena Choti d'University Research Company (URC, LLC), Jane Benbow, anciennement avec l'URC, LLC et autre personnel de l'URC, LLC et du Global Reading Network qui ont contribué à l'organisation et à l'accueil en 2015 de divers ateliers et séminaires d'évaluation des compétences fondamentales en lecture.

La mise au point de l'outil EGRA initiale n'aurait pu se faire sans l'appui d'organisations non gouvernementales et des équipes d'évaluation EGRA des

ministères de l'éducation d'Afghanistan, du Bangladesh, d'Égypte, de Gambie, du Guyana, d'Haïti, du Honduras, de Jamaïque, du Kenya, Libéria, du Mali, du Nicaragua, du Niger, du Pérou, du Sénégal et d'Afrique du sud. Nous devons une profonde gratitude aux enseignants, aux élèves et à leurs familles pour leur participation et leur foi soutenue dans les bénéfices de l'éducation. Pour les en remercier, nous nous employons à améliorer les résultats en lecture pour tous les enfants du monde entier.

Nous devons à Amber Gove la principale paternité du manuel original, avec des contributions de Luis Crouch, Amy Mulcahy-Dunn et Marguerite Clarke. La deuxième édition a bénéficié de la contribution de nombreux responsables, relecteurs et participants à des séminaires.[1]

Les opinions exprimées dans ce document sont celles des auteurs et ne reflètent pas nécessairement le point de vue de l'Agence américaine pour le développement international. Merci d'adresser toutes questions ou commentaires à Penelope Bender à pbender@usaid.gov.

[1] Les individus et organisations ayant contribué directement à la deuxième édition sont cités en **Annexe A.**

CONTENTS

LISTE DES FIGURES

ABRÉVIATIONS

ACER	Conseil australien pour la recherche en éducation
AIR	Instituts américains de recherche
ASER	Rapport annuel sur l'état de l'éducation (Pratham)
CFR	Code américain des réglementations fédérales
IC	intervalle de confiance
CLP	Programme de développement des moyens de subsistance communautaires (Yémen)
LCPM	lettres correctes par minute
SLCPM	sons de lettre corrects par minute
NMCPM	non-mots corrects par minute
CONFEMEN	Conférence des ministres de l'éducation des pays ayant le français en partage
SSCPM	sons de syllabe corrects par minute
CTOPP-2	Test détaillé du traitement phonologique, deuxième édition
TCT	théorie classique des tests
MCPM	mots corrects par minute
DDL	bibliothèque de données de développement
DIBELS	indicateurs dynamiques des compétences fondamentales précoces en alphabétisation
DID	différences de différences
DIFF	différence supposée
EDC	Education Development Center, Inc.
EdData II	Education Data for Decision Making (Données sur l'éducation pour la prise de décisions, programme de l'USAID)
EPT	Éducation pour tous
EGMA	Évaluation des compétences fondamentales en mathématiques
EGRA	Évaluation des compétences fondamentales en lecture

FLAT	Outil d'évaluation de l'alphabétisation fonctionnelle (World Vision)
GIZ	Agence d'aide allemande, Deutsche Gesellschaft für Internationale Zusammenarbeit
CGP	correspondance graphème–phonème
GPS	système de positionnement global
CCI	coefficient de corrélation interne
API	alphabet phonétique international
CEI	Comité d'examen institutionnel
IRR	fiabilité inter-évaluateurs
TRI	théorie de la réponse d'item
KNEC	Conseil kényan des examens nationaux
L1, L2	première langue, deuxième langue
LCD	écran à cristaux liquides
LLECE	Laboratorio latinoamericano de evaluación de la calidad de la educación (laboratoire latino-américain d'évaluation de la qualité de l'éducation)
LQAS	échantillonnage par lots pour l'assurance de la qualité
EDM	effet décelable minimal
OMD	Objectif du millénaire pour le développement
MoEST	Ministère de l'éducation, des sciences et de la technologie (Kenya)
MSI	Management Systems International
NCFL	Centre national pour l'alphabétisation familiale
NICHD	Institut national américain de la santé de l'enfant et du développement humain
MCO	moindres carrés ordinaires
ORF	fluence de lecture à haute voix
PASEC	Programme d'analyse des systèmes éducatifs de la CONFEMEN
PIRLS	Progrès dans l'étude internationale d'instruction de lecture
PISA	Programme international pour le suivi des acquis des élèves
PPT	probabilité proportionnelle de la taille

EVIP	échelle de vocabulaire en images Peabody
PRIMR	Mathématiques et lecture dans le primaire (Kenya)
FUP	fichier à usage public
MQE	modèle quasi-expérimental
ECR	essai contrôlé randomisé
RTI	RTI International (marque et nom déposés du Research Triangle Institute)
SACMEQ	Consortium d'Afrique australe pour l'analyse de la qualité de l'éducation
SART Ed	analyse secondaire pour projet de suivi des résultats, portail de l'éducation
TIMSS	Enquête international sur les mathématiques et les sciences
TOPA-2+	Test de conscience phonologique, deuxième édition plus
PNUD	Programme des Nations Unies pour le développement
UNESCO	Organisation des Nations Unies pour l'éducation, la science et la culture
URC	University Research Co., LLC
USAID	Agence américaine pour le développement international
YEGRA	Stratégie de lecture au primaire au Yémen

GLOSSAIRE

Les expressions en italique gras dans une notice renvoient à une entrée spécifique du glossaire

Terminologie reliée à la lecture

Attaque. Partie de la ***syllabe*** précédant la voyelle (*str-* dans le mot *strict*). Certains mots n'ont pas d'attaque (*on*, *arc*…).

Automaticité des procédures de lecture. Niveau de maitrise de la lecture à partir duquel la reconnaissance des mots écrits est devenue un acte quasi reflexe. Le fait que les mots écrits sont automatiquement identifiés se manifeste, par exemple, par l'effet dit *stroop* : le sujet ne peut pas s'empêcher de lire, même quand la tâche ne requiert pas la lecture. Ainsi, quand on demande de nommer la couleur de l'encre dans laquelle est écrit un nom de couleur, sans lire le mot, et quand *vert* est écrit en rouge, le temps de réponse est plus long que lorsque *vert* est écrit en vert, ce qui est le signe que les mots écrits sont automatiquement identifiés. C'est cette capacité qui permet au lecteur débutant d'atteindre un niveau de compréhension écrite égal à celui de sa compréhension orale.

Capacités métaphonémiques. Voir ***Conscience phonémique.***

Capacités métaphonologiques. Voir ***Conscience phonologique.***

Connaissance de l'alphabet. Voir ***Procédure alphabétique.***

Conscience phonémique. Capacité d'identifier et de manipuler les plus petites unités sans signification de la langue orale, les ***phonèmes***. Elle est évaluée, par exemple, par des exercices de comptage ou de suppression de phonèmes (compter le nombre de sons différents contenus dans le mot oral *tour* ou prononcer ce mot en enlevant son premier son).

Conscience phonologique. Capacité d'identifier et de manipuler les unités sans signification de la langue orale : de la ***syllabe,*** et ses composants (***attaque*** et ***rime),*** au ***phonème.***

Décodage. Voir ***Procédure alphabétique.***

Dérivation. Mot formé à partir d'un autre mot (*parfait* versus *imparfait* et *parfaitement*), ou de plusieurs mots (*autoroute*). Voir aussi ***Formes fléchies*** et ***Morphème.***

Diacritique. Signe ajouté à un ***graphème*** le plus souvent pour indiquer une prononciation différente, par exemple, *e* versus *é/*è (élève) ou *c* versus ç (*leçon de calcul*) ou encore l'accent circonflexe, qui est supposé indiquer un allongement de la voyelle (être, hôtel).

Digraphe. Groupe de lettres qui se suivent: *bigraphe* (2 lettres) et *trigraphe* (3 lettres). Par exemple, le mot *char* contient 3 bigraphes (*ch*, *ha* et *ar*) et 2 trigraphes (*cha* et *har*). Quelques digraphes sont des ***graphèmes*** (*ch*).

Fluence. Pont entre décodage et compréhension. La fluence se caractérise par la capacité de lire les mots avec précision et rapidité. Cette capacité permet au lecteur de consacrer ses ressources cognitives à la compréhension de ce qu'il lit, et donc de lire un texte de façon expressive (prosodie), Elle est le signe de l'***automaticité des procédures de lecture***.

Fluence (mesure de la). Mesure utilisée pour évaluer la précision et la rapidité en lecture à haute voix de mots inventés ou de mots connus présentés en isolat ou en contexte. Une bonne fluence est un indice du fait que le lecteur a automatisé les procédures de lecture qu'il utilise. Voir aussi ***Automaticité des procédures de lecture***.

Forme fléchie. Principalement, en français, marques du genre et du nombre pour les noms ou les adjectifs (*un joli petit blond* versus *une jolie petite blonde*), et marques de la personne, du nombre, et du temps pour les verbes (*je chante* versus *tu chantes* et *il chante* versus *ils chantent* ou *il chantera*). En français, de nombreuses marques écrites ne sont pas prononcées. Par exemple, l'énoncé *les garçons portent des pantalons longs* contient six marques à l'écrit mais deux à l'oral.

Graphème. Unité de base d'une écriture alphabétique qui transcrit un ***phonème.*** Un graphème peut avoir une ou plusieurs lettres (*ch*, *an*, *s* et *on* dans *chanson*) et il peut inclure un signe ***diacritique*** (cf. é/è dans élève).

Groupe consonantique. Groupe de deux consonnes (ou plus) en début ou en fin de ***syllabe :*** par exemple, *str-* et *-ct* dans le mot *strict.* À ne pas confondre avec ***graphème***.
Méthode phonique (ou phonographique). Méthode d'enseignement qui utilise de façon systématique les correspondances graphème-phonème.

Morphème. Le morphème est la plus petite unité linguistique ayant un sens ou une fonction. Les mots contiennent souvent plusieurs morphèmes : par exemple, les mots *inconfortable* et *incassable* ont, en plus des bases *confort* et *casser,* le préfixe privatif *in-* et le suffixe *-able* (ce dernier sert à former des noms à partir de noms ou de verbes). Voir aussi ***Dérivation*** et ***Forme fléchie.***

Morphologie dérivationnelle. Voir ***Dérivation*** et ***Morphème.***

Morphologie flexionnelle. Voir ***Forme fléchie*** et ***Morphème.***

Opacité de l'orthographe. Voir ***Transparence de l'orthographe.***

Orthographe. Ecriture correcte des mots, selon des normes dictées généralement par une académie. Ces normes suivent certains principes, souvent contradictoires (représenter la phonologie et/ou l'étymologie des mots), mais elles n'ont parfois ni rime ni raison.

Phonème. Unité de base de la langue orale qui permet de différencier, dans une langue donnée, deux mots. Le répertoire des phonèmes varie d'une langue à une autre. Ainsi, *b* et *v* sont deux phonèmes différents en français, permettant de différencier *bol* de *vol*, mais pas en espagnol.

Procédure alphabétique. Connaissance de l'alphabet et du principe que les unités de base d'une écriture alphabétique, les ***graphèmes*** (lettres ou groupes de lettres, par exemple *t-ou-r,* dans le mot *tour*) codent les unités de base de l'oral, les ***phonèmes*** (les sons *t-ou-r* du mot *tour*). Cette procédure, appelée *décodage* dans la tradition pédagogique, est également appelée *procédure phonologique* (ou *sublexicale*). Les unités de traitement utilisées par cette procédure n'ont pas de sens, à l'inverse de celles utilisées par la ***procédure orthographique****.*

Procédure orthographique. Procédure qui utilise le principe orthographique, à savoir, les mots écrits codent non seulement les unités de base de la langue orale (les ***phonèmes***), mais aussi des marques morphologiques (*ment* dans *sagement*) ou étymologiques (le *th* de *théâtre* représentant le *phi* [θ] grec, qui se prononce *t* en français). Cette procédure, également appelée *procédure lexicale de lecture* (parce qu'elle s'appuie sur le traitement d'unités du lexique qui ont un sens), se met très vite en place au début de l'apprentissage, mais un peu après la ***procédure alphabétique****.*

Procédure lexicale de lecture. Voir ***Procédure orthographique.***

Procédure phonologique de lecture. Voir ***Procédure alphabétique.***

Procédure sublexicale de lecture. Voir ***Procédure alphabétique.***

Rime. Partie de la ***syllabe*** qui suit l'***attaque*** consonantique : la rime du mot du mot *strict* est *-ict, str-* étant son attaque. Certaines rimes n'ont pas d'attaque (*arc*), et peuvent se limiter à une voyelle (*a, ou*, *on*, *en*…).

Syllabe. Une syllabe comporte au minimum une voyelle (*a*, en français, comme en anglais). D'autres syllabes simples contiennent une consonne et une voyelle (*fou*) ou une voyelle et une consonne (*il),* ou encore une suite consonne-voyelle-consonne (*bol*). Les autres structures syllabiques, dites complexes, commencent et/ou se terminent par des groupes de consonnes (*arc, truc, strict*).

Transparence/opacité de l'orthographe (Lecture/Ecriture). La transparence de l'orthographe pour la lecture dépend du nombre de fois qu'un graphème donné (par exemple, *ch*) se prononce d'une certaine façon ('*ch'* comme dans *vache*) par rapport au nombre total d'occurrences de ce graphème, quelle que soit sa prononciation ('*ch'* comme dans *vache* ou 'k' comme dans *technique*). Une orthographe est dite

transparente (ou consistante) quand les relations entre graphèmes et phonèmes, ou entre phonèmes et graphèmes, sont quasi-univoques. Dans les orthographes moins transparentes (également appelées orthographes opaques), un graphème peut transcrire plusieurs phonèmes et un phonème peut être représenté par plusieurs graphèmes.

Termes statistiques

% brut d'accord intra-assesseur. Rends compte du degré de similitude de l'évaluation d'une même réponse par deux assesseurs ou plus. Etant donné que cette statistique est peu précise, il n'est pas possible d'en dériver un critère de référence. Il est souhaitable que le % d'accord soit le plus élevé possible (le plus proche de 100%) lors de l'évaluation des élèves. Quelle qu'en soit la valeur, il faudra y ajouter le calcul du ***Kappa*** afin d'interpréter la qualité du % brut d'accord intra-assesseur.

Approche de discontinuité par régression. Méthodologie de recherche quasi expérimentale, cette approche est utilisée pour estimer les effets d'une intervention en comparant les individus dont les performances chevauchent un seuil de valeur donné. Dans le cadre de l'étude d'impact d'un programme d'amélioration de la lecture, les chercheurs administrent ce programme à un ***groupe expérimental***, constitué d'élèves qui, par exemple, lisent un peu moins de 50 mots par minute (seuil de valeur). L'approche de discontinuité par régression reposera sur la comparaison des performances du groupe expérimental à celles d'un groupe n'ayant pas reçu le programme expérimental (***groupe témoin***), et qui lit un peu plus de mots que le seuil de valeur fixé. Cette méthode présuppose que les deux groupes sont à l'origine (avant l'intervention), identiques, dans la mesure où une différence de quelques points ne suffit pas à les rendre statistiquement différents. Après l'intervention, l'efficacité du programme expérimental sera démontrée si les scores de lecture du groupe expérimental sont significativement supérieurs à ceux du groupe témoin.

Base d'échantillonnage. Liste de tous les membres composant la ***population*** qui peut être incluse dans l'échantillon. Les sujets qui composent l'échantillon sont choisis parmi les membres de cette liste.

Calcul de la puissance. Calcul pouvant être utilisé pour obtenir la taille minimale de l'échantillon requit afin de pouvoir détecter un effet ou une différence de taille donnée. Peut également être utilisé pour calculer à priori la taille de l'effet décelable minimal sur la base de la taille de la population de l'étude (échantillon).

Coefficient de corrélation interne. Statistique descriptive utilisée pour les données regroupées en grappes. Ce coefficient indique le degré de similarité entre les membres d'une même grappe et permet également de mesurer la consistance entre assesseurs dans leur manière de coder les mesures. Sa valeur est comprise entre 0 et 1.

Coefficient Kappa. Voir **Kappa**

Echantillon. Groupe de sujets (individus ou unités) qui, parmi une ***population*** cible, a été sélectionné pour participer à une étude.

Echantillonnage à l'aveuglette. Cette méthode de sélection repose sur l'échantillonnage de participants disponibles au moment de l'enquête. Elle fait partie des échantillonnages non-***aléatoires*** dont la collecte de données est conduite auprès des membres de la population qui sont faciles à recruter. Elle ne permet pas de généraliser les résultats de l'enquête et sa pertinence est très limitée pour la recherche en sciences sociales.

Echantillonnage aléatoire. Terme générique qui désigne tout échantillon sélectionné selon les principes de la théorie de la probabilité. Les méthodes de probabilité d'inclusion proportionnelle à la taille ou l'échantillonnage aléatoire simple en sont des exemples courants.

Echantillonnage aléatoire simple. Méthode de sélection d'un échantillonnage aléatoire qui donne à chaque individu de l'échantillon (la population de l'étude), la même probabilité d'être inclus dans l'échantillon.

Echantillonnage en boule de neige. ***Echantillonnage non-aléatoire*** dans lequel le premier échantillon de participants fourni les informations nécessaires à la sélection de participants supplémentaires.

Echantillonnage en grappes. Méthode d'échantillonnage par laquelle la ***population*** est divisée en groupes, appelés grappes. Les grappes sont sélectionnées dans un premier temps, puis les éléments faisant parti des grappes sont évalués. Par exemple, 20 écoles sont sélectionnées parmi toutes les écoles primaires existantes, puis tous les élèves de 3e année de ces 20 écoles sont évalués.

Echantillonnage non-aléatoire. Désigne toutes procédures d'échantillonnage qui ne reposent pas sur les principes de la probabilité. Inclus l'échantillonnage à l'aveuglette, l'échantillonnage en boule de neige et l'échantillonnage par quota.

Echantillonnage stratifié. Méthode utilisée en statistique pour s'assurer que l'échantillon final représente un nombre adéquat d'individus faisant partie de sous-groupes nécessaires à la représentation fidèle de la population cible.

Effet décelable minimal. Taille (amplitude) de l'effet d'une intervention la plus faible qui puisse être obtenu étant donné la taille de la population sélectionné.

Effet plafond. Lorsque les valeurs d'une variable d'un test sont concentrées autour des valeurs les plus élevées possibles avec un large nombre de participants ayant des scores à cette limite, ou proche de cette limite. Si une tache d'EGRA est trop facile pour l'ensemble des élèves, les scores se concentreront artificiellement autour des valeurs maximales du test ce qui restreindra la variabilité des scores et réduira la validité de l'outil. C'est l'inverse de l'***effet plancher***.

Effet plancher. Terme statistique indiquant que les scores d'un sujet, ou d'un groupe de sujets, sont au niveau le plus bas: zéro, pour la précision de la réponse, par exemple. Les scores sur une capacité testée dans EGRA seront biaisés par des effets plancher si l'épreuve est trop difficile pour la plupart des enfants des premières années du primaire, qui ont des notes à zéro. Par contre, si l'épreuve est trop facile, les scores peuvent être au plafond (100% de réponses correctes). La présence d'effets plancher ou d'***effets plafond*** ne permet pas d'utiliser correctement certaines analyses statistiques, en particulier celles qui sont basées sur les corrélations (qui évaluent les relations entre deux variables).

Estimation ponctuelle. Statistique (valeur ou taille de l'effet) dérivée de l'échantillon qui fournit une estimation du paramètre (vraie valeur) pour la ***population*** sous-jacente.

Fiabilité d'un instrument de mesure. Cohérence ou régularité des réponses données par les participants d'une étude testés au moyen d'un instrument de mesure. Un instrument de mesure est fiable lorsqu'il fournit les mêmes résultats chaque fois qu'il est utilisé (dans la mesure où les conditions sont comparables). La fiabilité n'est pas une mesure exacte et il n'est possible d'en dériver qu'une estimation. Il existe plusieurs grands types de fiabilité :

- **La fiabilité inter-évaluateur** (ou inter-assesseur), qui est élevée lorsque la réponse d'un même élève est identique, quelle que soit la personne qui l'évalue.
- **La fiabilité test-retest**, qui est fiable lorsqu'un instrument utilisé de manière successive donne des résultats identiques à chaque fois.
- **La fiabilité interne**, qui rend compte de la cohérence des réponses aux différents items d'un même test.

Intervalle de Confiance (IC). Fourchette de valeurs à l'intérieur de laquelle l'estimation de la vraie valeur recherchée est comprise. Il donne une visualisation de l'incertitude de l'estimation obtenue auprès d'un échantillon de la population. Plus l'intervalle de confiance est large, plus le degré d'incertitude est grand, donc moins l'estimation est précise. Par exemple, dans le cas où l'estimation de l'âge moyen de l'échantillon est de 36 ans et l'IC est faible (de 35 à 37), l'âge moyen de l'échantillon estime l'âge moyen réel de la ***population*** de manière plus précise que si l'IC est plus fort (de 34 à 38).

Kappa. Mesure la probabilité que la concordance entre assesseurs indépendants évaluant un même participant soit due à la chance. Ses valeurs sont comprises entre -1 et 1. Plus la valeur est faible, plus il est probable que la concordance soit due à la chance et indique que les données sont peu fiables.

(d'après Fleiss, 1981)

Coefficient Kappa	Degré de concordance
Inférieur à 0.40	Faible
Entre 0.40 et 0.75	Moyen
Supérieur à 0.75	Fort

Mesure de la précision. Indice de la similitude entre des ***estimations ponctuelles*** obtenues auprès de plusieurs échantillons de la même ***population***. Plus les valeurs sont proches, plus les estimations ponctuelles sont précises.

Méthode d'appariement sur les coefficients de propension. Cette procédure permet d'assigner les participants d'une étude soit à un ***groupe expérimental***, soit à un ***groupe contrôle*** (ou ***groupe témoin***) sur la base de la probabilité de leur participation à une intervention (par exemple, aide à l'apprentissage de la lecture ciblée sur le décodage ou sur la compréhension). Elle permet de s'assurer que les participants assignés aux dex groupes sont comparables, en particulier dans le cas où une assignation aléatoire n'est pas possible ou n'a pu être conduite.

Population. Groupe de sujets cible (individus ou unités) que les résultats de l'étude sont censés représenter. L'échantillon d'une étude doit être sélectionné de manière à partager les mêmes caractéristiques que la population sous-jacente dont il fait partie.

Puissance statistique. Mesure la probabilité qu'un test statistique détecte une différence significative au sein de l'échantillon étudié.

Recensement. Lorsque les participants d'une étude constituent l'ensemble de la ***population*** (aucun échantillonnage n'est conduit). Ce type de recensement est différent de ce qui est appelé en France « Recensement », qui concerne la population à un niveau national.

Significativité statistique. Mesure la probabilité qu'une différence statistique détectée n'est pas obtenue par chance. Plus la valeur est proche de zéro, plus il est probable que la différence détectée ne soit pas due à la chance.

Sondage complexe. Méthode similaire à celle de l'échantillonnage en grappes, mais les éléments faisant partie des grappes sélectionnées sont à leur tour eux aussi échantillonnés. Par exemple, après la sélection de 20 écoles parmi les écoles primaires existantes, 10 élèves de 3e année sont sélectionnés dans chacune de ces 20 écoles.

Unité d'enquête. Individus faisant partie de l'échantillon, c'est-à-dire les sujets auprès desquels les données de l'étude seront collectées. L'unité d'enquête d'une étude peut être un foyer fiscal ou une personne par exemple.

Termes méthodologiques

Aperçu sommaire. ***Etude transversale*** conduite une seule fois, sans comparaison dans le temps.

Attrition. Perte graduelle de participants au cours d'une étude (ceux qui déménagent ou ne souhaitent plus participer à l'étude) ; ce phénomène concerne surtout les études longitudinales.

Biais d'effet séquentiel. Décrit le fait que lors de la collecte de données, la perception par les assesseurs de ce qu'est une réponse correcte (ou incorrecte) tend à changer avec le temps. Pour éviter ce biais, il faut mesurer régulièrement la qualité de l'accord intra-assesseur.

Etude longitudinale. Etude dans laquelle les mêmes individus sont suivis pendant un certain temps (plusieurs mois ou plusieurs années), les données étant collectées de façon répétée, à la différence d'une é***tude transversale.*** Par exemple, une cohorte d'enfant peut être suivie longitudinalement de la maternelle à la fin du primaire pour déterminer la proportion d'individus qui développeront des troubles d'apprentissage de la lecture.

Etude transversale. Comparaison de groupes d'individus indépendants dont les données ne sont recueillies qu'une seule fois. Par exemple, les scores obtenus en lecture par des élèves de 3ème année du primaire fin 2000 sont comparés à ceux d'élèves de la même année fin 2015. L'étude transversale diffère de l'étude longitudinale, qui repose sur la comparaison du même groupe d'individus au cours du temps.

Groupe contrôle (également appelé ***groupe témoin***)**.** Sujets assignés aléatoirement au groupe qui n'est pas exposé au programme expérimental, le pendant de ce groupe est le ***groupe expérimental*** qui, lui, bénéficie d'un programme expérimental. Les participants du groupe contrôle sont choisis de manière à être le plus similaire possible à ceux du groupe expérimental quant à un certain nombre de caractéristiques prédéfinies.

Groupe expérimental. Sujets qui bénéficient d'une intervention.

Groupe témoin. Sujets assignés aléatoirement au groupe qui n'est pas exposé au programme expérimental. Le groupe témoin est également appelé ***groupe contrôle*** et son pendant est le ***groupe expérimental***.

Situation contrefactuelle. Mesure qui évalue, dans le cadre d'une étude portant sur l'impact d'une intervention, ce qui serait advenu aux participants du groupe expérimental sans l'intervention. Parce que la vraie situation contrefactuelle est inobservable (les individus ayant reçu l'intervention ne peuvent pas être, dans le même temps, observés en l'absence de l'intervention), une variété de méthodes statistiques ont été développées pour construire un groupe contrefactuel qui

représente ce qui serait probablement arrivé au groupe expérimental en l'absence d'intervention. Le ***groupe expérimental*** est ensuite comparé à ce groupe pour obtenir une estimation de l'effet de l'intervention.

Tests équivalents. Série de tests dont le contenu a été développé afin que leur niveau de difficulté soit comparable. Pour ce faire, les tests sont étalonnés en collectant les scores auprès des mêmes individus et en les transformant statistiquement pour les représenter sur une échelle unique. Il en résulte des scores ajustés qui permettent de s'assurer que, bien que les tests soient différents, un même individu se verra attribuer le même score, quelle que soit la version du test utilisée. L'utilisation de tests équivalents est cruciale lorsque l'impact d'un programme est étudié, car elle permet d'attester que les différences entre scores avant et après l'intervention ne sont pas dues au contenu du test mais bien à l'intervention.

Tests étalonnés. Tests dont les résultats ont été ajustés statistiquement de manière à ce que leurs scores soient équivalents.

Validité apparente. Test dont les items sont en rapport avec ce qu'ils sont censés mesurer.

Validité conceptuelle. Terme utilisé pour indiquer que l'instrument mesure bien ce qui est attendu. Plus particulièrement, la ***validité apparente*** caractérise un test dont les items sont en rapport avec ce qu'ils sont censés mesurer.

- **Validité du contenu.** Correspond à une vision d'ensemble des items qui composent le test. Le contenu du test est valide lorsque les items qui le composent peuvent être considérés comme de bons indicateurs de ce que l'on cherche à mesurer.
- **Validité critériée.** Mesure la validité de l'instrument sur la base d'hypothèses et de prédictions précises. Il existe différents types de validité qui diffèrent par le type de critère utilisé pour prédire la validité de l'instrument :
    - Validité prédictive. Rend compte de l'aptitude de l'instrument à prédire le résultat qu'il est censé prédire.
    - Validité concourante. Rend compte de l'aptitude de l'instrument à distinguer des groupes d'individus dont les performances sont censées être distinctes.
    - Validité convergente. Rend compte du degré de convergence, ou similitude, avec d'autres outils qui examinent les mêmes construits et que l'on considère similaires.
    - Validité discriminante. Rend compte de la différence avec d'autres outils qui sont censés mesurer, par exemple, des compétences opposées ou radicalement différents.

Versions de test comparables. Tests qui sont créés et utilisés à des fins de comparaison et dont la conception est guidée par les mêmes principes, inclus les mêmes sous-tâches, etc.

1 INTRODUCTION

1.1 Pourquoi avons-nous besoin d'une évaluation pour le début de l'apprentissage de lecture ?

Les pays du monde entier ont favorisé une scolarisation de masse pour le primaire. Grâce aux efforts ciblés sur l'Education Pour Tous (EPT) de l'Organisation des Nations Unies et aux objectifs du millénaire pour le développement (OMD), supposés être atteints en 2015, le monde a connu des améliorations spectaculaires des taux de scolarisation dans le primaire. Dans certains endroits, ces taux sont maintenant à peu près les mêmes dans les pays à faible revenu que dans ceux à revenu élevé. Le taux net de scolarisation à l'école primaire dans les régions en développement a atteint environ 91 % en 2015, contre 83 % en 2000. Autre fait marquant, le nombre d'enfants non scolarisés en âge de fréquenter l'école primaire dans le monde a chuté de près de la moitié dans le même laps de temps (Nations Unies, 2015).

Les données sur les résultats des pays à faible revenu ayant participé à diverses évaluations, y compris celles administrées aux niveaux scolaires 1 à 3, sont maintenant disponibles sur le site en ligne de la Banque mondiale (EdStats, Banque mondiale, 2015a). Les résultats indiquent que, si le pourcentage d'enfants scolarisés a augmenté dans les pays à faible revenu, les scores des enfants restent encore faibles dans la plupart de ces pays. La Banque mondiale a récemment résumé la situation ainsi : « Tout le monde au sein de la communauté internationale s'accorde pour reconnaître que la réalisation de l'objectif du millénaire pour le développement de l'éducation (OMD) exige l'amélioration des résultats de l'apprentissage » (Banque mondiale, 2015b); l'éducation de qualité a été adoptée à l'échelle mondiale comme étant le quatrième objectif du plan post-2015 pour le développement durable (Programme Développement des Nations Unies [PDNU], 2015). L'importance de la qualité de l'éducation pour le développement économique national est un autre domaine qui fait l'objet d'un large accord : « La recherche récente révèle que c'est la qualité de l'apprentissage plutôt que le nombre d'années de scolarité qui contribue à la croissance économique d'un pays : une augmentation de 10 % de la proportion des enfants atteignant un niveau d'alphabétisation de base entraine une augmentation du taux de croissance annuel de 0,3 % (Hanushek & Woessman 2009, cité dans Gove & Wetterberg, 2011, pp.1-2).

Au moment de la première édition du Manuel, en 2009, les évaluations les plus couramment utilisées permettaient de connaitre ce que les élèves de pays à faible revenu ne savaient pas, mais pas ce qu'ils savaient. Cela s'expliquait surtout par la faiblesse de leurs scores, qui ne permettait pas de situer leurs connaissances et compétences sur un continuum. En outre, la plupart des évaluations nationales et internationales ont, par le passé, été administrées en version papier-crayon à des

élèves scolarisés au mieux en 4ème année du primaire, et donc supposés savoir lire et écrire. Les résultats de ces tests ne permettaient pas de dire si un échec était la conséquence de lacunes dans les connaissances testées, ou d'une faiblesse du niveau de base de lecture. Depuis 2010, un virage vers l'évaluation, dès les premiers niveaux du primaire, des compétences de base en lecture a été pris, virage dû en grande partie à l'influence d'USAID et la Banque mondiale. Cela a entrainé une prise de conscience, parmi les acteurs de l'éducation au niveau international, du besoin d'informations plus empiriques sur la capacité des jeunes enfants à lire avec compréhension.

La capacité de lire un texte simple est l'une des compétences les plus fondamentales qu'un enfant doit apprendre. Sans un niveau d'alphabétisation de base, il a peu de chance d'échapper au cycle intergénérationnel de la pauvreté. Pourtant, dans de nombreux pays, les élèves inscrits à l'école pendant au moins six ans s'avèrent incapables de lire un texte simple. Les résultats des recherches indiquent qu'apprendre tôt les principes à la base de la lecture, c'est-à-dire être capable de décoder les mots avec précision et rapidité, va faciliter le futur niveau de compréhension écrite.

1.1.1 Pourquoi évaluer la **lecture** ?

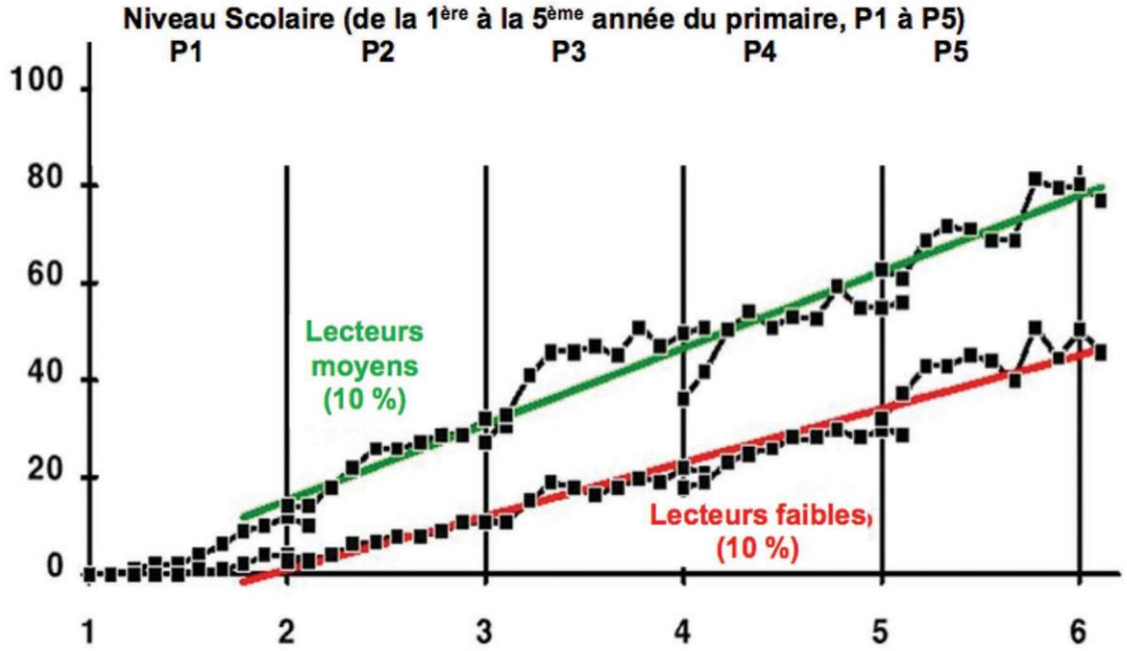
Un niveau d'alphabétisation de base est le fondement dont les enfants ont besoin pour réussir dans tous les domaines de l'éducation. Les enfants doivent d'abord « apprendre à lire » afin de pouvoir « lire pour apprendre ». Autrement dit, alors qu'ils passent d'un niveau scolaire à l'autre, de plus en plus de contenus académiques leur sont transmis à travers des textes, et leur capacité à acquérir de nouvelles connaissances et compétences dépend en grande partie de leur capacité à construire le sens des textes lus. Par exemple, il faut savoir lire pour utiliser un livre de mathématiques. Les élèves sont également de plus en plus tenus de démontrer ce qu'ils ont appris par l'écriture, une compétence fortement liée à la lecture. Par ailleurs, un faible niveau d'alphabétisation limite considérablement la capacité qu'a une personne d'apprendre par elle-même, et de façon continue, ce qui est très important au-delà des murs de l'école et dans le monde des responsabilités des adultes.

1.1.2 Pourquoi l'évaluer dès les **premières** étapes de l'apprentissage ?

Alors que les élèves grandissent, il devient de plus en plus difficile d'acquérir les compétences de lecture; les élèves n'ayant pas appris à lire au cours des premières années sont plus susceptibles de redoubler et, à terme, d'abandonner l'école. De plus, comme l'indiquent les résultats présentés à la **Figure 1**, les différences entre les élèves qui sont arrivés à maitriser les compétences de base en lecture et ceux qui n'y sont pas arrivés augmentent au fil du temps (voir aussi Adolf, Catts, & Lee, 2010 ; Daniel et al., 2006 ; Darney, Reinke, Herman, Stormont, & Lalongo, 2013; Scanlon, Gelzheiser, Vellutino, Schatschneider, & Sweeney, 2008, Torgesen 2002 ; pour des résultats en français, voir Billard, Bricout, Ducot, Richard, Ziegler, & Fluss,

2010). La métaphore « les riches s'enrichissent et les pauvres s'appauvrissent » est souvent citée dans les discussions sur l'augmentation dans le temps des différences entre bons et faibles lecteurs en fonction de leur niveau de lecture initial (Stanovich, 1986 ; voir aussi Gove & Wetterberg, 2011).

Figure 1. Evolution des scores en lecture d'enfants lecteurs faibles (ligne du bas) ou moyens (ligne du haut) : nombre de mots corrects prononcés en une minute

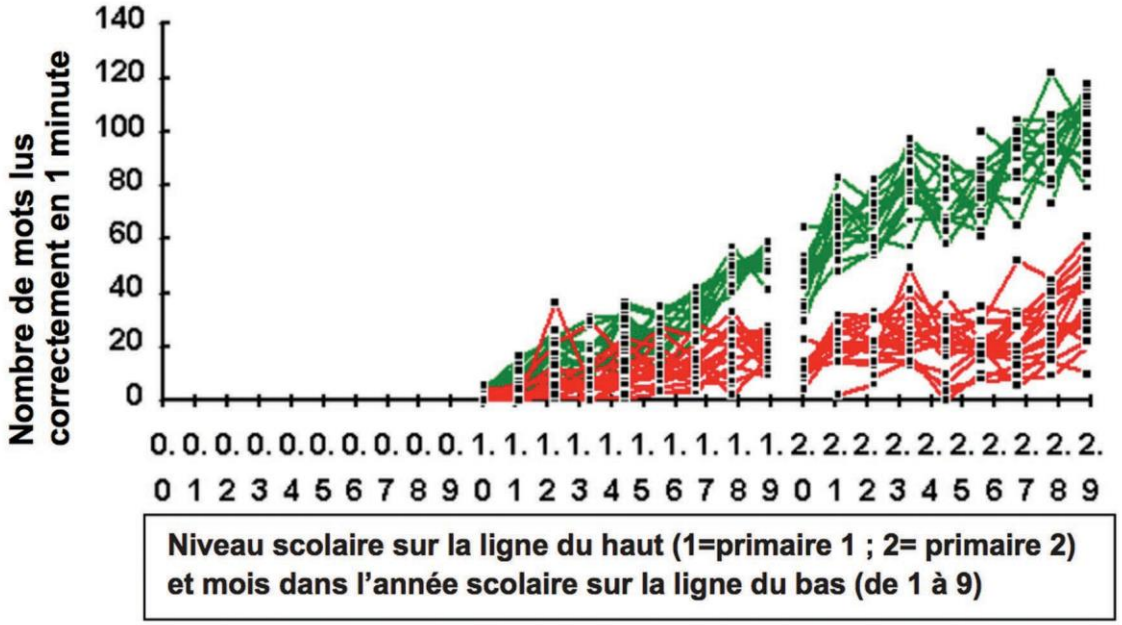


(D'après Good, Simmons, & Smith, 1998, Figure 1).

Contrairement au langage oral, le langage écrit ne s'acquiert pas naturellement, sans instruction. Comme le signalent certaines études (cf. Juel, 1988 ; Torgesen, 1998), sans un enseignement de qualité, les enfants qui lisent mal dans les premières années continueront à lire mal dans les classes supérieures. En outre, il leur faudra beaucoup plus d'aides pédagogiques pour rattraper leur retard.

La **figure 2** indique l'évolution des performances en lecture d'élèves des USA scolarisés en 1ère et 2ème année du primaire n'ayant pas bénéficié d'un enseignement supplémentaire visant à améliorer la lecture. Les lignes de la partie supérieure et inférieure du côté gauche de cette figure montrent les résultats mois par mois des élèves qui, à la fin de la 1ère année, pouvaient lire en une minute soit au moins 40 mots, soit moins de 40 mots. On peut constater que l'écart entre les lecteurs les plus compétents et les moins compétents s'est creusé par la fin de 2ème année. Ces résultats signalent que, en l'absence d'intervention en temps opportun, les élèves qui, en début d'apprentissage, avaient un faible niveau de lecture progressent moins que ceux qui, à la même époque, avaient un meilleur niveau de lecture ; cet écart devient de plus en plus difficile à combler.

Figure 2. Evolution des scores en lecture d'enfants plus ou moins bons lecteurs en 1ère année du primaire (ceux qui, au début de l'étude, lisaient plus ou moins de 40 mots par minute : lignes en vert du haut de la figure comparées aux lignes en rouge du bas)



D'après Good et al., 1998

On sait maintenant que, plus les enfants ont des difficultés scolaires dès le début de leur scolarisation, plus grand est le risque qu'ils se découragent et abandonnent l'école, et perdent ainsi les avantages potentiels que l'éducation aurait pu leur offrir. Le résultat inverse s'observe pour les enfants qui ont de bons résultats dès le début de leur scolarisation (Patrinos et Velez, 2009). Un autre étude a révélé que le plus fort prédicteur du niveau de réussite scolaire en fin de primaire est le niveau de lecture en deuxième année (par exemple, l'étude de Glick et Sahn, 2010, effectuée au Sénegal). Enfin, soit pour un enfant individuel, ou pour un système d'éducation entier, il est également maintenant connu qu'il est plus facile de compenser un déficit en lecture dans les premières années que plus tard (Ehri, Nunes, Stahl, & Willows, 2001). Ces différents résultats signalent l'importance d'une évaluation précoce du niveau de lecture des enfants.

1.1.3 Pourquoi utiliser des épreuves de lecture à **haute voix** ?

Il faut avoir acquis un niveau de lecture et d'écriture de base pour pouvoir répondre correctement aux tests traditionnels papier-crayon. Quand un enfant s'avère incapable, par exemple, de lire la question ou d'écrire la réponse, il n'est pas possible de connaitre l'origine de son échec. En effet, cet échec peut provenir de difficultés en lecture ou en écriture, les premières pouvant avoir trois origines : (1) la non-maitrise des mécanismes de base de la lecture (le décodage) ; (2) une maitrise

insuffisante de la langue du test ; (3) des difficultés de compréhension en général (à l'oral, comme à l'écrit).

Dans de nombreux pays, les élèves doivent passer un examen national à la fin de la 6ème année du primaire, examen qui permet de valider leurs acquis (et obtenir leur certificat d'études primaires) et d'entrer à l'école secondaire (Braun & Kanjee, 2006). En outre, dans certains pays (principalement des pays développés), des évaluations internationales sont proposées aux enfants : PIRLS (Progress in International Reading Literacy Study, pour les enfants de 4ème année du primaire) ou PISA (Programme for International Student Assessment, pour les élèves âgés de 15 ans). Lors de ces évaluations, les élèves doivent lire (souvent des passages courts) et montrer ce qu'ils ont compris par le biais de réponses à des questions (souvent des questions à choix multiples). Comme indiqué précédemment, quand un élève échoue, il est impossible de savoir si son échec s'explique par la non-maitrise du décodage et/ou de la langue du test, ou encore par l'absence des connaissances nécessaires pour répondre aux questions.

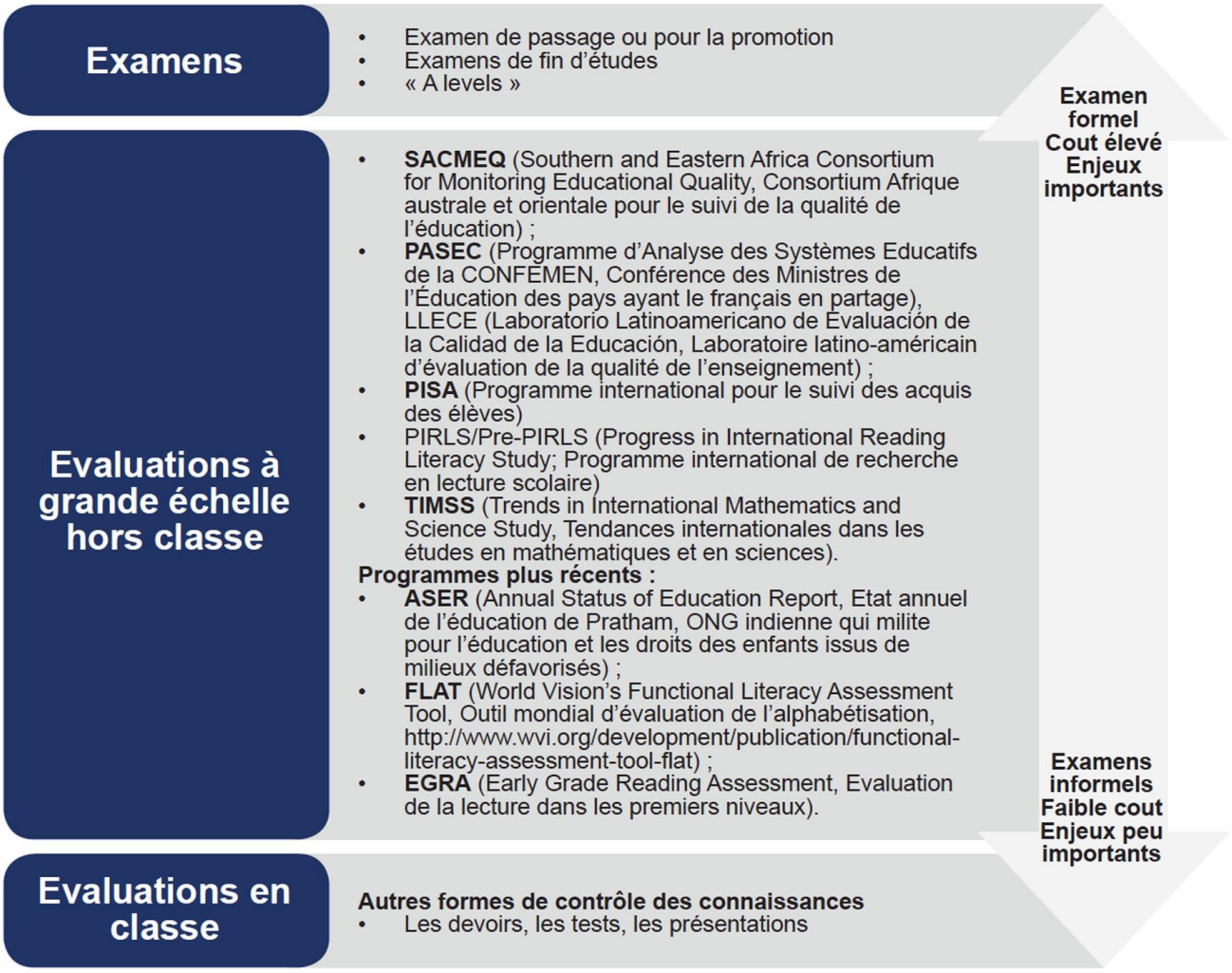
La compréhension et la fluence en lecture, comme la compréhension orale, sont des capacités de haut niveau qui s'opposent à la capacité de bas niveaux impliquée spécifiquement dans la lecture, le décodage. Ce dernier se défini comme étant la capacité de mettre en relation, dans une écriture alphabétique, les unités de base de cette écriture, les graphèmes (lettres ou groupes de lettres, par exemple t-ou-r, dans le mot tour) avec les unités de base de l'oral, les phonèmes (les sons t-ou-r du mot tour). C'est l'automaticité du décodage (niveau de maitrise de la lecture à partir duquel la reconnaissance des mots écrits est devenue un acte quasi réflexe) qui permet au lecteur débutant d'atteindre progressivement un niveau de compréhension écrite égal à celui de sa compréhension orale.

Le niveau de décodage peut être évalué par la lecture à haute voix de mots ayant des correspondances graphème-phonème régulières, ce qui est le cas de la plupart des mots en français (voir le chapitre 4). Il est possible de tenir compte de la précision et de la rapidité dans ces tests, la rapidité d'un décodage précis étant un indicateur du niveau d'automaticité. La capacité à décoder les mots correctement dépend elle-même d'autres capacités, en particulier du niveau de conscience phonémique. Une évaluation orale des capacités de décodage et de conscience phonémique peut donner des informations précises sur ce que les élèves savent et où ils en sont de leurs acquisitions. Ce type d'évaluation permet aussi de détecter l'évolution dans le temps des performances des élèves et de détecter des changements qui ne sont pas détectables en utilisant les tests traditionnels papier-crayon.

1.1.4. Place de l'évaluation des compétences fondamentales en lecture (EGRA) par rapport à d'autres outils

Pour situer EGRA parmi les autres options d'évaluation, il est utile de placer les différents types d'évaluations sur un continuum tel qu'il apparait sur le **Figure 3**.

Figure 3. Différents types d'évaluations : Un continuum



Adapté de Kanjee, 2009

Ce continuum est divisé en trois grandes catégories : les examens, les évaluations hors classe et les évaluations en classe. Kanjee (2009) définit les examens comme étant les procédés utilisés pour tester les qualifications des candidats (par exemple, examens de promotion et examens de fin d'études). Ces évaluations sont généralement plus longues et plus formelles que les autres évaluations, les tests – normalisés – étant administrés à tous les élèves (les rendant par conséquent plus couteuxet exigeant plus de temps, entre autres). A l'autre extrémité du spectre se situent les évaluations en classe, qui sont définies comme étant des mesures utilisées pour obtenir des données sur les connaissances, les compétences et les attitudes des apprenants dans le but d'informer et d'améliorer l'enseignement et l'apprentissage (Kanjee, 2009). Ces évaluations, moins formelles que les précédentes, se présentent souvent sous la forme de tests qui sont effectués en salle de classe, incluant des devoirs et des présentations. De par leur conception, les évaluations en classe sont peu couteuses, prennent moins de temps, faibles que les examens, en particulier.

D'autrLes évaluations, à grande échelle, sont conçues dans le but explicite d'obtenir des informations sur les performances des élèves, ainsi que sur les systèmes d'éducation. En plus de PIRLS et PISA, de nombreuses autres évaluations internationales, nationales ou régionales entrent dans cette catégorie. On peut citer celles de la SACMEQ, de la CONFEMEN[2] (PASEC), du LLECE ainsi que TIMSS (voir le tableau 1 pour l'explication des sigles).

Les tests associés à ces programmes sont destinés à mesurer les tendances en matière d'alphabétisation dans le but de permettre des comparaisons entre pays ou de cerner les évolutions dans le temps pour un même pays. Ils doivent donc prendre en compte les complications qui résultent des différences entre les langues, et utiliser des procédures de notation et de mise à l'échelle complexes. En outre, ces évaluations nécessitent la maitrise des capacités de base en lecture (étant donné qu'elles utilisent la lecture de passage), ce qui peut limiter leur pertinence pour l'évaluation des compétences en lecture aux premiers niveaux de scolarisation, en particulier dans les pays en voie de développement (en raison des effets « plancher » majeurs).

De nouvelles évaluations des compétences précoces en lecture ont été développées pour combler cette lacune, par exemple, ASER, FLAT[3] et EGRA (voir le tableau1 pour l'explication des sigles). Ces évaluations administrées individuellement sont présentées comme étant plus rapides et moins couteuses que les autres évaluations internationales (Wagner, 2011).

1.2 Développement de l'outil EGRA

Dans le contexte des questions précédemment évoquées sur l'apprentissage et sur les investissements pour l'éducation pour tous, les départements de l'éducation et du développement professionnel de la Banque mondiale et de USAID, ainsi que d'autres institutions, ont sollicité la création de mesures simples, efficaces et à faible coût permettant d'évaluer les résultats en lecture des élèves, en particulier dans les débuts de l'apprentissage (Abadzi, 2006 ; Center for Global Development [Centre pour le développement mondial], 2006 ; Chabbott, 2006 ; World Bank : Independent Evaluation Group [Banque mondiale : Groupe indépendant d'évaluation], 2006).

Pour répondre à cette demande, un projet ayant pour objectif premier de développer un outil simple pouvant servir à évaluer précisément les premières étapes de l'apprentissage de la lecture a été mis en place en octobr 2006. A cette date, USAID a passé un contrat avec RTI International par le biais de « Education Data for Decision Making » (EdData II, données pour la prise de décision en éducation). Ce projet visait à développer un outil qui pourrait aider les pays partenaires d'USAID à mesurer d'une manière précise et systématique la façon dont les enfants acquièrent les compétences de base de la lecture au cours des premières années de l'école primaire. Cet outil devait être facilement adaptable à de nouveaux contextes et à

[2] CONFEMEN: Conférence des ministres de l'Éducation des pays ayant le français en partage.
[3] Functional Literacy Assessment Tool developed and used by World Vision: http://www.wvi.org/development/publication/functional-literacy-assessment-tool-flat

différentes langues, les évaluations devaient être de courte durée et avoir des enjeux faibles. De plus, le système de notation devait être simple. Un autre objectif de ce projet était de permettre d'améliorer les compétences de base impliquées dans la lecture.

Sur la base d'un examen des résultats des travaux de recherche et des évaluations existantes, RTI a élaboré un protocole d'évaluation individuelle, à l'oral, des compétences de base en lecture. Un séminaire, organisé en novembre 2006 par USAID, la Banque mondiale et RTI, a rassemblé des chercheurs en sciences cognitives travaillant dans le domaine des premiers apprentissages de la lecture, des méthodologues ainsi que des experts en évaluation (une douzaine d'experts de différents pays et 15 observateurs d'institutions telles que USAID, la Banque mondiale, la Fondation William et Flora Hewlett, l'Université George Washington, le ministère sud-africain de l'éducation et Plan international, entre autres). Ces experts, ainsi que les observateurs, ont fourni des commentaires sur le projet EGRA et, en particulier, ont confirmé sa validité.

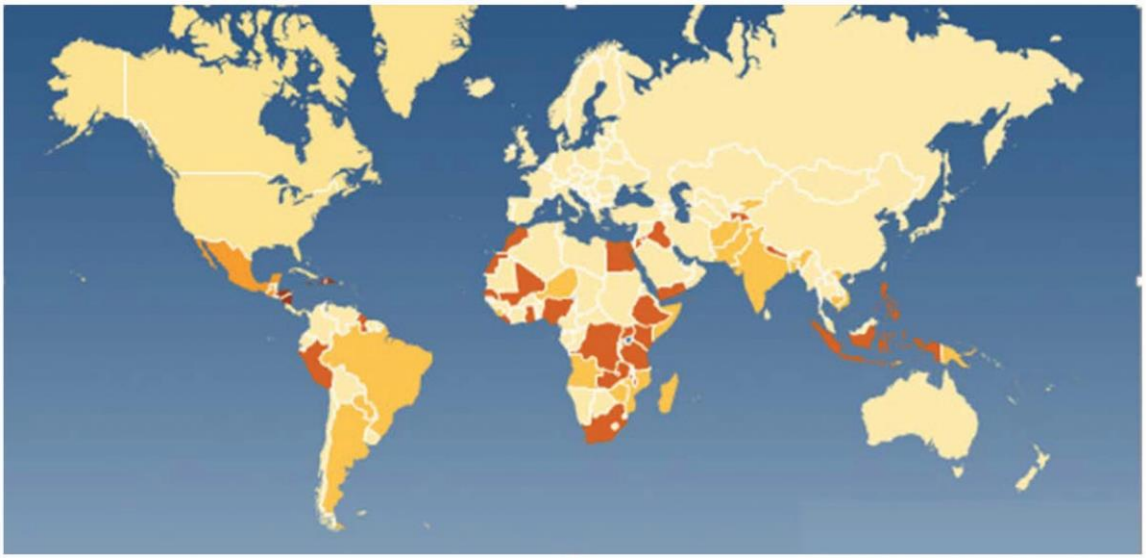
EGRA, qui est en libre accès et facilement disponible, permet d'obtenir des données solides sur les résultats de l'apprentissage de la lecture, et de bien diffuser ces résultats. Les concepteurs de cet outil souhaitaient qu'il soit utilisable par différents professionnels, et permette d'obtenir, dans un laps de temps court, des données précises concernant le niveau d'alphabétisation des enfants, ces données étant destinées à faciliter les prises de décisions en matière de politique éducative (à noter, le projet développé par RTI intègre également une évaluation des compétences de base en calcul).

1.3 EGRA en action

En 2007, deux institutions (la Banque mondiale et USAID) ont subventionné des projets pilote pour le développement et l'évaluation d'EGRA. Cet outil a été évalué en Gambie (en anglais), au Sénégal (en français et en wolof) et au Nicaragua (en espagnol). Ces évaluations ont donné lieu à des rapports adressés à la Banque Mondiale, entre autres, un pour le Sénégal (Sprenger-Charolles, 2008). A la suite de ces premiers pilotes, EGRA s'est largement développé, avec l'aide de différents bailleurs, dans différents pays et en différentes langues. USAID a été l'un des plus grands sponsors de ces développements par le biais du contrat EdData II. Entre 2006 et le milieu de 2015, EdData II a, à lui seul, subventionné des études EGRA dans 23 pays et en 36 langues (voir la **Figure 4**)

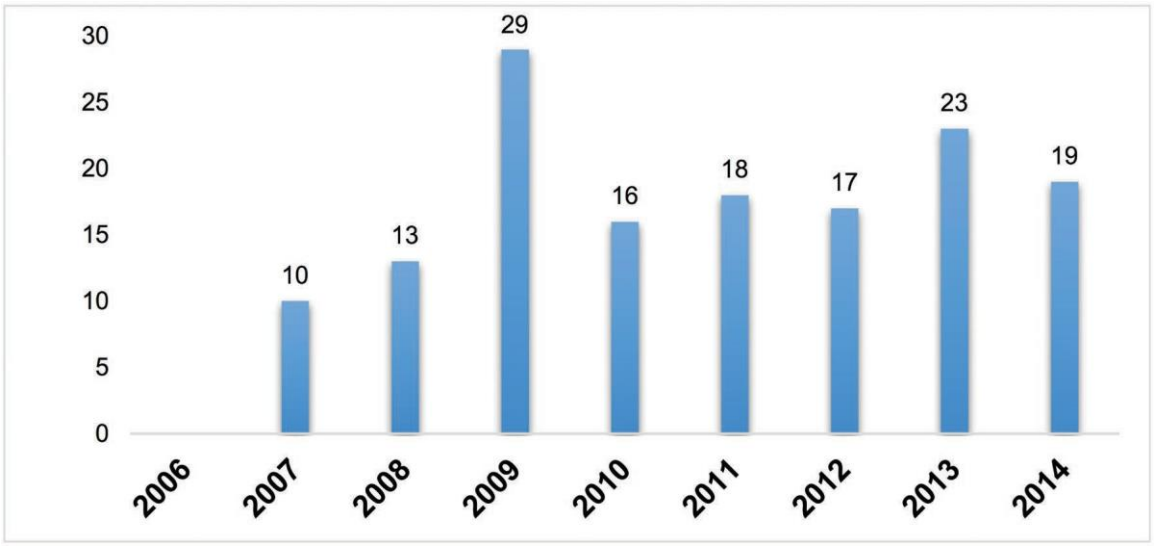
En septembre 2015, près de 10 ans après le développement initial d'EGRA, cet outil a été utilisé par plus de 30 organismes dans plus de 70 pays. L'approche sur les premiers apprentissages de la lecture s'est aussi déplacée pour se focaliser sur l'enseignement en langue maternelle (la langue première des enfants, celle qu'ils parlent à la maison). A cette fin, EGRA a été adapté pour pouvoir être administré dans de nombreuses langues (en fait, plus de 120 langues différentes). EdData II a suivi ces applications, au nom d'USAID (voir les **figures 4 et 5**).

Figure 4. Carte des pays dans lesquels EGRA a été administré



Source: RTI International pour le projet de site WEB « EdData II » (https://www.eddataglobal.org/countries/index.cfm)

Figure 5. Utilisation d'EGRA à travers le monde (Nombre de pays par année)



Source : RTI International, 2015.

1.4 La version originale de l'outil EGRA et la seconde édition

Afin de consolider les diverses expériences qui ont été effectuées, et de diffuser un outil raisonnablement normalisé permettant d'évaluer les premières étapes de l'acquisition de la lecture, la Banque mondiale a, en 2009, demandé à RTI de créer une « boite à outils » (un « mode d'emploi »). Cette boite à outils, qui doit servir de guide pour les pays qui commencent à utiliser EGRA, devait contenir des

informations dans plusieurs domaines : adaptation d'EGRA au contexte local, au travail sur le terrain, et à la façon d'analyser les résultats.

Dans la mesure où EGRA est devenu de plus en plus utilisé par différents organismes et pays, en grande partie en raison du libre accès de cet outil, USAID a demandé à RTI de réviser le Manuel et de le mettre à jour. En vertu d'une demande dans le cadre du projet EdData II (Measurement and Research Support to Education Stratégies Goal 1), RTI a dirigé l'élaboration de la seconde édition de la boite à outils EGRA. Cette nouvelle version reflète les progrès accomplis depuis la version originale, permettant en particulier d'améliorer la qualité des données qui seront disponibles et de faire avancer ainsi certains des buts de l'agenda pour le « Développement durable ».

La révision a débuté en décembre 2014. RTI a commencé par compiler les résultats des expériences et d'autres informations sur EGRA, sur la base des études exécutées à la demande d'EdData II (ainsi que d'autres études également financées par USAID). De nouvelles informations ont été recueillies dans les rapports, les travaux de recherche récents ainsi qu'auprès de chercheurs dans le domaine de l'éducation. En plus des données provenant des rapports présentant les résultats obtenus par différents pays, les nouvelles informations recueillies concernent les meilleures pratiques et les nouvelles avancées technologiques en matière de planification et de mise en œuvre de cet outil, entre autres. L'information a été revue et condensée dans différentes présentations et dans de la documentation qui a été distribuée. Ces présentations et documents ont été utilisés au cours du séminaire EGRA intitulé « Conception et implémentation d'EGRA : Principes de base », qui a été organisée par le Global Reading Network (Réseau Mondial sur la Lecture) en tant que séminaire et webinaire mondial en mars 2015.

Après les réunions de mars, de nouvelles réunions, également destinées à améliorer la qualité des données d'EGRA, ont eu lieu en mai 2015, avec, comme précédemment, un séminaire et un webinaire sous l'égide du Global Reading Network (financé par USAID). Des experts de divers organismes ont fait des présentations sur la conception d'EGRA, son administration, l'analyse et la présentation des données. Ces présentations ont été suivies par des discussions animées entre eux et les autres participants.

Le but de ces séminaires était double : présenter des informations sur la façon de mener une étude avec EGRA (afin de permettre son utilisation à grande échelle) ainsi que sur les processus d'analyse des données (afin d'améliorer leur qualité). Pour obtenir davantage de détails au sujet de ces séminaires, voir l'**annexe A.** L'étape suivante du processus de mise à jour du Manuel après la conclusion des deux précédents séminaires, a été la constitution de groupes de travail avec des experts techniques issus de plusieurs organismes en charge de l'implémentation d'EGRA. Ces groupes de travail devaient présenter, après discussion, un ensemble final de recommandations consensuelles sur les méthodologies à mettre en œuvre pour planifier, implémenter et analyser les données EGRA.

La mise à jour du Manuel est le produit de ces différents séminaires et groupes de travail. Elle est le résultat de collaborations entre plusieurs organismes et individus issus des domaines du développement et de l'éducation au niveau international.

1.5 Comment utiliser le Manuel ?

La présente boite à outils est destinée à être utilisée par les personnels des ministères de l'éducation, ainsi que par les bailleurs de fonds, les praticiens et les professionnels dans le domaine de l'éducation. Le document écrit, qui comprend 12 chapitres, tente de résumer un grand nombre de recherches d'une manière accessible. Les procédures décrites doivent être utilisées dans toutes les mises en place d'EGRA financées par USAID et, on l'espère, dans toutes les autres administrations de ce protocole.

Le manuel ne présente pas une synthèse exhaustive des recherches sur la lecture. Même avec les apports d'autres organisations et des individus, par souci de brièveté, le Manuel ne reprend pas l'ensemble des moyens d'évaluer la lecture. Il convient également de noter que ce n'est pas un guide pouvant être utilisé sans modifications. En effet, chaque évaluation dans un nouveau pays requiert la mise en place d'un vocabulaire ainsi que d'énoncés et de textes adaptés au contexte local. Les personnes qui cherchent des conseils spécifiques sur la planification et la mise en œuvre d'EGRA peuvent consulter Guidance Notes for Planning and Implementing EGRA ([Notes d'orientation pour la planification et la mise en œuvre EGRA], RTI International & International Rescue Committee, 2011). A la suite de ce chapitre introductif, le chapitre 2 est centré sur la question de la protection des sujets humains dans la recherche. Le chapitre 3 présente une vue d'ensemble des objectifs et de l'utilisation d'EGRA. Le chapitre 4 aborde les cadres de référence issus de la recherche (les fondements théoriques de l'évaluation). Le chapitre 5 discute des options pour la conception de l'étude. Le chapitre 6 décrit les étapes préparatoires à l'administration de cette évaluation, y compris pour la construction d'un outil adapté à chaque contexte. Le chapitre 7 est une vue d'ensemble sur la collecte électronique des données. Le chapitre 8 fournit des informations sur les procédures à suivre pour la formation des évaluateurs. Le chapitre 9 donne des conseils sur la collecte de données pour les études pilotes et les autres études. Le chapitre 10 aborde les protocoles appropriés pour le nettoyage et la préparation des données. Le chapitre 11 présente une vue d'ensemble des analyses à effectuer. Enfin, le chapitre 12 fournit des indications sur différents points : l'interprétation des résultats, l'établissement de points de repère, les implications pour les politiques liées à l'amélioration de l'instruction, et la façon de présenter les résultats aux écoles. Une série d'annexes développe certains des points présentés dans le texte avec des exemples, des détails techniques et des conseils statistiques.

2 ÉTHIQUE DE RECHERCHE ET RÉVISION OBLIGATOIRE PAR UN CONSEIL D'EXAMEN INSTITUTIONNEL (CEI)

Les organismes de recherche bénéficiant d'un financement fédéral sont tenus de se conformer aux réglementations fédérales régissant la menée de recherches éthiques et au Principe fondamental de la statistique officielle de l'ONU qui stipule que « Les organismes de recherche bénéficiant d'un financement fédéral sont tenus de se conformer aux réglementations fédérales régissant la menée de recherches éthiques et au Principe fondamental de la statistique officielle de l'ONU qui stipule que « Les données individuelles collectées pour l'établissement des statistiques par les organismes qui en ont la responsabilité, qu'elles concernent des personnes physiques ou des personnes morales, doivent être strictement confidentielles et ne doivent être utilisées qu'à des fins statistiques »[5] Toutes les organisations américaines menant des recherches portant sur des sujets humains sont tenues de consulter un Conseil d'examen institutionnel avant de procéder à une étude. Conseils d'examen institutionnels et de protections des sujets humains et des réglementations américaines portant sur la protection de sujets humains ont été mises en place en 1974.

Toutes les organisations américaines menant des recherches portant sur des sujets humains sont tenues de consulter un Conseil d'examen institutionnel avant de procéder à une étude

2.1 Qu'est-ce qu'un CEI ?

Les CEI ont recours aux principes de base définis dans le « rapport Belmont », établi aux États-Unis par la Commission nationale pour la protection des sujets humains dans le cadre de la recherche biomédicale et béhavioriste (1978), dans le souci de les éclairer dans leur révision de protocoles de recherche proposés. Le rapport Belmont énonce trois principes fondamentaux :

[5] La politique fédérale pour la protection de sujets humains est requise par le Code américain des réglementations fédérales, 22 CFR, partie 225.

- **Respect de la personne.** Les sujets pressentis doivent être traités comme des agents autonomes capables d'envisager des alternatives, de faire des choix et d'agir sans pression ou interférences indues d'autres.
- **Bienfaisance.** Les deux principes de base de la bienfaisance sont : (1) ne pas faire de tort et (2) protéger des nuisances en maximisant les avantages et minimisant les dommages possibles.
- **Justice**. Ce principe éthique repose sur l'équité dans la distribution des fardeaux et des avantages de la recherche.

Des recommandations additionnelles pour l'évaluation des sujets humains ont été établies en 1981 par la Food and Drug Administration des États-Unis et le Ministère de la santé et des services sociaux des États-Unis. Ces deux agences ont reposé leurs déterminations sur les critères suivants :

1. Le protocole doit être évalué pour déterminer s'il est scientifiquement fiable et utile
2. Les risques doivent être minimisés dans la mesure du possible
3. Les sujets doivent être sélectionnés de manière équitable
4. Un consentement éclairé est requis
5. La vie privée et la confidentialité doivent être protégées
6. L'étude doit être adéquatement supervisée

On définit la recherche comme étant « une investigation systématique, notamment développement de la recherche, mise à l'essai et évaluation, conçue pour développer ou contribuer à des connaissances généralisables ».

–Code américain des réglementations fédérales, 22 CFR 225

2.2 En quoi l'approbation d'un CEI s'applique-t-elle aux études EGRA ?

Comme nous l'avons vu plus haut, toutes les organisations bénéficiant d'un soutien financier par le biais de fonds du gouvernement fédéral des États-Unis ou qui sont par ailleurs sujettes aux réglementations d'un organisme ou d'un service fédéral et qui mènent des recherches impliquant des sujets humains sont tenues de consulter un CEI et de recevoir l'approbation d'un CEI avant de procéder à des recherches.

et de déterminer le degré de risque auquel les sujets peuvent être exposés suite à leur participation à la recherche. Les activités de recherche sont approuvées ou refusées en conséquence par le CEI après examen approfondi des protocoles de recherche et des circonstances dans lesquelles la recherche est menée.

Les recherches comportant des tests de connaissances sont souvent dispensées des obligations d'un CEI (selon le principe que les tests administrés ne diffèrent pas énormément de ce que à quoi les enfants sont exposés dans leur environnement scolaire nature). Seul un CEI est néanmoins à même de décider du statut d'exemption d'une étude EGRA. Si un CEI décide d'accorder une exemption à une étude dans son ensemble, certaines questions de l'enquête peuvent toujours faire l'objet d'une approbation préalable, notamment si les renseignements recueillis dans le cadre du sondage risquent d'exposer les élèves ou les enseignants à un danger quelconque.

Dans le cas d'évaluations EGRA et d'études similaires menées auprès de jeunes enfants, chaque pays autorisant une étude EGRA doit également pouvoir permettre à son propre organisme d'éthique d'examiner les conditions de l'étude et d'accorder son approbation avant d'entamer celle-ci ou de demander l'apport de toutes modifications nécessaires avant d'accorder cette approbation (22 CFR 225).

2.3 Assentiment et consentement éclairé individuel des participants

Le modèle d'étude EGRA et ses instruments de soutien commencent toujours par une section expliquant aux évaluateurs comment demander aux participants retenus leur consentement (pour les adultes) ou leur assentiment (pour les enfants). Avant d'administrer les tests EGRA aux enfants, les évaluateurs décrivent les objectifs de l'étude et informant les élèves que l'évaluation est anonyme, n'aura pas d'incidence sur leurs résultats scolaires et servira à apporter des améliorations aux méthodes d'apprentissage de la lecture employées dans leur pays. Chaque enfant peut verbalement accepter d'être évalué ou refuser de participer sans conséquences d'aucune sorte. Si des sondages sont menés auprès de directeurs d'école ou d'enseignants dans le cadre de l'étude, un processus de consentement similaire—écrit plutôt que verbal—est obtenu.

Bien que ce processus d'assentiment / consentement puisse ne pas être connu des homologues des pays hôtes, il est souvent bien accepté par les élèves et les enseignants qui reconnaissent avoir un rôle dans la décision de participer. Dans plusieurs mises en œuvre d'études EGRA, l'expérience montre à ce jour que peu d'élèves et d'enseignants refusent de participer. Si un admissible participant refuse de participer, une autre personne interrogée est choisie au hasard. Pour en savoir plus sur les CEI et la recherche éthique auprès de sujets humains, y compris les enfants, veuillez consulter le site Web du Ministère de la santé et des services sociaux des États-Unis, http://www.hhs.gov/ohrp.

3 OBJECTIF ET UTILISATIONS D'EGRA

3.1 Historique et aperçu

Bien que dès le départ il était clair qu'EGRA se concentrerait sur les premières années primaires et sur les compétences fondamentales en lecture, l'utilisation des résultats était encore à débattre.

L'instrument EGRA original était principalement conçu pour constituer une analyse « diagnostique du système ». Son principal objectif était de documenter la performance des élèves en ce qui concerne les compétences fondamentales en lecture afin d'informer les gouvernements et les donateurs sur les besoins systémiques en matière d'amélioration de l'instruction. Ses utilisations englobent maintenant toutes les suivantes, différents emplois s'appliquant dans différents contextes :

- Récolter des données de référence sur l'acquisition de compétences de base en lecture dans des classes et / ou des géographies particulières
- Éclairer la mise au point de programmes d'enseignement en déterminant des compétences ou des domaines d'enseignement clés ayant besoin d'être améliorés
- Déterminer dans le temps les changements dans les niveaux de lecture
- Évaluer les résultats ou l'impact de programmes conçus pour améliorer la lecture dans le primaire
- Explorer la rentabilité de différents modèles de programmes
- Mettre au point des indicateurs et références en matière de lecture
- Servir de diagnostic de système (voir Section 3.2) pour informer la politique, la planification stratégique et l'allocation des ressources du secteur de l'éducation

De plus, « les sous-tâches incluses dans EGRA peuvent être adaptées pour permettre aux enseignants d'informer leur instruction.[6] Comme évaluation formative, les enseignants peuvent soit employer EGRA dans son intégralité soit sélectionner des tâches pour surveiller le progrès de la classe, déterminer les tendances dans les résultats et adapter l'instruction pour répondre aux besoins en éducation des enfants » (Dubeck & Gove, 2015, p. 2).

[6] L'emploi de l'EGRA comme évaluation formative en salle de classe ne peut se faire qu'après l'apport obligatoire de modifications particulières à l'instrument et aux procédures d'échantillonnage. Des évaluations en salle de classe exigeraient également le perfectionnement professionnel des enseignants comportant notamment des instructions particulières sur l'administration et l'interprétation des tâches.

Cependant, pour être clair, tel qu'actuellement conçu, EGRA a des limitations. Il n'a pas pour but de constituer une mesure de responsabilisation d'envergure pour determiner le passage en classe supérieure des élèves ou pour évaluer individuellement les enseignants. EGRA est conçu pour compléter, et non pas remplacer, les évaluations classiques sur papier existantes fondées sur les attentes du programme scolaire.

EGRA est constitué d'un ensemble de tâches qui mesurent les compétences fondamentales et qui se sont avérées prédictives d'une réussite ultérieure en lecture. Cependant, du fait des contraintes imposées par l'énergie et la durée d'attention limitées des enfants, ni EGRA ni aucun autre instrument n'est capable à lui seul de mesurer toutes les compétences requises pour permettre aux élèves de lire avec compréhension. EGRA n'a pas pour but de constituer un programme d'éducation ; il permet plutôt d'informer des programmes d'éducation. EGRA ne peut pas entièrement déterminer des antécédents comportementaux ou des attitudes en matière d'alphabétisation qui pourraient avoir une incidence sur la capacité de lecture d'un élève (Dubeck & Gove, 2015). De plus, les mesures d'EGRA se limitant aux compétences sujettes à une influence de l'instruction, les résultats seront réalisables.

3.2 EGRA comme diagnostic systémique

Le diagnostic systémique d'EGRA, tel qu'il est présenté dans ce manuel, est conçu pour faire partie d'un cycle complet d'appui et d'amélioration de l'apprentissage. Comme il est décrit dans la **Figure 6**, EGRA peut être utilisé dans le cadre d'une méthode exhaustive d'amélioration des compétences en lecture des élèves, la première étape étant une identification systémique et générale des domaines nécessitant une amélioration. EGRA est capable de produire des données de référence sur l'apprentissage de la lecture (Dubeck & Gove, 2015). Un étalonnage des performances générales et la création d'objectifs pour de futures applications (voir Section 12.2) peuvent également être réalisés durant l'application initiale d'EGRA. Selon les résultats d'EGRA, les ministères de l'éducation ou les systèmes d'éducation locaux peuvent alors intervenir pour éclairer le contenu de nouveaux programmes ou de programmes existants en mettant en œuvre des méthodes didactiques fondées sur des données factuelles, afin de soutenir les enseignants dans l'amélioration des compétences fondamentales en lecture. Les résultats provenant d'EGRA peuvent donc informer l'élaboration de programmes de formation de futurs enseignants et de perfectionnement d'enseignants déjà en service.

Une fois ces recommandations mises en place, des formulaires parallèles d'EGRA peuvent être utilisés pour peu à peu suivre les progrès et les bénéfices de l'apprentissage chez les élèves grâce à un suivi continu, tout en s'attendant à ce qu'un tel processus promeuve la responsabilité de l'enseignant et de l'administrateur en éducation d'assurer les progrès des élèves en matière de compétences fondamentales.

Figure 6. Le cycle continu d'amélioration de l'apprentissage des élèves

EGRA et les évaluations basées sur EGRA peuvent permettre d'identifier les besoins d'intervenir et de suivre les progrès afin d'améliorer les résultats d'apprentissage de l'élève.

Lorsqu'ils travaillent au niveau systémique, les chercheurs et les administrateurs en éducation commencent généralement par une analyse des données récoltées sur les élèves, sur base d'exemples, afin de tirer des conclusions sur le fonctionnement du système (ou des élèves au sein du système). Cette démarche s'appuie sur le fait qu'il est entendu que la façon dont les élèves apprennent reflète directement l'instruction qu'ils reçoivent. En utilisant la performance moyenne des élèves, par année primaire au sein du système, les administrateurs peuvent évaluer à quel niveau les élèves éprouvent généralement des difficultés et peuvent utiliser ces informations pour développer des méthodes didactiques adéquates. Comme toute évaluation dont le but est de diagnostiquer des difficultés et d'améliorer les performances d'apprentissage, les éléments suivants sont nécessaires pour que l'analyse soit utile : (1) l'évaluation doit être liée aux attentes et aux points de référence existants, (2) elle doit être en corrélation avec les compétences ultérieurement désirées et (3) il doit être possible de modifier ou d'améliorer les compétences grâce à une instruction supplémentaire (Linan-Thompson & Vaughn, 2007). EGRA répond comme suit à ces exigences suivantes.

Premièrement, dans de nombreux pays à revenu élevé, les enseignants (et les administrateurs en éducation) peuvent observer les distributions nationales et les normes de performance existantes afin de comprendre comment la performance de leurs élèves se compare à celle d'autres élèves.

En comparant la performance des sous-groupes d'élèves aux distributions nationales et aux normes de performance, les administrateurs des systèmes d'éducation américains et européens peuvent décider si les écoles et les enseignants ont besoin de soutien supplémentaire. EGRA peut également être utilisé par les pays à faibles revenus pour déterminer quelles régions (ou, si l'échantillonnage le permet, quelles écoles) ont besoin d'un soutien supplémentaire, notamment formation des enseignants ou autres interventions. Quand EGRA a été mis au point pour la première fois, le problème pour les pays à faibles revenus était que de tels étalonnages de performance, basés sur des résultats générés localement, n'étaient pas (encore) disponibles. Entre temps, des travaux ont été entrepris dans au moins 12 pays pour établir des références nationales ou régionales à l'aide de données EGRA. Ceci est traité en détail à la Section 12.2.

De plus, les tests d'EGRA ont été mis au point à dessein pour être prédictifs en matière de performances en lecture ultérieures et de nombreuses administrations d'EGRA dans plusieurs pays et plusieurs langues ont confirmé les corrélations attendues. Bien que les variations phonologiques et orthographiques d'une langue à une autre influencent le taux et la période d'acquisition de la lecture, toutes les compétences mesurées par EGRA se sont avérées être en corrélation avec les compétences en lecture dans des orthographes alphabétiques. Par exemple, connaître le rapport entre les sons et les symboles qui les représentent est en rapport prédictif avec la réussite dans la lecture de mots. Il a été montré que la fluence en lecture à haute voix était prédictive de la compréhension en lecture. Ces compétences sont mesurées dans EGRA et nous pouvons donc affirmer sans trop risquer de nous tromper que les résultats d'EGRA font état d'une situation représentative de la direction suivie par les enfants dans le processus d'acquisition de la lecture.

Troisièmement, EGRA peut non seulement nous apporter des prédictions représentatives de performances ultérieures, mais aussi attirer notre attention sur les changements didactiques nécessaire. Il n'est pas très logique de mesurer quelque chose qu'il n'y a aucun espoir de pouvoir changer par le biais d'ajustements de l'enseignement. EGRA est utile comme outil diagnostique précisément parce qu'il comprend des analyses de compétences qui peuvent être améliorées.

4 CADRE CONCEPTUEL ET BASES DES RECHERCHES

4.1 Compétences nécessaires pour comprendre un texte écrit

Le cadre conceptuel à la base d'EGRA est issu de nombreux travaux de recherche qui ont été synthétisés dans plusieurs publications, entre autres, celles du National Reading Panel ([NRP], National Institute of Child Health and Human Development, 2000) et du National Early Literacy Panel ([NELP], 2008) ; voir aussi Abadzi (2006), August & Shanahan (2006). Ces synthèses, et d'autres travaux, ont été intégrés dans l'Expertise collective de l'Institut National de la Santé et de la Recherche Médicale (INSERM, 2007), ainsi que dans Dehaene [Ed] et al. (2011) et Sprenger-Charolles & Colé (2013).

Ces synthèses mettent en exergue le fait que la compréhension de l'écrit, qui est la finalité de la lecture, dépend du niveau de compréhension orale de celui qui lit et de sa maitrise de mécanismes spécifiques à la lecture. L'exemple de la lecture d'une partition de musique peut permettre de comprendre ce que sont ces mécanismes. En effet, l'incapacité de lire une partition de musique est généralement due à la non-maitrise des mécanismes qui permettent au musicien expert d'associer automatiquement dans sa tête une suite de notes écrites à un bout de mélodie, et non à des difficultés de compréhension de la musique. Il en va de même pour la lecture. Les enfants qui lisent dans une langue qu'ils maitrisent à l'oral ne peuvent comprendre ce qu'ils lisent que s'ils ont automatisé les mécanismes qui permettent de décoder les mots écrits (pour l'anglais, Adolf, Catts & Lee, 2010 ; Hoover & Gough, 1990 ; Spencer, Quinn & Wagner, 2015 ; pour le français, Gentaz, Sprenger-Charolles & Theurel, 2015). Plus généralement, la compréhension écrite dépend d'une compétence de haut niveau (la compréhension orale, incluant la maitrise du vocabulaire (Adolf, Perfetti & Catts, 2011 ; Perfetti, 2007 ; Tunmer & Chapman, 2012) et d'une compétence de bas niveau, le décodage. Dans une écriture alphabétique, le décodage implique la maitrise des relations graphème-phonème, qui se mesure par la précision et la rapidité (c'est-à-dire par la fluence) en lecture de mots réguliers (comme table) ou de mots inventés (comme mapre), le décodage dépendant du niveau de conscience phonémique de l'enfant. Ce sont ces compétences qui sont évaluées dans EGRA parce qu'elles sont non seulement les meilleurs prédicteurs du futur niveau de lecture (voir ci-dessous 4.3.1 et 4.4.1) mais également les composantes les plus efficaces de l'enseignement de la lecture. Cela a été largement souligné dans les synthèses du NRP (2000 ; voir aussi Ehri, Nunes, Stahl & Willows, 2001a-b) et du NELP (2008) :

a. **Conscience phonémique.** Aider les lecteurs débutants à développer cette conscience avec, pour objectif, de faciliter la maitrise des CGP ;

b. **Correspondances graphème-phonème** (CGP). Aider les lecteurs débutants à bien comprendre et à bien maitriser les CGP ;

c. **Vocabulaire.** Aider les lecteurs débutants à développer leur vocabulaire ;

d. **Fluence.** Aider les lecteurs débutants à être précis et rapides en lecture ;

e. **Compréhension**. Aider les lecteurs débutants à développer leurs capacités de compréhension, à l'oral comme à l'écrit.

Ces différents points, et leurs implications pour EGRA, sont abordés plus loin dans ce chapitre : le point a dans la section 4.2 (Conscience phonémique), et le point b dans la section 4.3 (Connaissances alphabétiques et procédures de l'identification des mots écrits). Le point c est intégré dans la sections 4.4 (Vocabulaire et compréhension orale), le point d dans la section 4.5 (La fluence), et le point e dans la section 4.6 (Compréhension écrite).

4.2 Conscience phonémique

4.2.1 Introduction

La *conscience phonémique* est la capacité d'identifier et de manipuler les plus petites unités sans signification de la langue orale, le phonème. Elle est une composante de la conscience phonologique, qui est la capacité d'identifier et de manipuler les unités sans signification de la langue orale, quelle que soit leur taille : de la syllabe, et ses composants (attaque et rime), au phonème.

Le rôle de la conscience phonémique dans l'apprentissage de la lecture a fait l'objet de débats intenses autour des années 80. Ces débats ont opposé des chercheurs qui soutenaient que le développement de cette conscience est une conséquence de l'apprentissage de la lecture dans une écriture alphabétique, à d'autres chercheurs qui défendaient la position inverse : à savoir qu'elle est un prérequis de cet apprentissage (d'un côté, Morais, Cary, Alegria & Bertelson, 1979, de l'autre, Bradley & Bryant, 1983). Les arguments des premiers viennent d'observations qui ont montré que des adultes illettrés ne pouvaient pas analyser des mots oraux en phonèmes. L'autre position a eu pour point de départ une étude d'entrainement qui a utilisé des tâches de *chasse à l'intrus* au niveau phonémique (quel est, parmi trois mots, celui qui n'a pas le même son au début : *balle, bulle, colle*) ou sémantique (quel est le mot qui n'est pas de la même famille : *bol, tasse, bras*). Seuls les entrainements à base phonémique se sont avérés avoir une incidence positive sur l'apprentissage de la lecture.

Depuis ces publications, de nombreuses études longitudinales, dans lesquelles les enfants sont suivis le plus souvent depuis une période qui précède l'apprentissage de la lecture, ont montré que les relations entre conscience phonémique et apprentissage de la lecture sont bidirectionnelles. D'un côté, l'apprentissage de la lecture favorise le développement de la conscience phonémique ; de l'autre, le niveau de conscience phonémique avant cet apprentissage est un bon prédicteur du succès ou de l'échec de cet apprentissage (Perfetti, Beck, Bell & Hughes, 1987

; en français, voir Casalis & Louis-Alexandre, 2000). Il a aussi été montré que les entrainements à la conscience phonémique ont un effet positif sur l'apprentissage de la lecture, à condition toutefois d'utiliser en même temps un support écrit (les graphèmes correspondant aux phonèmes manipulés, cf. Ehri et al., 2001b ; Bara, Gentaz, Colé & Sprenger-Charolles, 2004). Ces résultats ont été relevés chez des apprenants de langue seconde, comme chez ceux de langue maternelle.
Une forte incidence de la conscience phonémique sur la lecture a été observée dans différentes écritures alphabétiques, cette incidence étant toutefois plus marquée dans celles qui ont une orthographe opaque (Ziegler, Bertrand, Toth et al., 2010). Enfin, quelques études signalent que les capacités de discrimination phonémique (être capable de distinguer vol de bol, par exemple) permettent de prédire le devenir en lecture des enfants de façon fiable et de distinguer les faibles lecteurs des bons lecteurs (Ziegler, Pech-Georgel, George & Lorenzi, 2009). Il est donc crucial d'évaluer ces différentes capacités dans un bilan des premières étapes de l'apprentissage de la lecture.

4.2.2 Mesures utilisées dans EGRA

EGRA contient typiquement en option une des trois tâches de conscience phonémique suivantes (cf. chapitre 6 pour des exemples). La première est une épreuve de **discrimination de son initial** d'un mot. Dans ce type d'épreuve, on demande à l'enfant de choisir le mot qui, parmi trois, ne commence pas par le même son initial (chasse à l'intrus). Une autre épreuve requiert **l'identification du son initial** d'un mot. On demande à l'enfant soit de prononcer le premier son d'un mot qu'on lui présente à l'oral (/k/ pour car). L'épreuve de **segmentation phonémique (ou syllabique)**, dans laquelle on demande de prononcer les différents sons d'un mot, s'est révélée difficile (effet plancher, Linan-Thompson & Vaughn, 2007) et n'a pas été retenue dans la batterie de base. Elle peut toutefois être une alternative appropriée lorsque l'épreuve d'identification du son initial d'un mot donne lieu à des effets plafond.

4.3 Connaissances alphabétiques et procédures d'identification des mots

4.3.1 Introduction

Pour pouvoir lire dans une écriture alphabétique, il faut maitriser les correspondances graphème-phonème (CGP). Les recherches ont montré, d'une part, que ces connaissances sont, avec le niveau de conscience phonémique, un des meilleurs prédicteurs précoces des résultats ultérieurs en lecture et, d'autre part, qu'un enseignement basé sur le décodage est celui qui est le plus bénéfique (Adams, 1990 ; Ehri et al., 2001a ; Wagner, Torgesen & Rashotte, 1994 ; Yesil-Dagli, 2011). Ces résultats ont été relevés chez des apprenants de langue maternelle et seconde (August et Shanahan, 2006).

La connaissance du nom des lettres est également un prédicteur du futur niveau de

lecture (Foulin, 2005), au moins en anglais. Cependant, pour EGRA, l'identification des sons plutôt que des noms des lettres est souvent l'indice plus utile et plus utilisé, surtout dans les orthographes alphabétiques transparentes.

D'autres résultats signalent que, en français, les enfants prennent rapidement les graphèmes comme unité de base de l'écrit (Kandel, Soler, Valdois & Gros, 2006 ; Rey, Ziegler & Jacobs, 2000 ; Sprenger-Charolles et al., 2005). Cela peut s'expliquer par le fait que les graphèmes de plus d'une lettre sont fréquents dans cette langue : ils sont utilisés pour différencier ou de u (/u/, /y/) ou pour noter les voyelles nasales (an /â/, in /î/, o /ô/ et un /û/). La connaissance des graphèmes est donc évaluée dans la version française d'EGRA. Cette connaissance peut être considérée comme un premier indicateur du niveau de conscience phonémique de l'enfant.

D'après les modèles à double voie (Coltheart, Rastle, Perry, Langdon & Ziegler, 2001 ; Ziegler, Perry & Coltheart, 2003 ; Ziegler, Perry & Zorzi, 2014b), l'association entre mot écrit et oral s'effectue par une procédure *lexicale* (ou *orthographique*) ou par une *procédure phonologique sublexicale*, le *décodage* (ou *procédure alphabétique*). Dans une écriture alphabétique, le décodage implique la mise en relation des unités de base de la langue écrite (les *graphèmes*, comme *b, a, u, ou*…) avec les plus petites unités de la langue orale qui leur correspondent (les *phonèmes*[7], comme /b/, /a/, /y/, /u/…). Cette mise en relation est suivie par l'assemblage des unités résultant du décodage (b+a=/ba/, d+u=/dy/, f+ou=/fu/…).

De nombreuses études ont montré que les lecteurs débutants utilisent quasi-exclusivement le décodage (Backman, Bruck, Herbert & Seidenberg, 1984 ; Sprenger-Charolles, Siegel & Bonnet, 1998b). De plus, le niveau de décodage permet de distinguer les enfants en difficultés de lecture des autres enfants, quel que soit leur âge ou leur niveau scolaire (Rack, Snowling & Olson, 1992). Enfin l'apprentissage de la lecture dépend de la régularité des correspondances graphème-phonème dans la langue dans laquelle il s'effectue (voir l'encadré 1).

Les études sur l'apprentissage de la lecture ont également montré que les capacités initiales de décodage permettent de prédire le futur niveau de lecture, y compris à long terme (Juel, 1988). Le niveau de décodage, évalué par la lecture de pseudomots, prédit même le niveau de lecture de mots irréguliers comme sept ou femme. C'est ce qu'a montré une étude d'Ouellette et Beers (2010) qui a pris en compte des enfants de 1ère année du primaire. Des résultats identiques, sur un plus long terme, ont été relevés par Sprenger-Charolles, Siegel, Béchennec & Serniclaes (2003) dans une étude longitudinale au cours de laquelle les enfants ont été suivis pendant 4 ans, depuis le milieu de la 1ère année du primaire. Ces résultats s'expliquent si l'on admet que les enfants utilisent d'abord la procédure phonologique de lecture et que des connexions vont progressivement se créer entre unités orthographiques et phonologiques. L'établissement de ces connexions dépend de la régularité des CGP et de la fréquence des mots.

[7] Dans l'alphabet phonétique International [IPA], les phonèmes sont notés entre deux barres obliques (voir http://www.internationalphoneticassociation.org/content/ipa-chart, pour citation et utilisation de l'API ; Creative Commons Attribution-Sharealike 3.0 Unported License. Copyright © 2005 International Phonetic Association).

ENCADRE 1. DIFFERENCES ENTRE LANGUES : PHONOLOGIE ET ORTHOGRAPHE

Les langues varient dans la complexité de leur système phonologique. Ainsi, certaines ont beaucoup plus de phonèmes que d'autres : par exemple, 5 voyelles simples en espagnol et 10 à 16 en français (Delattre, 1965). D'autres langues ont des mots plus courts que d'autres en termes de nombre de syllabes, celles ayant beaucoup de mots courts ont également souvent des structures syllabiques complexes, avec, par exemple, des groupes de consonnes en position initiale et finale (comme en anglais).

De même, les langues varient dans la complexité de leur système orthographique, certaines ayant une orthographe plus transparente que d'autres, c'est-à-dire des correspondances graphème-phonème (CGP) régulières. La transparence des CGP facilite l'acquisition de la lecture : le nombre de CGP à apprendre est en effet moins élevé dans une orthographe transparente (comme en espagnol) que dans une orthographe opaque (comme en anglais), l'orthographe du français se situant entre ces deux extrêmes (Peereman, Sprenger-Charolles & Messaoud-Galusi, 2013).

Ainsi un enfant qui apprend à lire dans une langue qui a une orthographe transparente, un nombre relativement faible de phonèmes et des structures syllabiques simples, arrivera mieux et plus vite à maitriser les CGP que celui qui est confronté à une langue ayant une orthographe opaque, de nombreux phonèmes et des structures syllabiques complexes. C'est ce que de nombreuses comparaisons inter-langues ont montré (Seymour et al., 2003 ; Ziegler & Goswami, 2005 et 2006).

Reconnaitre les mots écrits automatiquement permet de libérer des capacités de mémoire pour la compréhension, qui dépend de notre mémoire à long terme (qui stocke durablement nos connaissances), et de notre mémoire de travail (qui stocke les informations de façon provisoire, le temps de les traiter), cette dernière ayant une capacité limitée (Baddeley, 2012). Chez le débutant, qui décode lentement les mots, une bonne partie de la charge de la mémoire de travail est consacrée au décodage. De plus, pour trouver les mots qu'il n'arrive pas à décoder, ce lecteur va s'appuyer sur le contexte (Stanovich, 1980). En conséquence, il ne lui reste que peu de ressources cognitives disponibles pour la compréhension. En revanche, celui qui sait lire peut récupérer les mots écrits automatiquement (Perfetti & Zhang, 1995), sans effort cognitif et sans avoir besoin d'aides contextuelles. Il peut donc consacrer ses ressources cognitives au processus de compréhension (Adolf et al., 2011 ; Arrington, Kulesz, Francis et al., 2014 ; Stanovich, 2000).

Ces résultats ont deux implications majeures pour l'évaluation de la lecture. Les faibles lecteurs arrivent parfois à compenser leurs difficultés de décodage en utilisant des anticipations contextuelles, les tests de lecture utilisant des textes permettent donc de moins bien les détecter que la lecture de mots isolés et, surtout, celle de mots nouveaux (des pseudomots). Autre implication : la nécessité de tenir compte de la précision et de la rapidité. Cela est d'autant plus important que c'est surtout

la rapidité qui permet de distinguer les bons lecteurs des faibles lecteurs chez les sujets les plus âgés ou encore chez ceux qui apprennent à lire dans une langue qui a une orthographe transparente. Un indicateur de rapidité utilisé depuis longtemps en psychologie et en sciences de l'éducation, relativement fiable et facile à mettre en œuvre, est le nombre d'items (mots ou pseudomots) correctement lus en un temps donné, généralement 1 minute (Hudson, Lane & Pullen, 2005). C'est cette mesure qui a été retenue dans EGRA pour les épreuves de lecture de mots isolés ou en contexte.

4.3.2 Mesures utilisées dans EGRA

Pour les raisons signalées au paragraphe précédent, la version française d'EGRA évalue la connaissance de l'alphabet des enfants par une épreuve **d'identification du son des lettres** du français. Pour les mêmes raisons, alors que dans les premières versions d'EGRA la connaissance de l'alphabet était évaluée par une épreuve **d'identification du nom des lettres**, au fil du temps, l'épreuve connaissance du son des lettres est devenue l'option la plus fréquente, cette connaissance étant directement liée à la capacité de décodage.

Les enfants doivent dire comment, dans un mot, différentes lettres (ou graphèmes de plus d'une lettre) se prononcent. Ces items sont présentés dans une liste, dans un ordre aléatoire, en majuscule ou en minuscule. Pour éviter les problèmes de codage, il faut sélectionner des items fréquents qui se prononcent toujours (ou presque toujours) de la même façon (par exemple, en français, les voyelles *a, é, i, o, u, ou, on, an*, et les consonnes *b, d, f, j, l, m, n, p, r, t, v, z*).

Deux tâches de lecture ont été retenues pour EGRA : la lecture de mots familiers et de mots inventés (voir le chapitre 6 pour des exemples). La **lecture de mots inventés** (des pseudomots qui peuvent se lire sans ambigüité en utilisant les correspondances graphème-phonème [CGP]) donne un aperçu de la capacité des enfants à décoder des mots inconnus et donc à utiliser la procédure phonologique de lecture. Les enfants sont invités à lire à haute voix une liste de pseudomots, aussi précisément et rapidement qu'ils le peuvent. La tâche de **lecture de mots familiers** est similaire à celle de mots inventés, sauf que les items sont des mots que les enfants sont supposés connaitre. La lecture de ce type d'items, particulièrement ceux qui sont fréquents et non réguliers sur le plan des CGP (comme sept, femme ou automne), donne un aperçu de la capacité des enfants à reconnaitre des mots en utilisant la procédure lexicale (ou orthographique) de lecture.

Enfin, avec **la dictée**, les élèves doivent écouter les sons des lettres (ou des suites de lettres), soit isolées, soit dans des mots ou des phrases courtes, et les écrire. Cette épreuve mesure leurs connaissances des correspondances phonème-graphème (CPG). Lorsqu'il s'agit d'une phrase, l'épreuve mesure aussi leur maitrise des conventions de l'écrit, telles que l'usage des lettres majuscules et la ponctuation. Cette épreuve peut être difficile à noter d'une manière normalisée dans certains contextes ; elle ne fait plus partie de l'instrument de base, mais elle a été utilisée dans certains pays, qui l'ont trouvée néanmoins utile.

4.4 Vocabulaire et compréhension orale

4.4.1 Introduction

Comme expliqué au début de ce chapitre, pour comprendre ce qu'on lit, il faut non seulement savoir décoder les mots, il faut aussi les comprendre (Perfetti, 2007 ; Rayner et al., 2001 ; Tunmer & Chapman, 2012). Cette compréhension est dite réceptive quand elle désigne la capacité de comprendre le sens des mots entendus ou lus ; elle est dite productive quand elle désigne la capacité d'utiliser ces mots quand nous parlons ou écrivons. Selon certains chercheurs, il faut connaitre 90 à 95 % des mots d'un texte pour le comprendre (Nagy & Scott, 2000). Il n'est donc pas étonnant que le vocabulaire (le plus souvent testé sur le volant réceptif) soit un facteur prédictif de la compréhension écrite (Muter et al., 2004 ; Tunmer & Chapman, 2012 ; pour le français, voir Gaonac'h & Fayol [Eds], 2003 ; Gentaz et al., 2015).

La compréhension de l'écrit nécessite aussi la capacité de comprendre les relations entre les mots dans des énoncés. Cette capacité est requise quelle que soit la taille de ces énoncés (de la phrase au texte) ou leur nature (par exemple faire les mouvements que le professeur de gymnastique demande d'effectuer ou répondre à des questions après avoir entendu une histoire…). Le niveau de compréhension orale d'énoncés, comme celui de vocabulaire, est un facteur prédictif du niveau de compréhension écrite en anglais (Muter et al., 2004 ; Tunmer & Chapman, 2012), comme en français (Gentaz et al., 2015).

4.4.2 Mesures utilisées dans EGRA

Pour évaluer le **vocabulaire**, il est difficile, dans un protocole EGRA, de recourir à des tests standardisés dans lesquels l'enfant doit, la plupart du temps, désigner l'image correcte (parmi plusieurs) qui correspond aux mots que l'expérimentateur lui présente successivement. En effet, ces tests ne sont, le plus souvent, pas adaptés au contexte des pays en voie de développement.

L'évaluation de la **compréhension orale** est une des épreuves centrales d'EGRA, qui évalue aussi indirectement le vocabulaire oral (celui utilisé dans l'épreuve). Dans cette épreuve, les évaluateurs lisent à l'enfant une histoire courte sur un sujet familier et lui posent ensuite trois à cinq questions sur cette histoire. Cette épreuve est similaire à celle utilisée pour évaluer la compréhension en lecture (voir 4.6 ci-dessous) afin de permettre d'estimer au mieux l'origine des difficultés de compréhension.

4.5 La fluence

4.5.1 Introduction

La fluence est « la capacité de lire un texte rapidement, précisément et de façon expressive » (NICHD, 2000, pp. 3–5). D'après Snow et le RAND Reading Study Group (2002) :

> La fluence peut être conceptualisée comme étant à la fois un prérequis et une conséquence de la compréhension. Certains aspects de la fluence, par exemple la lecture expressive d'un texte, peuvent être le signe d'une compréhension approfondie. Cependant, d'autres aspects de la fluence, tels que la reconnaissance précise et rapide des mots ainsi que, au moins en partie, certains aspects des traitements syntaxiques […] sont des prérequis pour la compréhension. (p.13)

La fluence peut être considérée comme un pont entre reconnaissance des mots et compréhension de l'écrit (Hudson et al., 2005). Comme expliqué au début de ce chapitre, pour comprendre ce qu'on lit, il faut pouvoir reconnaitre les mots écrits de façon automatique. En effet, les lecteurs compétents, qui sont en mesure de récupérer les mots écrits automatiquement, sans effort cognitif et sans avoir besoin d'aides contextuelles, peuvent consacrer leurs ressources cognitives au processus de compréhension (Perfetti, 1995 ; Stanovich, 1980 ; Stanovich, 2000). Si le lecteur décode les mots trop lentement, il n'y aura pas assez d'espace dans sa mémoire de travail pour traiter un énoncé, même court, et il risque fort d'avoir oublié son début quand il arrive à sa fin. Celui qui ne peut pas tenir un énoncé dans sa mémoire de travail, n'est pas en mesure de le comprendre (Arrington et al., 2014).

Les recherches ont montré que la compréhension en lecture est corrélée à la fluence, en particulier dans les débuts de l'apprentissage (Fuchs, Fuchs, Hosp & Jenkins, 2001). Par exemple, la fluence mesurée par le nombre de mots corrects lus par minute, est fortement corrélée à l'épreuve de compréhension écrite du Stanford (0,91, Fuchs et al., 2001). Les données provenant de nombreuses administrations d'EGRA dans différents contextes et différentes langues ont confirmé ce résultat (par exemple, Boulat et al, 2014).

L'importance de la fluence comme prédicteur de la compréhension décline toutefois avec l'âge (Yovanoff, Duesbery, Alonzo & Tindall, 2005) et le niveau de décodage. Par exemple, dans une étude de Gentaz et collaborateurs (2015), les enfants ont été séparés en trois groupes en fonction de leur niveau de décodage évalué en fin de 1ère année du primaire par la fluence en lecture de pseudomots : des décodeurs bons, moyens ou faibles. Le meilleur prédicteur de la compréhension écrite est, chez les bons décodeurs, la compréhension orale et, chez les décodeurs faibles et moyens, la fluence du décodage (voir aussi Stanovich, 1980).

La fluence du décodage et de la reconnaissance des mots écrits varie aussi en fonction de la longueur des mots. Une langue avec des mots courts (l'anglais), permet aux élèves de lire plus de mots par minute qu'une langue ayant des mots plus longs (le finnois). Il faut donc, dans les comparaisons utilisant la fluence, vérifier que les items ont la même longueur (nombre de lettres, de graphèmes et de syllabe).

4.5.2 Mesures utilisées dans EGRA

Etant donné l'importance de la fluence pour la compréhension, il est tenu compte de cette mesure dans trois des épreuves principales d'EGRA : en plus des épreuves de lecture à haute voix de mots familiers et de mots inventés, qui sont des épreuves en temps limité (1 minute), celle de **lecture à haute voix d'un passage court**, sur un sujet familier, que l'enfant est invité à lire en étant rapide et précis. Cette épreuve est suivie par un questionnaire évaluant la compréhension.

Les enfants devenant de plus en plus fluents au cours de l'apprentissage de la lecture, les évaluations chronométrées permettent de suivre leurs progrès à travers différentes mesures. Ces évaluations permettent ainsi de les situer sur le chemin de la lecture experte.

4.6 Compréhension écrite

4.6.1 Introduction

Comprendre ce qu'on lit est le but de la lecture. Cette capacité permet aux élèves de donner un sens à ce qu'ils lisent et d'utiliser ce sens non seulement pour le plaisir de lire, mais aussi pour apprendre des choses nouvelles, en particulier de nouvelles connaissances académiques. La compréhension de l'écrit est cependant une capacité complexe. Elle repose sur une interaction réussie entre différents facteurs : d'un côté, la motivation, l'attention, la mémoire, les connaissances linguistiques et les autres connaissances nécessaires pour comprendre la thématique d'un texte ; de l'autre, un accès précis et rapide aux mots écrits. Il n'est donc pas facile de mesurer la compréhension de l'écrit (Snow et le RAND Reading Study Group, 2002 ; Gaonac'h & Fayol [Eds], 2003).

4.6.2 Mesures utilisées dans EGRA

La mesure de la **compréhension écrite** choisie en premier pour EGRA est de même nature que celle utilisée pour évaluer la compréhension orale. Après avoir lu à haute voix un court récit, l'enfant doit répondre à des questions (3 à 5). Certaines questions portent sur des informations présentes explicitement dans le texte, d'autres nécessitent de faire des **déductions** et il est aussi possible de poser des questions de vocabulaire (voir le chapitre 6 pour des exemples). Pour répondre, l'enfant peut relire le texte, ce qui réduit la charge de la mémoire. Cette possibilité n'a toutefois été que rarement prise en compte dans EGRA.

D'autres options, telles que le rappel d'une histoire peuvent être utilisées, mais les réponses ne sont pas faciles à coder. La compréhension peut aussi être évaluée par un exercice à trou (test de closure) dans lequel l'enfant doit identifier le mot qui, parmi plusieurs, peut compléter un énoncé. Cette possibilité n'a cependant été utilisée que rarement dans les évaluations EGRA. Cette option, et d'autres, sont à l'étude.

EGRA a été conçu pour être utilisé à grande échelle, avec des données normalisées. La conception des tests de compréhension reflète le fait que la recherche n'a pas encore produit un moyen éprouvé permettant d'évaluer la compréhension en lecture avec des épreuves standardisées qui pourraient être acceptées par tous comme étant valides et fiables.

5 CONCEPTION D'UNE ETUDE EGRA

5.1 Conception d'une étude EGRA : considérations

Ce chapitre décrit la conception et les principes qui guident le développement des recherches et des évaluations reposant sur l'utilisation d'EGRA. Comme pour toute étude scientifique, la conception d'une étude EGRA est principalement guidée par l'objectif de l'évaluation, ainsi que par les questions de recherche qui en émanent. Sa conception nécessite la collaboration des bailleurs de fond et du gouvernement pour s'assurer que l'étude est faisable et les objectifs appropriés.

ELEMENTS ESSENTIELS A LA SELECTION D'UN ECHANTILLON POUR TOUT PROJET DE RECHERCHE

La conception d'une étude EGRA et de son échantillon sont à la fois liées et indépendantes. Toutefois, quel que soit l'objectif de l'étude, les rapports EGRA doivent inclure dans les analyses et la rédaction du rapport les éléments suivants :

- **Une définition claire et précise de la population cible** décrivant toutes les exclusions opérées avant l'échantillonnage.
- Le calcul approprié des **poids d'échantillonnage** (utilisé pour pondérer l'échantillon et le rendre représentatif de la population dont il a été extrait).
- L'emploi d'un **module d'analyse de données de sondage complexe** (voir glossaire) tel que celui inclus dans les logiciels SPSS, Stata ou SAS. Ce type de module permet de tenir compte de la méthodologie d'échantillonnage et de l'effet d'échantillonnage en grappes (voir glossaire).
- **L'utilisation d'analyses statistiques** qui reposent sur l'utilisation d'un module d'analyse de données de sondage complexe (voir glossaire).

Pour concevoir une étude EGRA, il faut dans un premier temps se poser la question suivante : **« Quel est le but de l'étude EGRA ? »** La plupart des études EGRA ont pour but un des trois objectifs suivants :

1. **Un aperçu sommaire** dont l'objectif est d'obtenir un diagnostic des performances des élèves à un moment donné (et unique) dans le temps.

2. **L'évaluation des performances** dont l'objectif est de mesurer l'amélioration des résultats des élèves au cours d'une période donnée. Cette évaluation est basée sur une succession de mesures dans le temps.

3. **L'évaluation de l'impact** dont le but est d'évaluer l'effet d'une intervention ou d'un programme sur les performances des élèves au cours d'une période donnée. Cette évaluation est basée sur une comparaison entre un groupe expérimental et un groupe témoin.

La section 5.2 décrit de manière plus détaillée le but de chacune de ces évaluations et la façon dont chacune répond aux besoins de différents types d'étude EGRA. Chaque objectif est décrit et références sont faites aux annexes de ce document qui présentent les méthodologies d'échantillonnage propres à chacun d'entre eux.

5.2 Quel type d'étude, pour quel objectif ?

5.2.1 Conception d'un aperçu sommaire ou d'une étude d'évaluation des performances

L'aperçu sommaire a pour but de fournir, à un instant donné, des informations sur une variable ou une compétence d'intérêt, fluence en lecture par exemple (voir glossaire et chapitre 4). L'étude d'évaluation des performances a le même objectif, mais collecte les informations à plusieurs reprises. Aucune de ces deux méthodes ne permet d'attribuer un résultat à une cause ou une intervention spécifique.

ECHANTILLONAGE POUR UN APERCU SOMMAIRE ET UNE ETUDE D'EVALUATION DES PERFORMANCES

Dans la plupart des études EGRA, l'objectif d'un aperçu sommaire est de mesurer les compétences en lecture d'une population donnée. Ce type d'évaluation utilise une méthode d'échantillonnage en grappes et / ou celle du sondage complexe (échantillonnage décrit dans les **Annexes B et C**).

Au contraire de l'aperçu sommaire, la méthode d'échantillonnage d'une évaluation des performances doit être basée sur les questions de recherche posées par l'étude ainsi que sur les ressources disponibles. Si, par exemple, l'étude souhaite comparer les scores des élèves bénéficiaires d'une intervention à des scores normés au niveau national, un échantillonnage aléatoire simple ou en grappes / complexe sera nécessaire pour s'assurer que cet échantillon est représentatif de la totalité de la population concernée. Toutefois, si les ressources sont limitées, il sera possible de choisir un échantillon plus petit et non-représentatif de la population en expliquant ensuite les importantes limitations des données obtenues quant aux inférences statistiques possibles sur la population cible. Avec de telles limitations, les résultats pourront être utilisés pour une étude interne au programme mais ne pourront pas être généralisés à la population cible.

5.2.2 Evaluation de l'impact comme plan de recherche

Une évaluation de l'impact diffère d'une évaluation des performances en ce qu'elle a pour objectif de mesurer l'impact d'une intervention sur une variable dépendante clé en comparant les résultats d'un groupe expérimental (ou de plusieurs groupes recevant différentes versions du programme expérimental) à ceux d'un groupe témoin afin d'isoler la contribution de l'intervention de celle d'autres variables d'influence. En d'autres termes, une évaluation de l'impact utilise une situation contrefactuelle (voir glossaire) pour déterminer si l'intervention d'intérêt est la cause de l'amélioration des performances ou scores observés. Un nombre croissant d'évaluations visent à mesurer l'impact d'interventions pédagogiques sur les scores EGRA. Il existe deux types d'évaluation d'impact : les évaluations expérimentales et les évaluations quasi- expérimentales (voir encart ci-dessous). On trouvera en **Annexe D** de plus amples informations sur l'échantillonnage nécessaire à une évaluation de l'impact.

DEUX TYPES D'EVALUATION D'IMPACT

Plan expérimental

L'élaboration du plan expérimental, parfois appelé essai contrôle à répartition aléatoire, doit débuter avant que ne commence l'intervention. Ce type d'étude repose sur la collecte de données de référence et la répartition aléatoire des participants à l'intervention (écoles, zones ou toute autre type d'unité) en deux groupes : un groupe expérimental et un groupe témoin (aussi appelé groupe de comparaison). La probabilité d'inclusion dans un de ces deux groupes doit être identique pour tous les participants (ou unité) et la taille de l'échantillon doit être suffisamment grande pour garantir que l'effet minimum décelable puisse être détecté lorsque les groupes seront comparés (voir **Annexe D**).

Plan quasi-expérimental

L'élaboration du plan quasi-expérimental peut débuter avant ne commence que ne commence l'intervention. Cela n'est toutefois pas nécessaire, tant que des données de référence sont disponibles pour le groupe expérimental et pour le groupe témoin dès le début de l'intervention. Les participants à une étude quasi-expérimentale ne sont pas affectés de manière aléatoire dans l'un ou l'autre de ces deux groupes. Ils peuvent, par exemple, choisir leur groupe ou être affectés à un groupe en fonction d'un critère donné. Dans les deux cas, un biais de sélection sera introduit dans l'étude, biais qui devra être minimisé ou contrôlé par différentes procédures statistiques.
Les études quasi-expérimentales permettent d'attribuer les résultats à l'intervention d'intérêt, mais sont moins rigoureuses et moins fiables que les études expérimentales. Quelques exemples d'études quasi-expérimentales utilisant EGRA incluent *l'approche de discontinuité par régression* et *la méthode d'appariement sur les coefficients de propension* (voir glossaire).

Niveaux d'affectation pour les évaluations de l'impact

Une fois qu'on aura décidé d'utiliser une évaluation d'impact pour mesurer les résultats d'une intervention, il faudra déterminer les critères d'affectation dans les groupes et la nature de l'étude, transversale ou longitudinale, voire semi-longitudinale. Une étude transversale évalue la progression dans le temps d'élèves différents alors qu'une étude longitudinale évalue celle des mêmes élèves. L'étude semi-longitudinale se situe entre les deux : elle évalue la progression d'élèves ayant les mêmes enseignants ou scolarisés dans les mêmes écoles. Les décisions dépendront du but de l'étude ainsi que de la manière dont l'intervention (programme, projet ou activité) évaluée sera mise en œuvre. L'objectif d'une intervention peut cibler plusieurs facteurs ou niveaux :

- **Le district, zone ou unité administrative** : formation des enseignants d'un district donné, par exemple.
- **La communauté** : par exemple, animation de programmes de sensibilisation de la communauté pour accroitre sa participation à la vie de l'école ou l'encourager à entreprendre la création d'un centre communautaire d'alphabétisation.
- **L'école** : par exemple, fourniture de livres, de matériel ou d'autres outils pédagogiques à certaines écoles (mais pas à d'autres) au sein d'une unité administrative donnée.
- **L'élève** : par exemple, donner à certains élèves au sein d'une école (mais pas à d'autres) accès à des bourses d'étude ou à des subventions conditionnelles en espèces.

Une intervention peut porter sur plusieurs niveaux ou facteurs à la fois. Il est donc important de les définir clairement afin que l'affectation aux groupes expérimental et contrôle se fasse sur la base du niveau le plus élevé ciblé par l'intervention. Par exemple, si une intervention vise à former les enseignants de différents districts et l'impact de la distribution de livres et de matériel pédagogique au niveau des écoles, l'affectation aux groupes se fera au niveau de chaque district étudié. C'est pourquoi il est très important que les équipes chargées de l'évaluation et celles chargées de sa mise en œuvre collaborent étroitement à la mise au point du plan d'évaluation.

Planification de la conception d'une évaluation d'impact

Il est important dans un deuxième temps de déterminer si les élèves seront suivis longitudinalement, semi-longitudinalement ou transversalement. Les informations suivantes sont nécessaires pour prendre une décision :

1. **Quel est le but de l'évaluation et quelles sont les questions de recherche qu'elle pose ?**

   Etude longitudinale. Si l'étude a pour objectif de déterminer et de comprendre les changements observables chez chaque participant à l'intervention, l'emploi d'une méthode d'évaluation longitudinale permettra de détecter avec certitude si

un changement s'est opéré pour chacun des participants à l'évaluation. Il ne sera toutefois pas possible de généraliser ces résultats à l'ensemble de la population faisant l'objet de l'intervention. Ce type d'étude convient mieux aux recherches pilotes ou aux évaluations continues informelles internes à un programme.

Etude semi-longitudinale. Elle est adéquate si le but de l'étude est d'expliquer les changements complexes au sein de chaque école. Dans ce cadre, les mêmes écoles sont évaluées à chaque collecte de données mais un échantillon aléatoire d'élève est sélectionné au sein de chacune. Ce type d'étude permettra d'étudier les changements opérés par l'intervention pour chaque école évaluée mais ne permettra pas de généraliser les résultats à l'ensemble des écoles de la population ciblée par l'intervention. Ce type d'étude convient mieux aux recherches pilotes ou aux évaluations continues informelles internes au programme.[8]

Etude transversale. Elle permet de déterminer comment une population (voir glossaire) d'écoles et d'élèves en leur sein peut changer suite à la mise en œuvre d'une intervention. On utilise une série d'échantillons d'écoles et d'élèves différents à chaque collecte de données et sont sélectionnés au sein de cette population ceux recevant l'intervention. Cette méthode ne permettra pas de déterminer les changements précis au sein de chaque école ou élève. Il sera néanmoins possible de généraliser tous changements notés à la population cible. Ce type d'étude convient aux évaluations externes d'une intervention donnée. Ces évaluations ont pour but de comprendre l'impact d'une intervention sur l'ensemble de la population.

2. **Dans quelle mesure est-il facile de suivre les mêmes élèves, enseignants ou écoles dans le temps ?** Par exemple, si l'équipe évalue une intervention dans un pays où chaque élève a un numéro identifiant qu'il garde même s'il déménage ou que chaque ménage est enregistré auprès du gouvernement pour les impôts ou le recensement, il n'est alors pas difficile de suivre les élèves individuellement dans le temps. Si ce n'est pas le cas et qu'en plus les individus ou les communautés tendent à être mobiles et le taux de décrochage scolaire élevé, il devient alors difficile et fastidieux de suivre les mêmes élèves dans le temps.

3. **Quelles ressources sont nécessaires pour suivre les élèves, enseignants, écoles longitudinalement ?** Même quand il est facile de suivre les élèves de manière longitudinale, le suivi reste plus coûteux et plus laborieux qu'un aperçu sommaire (voir glossaire). En effet, certains élèves peuvent être difficile à retrouver ce qui impose, par précaution, de sur-échantillonner lors de l'évaluation de référence.

[8] Les études longitudinales ou semi-longitudinales peuvent aussi être utiles pour les évaluations qui ne mesurent pas l'impact d'une intervention, étant donné que cette méthodologie permet de déterminer les changements dans le temps d'une unité d'analyse définie (ex : élèves) même si aucune intervention n'est mise en œuvre.

4. **Quel niveau de rigueur et de précision est nécessaire dans les résultats ?** Si un bailleur de fonds ou une personne chargée de la mise en œuvre d'un programme souhaite obtenir des résultats précis sur le taux de décrochage scolaire par exemple, une étude longitudinale sera sans doute nécessaire. Toutefois, si une estimation est suffisante (comme les informations relevées auprès des enseignants et des écoles), une étude transversale suffira.

6 CONCEPTION D'EGRA : ADAPTATIONS ET NOUVEAUX DÉVELOPPEMENTS

Ce chapitre présente la structure d'EGRA, et ce qui est nécessaire pour développer cet outil et l'adapter à différents contextes. Il décrit les épreuves qui peuvent être incluses dans une évaluation EGRA et précise comment elles ont été construites et la façon de les adapter.

6.1 L'Atelier

La première étape est d'organiser un atelier de 5 jours environ. Cet atelier, qui devra avoir lieu dans le pays où EGRA sera utilisé, sera le début de la démarche du développement (ou de l'adaptation) de l'instrument EGRA. A cet atelier, doivent participer des représentants officiels du gouvernement, des experts du curriculum et d'autres personnes habilitées à examiner l'adéquation des items et à évaluer les compétences de lecture des élèves.[9] Cet atelier permet ainsi de s'assurer d'une bonne validité du contenu (voir glossaire) et de l'adéquation de l'outil au curriculum et aux normes d'apprentissage du pays.

Les participants développent / adaptent l'instrument en préparant pour chaque épreuve des items appropriés au contexte du pays. Après l'atelier, l'instrument doit être prétesté dans une ou plusieurs écoles (la procédure de prétest et le travail sur le terrain sont discutés au chapitre 9).

La suite de la section 6.1 du chapitre 6 présente les différentes étapes nécessaires à l'organisation d'un atelier d'adaptation EGRA et en décrit le contenu. Des informations supplémentaires sur les qualités techniques et la fiabilité de l'instrument EGRA sont fournies dans l'**Annexe E** et le chapitre 9.1.2, qui comprennent des recommandations sur les analyses de fiabilité et de qualité à conduire.
Les objectifs d'un atelier EGRA sont les suivants :

- Informer les représentants officiels du gouvernement et les experts du curriculum, des travaux de recherches qui sous-tendent le développement de l'instrument.
- Adapter l'instrument au contexte local en utilisant, pour la construction des items, les recommandations incluses dans ce guide. Elles comprennent :

[9] The degree to which the items on the EGRA test are representative of the construct being measured is known as test-content-related evidence (i.e., early reading skills in a particular country).

- La traduction des instructions de chaque épreuve ;
- La création de différentes versions pour chacune des langues maternelles des élèves concernés par cette évaluation ;
- La modification des mots et textes utilisés afin de s'assurer de leur pertinence et de leur adéquation au contexte dans lequel ils seront utilisés.

- Réviser les procédures de consentement (pour les adultes et les élèves) et discuter des impératifs éthiques de la recherche impliquant la participation de sujets humains et plus particulièrement d'enfants.

Le tableau 6.1 détaille les différences entre un atelier de développement et un atelier d'adaptation. Un atelier de développement est organisé si EGRA est développé dans une langue ou un pays pour la première fois. Le cas échéant, c'est un atelier d'adaptation.

Figure 7. Différences entre un atelier EGRA de développement et un atelier d'adaptation

Développement d'un nouvel instrument	Adaptation d'un nouvel instrument
Analyse des propriétés statistique de l'orthographe	Analyse des propriétés statistique de l'orthographe (optionnel)
Choix des items	Modification de l'ordre des items (aléatoire)
Vérification des instructions	Vérification des instructions
Prétest	Prétest

6.1.1 Considérations pour l'organisation de l'atelier

Que l'atelier ait pour but le développement complet d'un instrument EGRA pour un pays donné ou l'adaptation d'un instrument existant, l'équipe en charge du projet doit s'assurer que l'instrument est approprié pour la(es) langue(s) concernée(s) par l'étude ainsi que pour les classes et les questions de recherche sélectionnées. Le développement de l'instrument nécessite la sélection d'épreuves et d'items appropriés.

De plus, le calendrier doit permettre de prétester l'instrument en cours de développement dans les écoles. Cela nécessitera que les participants à l'atelier (tous ou un petit groupe) se déplacent dans une école (ou plusieurs) pour utiliser l'instrument en situation réelle avec des élèves. Cette étape permettra aux participants de mieux comprendre le fonctionnement de l'instrument et servira de premier prétest de l'outil qui pourra mettre en relief les changements les plus nécessaires (par exemple, questions trop ambigües ou vocabulaire trop difficile).

- Les analyses statistiques et linguistiques nécessaires à la préparation des items peuvent être conduites en avance. C'est aussi le cas de la traduction des instructions, qui ont été standardisées et dont les principes doivent rester les mêmes dans tous les pays. Un panel d'expert et un comité d'éthique doivent

revoir les instructions pour s'assurer qu'elles respectent les règles d'éthique. Il est important de fournir à tous les élèves les mêmes instructions standardisées afin de s'assurer que tous les participants sont traités de manière identique, quels que soient le contexte, le pays et l'examinateur.

- Si l'atelier ne peut être organisé dans la région dans laquelle l'évaluation aura lieu, l'équipe en charge doit organiser la session de prétest dans les écoles de la région étudiée, ou doit identifier au cours de l'atelier un groupe d'écoles dont les élèves parlent la(es) langue(s) du test développé. Après la collecte de ces données pilotes, les résultats doivent être présentés et discutés avec tous les participants de l'atelier.
- L'écriture des textes pour la compréhension écrite constitue généralement la partie la plus difficile de l'atelier de développement/adaptation. Il est donc important de ne pas s'y atteler à la dernière minute. Ce travail requiert l'implication d'experts qui rédigeront des textes ayant un vocabulaire adéquat pour le niveau d'étude évalué. Il faudra également rédiger des questions de compréhension portant sur ces histoires. Il sera souvent nécessaire de traduire les histoires et les questions en français (et parfois en anglais) afin qu'elles puissent être examinées par d'autres experts en lecture, et révisées dans la langue de développement, avant la finalisation de l'outil.

6.1.2 Qui peut participer ?

Afin d'inclure une variété d'expériences pertinentes au développement / adaptation de l'outil, les personnes suivantes doivent participer à l'atelier : employés du gouvernement, enseignants en poste ou à la retraite, professeurs en éducation en charge de la formation des enseignants et chercheurs ou universitaires spécialistes des langues du test. La présence d'employés du gouvernement est importante et nécessaire à la viabilité de l'évaluation. Le nombre de participants dépendra aussi du nombre de langues dans lesquelles l'outil sera développé / adapté. En général, un maximum de 30 participants est approprié.

Les participants à l'atelier incluent toujours :

1. Des linguistes : ils vérifieront la traduction des instructions, guideront la révision des items et soutiendront la création et la modification des histoires.
2. Des praticiens : des universitaires (spécialistes en lecture en particulier) et des enseignants en poste ou retraités (avec une préférence pour des enseignants en lecture).
3. Des membres du gouvernement : des experts du curriculum et des évaluations.
4. Un psychométricien ou un expert du développement d'évaluations et tests.

Il est recommandé que les employés du gouvernement les plus concernés par l'évaluation participent à 'l'intégralité du processus d'adaptation, à la formation des examinateurs et au prétest (processus durant un mois environ, en fonction du nombre d'écoles testées). Il est important de préserver l'uniformité de l'évaluation

parmi les participants. Cela requiert une évaluation bien organisée, des équipes bien formées et une méthodologie pérenne pour le pays (ce qui signifie qu'elle doit pouvoir être réutilisable sans soutien extérieur).

L'atelier doit être animé par deux experts au minimum. Les animateurs doivent être experts de l'évaluation et capables d'expliquer et de justifier la composition de l'outil. Ils doivent avoir de l'expérience dans un certain nombre de contextes – et pays – différents :

- **L'expert en évaluation** sera en charge de l'adaptation / du développement de l'instrument. Dans un deuxième temps, il guidera la formation des évaluateurs et la collecte des données. Cet expert devra avoir une bonne formation en évaluation d'études en éducation, en développement de tests et sera formé en statistiques (utilisation de tableurs / logiciels de statistique tels qu'Excel, SPSS ou Stata).
- **L'expert en lecture** sera responsable de la présentation des recherches sur la lecture et l'enseignement sur lesquelles se basent EGRA. Cet expert aura une formation en évaluation et enseignement de la lecture.

6.1.3 Quel matériel préparer ?

Pour l'atelier, le matériel nécessaire inclus :

- Papiers, crayons et gommes pour les participants.
- Rétroprojecteur, tableau blanc et tableau papier (si possible, le rétroprojecteur doit pouvoir projeter les images sur le tableau blanc pour suivre en direct les exercices de notations du test pendant la formation).
- Des textes utilisés au plan national ou local, adaptés aux niveaux d'étude et aux langues concernés par l'évaluation. Ces textes peuvent fournir des informations sur la nature du vocabulaire et le niveau de difficulté des histoires utilisées dans EGRA.
- Copies des présentations pour l'atelier et la version provisoire de l'instrument.
- Documents et présentations des travaux de recherche sur EGRA et sur la lecture ainsi que sur le processus de développement, les buts, et l'utilisation de cet outil.
- Des exemples d'histoires utilisées pour mesurer la fluence en lecture, la compréhension écrite et la compréhension orale. Ces exemples peuvent être trouvés dans les outils utilisés dans d'autres pays et dans les versions utilisées précédemment dans le pays pour les cas où une simple adaptation est conduite.

Un exemple de calendrier pour l'atelier de développement / d'adaptation et de recherche est présenté dans le **Figure 8**.

Figure 8. Exemple de calendrier pour un atelier EGRA de développement ou d'adaptation

Jour et Heure	Jour 1	Jour 2	Jour 3	Jour 4	Jour 5
9:00-9:30	Accueil et Introduction	Résumé Jour 1	Résumé Jour 2	Résumé Jour 3	Visite des écoles pour le prétest de l'instrument et du questionnaire
9:30-10:30	Présentation du projet et d'EGRA	Révisions du premier jet d'EGRA (ex : mots inventés)	Développer / modifier le texte évaluant la compréhension orale	Développer / modifier les épreuves supplémentaires et les questionnaires, si nécessaire	
10h30-11:00	***Pause***				
11h00-12h30	Présentation d''EGRA (but, contenu de l'instrument, utilisation des résultats)	Développer / modifier le texte évaluant la compréhension écrite	Poursuite du travail sur le texte évaluant la compréhension orale et création des questions	Poursuite du travail sur les épreuves supplémentaires et les questionnaires, si nécessaire	Discussion sur la visite des écoles
12h30-13h30	***Déjeuner***				
13h30-15h00	Présentation du système d'écriture (orthographe et autres informations importantes pour développer EGRA)	Continuer le développement des histoires et préparer les questions	Révision et modification du questionnaire de l'élève	Révision de l'outil et entrainement pour l'administration des prétests	Finalisation des instructions
15h00-15h45	***Pause***				
15h45-17h00	Révisions du premier jet d'EGRA (ex : conscience phonémique et sons des graphèmes)	Finalisation de cette épreuve	Finalisation de cette épreuve	Révision de l'outil et entrainement pour l'administration des prétests	Clôture de l'atelier
Objectif du jour	*Comprendre le but et le contenu d''EGRA*	*Compréhension d'un texte écrit*	*Compréhension d'un texte oral et questionnaire de l'élève*	*Epreuves supplémentaires développées*	*Instrument finalisé*

Note. La durée de l''atelier d''adaptation et des sessions spécifiques dépendra de plusieurs facteurs : l''existence préalable d''une version 'd'EGRA développée pour la langue, le pays ou le niveau scolaire ; le nombre de sous-tâches à tester ; le nombre de langues à tester ; les besoins de questionnaires et d'instruments supplémentaires ; le but de l 'atelier, et son public.

6.2 Examen des composants de l'instrument de base

Comme expliqué dans la section 1, EGRA a été développé initialement avec des experts d'USAID, de la Banque mondiale, et de RTI. Grâce à l'aide d'autres experts, les différentes versions d'EGRA (en français comme en anglais) ont été progressivement mises en place, avec des révisions continuelles.
L'instrument de base pour les orthographes alphabétiques contient six épreuves, quatre étant essentielles :

1. Compréhension orale d'un texte : réponse à des questions de compréhension ;
2. Identification du son des lettres et de suites de lettres (voire du nom des lettres)[10] .
3. Fluence (précision et rapidité) de la lecture à haute voix de mots inventés ;

[10] L'identification du son des lettres est l'épreuve la plus couramment utilisée. Toutefois, selon le pays et la langue dans laquelle l'instrument est administré, une épreuve évaluant la connaissance du nom de lettres ou la lecture des syllabes peut être plus appropriée.

4. Compréhension écrite d'un texte : réponse à des questions de compréhension et fluence (précision et rapidité) de la lecture à haute voix de ce texte ;

Ces quatre épreuves clé de la batterie EGRA ont été utilisées dans des dizaines de langues à travers le monde. Deux autres épreuves (voir le **Figure 9**) sont souvent incluses dans les évaluations : celle de lecture à haute voix de mots familiers et celle de conscience phonémique.

D'autres épreuves sont optionnelles. C'est le cas pour les épreuves de dictée, de vocabulaire ainsi que pour certaines épreuves de compréhension, celles utilisant des exercices à trou (voir ci-dessus les sections 4.4.3, 4.5.2 et 4.6.2 ; voir aussi, pour une discussion sur ces différentes épreuves complémentaires, le chapitre 1 de Gove et Wetterberg, 2011, ainsi que la section 6.3 ci-dessous).[11]

En réponse à des demandes récurrentes des examinateurs ayant utilisé EGRA, le nombre de compétences évaluées a été limité. Comme indiqué dans ce qui précède, l'un des objectifs d'EGRA est d'évaluer les compétences de base de la lecture afin de faire ressortir celles que les enfants ne maitrisent pas, et qui nécessitent des instructions supplémentaires. Si EGRA ne testait que la fluence en lecture à haute voix de textes, d'importants effets « plancher » seraient observés dans de nombreux pays en voie de développement. Maintenir un nombre raisonnable d'épreuves (environ six) permet de noter les progrès des enfants dans au moins certaines capacités reliées à la lecture, y compris dans des pays où le niveau de lecture est très faible.

L'efficacité des épreuves prises en compte dans EGRA, en tant que point de départ raisonnable d'une évaluation des premières étapes de l'apprentissage de la lecture, a été démontrée (NICHD, 2000 ; Dubeck & Gove, 2015). Toutefois, si EGRA couvre un nombre important d'épreuves évaluant les compétences prédictives de l'apprentissage de la lecture, cette batterie n'évalue pas toutes celles qui contribuent au succès de cet apprentissage. Par exemple, EGRA ne mesure pas la motivation, l'attention, la mémoire, les stratégies de lecture, la compréhension de différents types de texte… Il est quasi impossible de couvrir l'ensemble des compétences impliquées dans l'activité de lecture. Et, même si une telle évaluation pouvait être mise en place, elle serait très longue, ce qui aurait pour conséquence de fatiguer les enfants, et de les rendre ainsi peu performants. Enfin, s'il ne faut pas sacraliser cet instrument, il est toutefois recommandé que les modifications, que ce soit dans les épreuves ou les procédures, soient justifiées et partagées avec la « communauté de pratiques ».

6.2.1 Compréhension orale

Après avoir entendu un court texte lu à haute voix par l'examinateur, l'enfant doit répondre à des questions de compréhension. Parce qu'elle est relativement facile et qu'elle oriente les enfants vers la langue cible de l'évaluation, cette épreuve peut se dérouler au début de l'évaluation.

[11] Optional subtasks such as dictation, maze, and cloze are occasionally used in addition to the common subtasks. See the discussion about these additional subtasks in Chapter 1 of Gove and Wetterberg (2011), as well as Section 6.3 below, "Review of Additional Instrument Components."

Figure 9. Examen des épreuves communes de EGRA

Epreuves	Compétences évaluées	Maitrise de la compétence démontrée par la possibilité de :
1. Compréhension orale	Compréhension du langage oral	Répondre correctement à différentes questions sur un texte lu par l'examinateur, certaines portant sur des informations présentées explicitement dans le texte, d'autres nécessitant de faire des inférences
2. Identification du son des lettres et suites de lettres (voire du nom des lettres	Connaissances alphabétiques	Donner le son de lettres (voire leur nom) présentées en minuscule ou en majuscule dans un ordre aléatoire ('i', 'a', 'L', 'r', 'ou', 's', 'f', 'O', 'u'…)
3. Lecture de pseudomots	Capacités de décodage	Utiliser les correspondances graphème–phonème (CGP) pour lire à haute voix des mots inventés (des pseudomots, comme mable)
	Fluence du décodage	Lire à haute voix, de façon fluente (avec précision et rapidité) ces pseudomots
4. Compréhension écrite	Compréhension du langage écrit	Répondre correctement à différentes questions sur un texte écrit, certaines portant sur des informations présentées explicitement dans le texte, d'autres nécessitant de faire des inférences
	Fluence en lecture de mots en context	Lire à haute voix, de façon fluente (avec précision et rapidité) ce texte
5. Discrimination du phonème initial ou final d'un mot Identification du phonème initial d'un mot Segmentation de mots en phonèmes, attaques et rimes, ou syllables	Conscience phonémique ou phonologique	Dire quel est le mot qui ne commence pas par le même son : par exemple, bal, bol, mal ? Dire quel est le premier son qu'on entend dans, par exemple, le mot bol Prononcer les différents sons ou syllabes d'un mot oral
6. Lecture de mots familiers	Reconnaissance des mots écrits	Lecture à haute voix d'une liste de mots fréquents présentés dans un ordre aléatoire
	Fluence de cette reconnaissance	Lire à haute voix, de façon fluente (avec précision et rapidité) ces mots

Evaluer la compréhension orale est important, cette évaluation permettant de savoir ce que les enfants sont capables de comprendre sans avoir à lire. Ceux qui ont des difficultés de décodage, ou qui n'ont pas encore appris à lire, ont des compétences langagières qui doivent être évaluées en dehors de la lecture. L'évaluation de la compréhension orale a longtemps été réservée aux enfants ayant un accès relativement limité à l'écrit (Orr & Graham, 1968). De mauvaises performances dans ce domaine peuvent s'expliquer, par exemple, par des connaissances lacunaires de la langue (en particulier, chez les apprenants de langue seconde) ou par des problèmes d'audition, les deux entravant l'apprentissage de la lecture.

Recueil des données. Nombre de bonnes réponses aux questions par rapport au nombre total de questions (voir **Figure 10**).

Construction de l'épreuve. La longueur du texte dépend du niveau de l'enfant et de sa langue maternelle. Toutefois, le texte doit avoir environ 30 mots afin de permettre

de poser trois à cinq questions de compréhension. Il doit raconter une activité ou un évènement adapté localement, familier à l'enfant. Les questions doivent être semblables à celles de l'épreuve de compréhension en lecture (décrite ci-dessous). La plupart d'entre elles portent sur des informations présentées explicitement dans le texte (questions littérales). Les autres exigent de faire des déductions à partir des connaissances de l'enfant et de ce qui est écrit dans le texte. Il faut toutefois éviter les questions auxquelles il est possible de répondre sans avoir entendu le texte. Il faut aussi éviter des questions avec seulement des réponses oui ou non.

Figure 10. Epreuve de compréhension orale (Epreuve n°1)

<table>
<tr><td colspan="3">Epreuve Compréhension orale</td><td>Page _</td><td>X</td></tr>
<tr><td colspan="4">Maintenant, je vais te lire une histoire UNE fois. Après cela, je vais te poser quelques questions sur cette histoire. Tu vas bien écouter, et ensuite tu répondras aux questions le mieux que tu peux. Tu peux répondre dans la langue que tu préfères. D'accord ? Commençons ! Ecoute bien :</td><td rowspan="8">Retirez le cahier de stimulus de la vue de l'élève. Ne laissez pas l'élève voir l'histoire ou les questions. Si l'élève dit «je ne sais pas », notez la réponse comme incorrecte.</td></tr>
<tr><td colspan="4">La poule blanche est tombée dans la mare.
Elle crie : « Aide-moi ! »
Un agneau noir vient à son secours.
Mais il tombe lui aussi dans la mare.
«Que faire ? » demande-t-il.
La poule dit : « Regarde ce tronc d'arbre qui flotte. Il peut nous sauver ! »
Les deux amis grimpent alors sur le tronc d'arbre et crient,
« Ouf, nous allons pouvoir retrouver la terre ferme ! »</td></tr>
<tr><td>Questions (Ne répétez pas les questions)</td><td>Corret</td><td>Incorrect</td><td>Pas de réponse</td></tr>
<tr><td>Où est tombée la poule ?
[dans la mare ; dans l'eau]</td><td></td><td></td><td></td></tr>
<tr><td>L'agneau est de quelle couleur ?
[noir]</td><td></td><td></td><td></td></tr>
<tr><td>Qui est tombé en dernier dans la mare ?
[l'agneau]</td><td></td><td></td><td></td></tr>
<tr><td>Quel objet important la poule a-t-elle vu ?
[un tronc d'arbre qui flottait]</td><td></td><td></td><td></td></tr>
<tr><td>Qu'est-ce que la poule et l'agneau ont fait par la suite ?
[grimper sur le tronc d'arbre]</td><td></td><td></td><td></td></tr>
<tr><td>A quoi cela va leur servir ?
[sortir de la mare ou de l'eau ; regagner la terre ferme]</td><td></td><td></td><td></td></tr>
<tr><td colspan="4">Merci bien ! On peut passer à la prochaine activité !</td></tr>
</table>

6.2.2. Identification des graphèmes (lettres et suites de lettres)

Savoir comment les graphèmes, qu'ils aient une lettre ou plus, se prononcent est une autre compétence critique que les enfants doivent maitriser pour devenir lecteurs. Une épreuve de ce type est souvent incluse dans les batteries d'évaluation des débuts de l'apprentissage de la lecture dans le monde anglo-saxon, y compris dans celles destinées aux enfants de maternelle (Lonigan, Wagner, Torgesen, & Rashotte, 2002).

Le plus souvent, l'évaluation porte sur la connaissance du son des lettres et des suites de lettres (les graphèmes). Cette évaluation permet de savoir si l'enfant

maitrise le principe à la base du décodage. En effet, pour prononcer un mot, il faut mettre en relation chaque graphème avec le phonème correspondant. Une étude dans laquelle cette connaissance a été évaluée chez des élèves ayant appris à lire dans différentes langues des pays d'Europe a permis de relever des scores supérieurs à 90 % de réponses correctes (Seymour, Aro, & Erskine, 2003 : 90, 91, 94 et 96 % de réponses correctes respectivement aux Pays-Bas, en France, en Grande Bretagne et en Espagne, cf. le Tableau 4, p.150). Dans des langues qui ont une orthographe peu transparente, comme l'anglais, dans la mesure où la recherche a montré que la connaissance du nom des lettres est également un bon prédicteur du futur niveau de lecture, cette connaissance peut -- en plus -- être évaluée (Loningan et al., 2002).

Première approche : Identification du son des lettres

Dans cette épreuve en temps limité (une minute), les enfants sont invités à produire le son des lettres (incluant celui des graphèmes de plus d'une lettre, en français, é, u, ou, an…). Les items sont présentés dans un ordre aléatoire, 10 par ligne, en mélangeant minuscules et majuscules. La police de caractères utilisée doit être lisible, grande, et familière aux enfants : par exemple, « Century Gothic » dans Microsoft Word, qui est similaire aux polices utilisées dans de nombreux manuels scolaires ou la police Andika de « SIL International », conçue pour les débutants (http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=andika).

Le nombre de répétitions d'un item doit être fonction de sa fréquence dans la langue (voir, pour le français, **Figure 11**).[12] Certaines tables sont disponibles pour différentes langues, par exemple, fréquence des lettres pour l'espagnol, l'allemand, le portugais, entre autres.[13] Pour les langues pour lesquelles ces tables ne sont pas disponibles, il faut les construire à partir d'exemples issus de l'analyse de 20 à 30 pages de livres scolaires ou d'un autre matériel adéquat.

Pour réaliser une table de fréquence des lettres, il faut intégrer les pages sélectionnées dans un programme de traitement de texte et, en utilisant la commande Rechercher, entrer chaque lettre pour avoir une indication de sa fréquence. Par exemple, d'après Microsoft Word, le début de la section 6.2 du présent chapitre contient un peu plus de 4600 lettres : la plus fréquente est 'e' (15,4 %), viennent ensuite 's' et 't' (8,3 et 7,3 %). Ces données reproduisent celles présentées dans le tableau 6.4 (issues du corpus pour le premier grade du primaire de Manulex, Peereman et al., 2013). Il y a toutefois quelques différences dues à la taille des corpus mais aussi à leur origine. Ainsi, la plus grande fréquence du 'a' dans Manulex (8,1 % contre 5,8 % dans la section 6.2 ci-dessus) provient d'une plus forte présence des verbes conjugués au futur et à l'imparfait dans Manulex (cf. les marques de flexion : 'a', 'as', 'ai', 'ais', 'ait'…). Ce point signale la nécessité de prêter attention au corpus choisi.

[12] issu de Manulex-Morpho, Peereman et al., 2013, http://lpnc.univ-grenoble-alpes.fr/resources/ronald_peereman/Manulex_morpho/indexfr.html).
[13] Université de Californie, Los Angeles, Online Computational Resource, http://wiki.stat.ucla.edu/socr/index.php/SOCR_LetterFrequencyData#SOCR_Data; dernier accès le 18 octobre 2016

Figure 11. Fréquence des lettres et des phonèmes du français

Pourcentages d'après les mots du corpus du 1er grade du primaire de Manulex-Morpho (Peereman et al, 2013):
- pour les lettres, fréquence lexicale des voyelles et consonnes
- pour les phonèmes, fréquence lexicale des voyelles, consonnes et semi-voyelles

Voyelles				
Lettre	**%**	**Phonème***	**%**	**Exemple**
a	8,14	a	7,54	plat
		A	0,50	tard
à	0,00			
â	0,13			
e	13,12	œ	0,54	neuf
		ø	0,31	deux
		ə	2,45	gare
é	3,14	e	6,43	évier
		E	4,81	lève
è	0,46			
ê	0,16			
ë	0,01			
i	7,53	i	5,77	Il
î	0,08			
ï	0,05			
y	0,30			
o	5,87	o	3,15	
		O	1,17	
ö	0,04			
u	4,72	y	1,86	lune
û	0,06			
ù	0,00			
ü	0,00			
		u	1,83	ou
		â	3,28	an
		ô	1,86	on
		û	0,03	un
		î	1,09	fin

Consonnes				
Lettre	**%**	**Phonème***	**%***	**Exemple**
p	3,13	p	3,91	poire
t	7,19	t	5,68	tour
c	3,94	k	3,94	col
k	0,08			
q(u)	0,49			
b	1,51	b	2,05	bal
d	2,17	d	2,76	dur
g	1,84	g	1,23	gare
f	1,35	f	1,74	four
s	7,17	s	5,07	sol
ç	0,09			
		ʃ	1,33	chic
v	1,55	v	2,16	vol
z	0,33	z	1,30	zone
j	0,27	ʒ	1,40	jeu
l	4,62	l	4,67	lac
m	2,82	m	3,06	mur
n	7,01	n	2,40	nid
r	8,86	r	10,70	rue
x	0,36	ks		taxi
		gz		examen
		ɲ	0,30	gagne
w	0,02			
h	1,39			
Semi-voyelles				
		j	2,41	yeux
		w	0,89	moi
		'u' de *lui*	0,41	lui

*La liste des caractères des alphabets phonétiques est consultable dans le site créé et géré par l'Association internationale de phonétique (http://westonruter.github.io/ipa-chart/keyboard/)

(d'après Peereman et al., 2013)

La présence de graphèmes autres que ceux correspondant à une lettre de l'alphabet dépend de la langue. Par exemple, comme le montre **Figure 11**, de nombreuses lettres ont un signe diacritique en français (é, è, à, ô…). Ce tableau permet aussi de noter que certains phonèmes sont transcrits par deux lettres (par exemple, le /u/ de 'ou', qui se différencie du /y/ de lu, ou les voyelles nasales s'écrivant 'an' (chante) et 'on' (tonton), qui se différencient des voyelles orales correspondantes /a/ et /o/). Ces items seront à considérer dans les évaluations, en fonction de leur fréquence, tout comme les diphtongues (qui n'existent pas en français). Des experts en linguistique seront consultés pour décider ce qu'il faut inclure dans les évaluations en fonction de la langue et de la fréquence des items considérés.

Pour les lettres correspondant à différents phonèmes ('c'/'g' dans *cerise* et *gel* comparés à *car* et *gare*), les deux prononciations sont acceptées. Les personnes en charge de la mise en place d'EGRA sur le terrain examineront attentivement les prononciations possibles de chaque item avec l'aide d'experts en linguistique. Il faudra obtenir un accord sur les réponses acceptables, en tenant compte des variations régionales. Ainsi, comme il est expliqué au chapitre 4, il n'y a que 10 voyelles en français si on ne tient pas compte du 'e muet' de *gare* (prononcé à Marseille, mais pas à Paris) ou de la différence 'un/in' (que les natifs du sud-ouest de la France distinguent, mais pas les parisiens ou les bourguignons, entre autres). Il en est de même pour les différences d'ouverture des voyelles 'a', 'o', 'eu' et 'é/è'… qui sont plus des variantes contextuelles (dépendant de la nature de la syllabe, ouverte ou fermée, cf. *bas* versus balle) que des variations régionales.

Les différences de prononciation doivent être manipulées avec délicatesse dans cette épreuve, comme dans les autres. L'objectif n'est pas d'évaluer la maitrise d'une norme, la prononciation *correcte*. Il est de savoir si l'enfant est capable d'oraliser les différents graphèmes en fonction de la prononciation qui est celle d'une région donnée, par exemple. Les accents régionaux sont donc acceptés.

Recueil des données. Le nombre d'items identifiés correctement en une minute est calculé. Quand la version papier-crayon est utilisée, l'examinateur note les réponses incorrectes par un slash (/), met un crochet (]) après la dernière lettre nommée en 1 minute. Quand l'enfant a passé l'épreuve en moins d'une minute, l'examinateur enregistre le temps restant sur le chronomètre. Dans la version électronique, les calculs sont effectués de façon automatique à partir de ce que l'examinateur a noté sur l'écran de sa tablette. Trois données sont utilisées pour calculer le nombre d'items correctement identifiés en une minute :

clspm = (nombre total de sons des lettres identifiés – nombre total de réponses incorrectes) / [(60 – temps restant) / 60].

Ces différentes données permettent de différencier les élèves qui nomment 50 sons en une minute, mais uniquement la moitié correctement, de ceux qui ne nomment

que 25 sons dans le même temps, mais tous correctement.

Cette épreuve, comme d'autres, est non seulement chronométrée, elle est aussi arrêtée après un temps donnée, qu'elle soit ou non terminée par l'enfant. La prise en compte du temps est nécessaire pour mesurer la fluence. Quant à la limite temporelle, elle rend l'évaluation plus courte et donc moins stressante pour l'enfant et l'examinateur.

Construction de l'épreuve. Cette épreuve comporte 100 items au total, qui sont distribués au hasard, sur 10 lignes avec 10 items par ligne, en mélangeant lettres majuscules et minuscules. La plupart des items sont présentés plusieurs fois, en fonction de leur fréquence dans la langue.[14] Attention, en français le 'h' ne se prononce pas, il ne devrait donc pas être inclus dans la présente évaluation, tout comme le 'q' qui est presque toujours suivi par 'u' (qu). De plus, le 'e' est ambigu : si cette lettre se nomme 'e', elle peut aussi se prononcer 'é/'è'. Il est donc préférable d'éviter cette lettre, pourtant très fréquente, et de ne la présenter qu'en majuscule. Dans ce cas, sauf si 'E' est suivi par 'm'/n', voire par 'u' (Europe, j'ai eu), il se prononce 'é/è'. Les encadrés qui suivent présentent des exemples : un en français (**Figure 12**) et un en wolof (**Figure 13**).

[14] While reordering within rows will limit significant changes in subtask difficulty, it is still recommended to test for order effects whenever possible.

Figure 12. Identification du son des lettres et groupes de lettres (graphèmes)

Epreuve __ Connaissance du son des lettres et groupes de lettres (graphèmes) | 🕮 **Page __** | 🕐 60 secondes

🗣 **Voici une page pleine de lettres et de groupes de lettres. Donne-moi le SON de ces lettres, pas le nom.**
Par exemple, cette lettre [Indiquer le "O" dans la ligne des exemples] **se lit "o" comme dans le mot "pot".**

Pratiquons maintenant : Lis-moi ce groupe de lettres [Montrez le "ou" sur la rangée des exemples] **:**

[Si l'élève répond correctement, dites :] **Très bien, ce groupe de lettres se lit "ou" comme dans le mot "cour".**

[Si l'élève ne répond pas correctement, dites :] **Non, ce groupe de lettres se lit "ou" comme dans le mot "cour".**

Essayons un autre exemple. Lis-moi cette lettre : [Montrez le "t" sur la rangée des exemples] :

[Si l'élève répond correctement, dites :] **Très bien, cette lettre se lit "t" comme dans le mot "table".**

[Si l'élève ne répond pas correctement, dites :] **Non, cette lettre se lit "t" comme dans le mot "table".**

Essayons encore un autre exemple. Lis-moi ce groupe de lettres : [Montrez le "ch" sur la rangée des exemples] :

[Si l'élève répond correctement, dites :] **Très bien, ce groupe de lettres se lit "ch" comme dans le mot "chat".**

[Si l'élève ne répond pas correctement, dites :] **Non, ce groupe de lettres se lit "ch" comme dans le mot "chat".**

Lorsque je dis "Commence", commence à lire ici [montrez lui le premier graphème] **et lis de gauche à droite** [faites glisser votre doigt vers la droite]**, ligne par ligne en parlant fort pour que je puisse t'entendre. Montre chaque lettre du doigt quand tu les lis. Essaye de lire rapidement et correctement. Si tu n'arrives pas à lire une des lettres, continue et lis celle qui suit. Mets ton doigt sur la première lettre. Tu es prêt(e) ? Commence.**

Faites démarrer le chronomètre quand l'élève lit la première lettre.

➲ Si l'élève ne répond pas et reste bloqué sur un graphème pendant plus de 3 secondes, dites-lui, "Continue", en lui montrant le prochain graphème.

✋ Quand le chronomètre sonne et atteint 0, demandez à l'élève de s'arrêter

✋ Si l'élève ne donne aucune réponse correcte parmi les dix premiers graphèmes (la première ligne), demandez-lui gentiment de s'arrêter, et cocher la case "auto-stop".

✍ (/) Barrez l'item si l'élève a donné une réponse incorrecte, ou n'a pas donné de réponse.
(Ø) Entourez les autocorrections lorsque vous avez déjà barré l'item.
( ]) Mettez un crochet après le dernier item lu par l'élève

Exemples **O** **ou** **T** **ch**

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
E	i	F	O	A	E	ch	S	z	ou	(10)
B	N	on	s	I	M	L	an	G	T	(20)
W	O	G	ou	L	T	j	C	p	M	(30)
V	K	A	R	U	F	é	J	s	b	(40)
S	L	C	an	D	Y	f	I	a	E	(50)
I	s	U	p	M	V	oi	T	n	P	(60)
Z	un	E	g	in	F	d	O	an	v	(70)
D	é	B	A	m	On	T	C	o	r	(80)
R	L	qu	B	E	N	i	A	p	ou	(90)
gn	E	ch	V	D	U	ç	oi	m	x	(100)

✍ Nombre exact de secondes restantes indiquées sur le chronomètre

✍ Cochez ici si l'épreuve a été arrêtée par manque de réponses correctes sur la première ligne (auto-stop) :

Merci bien ! On peut passer à la prochaine activité !

Figure 13. Identification des lettres et suites de lettres (graphèmes) en wolof (début de l'épreuve)

A	n	o	m	i	U	D	e	L	k
g	B	y	a	uu	X	T	w	N	f
S	r	c	J	P	Ñ	Oo	i	a	m
n	k	aa	R	ã	Ee	Ée	d	U	R

Une épreuve peut avoir plusieurs versions parallèles, par exemple, pour mesurer la progression des enfants entre différentes périodes de test. Dans ce cas, les items sont réorganisés (de façon aléatoire), pour créer une nouvelle épreuve ayant toujours 10 lignes et 10 items par ligne. Toutefois, et afin de maintenir un même niveau de difficulté entre les différentes versions de l'épreuve, les réorganisations ne doivent être effectuées que dans une même ligne. En d'autres termes, chaque item reste dans la même ligne quelle que soit la version de l'épreuve. Il est toutefois recommandé de vérifier, autant que faire se peut, les biais pouvant provenir d'un changement dans l'ordre des items.

Seconde approche : Identifier le nom des lettres

Dans la mesure où la recherche a montré que la connaissance du nom des lettres est également un bon prédicteur du futur niveau de lecture, tout au moins dans des langues qui ont une orthographe peu transparente, cette connaissance peut être également évaluée (Loningan et al., 2002). Il faut toutefois garder en mémoire le fait que, dans une étude avec des enfants français, le résultat précédent s'expliquait uniquement par la connaissance du nom des lettres pour les voyelles, pas pour les consonnes, donc quand le nom et le son des lettres coïncident (cf. la section 4.4.3 du chapitre 4 et Piquard-Kipffer & Sprenger-Charolles, 2013). De plus, le nom des lettres peut prêter à confusion (par exemple, il peut conduire à écrire *go* pour *géo* ou *pi* pour pays). Ce fait avait été relevé au 17ème siècle dans la *Grammaire de Port-Royal* (Arnaud & Lancelot, 1660). En outre, dans de nombreux pays dans lesquels EGRA a été administré, cette épreuve a donné lieu à des effets *plafond* (presque tous les enfants ont des scores élevés). En conséquence, elle a été souvent remplacée par d'autres.

Recueil des données et construction de l'épreuve. Cette épreuve est semblable dans sa structure et son administration à celle qui évalue la connaissance du son des lettres. La principale différence est qu'il faut donner le nom des lettres : les items sont donc uniquement des lettres de l'alphabet.

Les données recueillies sont, comme pour l'épreuve précédente, le nombre de lettres correctement nommées en une minute. Egalement comme pour l'épreuve précédente, la feuille de passation contient 100 items, 10 par ligne sur 10 lignes, en minuscule et en majuscule. Les lettres sont présentées une ou plusieurs fois, en fonction de leur fréquence dans la langue (cf. pour la fréquence des lettres en français sur le **Figure 12**). Les lettres d'une même ligne peuvent être réorganisées en vue de la préparation de différentes versions équivalentes de cette épreuve. Il faut toutefois contrôler les effets de l'ordre des items.

6.2.3 Lecture de mots inventés (pseudomots)

La lecture de pseudomots permet d'évaluer les capacités de décodage, celles qui correspondent à l'utilisation de la procédure sublexicale (ou phonologique) de lecture. Cette procédure de lecture se distingue de la procédure lexicale, qui est utilisée pour lire les mots familiers. Les enfants peuvent, dans les premières années,

reconnaitre globalement quelques mots. Cette stratégie, qui repose sur l'utilisation d'un vocabulaire global appris par cœur, a cependant des limites (Hirsch, 2003 ; voir aussi les synthèses du NELP, 2008 et de l'INSERM, 2007). De plus, de nombreuses études ont montré l'importance du décodage dans les débuts de l'apprentissage de la lecture (voir ci-dessus la section 4.3). En effet, les lecteurs débutants utilisent quasi-exclusivement le décodage, cette compétence permettant de distinguer les enfants en difficultés de lecture des autres enfants, quel que soit leur âge ou leur niveau scolaire.

Les travaux de recherche ont également montré que les capacités initiales de décodage permettent de prédire le futur niveau de lecture, y compris à long terme (Share, 1995). Elles permettent même de prédire la lecture de mots irréguliers comme sept ou femme (Ouellette & Beers, 2010 ; Sprenger-Charolles et al., 2003). Ces résultats s'expliquent si l'on admet que les enfants utilisent d'abord la procédure phonologique de lecture et que des connexions vont progressivement se créer entre unités orthographiques et phonologiques. L'établissement de ces connexions dépend de la régularité des correspondances graphème-phonème et de la fréquence des mots (Ziegler et al., 2014). C'est la raison pour laquelle l'évaluation des capacités de décodage est une des quatre épreuves principales d'EGRA.

Données recueillies. Le score final pris en compte est le nombre de pseudomots lus correctement en une minute. Comme pour les autres épreuves de même nature, trois scores sont notés sur papier par l'examinateur (ou engrangés dans la tablette pour la version électronique du test) : le nombre total d'items lus par l'enfant ; le nombre de réponses incorrectes ; et, quand l'enfant a effectué cette épreuve en moins d'une minute, le temps restant.

Construction de l'épreuve. Cette épreuve contient 50 items d'une à deux syllabes (suivant la langue), cinq par ligne. Ces items doivent respecter les orthographes possibles dans la langue. Par exemple, pour les épreuves en français, il n'y aura pas de double lettre en position initiale (sauf le 'w', très rare dans cette langue et qui est une lettre de l'alphabet) et pas de 'j' ou de 'v' en fin de mot. De plus, les mots inventés seront surtout des items facilement décodables (comme sar, bir, moudir). En outre, les items respecteront les structures syllabiques les plus simples, et les plus fréquentes, de cette la langue (CV, CCV, CVCV…). Enfin, il ne faut pas utiliser des items qui se prononcent comme un mot (roze, rouje…).

Comme pour l'épreuve précédente, la police de caractère doit être bien lisible, avec des espacements corrects entre mots. Egalement comme pour l'épreuve précédente, les items d'une même ligne peuvent être présentés dans des ordres différents lorsque plusieurs versions de la même épreuve doivent être mises en place. Il faut toutefois bien vérifier les biais pouvant être introduits par une modification aléatoire de l'ordre des mots (**figure 14)**.

6.2.4 Compréhension écrite et fluence

Figure 14. Epreuve de lecture de pseudomots

Epreuve __ Lecture de mots inventés | **Page __** | 60 secondes

Voici une page avec des mots inventés qui ressemblent à des mots FRANCAIS. Essaye de lire autant de mots que tu peux. Il ne faut pas dire les lettres mais lire le mot. Par exemple, ce premier mot [montrez le mot "bi"] **se lit "bi".**

Essayons. Peux-tu lire ce mot ? [montrez le mot "tuk" avec le doigt.]
[Si l'élève lit correctement dites :] **Très bien, ce mot se lit "tuk".**
[Si l'élève ne lit pas correctement, ou après 3 secondes de non-réponse, dites :] **Ce mot se lit "tuk".**
Essayons. Peux-tu lire ce mot ? [montrez le mot "sar" avec le doigt.]
[Si l'élève lit correctement dites :] **Très bien, ce mot se lit "sar".**
[Si l'élève ne lit pas correctement, ou après 3 secondes de non-réponse, dites :] **Ce mot se lit "sar".**

Lorsque je dis "Commence", commence à lire ici [montrez lui le premier mot] **et lis de gauche à droite** [faites glisser votre doigt vers la droite]**, ligne par ligne en parlant fort pour que je puisse t'entendre. Montre chaque mot du doigt quand tu le lis. Essaye de lire rapidement et correctement. Si tu n'arrives pas à lire un des mots, continue et lis celui qui suit. Mets ton doigt sur le premier mot. Tu es prêt(e) ? Commence.**

Faites démarrer le chronomètre quand l'élève lit le premier mot.

Si l'élève ne répond pas et reste bloqué sur un mot pendant plus de 3 secondes, dites-lui, "Continue", en lui montrant le prochain mot.

Quand le chronomètre sonne et atteint 0, demandez à l'élève de s'arrêter

Si l'élève ne donne aucune réponse correcte parmi les 5 premiers mots (la première ligne), demandez-lui gentiment de s'arrêter, et cocher la case "auto-stop".

*Les items soulignés contiennent un graphème ayant une prononciation dépendant du contexte ('s', 'c', 'g')

(/) Barrez l'item si l'élève a donné une réponse incorrecte, ou n'a pas donné de réponse.
(Ø) Entourez les autocorrections lorsque vous avez déjà barré l'item.
( ]) Mettez un crochet après le dernier item lu par l'élève

Exemples :* bi tuk sar

1	*2*	*3*	*4*	*5*	
ja	zi	vaf	tal	ol	(5)
sar	vor	ul	da	iko	(10)
biga	neul	ima	plovi	bilba	(15)
tipa	osi	flir	blu	toche	(20)
saré	nur	duse	rané	pro	(25)
mouli	chane	bape	cla	doupé	(30)
til	tandé	doul	zopé	nube	(35)
donré	dreu	ibrau	réce	lorpe	(40)
oti	neau	bir	nogir	moudir	(45)
bair	zode	nour	lépa	fipe	(50)

Nombre exact de secondes restantes indiquées sur le chronomètre

Cochez ici si l'épreuve a été arrêtée par manque de réponses correctes sur la première ligne (auto-stop):

Merci bien ! On peut passer à la prochaine activité !

Comprendre un texte écrit est le résultat d'une interaction réussie entre différents facteurs (Snow et le RAND Reading Study Group, 2002 ; Gaonac'h & Fayol [Eds], 2003) : en plus de la motivation, de l'attention et de la mémoire, d'un côté les connaissances de la langue et du monde nécessaires pour comprendre la thématique du texte, de l'autre, un accès automatique aux mots écrits. La compréhension peut s'évaluer non seulement par des questions, mais aussi par la

fluence en lecture à haute voix d'un texte, souvent mesurée par le nombre de mots correctement lus en une minute. Cette dernière mesure est fortement corrélée à des tests de compréhension, par exemple à l'épreuve de compréhension du Stanford (0,91, Fuchs et al., 2001).

Données recueillies. Deux scores principaux sont recueillis : le nombre de réponses correctes aux questions posées après la lecture (le score final étant le pourcentage de réponses correctes) et un score de fluence. Pour la fluence, comme pour l'épreuve de lecture de pseudomots (et les autres épreuves en temps limité) trois sous-scores sont notés : le nombre total de mots lus par l'enfant ; le nombre de réponses incorrectes ; et, quand l'enfant a effectué cette épreuve en moins d'une minute, le temps restant.

Construction de l'épreuve (voir l'encadré 6.4 pour un exemple). L'épreuve utilise le plus souvent un récit, c'est-à-dire un texte structuré, et non une suite de phrases vaguement connectées. Le récit contient généralement trois sections : la première présente les personnages de l'histoire, la seconde introduit une situation problématique et la troisième dévoile la solution qui a permis de résoudre le problème. Il faut éviter d'utiliser des histoires connues par les enfants, par exemple, celles fréquemment reprises dans les manuels scolaires. Cela s'explique par le fait qu'il est alors possible de répondre sans avoir lu le texte. Les noms des personnages, tout comme celui des lieux, doivent être typiques de la langue et de la culture locale mais, pour éviter les problèmes de mémoire, il ne faut que peu de personnages. De plus, la longueur du texte doit être d'environ 60 mots afin de permettre de poser trois à cinq questions de compréhension. Enfin, la police de caractères doit être familière aux enfants avec un espace correct entre les mots et les lignes du texte. Attention ! Le texte ne doit pas être accompagné par des images.

Les questions doivent être semblables à celles posées pour évaluer la compréhension orale. Pour la plupart d'entre elles, elles portent sur des informations présentées explicitement dans le texte (questions littérales). Au moins une de ces questions exige de faire des déductions (des inférences) basées sur les connaissances de l'enfant et sur ce qui est écrit dans le texte. Pour ces dernières, plusieurs réponses sont parfois possibles : le protocole précisera celles qui peuvent être marquées comme étant correctes. A noter, il faut éviter les questions auxquelles il est possible de répondre sans avoir lu le texte. Il faut aussi éviter des questions avec seulement des réponses *oui* ou *non.* (**Figure 15**)

Lorsque des formes équivalentes de cette épreuve doivent être créées pour plusieurs utilisations du même instrument dans la même langue (par exemple, au début, au milieu et à la fin d'une étude longitudinale), il est recommandé de faire des changements ne modifiant pas le niveau de difficulté du texte. Par exemple, les noms des personnages et ceux des lieux ou des actions peuvent être changés, ainsi que les adjectifs les décrivant. Ce qui est remplacé doit toutefois l'être par des alternatives de niveau de difficulté similaire.

Figure 15. Epreuve de compréhension écrite et de fluence

Epreuve __ Lecture et compréhension du texte (petite histoire) | Page __ | 60 secondes

Voici une petite histoire. Lis la à haute voix en essayant de lire rapidement et correctement ; après, je vais te poser quelques questions sur l'histoire. Lorsque je dis "Commence", tu commenceras à lire. Si tu vois un mot que tu ne sais pas lire, essaye le prochain. Mets ton doigt sur le premier mot. Tu es prêt(e)? Commence

Faites démarrer le chronomètre quand l'élève lit le premier mot.

(/) Barrez l'item si l'élève a donné une réponse incorrecte, ou n'a pas donné de réponse.
(Ø) Entourez les autocorrections lorsque vous avez déjà barré l'item.
( ]) Mettez un crochet après le dernier item lu par l'élève

Quand l'élève a fini de lire, RETIREZ le passage de sa vue.

Posez les questions qui correspondent aux lignes du texte jusqu'à la ligne à laquelle se trouve le crochet ( ]), c'est-à-dire, jusqu'à l'endroit où l'élève a cessé de lire. Si l'élève ne donne aucune réponse après 10 secondes, passez à la question suivante. Ne répétez pas les questions.

Maintenant, je vais te poser quelques questions sur l'histoire. Essaye de répondre aux questions du mieux possible. Tu peux répondre dans la langue que tu préfères.

Quand le chronomètre sonne et atteint 0, demandez à l'élève de s'arrêter

Si l'élève ne répond pas et reste bloqué sur un mot pendant plus de 3 secondes, dites-lui, "Continue", en lui montrant le prochain mot.

Si l'élève ne donne aucune réponse correcte sur la première ligne ne posez aucune question.

Si l'élève dit "Je ne sais pas", considérez la réponse comme incorrecte.

	Nombre de mots cumulés	Questions [Réponses]	Correct	Incorrect	Pas de Réponse
Flore est à la maison. Elle termine ses devoirs.	09	Où est Flore ? [A la maison]			
Son petit frère Luc arrive. Il a trois ans.	18	Quel est le nom du frère de Flore ? [Luc]			
Flore et Luc jouent aux billes.	24	Que font les enfants ensemble ? [Jouer aux billes]			
Soudain, Luc met une bille dans sa bouche.	32	Luc met quoi dans sa bouche ? [une bille]			
Mais Flore enlève vite l'objet de la bouche de son frère. Le petit Luc est sauvé ! Leur mère vient auprès d'eux.	53	Qui vient auprès des enfants ? [leur mère]			
Elle prend Luc dans ses bras et remercie Flore.	62	Pourquoi la mère de Flore la remercie ? [Elle a sauvé Luc]			

Nombre exact de secondes restantes indiquées sur le chronomètre

Cochez ici si l'épreuve a été arrêtée par manque de réponses correctes sur la première ligne (auto-stop):

Merci bien ! On peut passer à la prochaine activité !

6.2.5 Epreuves de conscience phonémique

Comme il est expliqué au chapitre 4 (Cadre conceptuel), la **conscience phonémique** est la capacité d'identifier et de manipuler les plus petites unités sans signification de la langue orale, les phonèmes. C'est une composante de la conscience phonologique, qui est la capacité d'identifier et de manipuler les unités sans signification de la langue orale, quelle que soit leur taille : de la syllabe, et ses composants (attaque et rime), au phonème. La conscience phonémique est nécessaire pour apprendre à lire dans une écriture alphabétique qui nécessite de décoder les mots écrits en utilisant les correspondances entre les plus petites unités de la langue écrite (les graphèmes) et celles de la langue orale (les phonèmes).

De nombreuses études longitudinales, dans lesquelles les enfants sont suivis le plus souvent depuis une période qui précède l'apprentissage de la lecture, ont montré que les relations entre conscience phonémique et apprentissage de la lecture sont bidirectionnelles. D'un côté, l'apprentissage de la lecture favorise le développement de la conscience phonémique ; de l'autre, le niveau de conscience phonémique avant cet apprentissage est un bon prédicteur du succès ou de l'échec de cet apprentissage (Perfetti et al., 1987 ; en français, voir Casalis & Louis-Alexandre, 2000). Ces résultats ont été relevés chez des apprenants de langue seconde, comme chez ceux de langue maternelle (August & Shanahan, 2006).

Une forte incidence de la conscience phonémique sur la lecture a été observée dans différentes écritures alphabétiques, cette incidence étant toutefois plus marquée dans celles qui ont une orthographe opaque (Ziegler et al., 2010). Enfin, quelques études signalent que les capacités de discrimination phonémique (être capable de distinguer vol de bol, par exemple) permettent de prédire le devenir en lecture des enfants de façon fiable et de distinguer les faibles lecteurs des bons lecteurs (Ziegler et al., 2009). Il est donc crucial d'évaluer ces différentes capacités dans un bilan des premières étapes de l'apprentissage de la lecture.

Les mesures de conscience phonémique le plus souvent utilisées dans EGRA sont l'identification ou la discrimination des sons (phonèmes) à l'initiale d'un mot d'une seule syllabe (pour éviter les découpages syllabiques). Ces mesures sont présentes dans de nombreux tests destinés à évaluer les capacités précoces de lecture, par exemple :

- Indicateurs dynamiques de compétences de base d'alphabétisation précoce (DIBELS) https://dibels.uoregon.edu/ ; voir aussi Measurement Group Dynamic : https://dibels.org/
- Test conscience phonologique, Second Edition Plus (TOPA-2+) : https://www. linguisystems.com/products/product/display?itemid=10293
- Batterie évaluant les traitements phonologiques du CTOPP-2 : http://www. pearsonclinical.com/language/products/100000737/comprehensive- test-of-phonological-processing-second-edition-ctopp-2-ctopp-2.html#tab-details.

Première approche : Identification du son initial d'un mot

Un moyen d'évaluer la conscience phonémique est de demander aux élèves d'identifier le premier (ou dernier) son de mots familiers. L'exemple présenté dans l'**Figure 16** utilise 10 mots monosyllabiques, les enfants devant identifier le son initial de chacun d'eux. L'examinateur lit chaque mot à haute voix deux fois avant de demander à l'élève d'identifier son premier son. Ce type d'épreuve pouvant être inconnu des enfants, le protocole comprend des essais.

Données recueillies. L'examinateur recueille le nombre de réponses correctes. L'épreuve n'est pas chronométrée et il n'y a pas de feuille de passation destinée à l'élève.

Construction de l'épreuve. Des mots simples sont choisis parmi ceux trouvés dans les manuels des deux premiers grades du primaire. Autant que faire se peut, il est préférable d'utiliser des monosyllabiques (afin d'éviter les découpages syllabiques).

Seconde approche : Discrimination du son initial d'un mot

Dans ce type d'épreuve on demande à l'élève d'écouter une série de trois mots et d'identifier celui qui commence avec un son différent des deux autres (épreuve aussi dite de chasse à l'intrus). Une épreuve typique implique 10 ensembles de trois mots.

Dans chacun, deux mots commencent par le même son, le troisième étant différent. La position du mot différent varie d'une série à l'autre (par exemple, mur, pur, par ; pour, four, peur ; bal, bol, mal). L'examinateur lit chaque ensemble de trois mots à haute voix lentement, en les répétant une fois, et il demande à l'enfant de choisir le mot qui commence par un son différent. Parce que ce type d'épreuve peut être inconnu des enfants, le protocole comprend des essais.

Données et construction de l'épreuve. L'examinateur enregistre le nombre de réponses correctes (l'épreuve n'est pas chronométrée). Des mots simples sont choisis parmi ceux trouvés dans les manuels des deux premiers grades. Il est préférable d'utiliser des monosyllabiques afin d'éviter des découpages syllabiques et pour ne pas surcharger la mémoire.

Figure 16. Conscience phonémique – Identification du premier son d'un mot

Epreuve _ IDENTIFICATION DU SON INITIAL 🕮 X | 🕐 X

Cette épreuve est une épreuve orale. Je vais te dire un mot deux fois, puis je veux que tu me dises le tout premier son du mot que tu entends, d'accord ?
Par exemple : Le mot "soupe" commence avec le son "sssss", n'est-ce pas ? Je dirai chaque mot deux fois et tu me diras le tout premier son de chaque mot.

Essayons encore quelques exemples :
Quel est le tout premier son dans le mot "chic" ? "chic" ?
[Si l'élève répond correctement, dites-lui] **Très bien ! Le premier son dans le mot "chic", c'est "ch"**
[Si l'élève ne répond pas correctement, dites-lui] **Le premier son dans le mot "chic", c'est "ch"**
Quel est le tout premier son dans le mot "poule" ? "poule" ?
[Si l'élève répond correctement, dites-lui] **Très bien ! Le premier son dans le mot "poule", c'est "p"**
[Si l'élève ne répond pas correctement, dites-lui] **Le premier son dans le mot "poule", c'est "p".**

Tu es prêt(e) ? Commence.

Seul le son prononcé isolément est correct. Cochez la case correspondant à la réponse de l'élève. En cas de non-réponse, après 3 secondes cochez la case "Pas de réponse" et passez au prochain item.

Quel est le tout premier son dans le mot "__" ? "__" ? *[Répétez le mot deux fois*

dur	/d/	o Correct	o Incorrect	o Pas de réponse
lac	/llllll/	o Correct	o Incorrect	o Pas de réponse
car	/k/	o Correct	o Incorrect	o Pas de réponse
sac	/ssssss/	o Correct	o Incorrect	o Pas de réponse
jour	/jjjjj/	o Correct	o Incorrect	o Pas de réponse
fil	/ffffff/	o Correct	o Incorrect	o Pas de réponse
tour	/t/	o Correct	o Incorrect	o Pas de réponse
balle	/b/	o Correct	o Incorrect	o Pas de réponse
par	/p/	o Correct	o Incorrect	o Pas de réponse
vol	/vvvv/	o Correct	o Incorrect	o Pas de réponse

Cochez ici si l'épreuve a été arrêtée par manque de réponses correctes parmi les 5 premiers items (auto-stop):

Cette épreuve n'est pas chronométrée. Retirez le cahier de stimuli de la vue de l'élève.

➲ Lisez les instructions à l'élève et donnez-lui les exemples.

Lisez les instructions et prononcez les items deux fois. Prononcez chaque item lentement.

Si l'élève ne donne aucune réponse correcte parmi les 5 premiers mots, demandez-lui gentiment de s'arrêter, et cocher la case "auto-stop". Passez à l'épreuve suivante.

6.2.6 Lecture de mots familiers

La lecture de mots isolés est une mesure plus précise des capacités de reconnaissance des mots et des compétences de décodage que la lecture de mots en contexte (les enfants ne pouvant pas deviner le mot suivant à partir du contexte, cf. la section 4.3 du chapitre 4 pour une discussion). Les mots doivent être fréquents et présents dans les manuels (ou autres livres) destinés aux enfants des premiers grades du primaire.

Données recueillies. Comme pour les autres épreuves chronométrées, trois scores sont notés sur papier par l'examinateur (ou sauvegardés dans la tablette pour la version électronique du test) : le nombre total d'items lus par l'enfant ; le nombre de réponses incorrectes ; et, quand l'enfant a effectué cette épreuve en moins d'une minute, le temps restant.

Construction de l'épreuve. Les mots pour cette épreuve sont sélectionnés dans des manuels scolaires destinés aux grades qui seront l'objet de l'évaluation, en fonction de leur fréquence et des caractéristiques de l'orthographe de la langue dans laquelle cette épreuve est passée. La prononciation des mots sélectionnée ne doit pas être ambigüe. Les mots choisis ne doivent pas faire partie du vocabulaire d'une autre langue connue des enfants. Cette épreuve ne doit également pas inclure des mots d'une lettre, ceux-ci étant examinés dans l'épreuve d'identification des lettres.

La liste des mots doit inclure 50 mots familiers qui peuvent représenter les différentes parties du discours (par exemple, noms, verbes, adjectifs). Les mots sont disposés sur 10 lignes (cinq par ligne) avec un espacement approprié, en minuscule. Comme pour les pseudomots, les mots d'une même ligne peuvent être présentés dans différents ordres quand plusieurs versions de la même épreuve doivent être mises en place. Il faut toutefois bien vérifier les biais pouvant être introduits par une modification aléatoire de l'ordre des mots. Trois mots supplémentaires (qui doivent être semblables en niveau de difficulté aux mots de la liste) servent d'exemple. (**Figure 17**)

La police utilisée pour cette épreuve sera similaire à celle utilisée dans les autres épreuves de lecture (voir aussi la discussion sur cette question dans la section 6.2 dédiées à l'identification des lettres).

Figure 17. Epreuve de lecture de mots familiers

Epreuve _. Lecture de mots familiers | Page __ | 60 secondes

Voici une page avec des mots en FRANCAIS. Essaye de lire autant de mots que tu peux. Il ne faut pas dire les lettres mais lire le mot. Par exemple, ce premier mot [montrez le mot "ta"] **se lit "ta".**

Essayons. Peux-tu lire ce mot ? [montrez le mot "par" avec le doigt.]
[Si l'élève lit correctement dites :] **Très bien, ce mot se lit "par".**
[Si l'élève ne lit pas correctement, ou après 3 secondes de non-réponse, dites :] **Ce mot se lit "par"**

Essayons. Peux-tu lire ce mot ? [montrez le mot "lune" avec le doigt.]
[Si l'élève lit correctement dites :] **Très bien, ce mot se lit "lune".**
[Si l'élève ne lit pas correctement, ou après 3 secondes de non-réponse, dites :] **Ce mot se lit "lune"**

Lorsque je dis "Commence", commence à lire ici [montrez lui le premier mot] **et lis de gauche à droite** [faites glisser votre doigt vers la droite], **ligne par ligne en parlant fort pour que je puisse t'entendre. Montre chaque mot du doigt quand tu le lis. Essaye de lire rapidement et correctement. Si tu n'arrives pas à lire un des mots, continue et lis celui qui suit. Mets ton doigt sur le premier mot. Tu es prêt(e) ? Commence.**

Faites démarrer le chronomètre quand l'élève lit le premier mot.

Si l'élève ne répond pas et reste bloqué sur un mot pendant plus de 3 secondes. dites-lui, "Continue", en lui montrant le prochain mot.

Quand le chronomètre sonne et atteint 0, demandez à l'élève de s'arrêter

Si l'élève ne donne aucune réponse correcte parmi les 5 premiers mots (la première ligne), demandez-lui gentiment de s'arrêter, et cocher la case "auto-stop".

(/) Barrez l'item si l'élève a donné une réponse incorrecte, ou n'a pas donné de réponse.
(Ø) Entourez les autocorrections lorsque vous avez déjà barré l'item.
( ]) Mettez un crochet après le dernier item lu par l'élève

Exemples:* ta par lune

1	2	3	4	5	
tu	il	vol	sa	ma	-5
ou	or	lire	ami	par	-10
sol	peur	papa	sage	bébé	-15
tarte	cri	vache	blé	fleur	-20
sur	chaise	peau	vole	bleu	-25
mil	mur	table	cil	monde	-30
fin	date	tour	posé	kilo	-35
ronde	pré	garde	faire	porter	-40
été	beau	pain	rougir	moto	-45
mal	douze	six	vélo	vide	-50

*Les items soulignés contiennent un graphème ayant une prononciation dépendant du contexte ('s', 'c', 'g') ou irrégulière (six)

Nombre exact de secondes restantes indiquées sur le chronomètre

Cochez ici si l'épreuve a été arrêtée par manque de réponses correctes sur la première ligne (auto-stop) :

Merci bien ! On peut passer à la prochaine activité !

6.3 Evaluations complémentaires et épreuves supprimées

Comme indiqué ci-dessus, d'autres épreuves moins couramment utilisées ont été créées et évaluées dans des études pilotes d'EGRA. Ces épreuves répondent à des besoins spécifiques de certaines langues, à des questions de recherche spécifiques, ou encore à des questions liées au curriculum. Elles sont décrites brièvement ci-dessous.

6.3.1 Dictée

Des épreuves de dictée faisaient partie de la batterie de base dans les premières versions d'EGRA. Elles n'ont pas été maintenue dans la batterie de base en raison de difficultés de mise en œuvre (voir ci-dessus la section 4.4.1). Comme il est indiqué dans le chapitre 4 (section 4.4.1) la dictée peut être utilisée pour évaluer l'orthographe lexicale (par la dictée de mots) et grammaticale (par la dictée de phrases), mais évaluer également la capacité de discriminer les phonèmes. Nous présentons donc ci-dessous une épreuve d'écriture de mots (proche de celle qui était dans la première version d'EGRA en français, dans la section 9 de l'**annexe B**) ainsi que deux autres suggestions : une pour la dictée de phrase et une qui peut constituer une alternative pour l'évaluation des capacités de discrimination des phonèmes (cf. 6.3.2).

Une épreuve d'orthographe lexicale peut figurer parmi les épreuves supplémentaires. Celle présentée dans l'**Figure 19** en créole haïtien comporte cinq lettres et trois mots.

Figure 19. Epreuve de dictée de lettres et de mots en créole haïtien

K-Seksyon 8 : Dictée	Fèy papye ak kreyon	×
	×	×

(✓) Korèk / pa korèk / pa gen repons ditou

A. Lèt

Bay timoun nan yon kreyon ak yon fèy papye. Pa kite l gade lèt yo. Si timoun nan di : « Mwen pa konnen, » make repons sa a kòm enkòrèk.

Mwen pra l di w kèk lèt. Se pou koute m avèk atansyon. Apre chak lèt mwen fin di w, m ap repete l yon lòt fwa pou ou, e w ap ekri lèt ou tande a sou papye a pou mwen. Eske w konprann sa m mande w fè a ? Oke, koute epi ann kòmanse.

[Li chak let 2 fwa]			
b	○ Kòrèk	○ Pa Korèk	○ Pa gen repons
j	○ Kòrèk	○ Pa Korèk	○ Pa gen repons
m	○ Kòrèk	○ Pa Korèk	○ Pa gen repons
v	○ Kòrèk	○ Pa Korèk	○ Pa gen repons
z	○ Kòrèk	○ Pa Korèk	○ Pa gen repons

Kounye a, mwen pra l di w kèk mo. Koute m avèk atansyon. Apre chak mo mwen fin di w, m ap repete l yon lòt fwa pou ou, e w ap ekri mo ou tande a sou papye a pou mwen. Eske w konprann sa m mande w fè a ? Oke, koute epi ann kòmanse.

B. Mo

[Li chak mo 2 fwa]			
fil	○ Kòrèk	○ Pa Korèk	○ Pa gen repons
ten	○ Kòrèk	○ Pa Korèk	○ Pa gen repons
pay	○ Kòrèk	○ Pa Korèk	○ Pa gen repons

Anfòm! Ou pare pou w fè pwochen aktivite a. Trè byen !

Une épreuve de dictée de phrase est également possible en tant qu'évaluation complémentaire. En français, cette épreuve peut être importante en raison de la spécificité de la morphologie flexionnelle du français (souvent non marquée à l'oral), mais une telle épreuve ne serait appropriée qu'après quelques années d'enseignement en français. Elle pourrait aussi se présenter comme un exercice à trou (les parties en gras n'étant pas sur la feuille donnée à l'enfant) ; Je sais que tu **as** trois **petits frères,** le plus **grand** a cinq **ans** et le plus **petit** a deux **ans**.

6.3.2 Autres évaluations de la conscience phonémique

La capacité de segmenter les mots en phonèmes, qui est un des prédicteurs de la réussite future de l'apprentissage de la lecture, est cependant une compétence complexe (Linan-Thompson & Vaughn, 2007). Dans l'épreuve utilisée pour évaluer cette compétence dans la première version d'EGRA, l'examinateur lisait à haute voix une liste de 10 mots simples n'ayant qu'une syllabe, un à la fois. Les élèves devaient prononcer les sons contenus dans chacun d'eux. L'examinateur enregistrait le pourcentage des phonèmes correctement identifiés par l'élève. L'épreuve n'était pas chronométrée. Les premières évaluations ont montré que cette épreuve était difficile pour l'examinateur, tout comme pour les élèves (peu d'entre eux se sont avérés capables de répondre correctement). L'épreuve a donc été retirée de la liste de base de la batterie EGRA. (**Figure 20**)

Figure 20. Epreuve de segmentation phonémique

Consignes pour l'élève : Je vais dicter des mots inventés, des mots qui n'existent pas. Tu vas essayer de les écrire le mieux possible. Tu es prêt ? Mots à dicter : ti/ti, da/da, bir/bir, tal/tal, pra/pra.

Consignes pour l'examinateur : Vous devez dicter les mots un par un, lentement, en laissant à l'élève le temps d'écrire ce que vous dictez. Vous répétez chaque mot deux fois. Il faut arrêter l'épreuve si l'élève n'a rien écrit pour les deux premiers items. Vous corrigez cette épreuve à l'aide de la fiche ci-dessous. Vous devez noter les lettres correctes dans chacun des mots : 1 point par lettre correcte.

		Nombre de points
1. ti /2 :	t + i/y	/ 2
2. da /2 :	d + a	/ 2
Arrêt si l'élève n'a rien écrit et entourer la croix : x		
3. bir /3 :	b + i/y + r/re	/ 3
4. vra /3 :	v + r + a	/ 3
Total		_ / 10

6.3.3 Evaluation de la compréhension par un exercice à trou

Les exercices à trou sont souvent utilisés en classe pour évaluer la compréhension. Ils impliquent la création d'un texte court qui peut être narratif ou non. Dans certains d'entre eux, l'enfant doit retrouver le mot (ou la suite de mots) qui manque parmi trois qui lui sont proposés. Dans d'autres, il peut choisir lui-même le mot manquant (cette version est plus difficile à mettre en place et à évaluerque la précédente). L'enfant peut lire silencieusement ou à haute voix et doit sélectionner (ou trouver) le mot qui convient (qui, de préférence, ne doit pas être le premier d'une phrase). Ce type d'épreuve n'est pas chronométré. Toutefois, un temps limite est souvent proposé (généralement 3 à 5 minutes).

Données recueillies. Les réponses sont évaluées sur la base d'une clé indiquant celles qui sont acceptables.

Construction de l'épreuve. La première et la dernière phrase du texte sont complètes. Dans chaque phrase, seul un mot est supprimé (de préférence, pas le premier). Dans quelques cas, la partie supprimée est une suite de mots. Trois choix sont proposés pour remplacer ce qui manque. Les options proposées sont pour certaines plausibles, mais une seule est correcte. Par exemple, dans la phrase « Dans l'eau du lac, il a vu un poisson qui (nageait / creusait / nageoire) », le nom nageoire est à exclure et, bien qu'il soit possible que le poisson creuse dans le sable, la réponse la plus plausible dans ce cas est nageait.

6.3.4 Evaluation du vocabulaire et autres évaluations

Au cours du développement de l'instrument original, sur la base de la revue de la littérature et de suggestions d'experts, il a été proposé d'inclure une épreuve utilisant un choix d'image pour évaluer le vocabulaire (comme dans le Peabody, qui a été adapté au française (EVIP [Echelle de vocabulaire en image du Peabody], Dunn, Thériault-Whalen, & Dunn, 1993).

Les développeurs EGRA n'ont pas retenu cette proposition telle que formulée ci-dessus pour plusieurs raisons : (1) Le vocabulaire est indirectement mesuré dans les épreuves de compréhension orale et écrite ; (2) l'utilisation de tests comme l'EVIP pose des problèmes de droit d'auteur ; (3) il est difficile de créer des images qui seraient universellement appropriées aux différentes cultures et contextes ; (4) tout comme il est difficile de trouver des illustrateurs dans le pays susceptibles de créer des illustrations originales dans le bref laps de temps alloué à l'adaptation d'EGRA. De plus lorsque les images sont développées localement, l'expérience a montré qu'elles sont souvent de très faible qualité graphique, ce qui peut rendre leur interprétation difficile, même pour un adulte expérimenté. Pour ces différentes raisons, les épreuves basées sur des images sont difficiles à créer et à normaliser.

Une autre épreuve a été supprimée en raison d'effets plafond (voir glossaire) : celle qui évaluait les idées que se font les enfants de l'écrit (basée sur un travail de Marie Clay, 1993).

Dans cette épreuve, il était demandé aux enfants, par exemple, d'indiquer où commence et où se termine un texte et dans quel sens il faut le lire. Cette épreuve ne fait pas partie de la batterie de base.

6.4 Traduction et autres problèmes liés au langage

6.4.1 Traduction vs adaptation

Le consensus émergeant des experts qui avaient été invités à la réunion de Washington en novembre 2006 est que le meilleur choix pour l'adaptation d'EGRA n'est pas de traduire les mots, ou les textes, d'une langue à une autre. Cela peut en effet même conduire à des aberrations. C'est ce qu'avait noté Penelope Collins (née Chiappe) dans une communication personnelle envoyée à RTI en 2006 à propos d'une évaluation EGRA dont elle était en charge en Afrique du Sud.

> Il est crucial que les différents textes utilisés soient écrits indépendamment, en raison des différences linguistiques (orthographiques et morphologiques) qui caractérisent les langues de l'évaluation. Il n'est pas possible de considérer que le texte en anglais est équivalent à celui utilisé dans une autre langue en réalisant une simple traduction. Ceci est clairement démontré par les données des prétests conduits en isiZulu. Le texte utilisé était une traduction de l'anglais. Alors qu'on peut espérer que les scores de lecture de texte soient équivalents à ceux de la lecture de mots familiers présentés isolément, les locuteurs isiZulu pouvaient lire 20 à 30 mots isolés mais aucun mot du texte. Donc, le texte isiZulu était clairement trop difficile pour les élèves de première année de primaire.
>
> *Anglais* : "John had a little dog. The little dog was fat. One day, John and the dog went out to play. The little dog got lost. But after a while the dog came back. John took the dog home. When they got home John gave the dog a big bone. The little dog was happy so he slept. John also went to sleep."
>
> *IsiZulu* : "USipho wayenenja encane. Inja yakhe yayikhuluphele. Ngolunye usuku uSipho wayehamba nenja yakhe ukuyodlala. Inja yalahleka. Emva kwesikhathi inja yabuya. USipho waphindela ekhaya nenja yakhe. Emva kokufika ekhaya, uSipho wapha inja ekhaya ukudla okuningi. Inja yajabula kakhulu yaze yagcina ilele. NoSipho ngokunjalo wagcina elele."

6.4.2 Comparaisons entre les langues

La question de la comparabilité des langues est un problème difficile pour l'évaluation. EGRA, administré dans des contextes différents ou dans des langues différentes, doit utiliser des épreuves comparables, conçues sur les mêmes bases. Toutefois, ces comparaisons ne sont pas faciles même quand elles impliquent des écritures similaires (les écritures alphabétiques). Elles sont très difficiles lorsque les écritures sont de nature différente : par exemple, les écritures syllabiques (qui ont pour unité de base la syllabe, cf. les kanas au Japon, 2003) ou les écritures alpha-syllabiques (comme les aksharas de la langue Kannada : l'alpha-syllabaire de cette langue comprend plus de 400 symboles (Nag & Snowling, 2010). En conséquence,

4 à 5 ans sont nécessaires pour apprendre à lire dans cette écriture (Nag, 2007) contre 1 an dans les écritures alphabétiques dans lesquelles les correspondances graphème-phonème sont transparentes (Seymour et al., 2003). C'est la raison pour laquelle cette partie se limite aux écritures alphabétiques.

Les travaux de recherche indiquent que, dans les écritures alphabétiques, les différences de régularité des correspondances graphème-phonème (CGP) ont des conséquences sur l'apprentissage de la lecture. Par exemple, les CGP sont peu régulières en anglais (Share, 2008) alors qu'elles sont très régulières en espagnol. L'orthographe de l'allemand et du français se situe entre ces deux extrêmes, les CGP étant toutefois plus régulières en allemand qu'en français. Les études ont montré que les enfants anglais apprennent moins bien et moins vite à lire que les petits français, qui eux-mêmes apprennent moins bien et moins vite à lire que les petits espagnols (Seymour et al., 2003 ; Ziegler & Goswami, 2005 ; pour une synthèse en français, voir Sprenger-Charolles & Colé, 2013).

C'est la raison pour laquelle il est important de pouvoir donner des indications sur le niveau scolaire à partir duquel les enfants arrivent à lire dans différents pays (et différentes langues). Cela permettrait d'orienter les politiques éducatives, par exemple. La nécessité de ce type de données est l'une des raisons qui a motivé la création d'un Baromètre pour EGRA, outil interactif développé grâce à un financement de l'USAID (http://www.earlygradereadingbarometer.org/users/login). Cet outil, qui utilise des ensembles de données EGRA collectées dans des dizaines de pays, est accessible au public (accès gratuit après enregistrement). Il permet de générer des graphiques indiquant les performances des élèves en lecture, par pays. Afin de pouvoir comparer les résultats d'évaluation de la lecture entre plusieurs langues ayant le même système d'écriture, Il faut remplir au moins deux conditions :

1. Pour avoir un instrument adéquat (c'est-à-dire permettant la génération d'analyses valides et de résultats fiables, auxquels on peut faire confiance), il est nécessaire d'adapter les épreuves, et non de les traduire. L'adaptation prendra en compte les différences culturelles et linguistiques entre pays et langue (cf. la section 6.4.1 ci-dessus).
2. Quand il est nécessaire de comparer des langues, ceux qui adaptent les protocoles et analysent les résultats doivent, au minimum, examiner attentivement :
    - L'adéquation technique de l'évaluation en fonction de son objectif déclaré ;
    - Les caractéristiques des langues, telles que l'opacité ou la complexité de leur orthographique ;
    - Chaque épreuve, afin de bien cerner ce que chacune d'entre elles est supposée capturer.

Pour de plus amples directives et recommandations sur la façon d'adapter et de comparer les résultats d'EGRA à travers les langues, voir l'**annexe F**.

6.5 Création d'épreuves équivalentes

Comme mentionné précédemment dans cette section, l'adaptation peut impliquer la modification d'un instrument existant dans une langue donnée. Si les enseignants ont un accès limité aux évaluations EGRA, il est peu probable que les élèves connaissent une forme particulière de cette évaluation. Dans ce cas, les mêmes épreuves peuvent être utilisées à plusieurs reprises dans le temps. Cependant, s'il y a des craintes de « fuite », il devient nécessaire d'avoir plusieurs versions des épreuves pour mesurer les évolutions des performances dans le temps. Afin de s'assurer de la validité des comparaisons entre les résultats obtenus sur différentes versions des épreuves, ces dernières doivent être modifiés de manière à créer, autant que faire se peut, de nouvelles formes d'un niveau de difficulté équivalent à celui de la forme de départ (et donc substituables).

Dans les cas où le niveau de difficulté des épreuves diffère entre deux batteries, des procédures permettant d'homogénéiser le niveau de difficulté doivent être appliquées pour tenir compte des différences (voir la section 10.5). Le terme **Test équivalent** (cf. le glossaire), fait référence à des formes qui ont été ajustées par une procédure statistique après une administration afin de rendre les scores comparables. Il est cependant recommandé, quand il faut modifier l'instrument, de limiter le besoin d'administration d'analyses statistiques à posteriori. Les techniques de préparation des formes équivalentes peuvent inclure :

- De simples changements dans les noms des sujets de l'histoire et les actions qui sont accomplies, ainsi que par les différents qualificatifs utilisés pour les décrire. Les mots du discours correspondant à ces différents éléments (en particulier les noms, les verbes et adjectifs), seront remplacés par des équivalents.
- Pour les sous-tâches qui sont présentées aux apprenants sous forme de liste, le changement de l'ordre des items dans une ligne peut être utilisé.

Bien que ces techniques aient pour but d'équilibrer le niveau de difficulté d'une épreuve, elles ne garantissent pas l'équivalence des différentes formes de cette épreuve et ne suppriment pas la nécessité de tester cette équivalence postérieurement. Pour les situations dans lesquelles ces techniques sont utilisées, mais encore aboutissent à des formes qui ne sont pas équivalentes, l'utilisation de méthodes statistiques peut être nécessaire (voir la section 10.5).

6.6 Les meilleures pratiques

Comme EGRA a été développé dans des dizaines de pays et davantage de langues, de nombreuses leçons ont été tirées qu'il faut garder à l'esprit pour le développement de nouvelles adaptations ainsi que pour toute nouvelle modification.

- **Instructions.** Le débat sur le protocole EGRA, ou sur les instructions que les évaluateurs doivent suivre, n'est pas productif. Les instructions ont été soigneusement mises au point sur la base de données de la recherche et de

l'expérience préalable. Elles ne doivent pas être modifiées. En conséquence, le temps consacré à la traduction exacte de ces instructions est crucial pour la réussite de l'implémentation d'EGRA.

- **Prétest et étude pilote.** Ces deux étapes sont des éléments importants du processus (voir la première partie du chapitre 6, ainsi que le chapitre 9). Elles doivent donc être planifiées et budgétisées.
- **Contenu minimum.** Il faut au minimum évaluer la compréhension orale, la compréhension d'un petit texte écrit (avec une évaluation de la fluence dans la lecture de ce texte), la lecture de mots inventés et la relation entre les lettres et groupes de lettres (les graphèmes) et le son qui correspond. Les autres épreuves dépendent du contexte. **Utilisation des mêmes éléments, ou d'éléments quasi-identiques, dans les épreuves à travers les différentes formes que l'instrument peut prendre.** La meilleure pratique est de limiter, autant que faire se peut, le recours à des tests statistiques pour valider a posteriori l'équivalence des épreuves. Une bonne conception des épreuves doit produire des formes très comparables, sans avoir besoin le plus souvent d'utiliser ces procédures statistiques.

7 EMPLOI DE DONNÉES ÉLECTRONIQUES

Les chercheurs EGRA ont commencé en 2010 à passer d'une collecte de données sur papier à une collecte électronique de données. La collecte électronique de données réduit les risques d'erreurs ou d'omissions dans les données et rend les résultats plus rapidement disponibles.

Des comparaisons entre la collecte de données sur papier et la collecte électronique de données ont montré que cette dernière présentait des avantages en termes d'efficacité et d'efficience. La disponibilité croissante de dispositifs mobiles abordables et d'une connectivité Internet permettant aux chercheurs d'analyser les données en temps réel continuent à encourager l'acquisition électronique de données (Walther et al., 2011).

Une différence notoire entre une collecte de données sur papier et une collecte électronique de données est l'élimination de la saisie manuelle de données entre formulaires remplis et bases de données électroniques. Ceci permet de réaliser des économies de temps et de réduire le risque d'erreurs associées à la saisie manuelle de données et de fautes résultant d'annotations incorrectes ou illisibles des évaluateurs ou de questions sautées. Les résultats de la collecte de données électronique peuvent de plus être téléchargés sur place et traités et analysés plus tôt. Cet avantage permet de détecter et de rectifier des problèmes quand les évaluateurs sont encore sur le terrain. La collecte de données électroniques améliore et renforce donc le travail sur le terrain.

La collecte de données électroniques améliore et renforce le travail sur le terrain.

Il est important de ne pas oublier que la collecte de donnée électroniques ne change pas les procédures fondamentales de mise en œuvre de l'évaluation. L'enfant lit toujours sur une feuille de papier où sont imprimées des lettres et des mots ; l'évaluateur donne toujours les mêmes instructions. Les instructions pour la collecte de données électroniques ne changent que lorsqu'il est fait référence à la façon de marquer les réponses (par ex. « marquer » plutôt que « toucher l'écran »).

Les premiers exemples connus de collecte automatisée d'information sans fil et cellulaire conçue spécialement pour EGRA étaient iProSurveyor, développé par Prodigy Systems pour l'arabe au Yémen puis au Maroc en 2011[17] et le système logiciel Tangerine®, créé par RTI International en 2010 et mis à l'essai en 2012. Ces deux programmes logiciels ont adapté l'instrument EGRA, notamment les exercices chronométrés, à une interface sur tablette à écran tactile discrète, portable

[17] Aux termes d'un sous-contrat avec RTI International pour le projet EdData II de l'USAID (voir Collins & Messaoud-Galusi, 2012 ; Prodigy Systems, 2011).

et intuitive qui n'interfère pas avec la procédure d'administration individualisée d'EGRA.[18] L'initiative iProSurveyor EGRA au Yémen a porté sur 38 écoles dans trois gouvernorats et 735 entrevues d'élèves de 2e et 3 e années. Tangerine a été testé sur le terrain pour la première fois en janvier 2012 dans le cadre de l'initiative Mathématiques et lecture dans le primaire (PRIMR) au Kenya pour laquelle 176 000 points de données ont été saisis au travers d'un petit échantillon de 200 élèves de 10 écoles évalués à l'aide d'un EGRA en anglais, d'un EGRA en Kiswahili et d'une Évaluation des compétences fondamentales en mathématiques (EGMA ; Strigel, 2012). Ces essais sur place ont démontré la facilité d'emploi et l'efficience du système et la collecte de donnée électroniques a été confirmée comme constituant une démarche faisable pouvant se substituer à la collecte de données sur papier pour des évaluations du niveau de lecture à haute voix (et de mathématiques) faisant intervenir des éléments chronométrés.

7.1 Mises en garde et restrictions relatives à la collecte électronique de données

Il convient d'être conscient des limitations suivantes en ce qui concerne la collecte électronique de données :

- **Risque d'erreur.** La collecte électronique de données n'est pas à toute épreuve. Elle s'accompagne d'un certain degré de possibilité d'erreurs de saisie ou de perte de données.
- **Considérations de coût.** Des analyses de coûts réalisées pour l'USAID dans le cadre du projet EdData II ont indiqué que l'emploi d'une collecte électronique de données est plus efficient qu'une collecte sur papier quand le matériel informatique est utilisé pour plusieurs collectes de données. Des économies de coût peuvent ne pas être réalisées si le matériel informatique requis est employé uniquement pour une seule collecte de données.
- **Il est nécessaire de faire des sauvegardes sur papier.** Les équipes d'évaluation doivent conserver des copies de sauvegarde sur papier en cas de défaillance du matériel électronique sur le terrain. Durant la formation des évaluateurs, des instruments imprimés sont donc introduits en même temps que le logiciel électronique.
- **Limitation de l'accès à la technologie.** Quand ils envisagent une collecte électronique de données, les planificateurs doivent tenir compte du contexte national / régional et du niveau des connaissances technologiques des évaluateurs.
- **Problèmes de sécurité.** La perte, le vol et l'endommagement de dispositifs risquent d'entraîner des pertes financières ou des préjudices personnels ; il est donc important de planifier soigneusement la sécurité du matériel et des évaluateurs.

[18] Les ordinateurs portatifs n'ont pas été considérés comme constituant une technologie viable à cet effet du fait des effets potentiels de la visibilité de la technologie dans la salle de classe et des limitations de leur emploi dans certains contextes (manque d'alimentation électrique, poussière / humidité, transport à pied, en vélo, par bateau, etc.). Des systèmes de collecte et de saisie de données existent également pour ordinateurs portatifs et de bureau, eEGRA développé et employé par exemple par l'Education Development Center, Inc. (EDC) (http://eegra.edc.org/).

- **Infrastructure de communication limitée.** Dans certains pays ou certaines régions, il peut être difficile de trouver ou de créer des points Wi-Fi mobiles pour le téléchargement de données de terrain.
- **Capacité locale limitée.** L'adaptation de l'instrument dans des langues et des écritures locales et la représentation du contenu dans le logiciel de collecte de données choisi présentent des problèmes connexes. Une affiliation avec des partenaires locaux chevronnés est essentielle à l'exploration et à l'atténuation des limitations de capacité en ce qui concerne la saisie de données électroniques.

S'ils choisissent une collecte de données sous format électronique plutôt que sur papier, les chercheurs doivent également tenir compte de la nécessité du maintien de la sécurité des données numériques ; selon le logiciel employé pour le recueil de données, plusieurs personnes peuvent accéder aux résultats bruts. Même les points du système de positionnement global (GPS) ne doivent être employés qu'à des fins de vérification, et non pas pour identifier des écoles particulières. Comme pour les recherches basées sur papier, il conviendra de faire tous efforts possibles pour respecter la confidentialité des données et faire en sorte que les résultats n'aient pas de répercussions négatives sur les écoles, les enseignants ou les élèves.

7.2 Logiciels de collecte de données

Il existe de nombreux outils d'étude mobiles pouvant être adaptés pour l'administration d'EGRA. Le logiciel libre Tangerine est un outil largement employé et, à la mi-2015, était appliqué dans plus de 60 interventions dans 36 pays par 27 organisations (voir www.tangerinecentral.org). Une comparaison des fonctionnalités de Tangerine et de plusieurs autres outils de collecte de données électroniques—Magpi, SurveyToGo, doForms, Droid Survey, Open Data Kit (ODK) et Command Mobile—figure en **Annexe G** et un exemple de directives pour des formulaires imprimés comparé à une méthode électronique est présenté en **Annexe H**. À ce jour, iProSurveyor (pour iPad), Tangerine et SurveyToGo étaient les seules plateformes, outre les systèmes de saisie pour portables et ordinateurs de bureau (voir note de bas de page 18 à la page précédente) dont on sait qu'ils ont été adaptés à EGRA. Les réalisateurs de l'étude déterminent quel logiciel est le plus compatible avec le contexte et la nature des données recueillies—notamment le format unique minuté sous forme de grille de nombreux exercices EGRA et la nécessité de calculer le nombre total d'éléments entrepris (précision) et les éléments corrects par minute (fluence). Où entreposer les données, qui en assure la gestion et la capacité technique peuvent également constituer des considérations dans le choix d'un logiciel particulier.

7.3 Considérations relatives à la sélection et à l'achat de matériel informatique

Quand il s'agit de se procurer le matériel informatique pour la collecte de données électroniques EGRA, les réalisateurs de l'étude doivent tenir compte de divers facteurs, expédition, entreposage et réutilisation des matériels notamment. En 2015,

les tablettes étaient considérées préférables aux téléphones portables, smartphones ou ordinateurs portatifs du fait de la taille de l'écran, de leur facilité d'utilisation, de leur légèreté et surtout de la durée de leur batterie. Les accessoires additionnels doivent au minimum comprendre un stylet, un étui protecteur et un routeur Wi-Fi pour faciliter la collecte de données et pouvoir transmettre les résultats tous les jours.

Les réalisateurs de l'évaluation doivent peser le pour et le contre pour déterminer s'il convient d'acheter le matériel dans le pays où va s'effectuer la collecte de données ou de l'acheter à l'extérieur du pays de mise en œuvre. Pour un achat effectué à l'extérieur du pays, il faudra programmer des délais suffisants pour l'expédition et le passage à la douane. Le transport de dispositifs sur sa personne d'un pays à un autre est possible s'il s'agit d'un petit nombre de tablettes et d'accessoires à utiliser (ou à réutiliser) mais les personnes qui les transportent doivent être au fait des réglementations douanières et, selon le contexte, des frais éventuels d'importation. Pour lever les droits d'importation, certains pays exigent par exemple la preuve d'intention d'exportation des dispositifs après la collecte des données.

Les réalisateurs de l'évaluation doivent également planifier l'entreposage de tout le matériel et des accessoires avant et après la collecte de données et durant la formation. Tous les dispositifs et périphériques doivent être entreposés en lieu sûr pour éviter les vols. Le lieu d'entreposage doit être protégé de la poussière, de l'humidité et des écarts de température. Noter qu'une longue période d'inutilisation peut affecter la durée de batterie des appareils.

Dans le cadre du processus de mise en œuvre, il est essentiel d'établir des procédures claires en ce qui concerne l'appartenance, l'accès et l'emploi du matériel, des logiciels et des données. Il est commun (et rentable) que le réalisateur de l'étude ou l'organisation de financement réutilise le matériel ou que la propriété des articles achetés soit transférée à des organisations locales pour leur permettre de continuer à les utiliser.

7.4 Fournitures nécessaires pour la collecte de données électroniques et la formation

- Tablettes, toutes munies d'un chargeur
- Logiciel contenant la version électronique de l'évaluation
- Étuis de protection
- Stylets
- Sacoches pour permettre aux évaluateurs de transporter les tablettes sur les sites d'étude
- Routeurs Wi-Fi, clé de connectivité et service de connexion
- Plusieurs tablettes supplémentaires en cas d'endommagement ou de perte

8 FORMATION DES ÉVALUATEURS EGRA

On trouvera dans cette section des indications sur la planification et la réalisation d'une formation d'évaluateurs EGRA.

Il est à noter que cette section n'a pas pour but de constituer un manuel à l'intention des évaluateurs ou des superviseurs ; elle est plutôt une ressource pour les organisateurs de la formation. Les Notes d'orientation pour la planification et la mise en œuvre d'évaluations des compétences fondamentales en lecture contiennent des détails supplémentaires sur la formation des évaluateurs et il est recommandé de les consulter parallèlement à ce document (RTI International & Comité international de secours, 2011).

Les évaluateurs qui vont piloter l'instrument devront suivre une formation d'une semaine environ.[20] La durée de cette formation sera fonction du nombre d'instruments à administrer (par ex. évaluation du niveau de mathématiques en plus d'EGRA), du nombre de formateurs disponibles, du nombre de personnes à former, de l'expérience antérieure des formateurs et du budget et du temps disponibles. Par exemple, si certains formateurs ont une compréhension limitée de la langue employée pour la formation (au point où la présence d'un traducteur est nécessaire), il sera sage d'ajouter deux ou trois jours au calendrier.

Pour une formation EGRA qui s'est déroulée en Tanzanie en 2013 et a mis en jeu 150 évaluateurs, un instrument bilingue et des études additionnelles, l'équipe technique de formation comportait cinq personnes : un spécialiste dans la langue 1, un spécialiste dans la langue 2, un expert dans le logiciel de collecte de données, un chargé de la logistique et un coordinateur général qui se concentrait également sur les tests de performance des évaluateurs, la préparation aux ateliers, la conception de l'étude et les relations avec les donateurs.

Pour assurer que tous les stagiaires comprennent l'objectif du travail et accepte celui-ci, un élément essentiel du programme sera l'examen des principes sous-jacents d'EGRA et du raisonnement à l'appui des composants de l'instrument. Les autres principaux objectifs sont les suivants :

- Former un groupe d'évaluateurs à la bonne administration d'EGRA, sous format électronique et imprimé
- Identifier des individus compétents qui assureront les fonctions d'évaluateurs pour la collecte de données

[20] Voir Section 9.1.3 sur les avantages et les inconvénients de diverses programmations possibles de formation d'évaluateurs par rapport à la collecte de donnée (mise à l'essai et déploiement).

- Identifier et former des individus sélectionnés qui assureront les fonctions de superviseurs durant la collecte de données

8.1 Recrutement de participants à la formation

Il est essentiel de recruter et de former 10 à 20 % plus d'évaluateurs que le nombre indiqué dans le plan d'échantillonnage. Il est inévitable que certains ne répondent pas aux critères de sélection et que d'autres abandonnent après la formation pour des raisons personnelles ou autres.

Les équipes de collecte de données peuvent être composées de représentants du secteur de l'éducation et / ou d'évaluateurs indépendants recrutés pour la collecte de données en question. Les conditions et les préférences sont déterminées au cours de la phase de recrutement, avant la formation, selon les circonstances et les objectifs particuliers.

La candidature de représentants du gouvernement peut être envisagée pour les fonctions d'évaluateurs ou de superviseurs. Pour être sélectionnés pour un travail sur le terrain, ils devront cependant répondre aux mêmes normes de performance que tous les autres stagiaires. Les facilitateurs doivent souligner les normes de sélection au début de la formation. Un avantage possible de la participation de représentants qualifiés du gouvernement est que ce dernier sera risque peut-être d'être davantage réceptif à l'analyse des données une fois les résultats annoncés.

Un autre facteur dont il convient de tenir compte au stade du recrutement est la d'détermination de conflits d'intérêts éventuels pour les candidats—dans le secteur public ou privé—découlant de la scène politique dans le pays.

Les planificateurs devront tenir compte des importants critères suivant dans la détermination des participants à la formation d'évaluateur :'

- Les candidats devront lire et parler couramment les langues requises pour la formation et l'administration d'EGRA
- Ils devront bénéficier d'une certaine expérience dans l'administration d'évaluations ou la collecte de données
- Ils devront avoir déjà travaillé avec des enfants d'âge primaire
- Ils devront être disponibles au cours de la phase de collecte de données et pouvoir travailler dans les régions ciblées
- Ils devront savoir utiliser et maîtriser les ordinateurs ou dispositifs électroniques portatifs (tablette, smartphone).
- L'équipe de formation déterminera la liste finale d'évaluateurs en fonction des critères suivants. Ces conditions sont communiquées aux stagiaires dès le début pour qu'ils comprennent que la sélection finale sera fonction des aptitudes des candidats.
- Capacité à administrer EGRA de manière exacte et efficace. Tous les candidats

sélectionnés pour remplir les fonctions d'évaluateurs doivent parfaitement administrer EGRA. Il s'agira notamment de faire preuve de connaissances dans les réglementations et procédures de son administration, de savoir enregistrer exactement les réponses des élèves et employer tous les matériels requis—tablette par exemple—pour administrer l'évaluation.

Les évaluateurs doivent pouvoir gérer plusieurs tâches à la fois (écouter l'élève, noter les résultats et travailler sur la tablette).

- **Capacité à établir un rapport positif avec les élèves.** Il est important que les évaluateurs sachent communiquer de manière avenante avec les jeunes enfants. L'établissement d'un rapport positif et chaleureux avec les élèves leur
- permettra une meilleure performance possible. Si cet aspect de l'administration du test peut être appris, tous les évaluateurs ne vont pas nécessairement le maîtriser.
- **Capacité à bien travailler en équipe dans un environnement scolaire.** Les évaluateurs ne travaillent pas seuls ; ils font partie d'une équipe. Ils doivent à cet effet faire preuve d'un esprit de collaboration pour la réalisation de toutes les tâches à effectuer lors d'une visite dans une école. Ils doivent de plus démontrer qu'ils peuvent bien travailler dans un environnement scolaire exigeant l'adhésion à certains protocoles, le respect du personnel scolaire et de la propriété de l'école et une interaction appropriée avec les élèves.
- **Disponibilité et adaptabilité.** Comme nous l'avons souligné plus haut, les évaluateurs doivent être disponibles tout au long de la collecte de données et montrer qu'ils sont capables de travailler dans les sites désignés. Ils devront peut-être par exemple passer une semaine dans un environnement rural où les transports peuvent être difficiles et l'hébergement minimal.

Parmi les stagiaires, les facilitateurs sélectionneront également des superviseurs dont la tâche consistera à soutenir et coordonner les évaluateurs au cours de la phase de collecte. Les superviseurs (qui peuvent aussi être appelés coordinateurs de collecte de données ou autre appellation similaire) doivent au minimum répondre aux critères établis pour les évaluateurs. Ils devront de plus :

- Faire preuve de compétences de leadership, avoir de l'expérience dans la direction d'une équipe et se valoir du respect de leurs collègues
- Être organisés et soucieux du détail
- Connaître suffisamment bien les procédures d'administration d'EGRA pour pouvoir superviser d'autres personnes et vérifier l'absence d'erreurs dans la collecte de données
- Savoir suffisamment bien employer les tablettes pour aider les autres
- Interagir de manière appropriée avec les enfants et les représentants des écoles

Les facilitateurs doivent également communiquer ces qualifications à l'avance aux

stagiaires et à tous partenaires locaux dans la collecte de données. Les superviseurs ne seront pas nécessairement des cadres supérieurs de la fonction publique ni des personnes jouissant d'une certaine ancienneté. Les représentants qui ne répondent pas aux critères pourront occuper d'autres fonctions de supervision, visites de contrôle par exemple. Ces situations se présentent parfois quand les responsables scolaires souhaitent jouer un rôle dans l'observation et la supervision de la collecte de données, qu'ils puissent ou non participer à la formation d'évaluateurs ; tenir compte de ces desiderata peut les amener à mieux comprendre le processus EGRA et en accepter les résultats.

8.2 Planification de l'atelier de formation

Tâches essentielles à entreprendre avant l'atelier de formation :

- **Préparer l'instrument et les matériels de formation EGRA.** Finaliser le contenu des instruments qui vont être employés au cours de la formation—sous format électronique et imprimé, pour toutes les langues. Il conviendra également de préparer et de faire des copies d'autres documents et polycopiés de formation (programme, vessions imprimés de questionnaires et feuilles d'accompagnement, manuel de superviseur, etc.).
- **Se procurer le matériel**. L'équipement et les matériels anticipés et achetés bien à l'avance regroupent notamment les ablettes et étuis, tableau-papier, chronomètres, multiprises et cadeaux d'élèves. Dresser un inventaire pour
- assurer le suivi de tous les matériels tout au long de la collecte de données et de la formation EGRA.
- **Préparer l'équipement.** Les personnes responsables des aspects technologiques de la formation devront, une fois que l'on se sera procurer les tablettes, préparer celles-ci pour la collecte de données. Cela veut dire charger le logiciel et les versions électroniques des instruments sur les tablettes et les régler correctement.
- **Préparer le programme de l'atelier.** Créer un programme provisoire et le faire circuler dans l'équipe de mise en place de l'atelier. Pour une formation EGRA uniquement, les principaux éléments du contenu de l'agenda comprendront :
    - Aperçu de l'instrument EGRA (objectif et nature des compétences mesurées)
    - Administration des tâches EGRA (protocoles et processus, répétition d'exercices)
    - Emploi de tablette (fonctionnalité, enregistrement et téléchargement des évaluations)
    - Protocoles de travail sur le terrain et d'échantillonnage

      Un exemple de programme figure en **Annexe I.**
- **Finaliser l'équipe de facilitation.** La formation des évaluateurs est facilitée par au moins deux formateurs disposant des compétences nécessaires dans l'évaluation du niveau de lecture (et d'EGRA en particulier) et d'une certaine

expérience dans la formation de collecteurs de données. Les formateurs n'ont pas nécessairement besoin de connaître la langue testée dans l'instrument EGRA s'ils sont soutenus par un expert dans la langue locale qui peut vérifier la bonne prononciation des lettres et des mots et aider dans toute traduction éventuellement nécessaire pour faciliter la formation. Les formateurs doivent cependant parler couramment la langue dans laquelle l'atelier est présenté.

Si la formation va être enseignée en plusieurs langues, une équipe compétente de formateurs est préférable et on peut envisager des formateurs additionnels.

8.3 Éléments de la formation d'évaluateurs

Comme indiqué dans l'exemple de programme figurant en **Annexe I**, la formation des évaluateurs intégrera plusieurs éléments constants. Dans une séquence similaire à celle qui suit, les facilitateurs :

- Inviteront des cadres supérieurs de la fonction publique dont l'objectif est de déclarer publiquement leur engagement envers EGRA et l'intérêt qu'ils portent aux résultats.
- Présenteront le projet d'évaluation, l'importance de la lecture dans le primaire, la nature d'EGRA et les principes fondamentaux de l'administration de l'instrument.
- Expliqueront l'importance pour la recherche du suivi de la performance des évaluateurs et les critères selon lesquels ils seront évalués et sélectionnés.
- Donneront un aperçu des tâches et démontreront leur mode d'administration.
- Présenteront et expliqueront tous instruments supplémentaires à administrer parallèlement à EGRA.
- Donneront aux participants l'occasion de pratiquer en groupes de deux et plus sous la supervision des formateurs principaux. Au bout de plusieurs jours de formation, une pratique au minimum sera organisée avec des enfants dans un environnement scolaire.
- Observeront, assisteront et reprendront au besoin la formation. Ils veilleront à ce que les stagiaires soient à l'aise tant avec le contenu de l'enquête qu'avec le matériel et le logiciel.
- Évalueront officiellement l'exactitude des évaluateurs (voir Section 8.7) ; les résultats serviront au rattrapage et à la sélection d'un groupe d'évaluateurs pour la collecte des données.

8.4 Méthodes et activités de formation

Les recherches effectuées dans le domaine de l'apprentissage des adultes mettent en valeur certaines meilleures pratiques à employer dans la formation d'évaluateurs. Que la formation porte sur une équipe de 20 évaluateurs ou de 100, la création de sessions interactives dans lesquelles les participants travaillent ensemble avec la

technologie et l'instrument résultera en un apprentissage plus efficace.

L'expérience acquise dans la formation d'évaluateurs EGRA indique dans son ensemble que plus les participants ont l'occasion de pratiquer l'administration d'EGRA, mieux ils apprennent à bien administrer l'instrument. De plus, varier les activités d'un jour à l'autre permettront aux participants de mieux s'impliquer et d'améliorer les résultats. Les activités quotidiennes de formation sur tablette peuvent par exemple comprendre :

- Des démonstrations par les facilitateurs
- Des vidéos
- Une pratique en groupe entier
- Une pratique en petits groupes
- Une pratique en groupes de deux
- Des démonstrations par les stagiaires

Tout au long de la formation, les facilitateurs devront varier les groupes de deux et les petits groupes. Ils pourront par exemple mettre un évaluateur plus compétent ou plus chevronné avec quelqu'un ayant moins d'expérience.

On peut notamment avoir recours à une démarche « à la ronde » pour pratiquer les éléments de la formation exigeant une attention particulière (par ex., participants s'assoient en rond et énoncent tour à tour les lettres de l'instrument EGRA) ou encore à des simulations dans lesquelles une personne jouant le rôle d'un évaluateur commet des erreurs ou ne suit pas les bonnes procédures et il est demandé aux participants de déterminer ce qui s'est passé et ce que « l'évaluateur » aurait dû faire différemment .

S'il s'agit de plus d'une langue, il est conseillé de mener ces activités au sein des mêmes groupes de langues.

Les facilitateurs devront également demander aux stagiaires de consacrer du temps à s'exercer au fonctionnement de la tablette, aux menus déroulants, aux fonctions de saisie, etc.

La présentation de vidéos d'administration d'EGRA aux participants peut les aider à comprendre le processus et les protocoles avant qu'ils aient l'occasion de l'administrer eux-mêmes. Ces vidéos—qui vont exiger des permissions appropriées et devront être enregistrées avant la formation—peuvent servir à démontrer les meilleures pratiques et les scénarios qui se présentent fréquemment. Elles peuvent servir de point de départ de discussions et de pratique.

8.5 Visites d'écoles

La formation des évaluateurs implique toujours au minimum une visite d'école pour leur permettre de pratiquer l'administration d'EGRA aux enfants et d'employer la technologie dans des conditions similaires à celles dans lesquelles ils vont travailler durant la phase de collecte de données. Les visites d'écoles leur permettent également de s'exercer aux procédures de sondage des élèves et de remplir tous les formulaires nécessaires sur la visite d'école.

Pour que les visites d'écoles soient productives, l'équipe de direction de la formation devra :

- Programmer au moins une visite d'école durant la formation (deux ou plus seraient préférables :
  - En programmer une à la mi-formation et une vers la fin.
- Déterminer le nombre d'écoles nécessaires :
  - Baser le nombre d'écoles sur le nombre de stagiaires, la taille des écoles voisines et le nombre de visites.
  - Éviter de surcharger une école en y amenant trop de personnes.
  - Ne pas affecter plus de 35–40 personnes à une grande école, moins pour une école plus petite.
- Déterminer les écoles à visiter avant d'entamer la formation :
  - Obtenir la permission requise, alerter les directeurs d'école et planifier les transports ; s'assurer que les écoles ne font pas partie d'échantillon complet de collecte de données (si cela n'est pas possible, veiller à exclure les écoles où l'évaluation est pratiquée de l'échantillon final).
- Préparer les équipes un jour à l'avance pour qu'elles sachent à quoi s'attendre :
  - Logistique du départ, qui va où, superviseurs d'équipes, nombre d'élèves par évaluateur, évaluations à mener, etc.
- Au cours d'une deuxième ou troisième visite, il sera peut-être plus facile pour les participants de travailler seuls et plus bénéfique de pratiquer l'administration de l'évaluation avec autant d'enfants que possible durant la visite. Ils seront de plus à même de pratiquer les procédures de sondage des élèves et d'autres aspects de la collecte de données qu'ils peuvent ne pas avoir déjà appris avant la première visite d'école.
- Chaque évaluateur administrera les instruments à un nombre d'enfants allant de quatre à dix20 à chaque visite d'école.
- Il est extrêmement important de procéder à un compte-rendu avec les participants après la visite. C'est l'occasion pour les stagiaires de communiquer au groupe ce qu'ils considèrent comme s'étant bien déroulé et les domaines où ils ont été confrontés à des difficultés. La visite d'école soulève souvent de nouveaux problèmes et constitue l'occasion de répondre à des questions ayant pu être soulevées au cours de la formation.

RÉSUMÉ DES RESPONSABILITÉS DES FORMATEURS AU COURS DE VISITES SCOLAIRES DE PRATIQUE

- Déterminer les stagiaires qui vont remplir les fonctions de superviseurs
- Venir en aide aux équipes en assurant au besoin les présentations
- Observer les évaluateurs et leur venir éventuellement en aide
- Après avoir obtenu la permission appropriée, prendre des photos ou des vidéos des évaluateurs pour une formation et des discussions supplémentaires au cours de la phase de conte-rendu
- Remettre les classes / ressources dans l'état où les équipes les ont trouvées
- Remercier le directeur d'école de sa collaboration

Les participants auront besoin d'un endroit tranquille et distinct de l'école pour pratiquer l'administration des évaluations. Dans des conditions idéales, les évaluateurs devraient pouvoir s'asseoir à une table face à l'enfant. Si on ne disposer pas de tables, l'enfant peut s'asseoir sur une chaise placée légèrement à la diagonale de l'évaluateur.

Au cours de la première visite d'école, il est utile pour les participants de réaliser EGRA en groupes de deux, pour qu'ils puissent s'observer l'un l'autre et se faire part d'observations. Le travail par deux peut être également utile, les participants étant souvent nerveux la première fois qu'ils procèdent à une évaluation EGRA auprès d'un enfant.

8.6 Processus d'évaluation évaluateur-stagiaire

Un processus transparent d'évaluation et des critères d'évaluation clairs sont utiles tant pour les facilitateurs que pour les stagiaires. Le processus employé pour évaluer les évaluateurs au cours de la formation comporte des méthodes d'évaluation formelles et informelles. Dans le cadre d'une évaluation informelle, les facilitateurs observent soigneusement les stagiaires au cours de l'atelier et de visites d'écoles et mènent également si possible des entrevues individuelles.

Les stagiaires vont avoir besoin d'observations tant sur leurs points forts que sur leurs points faibles tout au long de l'atelier. La présence d'une équipe adéquate de formateurs qualifiés permettra l'apport d'observations régulières et spécifiques. De même, disposer d'un nombre suffisant de formateurs permettra de formuler des observations répondant aux besoins des stagiaires nécessitant une aide supplémentaire et la bonne sélection de superviseurs.

L'observation soigneuse des évaluateurs permet de réaliser l'objectif ultime, à savoir la collecte de données de qualité. En conséquence, toutes les fois que les évaluateurs s'exercent, les facilitateurs circulent, surveillent et prennent note de tous problèmes devant être aborder avec tout le groupe.

L'évaluation des évaluateurs présentent plusieurs aspects et tient compte de plusieurs facteurs, notamment :

- Administrer correctement et efficacement les instruments, notamment connaître et suivre toutes les règles de l'administration
- Exactement prendre note des données démographiques et des réponses
- Déterminer les réponses correctes et incorrectes
- Utiliser correctement et efficacement le matériel, les tablettes en particulier
- Bien travailler en équipe
- Adhérer aux protocoles de visites d'écoles
- Établir un rapport avec les élèves et le personnel scolaire

Tout au long de la formation, les participants eux-mêmes font le point et partagent leurs expériences en ce qui concerne l'emploi de l'instrument. Les dirigeants de la formation sont préparés à clarifier le protocole EGRA (c.-à-d. les instructions qui lui sont intégrées) en fonction de l'expérience des évaluateurs tant dans le contexte de l'atelier qu'au cours des visites d'écoles.

L'évaluation formelle des évaluateurs fait maintenant partie des normes dans de nombreux projets financés par des donateurs et constitue une issue attendue d'un programme de formation d'évaluateurs. La section suivante traite en détail la mesure de la précision des évaluateurs. Les formateurs évaluent le degré de concordance parmi plusieurs noteurs (évaluateurs) administrant le même test en même temps au même élève. Ce type de test ou mesure des compétences des évaluateurs détermine la capacité des stagiaires à administrer correctement EGRA.

8.7 Mesure de la fidélité des évaluateurs

Dans le cadre du processus de sélection, les animateurs de l'atelier mesurent la justesse des évaluateurs au cours de la formation en évaluant la mesure dans laquelle les évaluateurs sont d'accord dans leur notation de la même observation.

APERÇU DE L'ÉVALUATION FORMELLE DU DEGRÉ DE JUSTESSE DES ÉVALUATEURS AU COURS DE LA FORMATION

1. **Évaluation et sélection des évaluateurs.** Établir un niveau. Les évaluateurs incapables de d'atteindre ce niveau ne sont pas sélectionnés pour la collecte de données. Dans une formation EGRA, le niveau est établi à une concordance de 90 % avec l'évaluation correcte de l'enfant pour l'évaluation finale de la formation.
2. **Détermination des priorités pour la formation.** Ces évaluations formelles indiquent des tâches et des éléments qui posent des problèmes aux évaluateurs, ce qui constitue également d'importants domaines d'amélioration sur lesquels la formation doit se concentrer.
3. **Rapport sur l'état de préparation des évaluateurs.** La formation d'évaluateurs implique trois évaluations formelles pour déterminer et surveiller le progrès de leur justesse.

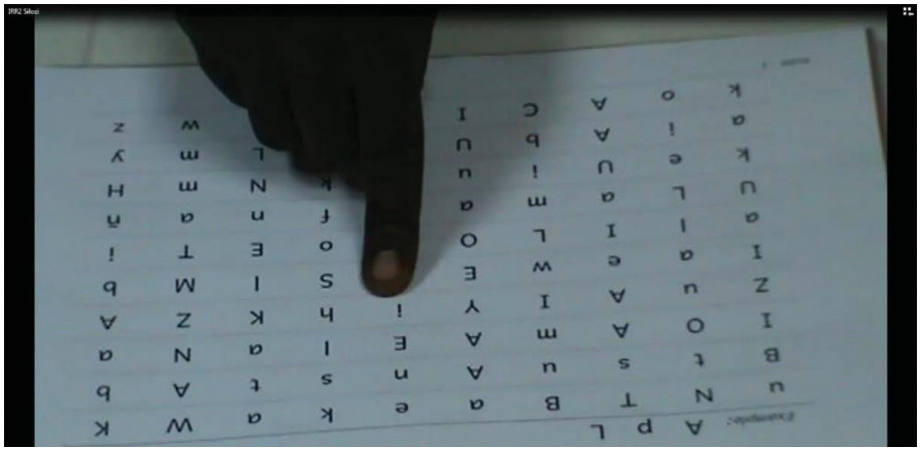
Ce type d'évaluation est particulièrement utile pour améliorer la performance des évaluateurs avant qu'ils se rendent sur le terrain. Elle doit également être employée pour sélectionner les évaluateurs les plus performants qui constitueront l'équipe finale pour la collecte de données à grande échelle, ainsi que des remplaçants et des superviseurs.

L'équipe de formation crée un instrument distinct sur les tablettes dans le but de procéder à la mesure de la justesse des évaluateurs.

Deux principaux moyens permettent de produire des données pour calculer la justesse des évaluateurs :

1. Si les dirigeants de la formation ont pu obtenir les permissions appropriées avant l'atelier et produire des enregistrements audio ou vidéo des élèves participant aux évaluations d'exercice ou pilotes (voir **Figure 22**), les enregistrements peuvent être présentés dans un contexte de groupe, **tous** les évaluateurs donnant une note à l'évaluation comme ils le feraient au cours d'une « vraie » administration d'EGRA. Un évaluateur EGRA compétent donne également une note à l'évaluation et ces résultats sont utilisés comme « étalon d'or ».

Figure 22. Image d'une vidéo employée pour l'évaluation



2. Les formateurs adultes ou les évaluateurs peuvent jouer le rôle de « l'élève » et de « l'évaluateur » dans le contexte d'un grand groupe (ou dans une vidéo) et les évaluateurs donnent tous une note à l'activité. L'avantage de ce scénario est que les adultes peuvent délibérément et sans ambiguïté commettre plusieurs erreurs sur une tâche donnée (par ex. sauter ou répéter des mots ou des lignes, varier le volume de la voix, marquer une longue pause pour obtenir une réplique, etc.). Le texte préparé à l'avance et accompagné des erreurs délibérées constitue l'étalon d'or.

Les formateurs vont alors télécharger toutes les évaluations des stagiaires dans Excel ou autre logiciel d'analyse et analyser comparativement les résultats. Des détails supplémentaires sur les statistiques et l'analyse de données pour mesurer la justesse des évaluateurs figurent en **Annexe J**.

Après une observation d'évaluateur, les données doivent être réduites aux tentatives d'évaluation des stagiaires et à l'évaluation étalon d'or.

Si pour une raison quelconque l'équipe de formation n'a pas établi d'étalon d'or ou durant l'évaluation des stagiaires, le formateur principal en prépare un après et en ajoute les résultats à la base de données. L'équipe de formation doit de plus passer en revue les réponses aux normes de l'étalon d'or pour s'assurer que ce qui est enregistré pour chacune reflète fidèlement le consensus sur les réponses correctes à l'évaluation. Une activité importante consiste à comparer l'étalon d'or à la réponse mode (la plus fréquente) des évaluateurs au niveau de l'élément.

Comme indiqué plus haut, il est important de mesurer la justesse des évaluateurs car elle permet au formateur de déterminer les évaluateurs dont les résultats de notation dévient de plus d'un écart type de l'étalon d'or et pouvant avoir besoin de pratique ou d'aide additionnelle.

Cette mesure peut également permettre de déterminer si le groupe tout entier doit faire l'objet d'un examen ou d'une formation supplémentaire portant sur certaines tâches ou si certaines compétences (arrêts précoces par exemple) exigent une pratique additionnelle.

Si l'analyse provenant de l'évaluation formelle révèle un niveau de performance demeurant insatisfaisant de la part d'un évaluateur donné et si sa performance ne s'améliore pas après une pratique et une aide additionnelles, cet évaluateur ne pourra pas participer au travail sur le terrain. Se reporter là encore à l'**Annexe J** où figurent des informations supplémentaires sur le mode d'évaluation des données sur la justesse des évaluateurs.

En plus du processus d'évaluation des évaluateurs au cours de la formation, ceux-ci doivent continuer à tester entre eux la fidélité et la cohérence de leurs résultats (fiabilité inter-évaluateurs [IRR]) une fois qu'ils procèdent à la collecte de données sur le terrain. L'IRR peut permettre d'améliorer la fidélité et la cohérence des données au fur et à mesure de leur collecte et d'éviter de plus le biais d'effet séquentiel (voir glossaire). Des informations supplémentaires sur la mesure de la fiabilité inter-évaluateurs au cours du processus de collecte de données sont présentées à la **Figure 23** ; l'**Annexe K** contient des graphiques illustrant plusieurs exemples de plans d'échantillonnage pour varier les jumelages d'évaluateurs.

Figure 23. Exemple de protocole pour la surveillance de la fiabilité inter-évaluateurs au cours du travail sur le terrain

Protocole pour la collecte de données de fiabilité inter-évaluateurs

Un important aspect de tout processus de collecte de données pour la réalisation d'évaluations particulières est de déterminer la mesure dans laquelle les évaluateurs sont d'accord entre eux et le degré de fiabilité dans leur notation des élèves. Dans des conditions idéales, les évaluateurs marqueraient chaque réponse exactement de la même façon. Il peut cependant arriver que les évaluateurs ne soient pas d'accord sur la détermination de l'exactitude de la réponse d'un élève. Il est à espérer que le processus de mise à l'essai et de formation permettra aux noteurs de se mettre systématiquement d'accord entre eux. Il est néanmoins important de mesurer continuellement le taux de concordance entre évaluateurs, ce que la procédure suivante permettra de réaliser.

Les évaluateurs travaillent tous les jours en équipe pour évaluer le premier élève de la journée. Par exemple, si une équipe d'évaluateurs comporte 4 individus, l'Évaluateur A et l'Évaluateur B forment une équipe. Les élèves sont sondés au hasard et l'Évaluateur A appelle le premier élève et l'amène au lieu / bureau de l'évaluation où attend l'Évaluateur B. L'Évaluateur B prend place de manière à ne pas pouvoir voir ce qu'écrit l'Évaluateur A. L'Évaluateur A mène l'évaluation normalement, pose les questions générales, administre les divers exercices de lecture et / ou de calcul et note les réponses de l'élève. L'Évaluateur B commence à noter séparément sa propre évaluation du même élève.

Au cours de l'évaluation, l'Évaluateur B ne pose aucunes questions mais se contente d'écouter et de noter les résultats. **Deux évaluations sont ainsi enregistrées pour le premier élève de chaque école.** Les Évaluateurs C et D vont suivre la même procédure avec un autre élève et procéder à deux autres évaluations pour le deuxième élève évalué dans l'école.

Les évaluateurs doivent soigneusement indiquer sur l'évaluation s'ils l'administrent ou s'ils écoutent et prennent note. Chaque évaluation doit comporter une section où ils notent cette information.

Une fois l'évaluation du premier élève terminée, l'Évaluateur A remercie l'élève de sa participation et le renvoie en classe. L'Évaluateur A et l'Évaluateur B comparent alors la façon dont ils ont noté l'élève. Les évaluateurs A et B devront aborder toutes sections où ils ne sont pas d'accord et déterminer la façon dont cette section aurait dû être notée. Si l'équipe de 2 personnes ne parvient pas à résoudre leur désaccord sur la notation, il conviendrait d'en prendre note et d'en parler avec tout le groupe d'évaluateurs à la fin de la journée. À noter : **une fois que les Évaluateurs A et B saisissent une réponse sur leur questionnaire, celle-ci ne devra jamais être modifiée, effacée ou corrigée une fois que l'élève a quitté la pièce.** Il est important de conserver ces points de désaccord car ils fourniront des informations sur l'accord et la fiabilité inter-évaluateurs. Il est parfaitement naturel qu'il existe un certain désaccord entre évaluateurs. Il est important d'en mesurer le degré dans le processus d'analyse de données car cela permettra d'obtenir des informations sur la mesure dans laquelle une erreur d'estimation d'évaluateur peut avoir une incidence sur les résultats de lecture observés chez les élèves.

Une fois que les évaluateurs ont discuté de leurs évaluations en équipe, ils doivent se séparer et appeler chacun leur élève suivant pour procéder à son évaluation individuelle.

Dans les écoles suivantes, il convient de modifier les équipes pout que différents membres assument le rôle d'évaluateurs qui parlent / écoutent et que chaque évaluateur soit jumelé tous les jours avec un évaluateur différent.
Un important aspect de tout processus de collecte de données pour la réalisation d'évaluations particulières est de déterminer la mesure dans laquelle les évaluateurs sont d'accord entre eux et le degré de fiabilité dans leur notation des élèves.

© 2015 Save the Children. Reproduit avec autorisation. Tous droits réservés.

9 COLLECTE DE DONNEES SUR LE TERRAIN : ETUDE PILOTE ET A GRANDE ECHELLE

9.1 Pilotage de l'instrument EGRA

Un test pilote est une étude préliminaire menée à petite échelle avant l'étude à grande échelle. Des études pilotes sont employées pour mener des évaluations au niveau des items pour comparer chaque tâche et mettre à l'essai la validité et la fiabilité de l'instrument EGRA et de tous questionnaires afférents. Les pilotes servent également à tester la logistique de mise en place de l'étude (coût, durée, efficacité des procédures et complications éventuelles) et permettre au personnel qui va déployer l'étude à grande échelle de s'exercer à l'administration dans des conditions réelles.

Pour ce qui est de l'évaluation des instruments qui vont être utilisés au cours de la collecte de données, le test pilote peut permettre de s'assurer que le contenu de l'évaluation est approprié pour la population cible (par ex. adapté aux spécificités culturelles et à l'âge des enfants, clairement énoncé, etc.). Il constitue également l'occasion de vérifier l'absence de fautes typographiques, d'erreurs de traduction ou le manque de clarté des instructions auxquelles il conviendra de remédier.

POURQUOI MENER UN TEST PILOTE D'EGRA ?

Un test pilote permet :

- De vérifier la fiabilité et la validité de l'instrument par le biais d'une analyse psychométrique
- D'obtenir des données sur plusieurs formulaires des instruments à des fins d'équivalence[21]
- De passer en revue les procédures de collecte de données, notamment la fonctionnalité des tablettes et des instruments électroniques, ainsi que les procédures de téléchargement des données depuis le terrain
- De confirmer l'état de préparation des matériels
- De passer en revue les procédures logistiques, notamment le transport et la communication, parmi les équipes d'évaluateurs, les coordinateurs sur le terrain et autre personnel

[21] S'il est nécessaire d'avoir plusieurs versions d'un instrument, pour mener des études de bases / études finales par exemple, la préparation et le pilotage de formulaires parallèles permettent à ce stade de déterminer et éventuellement de rendre moins nécessaire l'assimilation des données après une collecte à grande échelle ; se reporter à la Section 6.6 où figurent des conseils sur la création d'instruments équivalents et à la Section 10.5 f où figurent des conseils sur l'équivalence statistique.

La logistique d'un test pilote se rapproche autant que possible de celle anticipée pour la collecte complète des données, bien que toute les tâches puissent ne pas être testées et que les considérations d'échantillonnage générales (régions, districts, écoles, élèves par classe, etc.) vont probablement varier.

La **Figure 24** résume les principales différences entre le test pilote et la collecte de données à grande échelle.

Figure 24. Différences entre un test EGRA pilote et une collecte de données complète

	Test pilote	Collecte complète de données
Objectif :	Tester la fiabilité, la validité et l'état de préparation des instruments et permettre aux évaluateurs de s'exercer davantage	Procéder à l'évaluation complète des écoles et des élèves sondés
Calendrier :	Se déroule après l'adaptation	Tient compte de la période de l'année par rapport au calendrier académique ou de considérations saisonnières (jours fériés, conditions météorologiques) ; tient compte également d'ajustements après pilotage et de révisions apportées à l'instrument
Echantillon :	Echantillonnage à l'aveuglette basé sur la population ciblée pour la collecte à grande échelle	Est fonction de la population cible (niveau scolaire, langue, région, etc.)
Données :	Analysées pour réviser au besoin les instruments	Sauvegardées tout au long du processus de collecte de données (par ex. téléchargées sur une base de données externe) et analysées après la collecte de toute les données
Révisions de l'instrument :	Peuvent être apportées en fonction de l'analyse des données, avec un repilotage limité après l'apport de modifications	Aucune révision n'est apportée à l'instrument au cours de la collecte de données

9.1.1 Données de l'étude pilote et conditions d'échantillonnage

Pour s'assurer que les données pilotes suffisent à l'analyse psychométrique réalisée pour établir la validité et la fiabilité du test, il est nécessaire de recueillir un minimum de 150 scores qui ne manquent pas et qui ne sont pas nuls et ces scores non nuls doivent se situer dans une plage raisonnable et comparables aux scores non nuls anticipés dans l'étude complète. Bien que dans des conditions idéales l'échantillon pilote d'écoles et d'élèves serait sélectionné de façon aléatoire, il est le plus souvent obtenu au travers d'un échantillonnage à l'aveuglette (voir glossaire). Il y a à cela trois explications. Premièrement, le principal objectif du test pilote est de s'assurer que l'instrument fonctionne correctement ; deuxièmement, les données pilotes ne sont pas utilisées pour tirer des conclusions sur la performance générale des élèves au sein d'un pays, ce qui veut dire que l'échantillon ne doit pas nécessairement être représentatif ; troisièmement, la collecte de données à l'aide d'un échantillonnage à l'aveuglette est plus rapide et plus économique que la collecte de données par échantillonnage aléatoire.

Les écoles et les élèves retenus pour l'échantillonnage pilote doivent être semblables à la population cible de l'étude complète. Cependant, pour minimiser le nombre de scores nuls obtenus dans les résultats de l'étude pilote, les évaluateurs peuvent

intentionnellement sélectionner des élèves performants ou les planificateurs peuvent cibler spécifiquement et surreprésenter des écoles performantes.

Dans les pays où la majorité (70 à 80 %) des élèves du primaires font état de scores nuls, il serait nécessaire de sélectionner au hasard un échantillon pilote très important pour obtenir 150 scores non nuls. Par exemple, si l'on s'attend à obtenir des scores non nuls dans seulement 20 % des cas, un échantillon pilote de 750 élèves serait nécessaire pour obtenir les 150 scores non nuls requis pour l'analyse psychométrique. Une surreprésentation d'écoles performantes peut cependant sensiblement réduire la taille de l'échantillon pilote.

Pour voir comment fonctionne l'instrument EGRA quand il est administré à un groupe diversifié d'élèves, les données pilotes obtenues au travers d'un échantillonnage à l'aveuglette devra comprendre des élèves d'écoles moins performantes, de performance moyenne et plus performantes. Il convient de noter qu'en l'absence de données de performance des écoles il est conseillé de passer en revue les informations socioéconomiques relatives aux régions concernées et d'employer ces informations comme substitut aux niveaux de performance des écoles. Il n'est cependant pas recommandé que l'échantillon à l'aveuglette comprenne des classes supérieures à la population cible (par ex. une 5e année au lieu d'une 2e année), ces élèves ayant été exposés à des matériels pédagogiques différents de ceux des élèves des classes ciblées et la plage de scores nuls pouvant être très différente.

Enfin, l'échantillon pilote, au contraire de l'échantillon de l'étude EGRA complète qui limite le nombre d'élèves par classe et par école à 10 ou 12, tend à sonder des nombres plus importants d'élèves par école. Ce type de surreprésentation dans une école donnée permet de recueillir des données d'échantillonnage plus rapidement et avec moins d'évaluateurs. Là encore, cette pratique est acceptable, les données qui en résultent n'étant pas utilisées pour extrapoler les niveaux de performance générale d'un pays.

9.1.2 Etablissement de la validité et de la fiabilité du test

Fiabilité du test. La fiabilité peut être définie comme la cohérence générale d'une mesure. Cela peut par exemple porter sur le degré selon lequel les scores EGRA restent cohérents dans le temps ou pour des groupes d'élèves. En guise d'analogie tirée de la vie quotidienne, prenons une balance. Si un sac de riz est placé sur une balance cinq fois, et qu'on lit « 20 kg » chaque fois, alors la balance fournit des résultats fiables. Mais si la balance affiche un nombre différent (par exemple, 19, 20, 18, 22, 16) chaque fois que le sac est placé dessus, alors celle-ci n'est probablement pas fiable.

Validité du test. La validité porte sur la justesse des mesures et en fin de compte sur le caractère approprié des inférences ou des décisions basées sur les résultats des tests. Pour revenir une fois de plus à l'exemple de la balance, si un sac de riz pesant 30 kg est placé cinq fois sur la balance et qu'à chaque fois le poids indiqué est « 30 kg », la balance produit des résultats qui sont non seulement fiables mais également valables. Par contre, si la balance indique « 20 kg » chaque fois que le sac de 30 kg y est placé, elle fournit des résultats qui ne sont pas valables (mais qui demeurent fiables puisque la mesure, tout en étant fausse, reste constante).

La mesure la plus communément employée de la fiabilité des scores d'un test est le **coefficient alpha de Cronbach**, mesure de la cohérence interne d'un test (des progiciels statistiques tels que SAS, SPSS et Stata peuvent aisément calculer ce coefficient). S'il est appliqué à des items individuels aux sein des tâches, le coefficient alpha de Cronbach peut cependant ne pas constituer la mesure la mieux adaptée à la fiabilité de ces tâches du fait que certaines parties de l'instrument EGRA sont chronométrées. Des mesures chronométrées ou limitées dans le temps pour lesquelles les élèves doivent progresser de manière linéaire d'un item à un autre affectent le calcul du coefficient alpha de manière à en faire une estimation gonflée de la fiabilité des scores du test ; on ne connait cependant pas le degré auquel les scores sont exagérés. En conséquence, le coefficient alpha de Cronbach et des mesures similaires ne sont pas employés pour évaluer individuellement la fiabilité des tâches d'EGRA. Il ne conviendrait pas par exemple de calculer le coefficient alpha de Cronbach pour la lecture de mots inventés dans un test EGRA en considérant chaque mot inventé comme un item. Il est par contre nécessaire d'employer les scores récapitulatifs (par ex. pourcentage correct ou fluence) de tâches pour calculer le coefficient alpha général d'un test EGRA (pour toutes les tâches).[22]

Pour le coefficient alpha de Cronbach ou autres mesures de la fiabilité, plus le coefficient alpha ou la corrélation simple sont élevés, moins les scores EGRA sont sensibles à des changements quotidiens aléatoires de la condition des élèves testés ou de l'environnement de test. Une valeur de 0,7 ou plus est donc considérées comme acceptable, bien que la plupart des applications EGRA tendent à donner des scores alpha de 0,8 ou plus. D'autres types de test de fiabilité sont décrits en Annexe E.

Outre les mesures basiques de la fiabilité traitées ci-dessus, il est utile d'examiner si l'évaluation est ou non unidimensionnelle (c.-à-d. d. si elle mesure un seul concept, compétence en lecture dans le primaire par exemple). Une démarche de mesure de l'unidimensionnalité consiste à procéder à une analyse factorielle à caractère exploratoire. Ce type d'analyse suppose une structure sous-jacente (latente) dans les données de manière à déterminer le nombre total de constructs. Des valeurs propres associées peuvent permettre de déterminer si le premier facteur tient compte d'une variance suffisante pour que l'ensemble du test soit considéré comme étant unidimensionnel—c'est-à-dire qu'il mette à l'épreuve un seul construct général que l'on pourrait appeler « lecture dans le primaire ». S'il n'existe pas de limite spécifique aux valeurs propres, des scree plots servent de représentation visuelle pour déterminer s'il existe ou non plusieurs constructs (de manière à avoir une fin naturelle après le premier facteur avec un plateau de valeurs diminuées). La plupart des progiciels statistiques comportent des procédures pour une analyse factorielle à caractère exploratoire. Comme pour d'autres mesures, l'analyse n'est effectuée que sur des mesures récapitulatives des taches (par ex. pourcentage correct, fluence) et sur EGRA dans son ensemble, pas sur la justesse des items individuels au sein des tâches.

[22] Il convient de noter que ces mesures sont calculées d'abord sur les données pilotes pour s'assurer de la fiabilité de l'instrument avant l'administration complète ; mais elles sont recalculées sur les données opérationnelles (de l'évaluation complète) pour en reconfirmer le haut degré de fiabilité.

La plupart des applications EGRA comportent un premier facteur expliquant une variance suffisante pour suggérer que l'évaluation procède bien à l'estimation d'un seul construct général important.

Un autre problème se rapportant à la fiabilité est la cohérence de l'accord de plusieurs noteurs (IRR) au cours du processus de la collecte de données sur le terrain. Si deux évaluateurs sont à l'écoute du même enfant lisant une liste de mots du test EGRA, vont-ils enregistrer le même nombre de mots lus correctement ? Dans ce type de mesure de la fiabilité, les évaluateurs administrent une étude à deux, un administrant l'évaluation et l'autre se contentant d'écouter et de noter indépendamment les réponses. Une explication supplémentaire du mode d'administration IRR se trouve à la **Section 8**, notamment à la **Figure 23** « Exemple de protocole pour le contrôle de la fiabilité inter-évaluateurs ». La mesure de la concordance entre plusieurs évaluateurs peut alors être calculée en procédant à une estimation du coefficient kappa de Cohen (voir glossaire). Cette statistique (qui met en jeu un paramètre d'estimation) est considérée comme étant meilleure qu'un pourcentage de concordance, mais il conviendra de rendre compte de ces deux mesures. Si l'établissement de seuils significatifs pour le coefficient kappa de Cohen reste encore à débattre, on trouvera à la **Section J.4 de l'Annexe J** des informations sur les valeurs de référence pour la concordance des évaluateurs et des échelles communément citées pour les statistiques kappa

Pour vérifier la validité conceptuelle, il conviendra de produire des statistiques au niveau des items pour s'assurer que tous les items répondent aux attentes. Les analyses Rasch (qui reposent sur une hypothèse d'unidimensionnalité) fournit des informations sur la validité conceptuelle des plusieurs façons. Premièrement, le modèle Rasch place les items et les élèves sur la même échelle de mesure, par ordre, de facile (faibles compétences pour les élèves) à difficile (compétences élevées). L'ordre des items de moins difficiles à plus difficiles est donc la définition opérationnelle du construct. Si cette définition correspond à la conception prévue, cela indique une validité conceptuelle. S'il existe cependant des cas où les élèves n'ont pas d'items représentatifs évaluant avec exactitude leurs compétences, on dit qu'il y a sous-représentation du construct. Enfin, les analyses Rasch évaluent la performance des items par le biais de correspondances statistiques. Si les items ne mesurent pas les compétences avec exactitude ou produisent du « bruit », ils feront état de statistiques plus élevées (≥ 2,0) indiquant un manque de correspondance et devront être réévalués. On dit des évaluations comportant de nombreux items non correspondants qu'elles présentent une variance non pertinente du construct, qui est également au détriment de la validité conceptuelle. Les résultats d'un modèle Rasch peut permettre aux auteurs du test de déterminer si les items se comportent ou non selon les attentes et lesquels doivent (éventuellement) être éliminés ou révisés du fait d'un manque de correspondance. Il est essential que ces analyses soient menées tant sur les données pilotes (pour les données opérationnelles du test initial) que sur les données de l'étude complète (pour déterminer s'il convient ou non d'éliminer du score certains items particuliers).

Au cours de l'intervalle entre le test pilote et la collecte des données, les statisticiens et les psychométriciens analysent les données et proposent tous ajustements nécessaires, les linguistes et les traducteurs apportent des corrections, les versions électroniques des instruments

sont mises à jour et rechargées sur toutes les tablettes, tous les problèmes de matériel sont résolus et les évaluateurs et les superviseurs suivent une nouvelle formation sur les changements.

9.1.3 Considérations relatives au moment choisi pour le test pilote

Cette section traite des avantages et des inconvénients de deux options pour le moment choisi pour le test pilote par rapport au calendrier de formation des évaluateurs et de la collecte des données à grande échelle. *

L'essai pilote des instruments peut avoir lieu avant ou après la formation des évaluateurs. Ces démarches présentent toutes deux des avantages et des inconvénients et la décision revient souvent à des considérations d'ordre logistique et contextuel.

Si on ne dispose pas d'évaluateurs chevronnés (ayant acquis leur expérience lors d'une administration antérieure de l'évaluation), il peut être préférable de programmer le test pilote immédiatement après l'atelier de formation des évaluateurs. Un essai pilote prendra généralement un ou deux jours si l'on envoie tous les évaluateurs formés. Un avantage est que le test pilote, outre la production d'importantes données sur les instruments eux-mêmes, fournit de précieuses indications sur la performance des évaluateurs. Les personnes qui analysent les données pilotes peuvent chercher des indications que les évaluateurs commettent certaines erreurs, presser l'enfant ou lui donner plus de temps que prévu pour certaines tâches par exemple.
L'inconvénient de mener un essai pilote après la formation des évaluateurs est que les instruments utilisés au cours de cette formation ne sont pas tout à fait finalisés puisqu'ils n'ont pas été testés. Dans de nombreux cas, la mise à l'essai préalable moins formelle des instruments aura contribué à leur mise au point de sorte que le test pilote formel n'entraîne pas de révisions importantes. Toujours est-il que dans ce scénario les évaluateurs doivent être informés que de légers changements seront apportés aux instruments sur lesquels s'ils s'exercent durant la formation. La personne assurant la mise en œuvre de l'évaluation ne devra pas manquer de communiquer aux évaluateurs tous changements apportés après le test pilote avant qu'ils ne se rendent sur le terrain.

Quand l'essai pilote a lieu immédiatement après la formation des évaluateurs, il est recommandé de laisser s'écouler au moins deux semaines entre le test pilote et la collecte à grande échelle pour permettre l'analyse des données pilotes, les révisions de l'instrument, l'impression, la mise jour des interfaces de collecte électronique de données et la distribution des matériels aux équipes d'évaluation.

Il est préférable dans d'autres cas de mener l'essai pilote avant la formation des évaluateurs. Dans des contextes où une étude EGRA a déjà été récemment menée (pas plus de deux ans auparavant) et où on dispose donc d'évaluateurs formés, une brève formation de remise à niveau sur un ou deux jours peut suffire pour se préparer au test pilote. L'avantage de cette démarche est que les instruments peuvent être finalisés (en fonction de l'analyse des données provenant du test pilote) avant que commence la formation des évaluateurs. Comme pour la recommandation précédente, il est prudent de laisser s'écouler au moins deux semaines entre le test pilote et la formation des évaluateurs pour pouvoir préparer tous les matériels, non seulement pour la formation, mais aussi pour la collecte de données. Dans ce scénario, la collecte des données peut commencer dès que possible après la fin de la formation.

*La partie surlignée de ce paragraphe est tirée directement de Kochetkova et Dubeck (sous presse). © Institut de statistique de l'UNESCO. Reproduit avec autorisation. Tous droits réservés.

9.2 Procédures de collecte des données pour les études à grande échelle

Transport. Chaque équipe disposera d'un véhicule pour transporter les matériels et arriver aux écoles sondées avant que commence la journée scolaire.

Evaluation du travail. A ce jour, l'expérience a démontré que l'application de l'instrument EGRA nécessite près de 15 à 20 minutes par enfant. Au cours de la collecte complète de donnés, cela veut dire qu'une équipe de trois évaluateurs peut compléter neuf ou dix instruments par heure, soit environ sur une période ininterrompue de trois heures.

Contrôle de la qualité. Il est important d'assurer la qualité des instruments utilisés et des données recueillies. Les responsables de la mise en œuvre de l'étude doivent suivre des meilleures pratiques de recherche générales :

- S'assurer de la sécurité et du bien-être des enfants testés, notamment obtenir leur consentement.
- Préserver l'intégrité des instruments (c.-à-d. d. éviter de les communiquer au public).
- S'assurer que les données sont recueillies, gérées et communiquées de manière responsable (qualité, confidentialité et anonymat[23]).
- Contrôler les données IRR pour améliorer la qualité des données et réduire la « dérive » — également appelée biais d'effet séquentiel (voir glossaire).
- Suivre rigoureusement la planification de la recherche.

Matériel. Pour les deux phases de la collecte de données sur le terrain, il est important que les évaluateurs et les superviseurs disposent des fournitures nécessaires.

Pour la collecte des données, on aura besoin du matériel suivant :

- Tablette, pleinement chargée et comportant la version actuelle de l'instrument
- Un cahier de stimuli laminé par évaluateur (le même cahier laminé sera employé pour chaque élève testé par l'évaluateur)[24]
- Chronomètres ou minuteurs (au cas où la tablette tomberait en panne et qu'il faille recourt à des instruments imprimés)
- Crayons avec gommes et porte-bloc

[23] Anonymat : la réputation de l'étude EGRA et d'instruments similaires repose sur le consentement des enseignants / assentiment des élèves et la garantie de l'anonymat. S'il était fait mauvais usage des données—même des données pilotes— (par ex. si les écoles étaient identifiées et pénalisées), cela pourrait discréditer la démarche d'évaluation dans son intégralité pour la prise de décisions dans un pays ou une région donnés.
[24] Les feuilles de stimuli étant utilisées pour plusieurs élèves, leur plastification, bien que pas entièrement nécessaire, protège les formulaires de réponse (on peut aussi les placer dans un classeur sous protection plastique).

- Crayons ou autres fournitures scolaires à donner aux élèves en guise de remerciement de leur participation (si les planificateurs ont vérifié au préalable que cela est en conformité avec les réglementations relatives aux donateurs)

Supervision. Il est important de prévoir l'accompagnement par un superviseur de chaque équipe d'évaluateurs. Les superviseurs jouent un rôle de premier plan pour les évaluateurs et le processus de collecte. Ils sont de plus à même de gérer les relations avec le personnel des écoles, d'accompagner les élèves au lieu de l'étude et de les raccompagner en classe, de s'assurer que les évaluateurs disposent de fournitures suffisantes, de communiquer avec l'équipe de soutien et de remplir au besoin le rôle d'évaluateur.

Logistique. L'essai pilote est utile pour examiner les dispositions logistiques et le soutien prévus pour le processus de collecte de données. La collecte de données à grande échelle met cependant en jeu des aspects supplémentaires de l'étude qui sont réglés avant que les évaluateurs ne se rendent sur le terrain (vérification des écoles sondées, détermination des emplacements et dispositions relatives au transporte et au logement). Il est également essential d'établir un itinéraire qui comprendra toujours la liste des dates, des écoles, des numéros de téléphone des proviseurs et des noms des membres de l'équipe. Cette liste est mise au point par une personne qui connait bien la région. Le statisticien de l'étude établira de plus les critères et les protocoles d'échantillonnage statistique pour remplacer les écoles, les enseignants et / ou les élèves et l'équipe de formation veille à les communiquer aux évaluateurs. Enfin, pour la phase de collecte de données à grande échelle, les planificateurs organisent et coordonnent la livraison des matériels d'évaluation, y compris des copies de sauvegarde des instruments et des lettres d'autorisation des écoles et des tablettes de secours.

Avant de se rendre dans les écoles, les évaluateurs et les superviseurs devront :

- Revérifier tous les matériels
- Discuter des procédures et des stratégies d'administration pour mettre les élèves à l'aise
- Vérifier que tous les administrateurs utilisent sans difficulté un chronomètre ou leur propre montre au cas où la tablette tomberait en panne

A l'arrivée à l'école, le superviseur présente l'équipe d'évaluateurs au directeur de l'école. Dans la plupart des pays, une lettre signée du gouvernement sera nécessaire pour mener l'étude ; le superviseur explique également oralement le but et les objectifs de l'évaluation et remercie le principal de l'école de sa participation dans l'évaluation du niveau de lecture dans le primaire. Le superviseur doit souligner auprès du principal que l'objectif de la visite n'est **pas** d'évaluer l'école, le principal ou les enseignants et que toutes les informations resteront anonymes.

Le superviseur doit demander au principal si une salle de classe, une pièce ou autre endroit tranquille est disponible pour permettre à chacun des administrateurs de procéder à l'évaluation individuelle des élèves. Les évaluateurs se rendent sur le lieu

indiqué est mettent en place deux chaises ou bureaux, un pour l'élève et un pour l'évaluateur.

Il est également utile de demander si quelqu'un à l'école peut aider tout au long de la journée ; cette personne reste aussi avec les élèves sélectionnés au lieu prévu.

Tous les jours, pendant la première évaluation, le superviseur fait en sorte que les évaluateurs travaillent par deux pour administrer simultanément le test EGRA au premier élève sélectionné, un procédant activement à l'administration du test, l'autre observant en silence et prenant note des points. Cette double évaluation—qui permet d'assurer la qualité en mesurant en permanence la fiabilité inter-évaluateurs—est décrite plus loin à la Section 8.7 et en **Annexe K**.

Pendant la journée scolaire, l'accent est mis sur les élèves participant à l'étude. Les évaluateurs auront été formés sur l'établissement d'un rapport, mais le test pilote est souvent la première fois qu'ils auront travaillé avec des enfants. Les superviseurs observeront de près pour s'assurer qu'aucun enfant semble stressé ou malheureux et que les évaluateurs prennent le temps d'établir un rapport avec les élèves avant de leur demander leur assentiment. Toutes les principales observations des évaluateurs travaillant avec les enfants sont communiquées au cours du compte-rendu de l'étude pilote si bien qu'une fois que les équipes se rendent sur le terrain, elles sont mieux à même de travailler avec les élèves. Quelque chose de très simple peut faire une différence pour les élèves, s'assurer que les évaluateurs mettent leur téléphone portable en mode silencieux par exemple.

Le superviseur doit rappeler aux évaluateurs que si les élèves ne consentent pas à être testés. Ils seront gentiment remerciés et un remplacement sera sélectionné conformément au protocole établi.

Si le principal ne désigne pas un lieu réservé à l'activité, l'équipe d'évaluation trouvera un endroit tranquille (adapté à une interaction adulte / enfant) qui conviendra à l'évaluation. Ce lieu devra :

- Etre suffisamment bien éclairé pour pouvoir lire et permettre aux évaluateurs de voir les tablettes
- Comporter des bureaux disposés de sorte que les élèves ne puissent pas voir par une fenêtre ou une porte ou faire face à d'autres élèves
- Comporter des bureaux dégagés de tout papier et matériels (les matériels de l'évaluateur sont placés sur une autre table ou sur un banc pour ne pas distraire l'enfant)
- Etre hors de portée des élèves sélectionnés ; ceux qui attendent ne peuvent pas entendre ou voir le test

9.3 Sélection des élèves

Cette section présente deux options pour le sondage des élèves une fois que les

évaluateurs arrivent à l'école. La première est basée sur l'inscription scolaire et on appelle la deuxième intervalle d'échantillonnage.

9.3.1 Option 1 pour l'échantillonnage des élèves : table de nombres aléatoires

Si des données récentes et exactes sur l'inscription des élèves par école, niveau scolaire et classe sont disponibles au niveau central avant l'arrivée des équipes d'évaluation aux écoles, une table de nombres aléatoires peut servir à produire l'échantillon d'élèves. Une table de ce type peut être statistiquement plus exacte qu'un intervalle d'échantillonnage. Cette situation étant extrêmement peu probable dans la plupart des contextes, l'Option 2 est celle la plus souvent employée.

9.3.2 Option 2 pour l'échantillonnage des élèves : intervalle d'échantillonnage

Dans cette méthode d'échantillonnage, on établit un échantillon distinct pour chaque classe évaluée dans une école. L'idée est de déterminer un intervalle d'échantillonnage pour sélectionner des élèves au hasard, en commençant par le nombre d'élèves présents le jour de l'évaluation. Cette méthode comporte trois étapes distinctes.

Etape 1 : établir à partir du plan de recherche les groupes qui vont former la base de l'échantillonnage

Il est important de noter que l'Etape 1 doit être mise au point bien avant que les évaluateurs arrivent à l'école. Cette détermination est faite durant les phases initiales de planification et d'établissement du plan de recherche. Au cours de la formation des évaluateurs, il sera demandé aux candidats à la formation de s'exercer à la méthodologie d'échantillonnage en fonction du plan de recherche.

L'objectif de l'Etape 1 est de déterminer le rôle des données relatives aux enseignants, les niveaux scolaires et / ou les classes requises et les attentes en ce qui concerne la communication des résultats séparément pour les filles et les garçons. La **Figure 25** à la page suivante présente les facteurs à prendre en considération.

Etape 2 : déterminer le nombre d'élèves à sélectionner dans chaque groupe : n

La deuxième étape consiste à procéder à des calculs basés sur le nombre total d'élèves à sonder par école et le nombre de groupes concernés.[25]

Illustration : si le nombre total d'élèves à sonder est de 20 par école et que les élèves doivent être sélectionnés à partir d'une classe dans deux niveaux scolaires (par ex. une classe de 2e année et une classe de 3e année) en fonction du fait qu'il s'agit de filles ou de garçons, il faudra alors sélectionner comme suit quatre groupes et cinq

Figure 25. Déterminants des groupes d'échantillonnage

Plan de recherche—données relatives aux enseignants :	L'étude n'implique pas de données relatives aux enseignants qui seront liées aux élèves	L'étude implique des données relatives à un seul enseignant dans chaque classe qui seront liées aux données de performance des élèves	L'étude implique des données relatives à plusieurs enseignants dans chaque classe qui seront liées aux données de performance des élèves
Base de l'échantillonnage—niveau scolaire ou classe :	Niveau scolaire	Niveau de la classe – une classe par niveau scolaire	Niveau de la classe – plus d'une classe par niveau scolaire
Remarques : ï Les études peuvent porter sur un ou plusieurs niveaux scolaires. ï Outre la sélection par niveau scolaire / classe, le plan de recherché peut préciser que les élèves doivent être sélectionnés par sexe (voir ci-dessous). ï Les matériels scolaires des évaluateurs comprennent un jeu de dés pour procéder à la sélection au hasard d'une classe ou de plusieurs s'il y a plusieurs enseignants pour le niveau scolaire sondé. Le protocole d'échantillonnage précise comment utiliser les dés.			
Base de l'échantillonnage—niveau scolaire ou classe :	Niveau scolaire	Niveau de la classe – une classe par niveau scolaire	Niveau de la classe – plus d'une classe par niveau scolaire

élèves (20 ÷ 4) dans chaque groupe :

1. *5 garçons de la classe sélectionnée en 2e année*
2. *5 filles de la classe sélectionnée en 2e année*
3. *5 garçons de la classe sélectionnée en 3e année*
4. *5 filles de la classe sélectionnée en 3e année*

Etape 3 : sélectionner au hasard n élèves de chaque groupe

L'objectif de cette étape est de sélectionner les enfants précis à évaluer. La procédure recommandée est la suivante :

1. Mettre les enfants en rang à l'extérieur de la salle de classe.
    - Si l'évaluation porte sur des enfants de plus d'un niveau scolaire, commencer avec les enfants de la classe inférieure au début de la journée.
2. Compter les enfants du rang : m.
3. Diviser m par n (voir Etape 2) et arrondir au chiffre entier le plus proche : p.

[25] Voir Annexes B et C, ainsi que la Section 5, où figurent des informations supplémentaires sur le plan d'échantillonnage.

4. En commençant au début du rang, sélectionner au hasard tout enfant des premiers p enfants, puis sélectionner chaque pe enfant après cela.

Illustration : pour sélectionner n = 8 enfants d'un groupe donné :

1. *Il y a 54 enfants dans le rang (n=54)*
2. *Calculer p : 54 ÷ 8 = 6,75 ; arrondir : p = 7*
3. *Sélectionner au hasard un enfant parmi les premiers p = 7 enfants26 – par exemple, l'enfant numéro 3*
4. *Sélectionner chaque pe enfant à partir de l'enfant numéro 3 :*

   3 ; 10 ; 17 ; 24 ; 31 ; 38 ; 45 ; 52

Noter que cette procédure doit permettre d'obtenir la sélection de 9 enfants—le 9e est un remplaçant au cas où un enfant ne souhaite pas participer. Dans l'exemple ci-dessus portant sur 54 enfants, l'évaluateur doit continuer à compter jusqu'au bout du rang, puis revenir au début du rang pour sélectionner le 7e enfant suivant (qui serait le 5e enfant à partir du début du rang).

Une fois que les évaluateurs ont administré le test EGRA à tous les élèves du premier groupe (tel que déterminé à l'Etape 2), l'équipe d'évaluation reprend l'Etape 3 pour sélectionner les enfants du deuxième groupe. Le superviseur veille à ce que les évaluateurs aient toujours un élève à évaluer de manière à ne pas perdre de temps pendant l'administration du test.

9.4 Fin de la journée d'évaluation : récapitulation

Dans la mesure du possible, toutes les entrevues se déroulant pour une école sont terminées à la fin de la journée scolaire. Il conviendra cependant d'établir une alternative en début de journée et de parler à l'avance avec les évaluateurs et les superviseurs pour déterminer la pratique la mieux adaptée aux conditions locales. Si l'école n'a qu'une équipe et que certaines évaluations ne sont pas terminées à la fin de l'équipe, le superviseur ira trouver les élèves restants pour leur demander d'attendre après la fin de la journée scolaire. Le directeur de l'école ou les enseignants prennent dans ce cas les dispositions nécessaires pour informer les parents que certains enfants rentreront tard de l'école.

9.5 Téléchargement des données recueillies sur le terrain

Si les données sont recueillies électroniquement (meilleure pratique recommandée à l'heure actuelle—voir Section 7), les planificateurs mettent en place des moyens qui permettent aux évaluateurs de transmettre chaque jour les données à un serveur central pour éviter toute perte de données éventuelle (c.-à-d. si un dispositif portable est perdu ou cassé). Si cela ne s'avère pas possible, des procédures de sauvegarde sont mises en place. Les procédures permettant d'assurer que les données sont correctement téléchargées ou sauvegardées seront les mêmes pour l'essai pilote

et la collecte de données à grande échelle. Le test pilote joue un rôle important en ce qu'il constitue l'occasion de s'assurer que ces procédures fonctionnement correctement. Les évaluateurs transmettent leurs données au serveur central par Internet sans fil, soit en se connectant sur un réseau sans fil dans un lieu public ou un cybercafé, soit par service de transmission de données mobiles (3G). Lors de la planification de
la collecte de données, les planificateurs doivent tenir compte de facteurs tels que la disponibilité d'un réseau porteur, la compatibilité entre routeurs et modems et la capacité technique des évaluateurs pour trouver des solutions pratiques et fiables.

Durant le pilotage, les évaluateurs s'exercent à télécharger et à sauvegarder les données selon la méthode sélectionnée. Un analyste vérifie que les données sont bien téléchargées sur le serveur, puis examine la base de données pour s'assurer qu'elle ne comporte pas d'erreurs techniques (c.-à-d. redondance de noms de variables) avant de procéder à la collecte de données à grande échelle.

AVANTAGES DE TELECHARGER ET DE PASSER REGULIEREMENT EN REVUE LES DONNEES

Au cours de la collecte de données, leur téléchargement et examen réguliers permettent de repérer toutes erreurs avant la fin de la collecte des données, évitant ainsi d'avoir à renvoyer les collecteurs de données sur le terrain après plusieurs semaines de travail. Un téléchargement quotidien permet de plus d'éviter la perte de données si une tablette est perdue, volée ou cassée. Les données peuvent être vérifiées pour s'assurer que l'évaluation porte bien sur le niveau scolaire concerné, que les évaluateurs se rendent dans les écoles sondées et que le nombre correct d'élèves est évalué et pour vérifier l'absence de toutes autres incohérences. Une communication et des mises à jour constantes pour informer l'équipe du projet des dates de collecte des données, de soumission aux analystes des données téléchargées et de délais éventuels ou autres raisons entravant le téléchargement quotidien des données peuvent faciliter l'examen des données et permettre de savoir quels résultats attendre et à quel moment.

Les procédures de sauvegarde pour la collecte électronique de données comportent notamment la disponibilité de versions imprimées de l'instrument. Une fois chaque évaluation terminée sous format papier, le superviseur la passe en revue pour s'assurer que le formulaire est lisible est bien rempli (tous les codes d'école sont indiqués et les cases sont cochées sans ambigüité). Le superviseur ou autre individu désigné est chargé de l'organisation et de la protection des formulaires remplis qui ne devront être accessibles qu'aux personnes autorisées.

10 PREPARATION DES DONNEES EGRA

Cette section traite du processus de nettoyage et de préparation des données EGRA. Après leur collecte, il faudra les recoder et leur appliquer des formules pour créer des variables sommaires et super-sommaires. Noter qu'il est supposé ici que les erreurs d'échantillonnage provenant du plan de l'étude ont été correctement pondérées et ajustées.

Presque toutes les études EGRA consistent en une forme quelconque d'échantillon complexe stratifié à plusieurs degrés. Il faudra soigneusement procéder au contrôle, à la vérification, à la révision, à la fusion et au traitement des données avant leur dernière mise au point et analyse. Ces processus doivent être réalisés par deux statisticiens (extrêmement chevronnés) au maximum. Une personne procède à ces étapes tandis que l'autre vérifie le travail. Une fois les données traitées et finalisées, tout individu disposant d'une certaine expérience dans l'exploration d'échantillons complexes et de données hiérarchiques peut se familiariser avec les objectifs de la recherche, les questionnaires et les évaluations, la méthodologie d'échantillonnage et la structure des données, puis procéder à l'analyse des données.

Il est supposé dans cette section que le statisticien procédant au traitement des données jouit d'une longue expérience dans la manipulation d'échantillons complexes et les structures de données hiérarchiques ; on y également donne des précisions sur le traitement des données EGRA.

10.1 Nettoyage des données

Le nettoyage des données recueillies constitue une étape importante avant leur analyse. Il est rappelé que le nettoyage et le contrôle des données doivent être effectués par un statisticien disposant de l'expérience nécessaire dans ce type de traitement de données.

Le suivi de la qualité des données se fait au fur et à mesure de leur collecte. Le calendrier de collecte des données et les rapports provenant de l'équipe opérationnelle permettent au statisticien de relier les données qui sont téléchargées aux nombres d'évaluations prévus pour chaque école, à la langue, à la région ou autre unité d'échantillonnage sera alors à même de communiquer avec le personnel sur le terrain pour remédier à toutes erreurs ayant pu être commises au cours de la saisie des données et pour s'assurer que le nombre correct d'évaluations est bien mené dans les bonnes écoles et aux jours prévus. La triangulation des renseignements signalétiques est un important aspect de l'étude car elle permet de

confirmer que l'on dispose d'un échantillon suffisamment grand pour répondre aux objectifs de l'étude.

Pouvoir rapidement et correctement identifier ce manque de cohérence permettra de nettoyer les données mais également d'assurer que la collecte ne doit pas être retardée ou reprise du fait de légères erreurs.

La **Figure 26** présente une petite liste de vérification que les statisticiens suivront au cours du processus de nettoyage pour s'assurer que toutes les données EGRA sont entièrement et uniformément nettoyées aux fins de leur analyse.

Figure 26. Liste de vérification pour le nettoyage des données

- ☐ **Passer en revue les évaluations incomplètes.**

  Les évaluations incomplètes sont vérifiées pour déterminer leur caractère exhaustif et leur degré de convenance et estimer si elles doivent rester dans les données finales. Chaque projet aura convenu de critères à l'appui de ces décisions. Par exemple, les évaluations qui ne sont pas complètement terminées pourront être au besoin conservées—pour répondre aux conditions de taille de l'échantillon—pour en tirer des informations incomplètes ou les évaluations pourront être vérifiées comme étant exactes et ne manquant pas d'importants renseignements signalétiques.

- ☐ **Eliminer toutes évaluations « test » complétées avant la collecte officielle des données.**

  Vérifier que toutes les évaluations faisant partie de la version « nettoyée » des données employées pour l'analyse existent et ont bien été réalisées au cours de la collecte officielle des données.

- ☐ **S'assurer que toutes les évaluations sont liées aux renseignements signalétiques de l'école correspondante.**

  Eliminer toutes évaluations qui ne sont pas correctement identifiées ou collaborer avec l'équipe opérationnelle pour s'assurer que toutes les évaluations non étiquetées sont correctement identifiées et étiquetées.

- ☐ **S'assurer que l'enfant a donné son assentiment et que celui-ci a bien été consigné pour chaque observation.**

  Eliminer immédiatement toutes évaluations réalisées sans que l'évaluateur ait demandé ou consigné l'assentiment exprès de l'enfant à évaluer.

- ☐ **Calculer les scores de toutes les tâches chronométrées et non chronométrées.**

  On trouvera à la Section 10.2 des Informations sur la notation de tâches chronométrées et non chronométrées.

- ☐ **S'assurer que les scores de toutes les tâches chronométrées se situent dans une fourchette de notation acceptable et réaliste.**

  Au cours de la collecte de données, les évaluateurs peuvent faire des erreurs

ou un mauvais fonctionnement du logiciel de collecte de données peut et ne résultent pas d'une erreur quelconque. Eliminer toutes observations extrêmes qui sont déterminées comme étant des erreurs d'évaluation de manière à ne pas fausser l'analyse des données. Il n'est pas nécessaire d'éliminer toutes les observations correspondant à cet élève, ceci pouvant avoir une incidence sur la taille de l'échantillon pour l'analyse dans d'autres tâches.

Il suffit d'éliminer toute notation associée à la tâche concernée et s'avérant être une erreur

10.2 Traitement des tâches EGRA

On trouvera au début de cette section la nomenclature employée pour les variables et les tâches EGRA courantes, puis une explication du type d'information à recueillir au cours de l'évaluation et la façon d'extraire le reste des variables nécessaires des variables brutes collectées. Noter que l'**Annexe L** du manuel présente l'exemple de manuel de codes pour les variables d'un jeu de données EGRA.

Les noms des variables EGRA ont essentiellement la structure suivante :

<prefix>_<core><suffix>

Exemples :

e_letter_sound1
e_letter_sound2
e_letter_sound_time_remain

Pour préserver la cohérence dans toutes les études EGRA, il est important de donner les mêmes noms aux variables des tâches. La **Figure 27** présente une liste des noms de variables pour les tâches EGRA ainsi que les noms de variables pour les scores chronométrés (si la tâche est chronométrée).

10.2.1 <prefix>_

Si un élève est évalué dans plus d'une langue, il est important de différencier les langues à l'aide d'un préfixe. Les langues secondaires devront être indiquées par un préfixe, e_ pour l'anglais ou f_ pour le français par exemple.

Remarque sur les tests comportant plusieurs passages : dans de nombreuses études pilotes, il existe plus d'une version de la même tâche. On peut par exemple avoir trois versions différentes du passage employé pour la facilité de lecture à haute voix, ainsi que trois différentes séries de questions pour la compréhension. Les préfixes représentent dans ce cas la lettre correspondant à la langue et le numéro de la tâche. En anglais, les noms de variables seraient donc e1_oral_read<suffix>, e2_oral_read<suffix> et e3_oral_read<suffix> pour permettre de différencier les passages de lecture auquel ils se rapportent.

Figure 27. Nomenclature des variables des tâches EGRA et noms des variables pour les scores chronométrés

Nom de la variable de la tâche	Nom de la tâche	Nom de la variable chronométrée de la tâche	Nom de la tâche chronometrée
letter	Identification des lettres (noms)	clpm	Noms de lettres corrects par minute
letter_sound	Identification des lettres (sons)	clspm	Sons de lettres corrects par minute
fam_word	Lecture de mots familiers	cwpm	Mots corrects par minute
invent_word	Lecture de mots inventés	cnonwpm	Mots inventés corrects par minute
oral_read	Facilité de lecture à haute voix	orf	Facilité de lecture à haute voix
read_comp	Compréhension en lecture		
list_comp	Compréhension à l'écoute		
syll_sound	Identification des syllabes (sons)	sscpm	Sons de syllabes corrects par minute
oral_vocab	Vocabulaire oral		
vocab	Vocabulaire		
maze	Labyrinthe		
dict	Dictée		

10.2.2 <suffix>

Les tâches EGRA permettent la collecte de données pour chaque item auquel un élève aura répondu correctement, incorrectement ou pas du tout parce que le temps alloué s'est écoulé. Autrement dit, pour la tâche d'identification des lettres (sons) par exemple, les données comporteront une variable pour chaque item testé. Il est possible, à partir de ces informations, de calculer le récapitulatif de toutes les variables de scores non chronométrés. Les suffixes indiquent le numéro de l'item de la tâche et le récapitulatif des scores.

Le suffixe sera le numéro de l'item dans la tâche ou toutes variables additionnelles associées à celle-ci (par exemple : _auto_stop, _attempted, _time_remain). Le suffixe peut être le numéro de l'item qui se trouve dans la tâche. Par exemple, si la section portant sur la compréhension en lecture comportait cinq items, les noms des variables seraient e1_read_comp1, e1_read_comp2, e1_read_comp3, e1_read_comp4, e1_read_comp5, e1_read_comp_attempted.

Noter qu'il ne figure pas de tiret du bas « _ » dans ces noms de variables entre le radical et le numéro de suffixe de 1 à 5. Les variables ne seraient donc PAS : e_read_comp_1, e_read_comp_2, e_read_comp_3, e_read_comp_4, e_read_comp_5.

Dans le nom des variables non-item, un tiret du bas « _ » est placé entre le radical et le suffixe. Les variables EGRA non-item sont appelées e_read_comp_attempted and e_read_comp_score.

La **Figure 28** présente des exemples de la façon dont les variables EGRA sont nommées en fonction de la langue et du nombre de sections reprises dans l'instrument.

Figure 28. Nomenclature des suffixes pour les variables de score et d'item

Suffixe	Etiquette de suffixe de variable	Valeurs possibles
1-#	N° d'item	0 "Incorrect" 1 "Correct" . <missing> "Not asked/didn't attempt"
_score	Score brut	0 - # Items in Subtask
_attempted	Tentatives totales de réponse à des items	0 - # Items in Subtask
_score_pcnt	Pourcentage correct	0-100
_score_zero	Indicateur de score zéro	0 "Score>0" 1 "Score=0"
_attempted_pcnt	Pourcentage correct de tentatives	0-100

Les variables récapitulatives suivantes sont alors calculées :

- **_score.** Somme des réponses correctes aux items (codée 1).
- **_attempted.** Compte des réponses correctes et incorrectes aux items, codé 1 ou 0.
- **_score_pcnt.** Subtask_score divisé par le nombre d'items possibles dans la tâche.
- **_score_zero.** Oui (marqué 1) si l'élève a marqué zéro ; sinon, Non (codé 0).
- **_attempted_pcnt.** _score divisé par _attempted.

10.3 Tâches chronométrées

Dans l'instrument EGRA, une tâche chronométrée est conçue pour être calculée à la minute. Les réponses, lettres ou mots individuels par exemple, doivent être codées correctes, incorrectes ou pas de réponse/n'a pas répondu. L'évaluateur doit faire sur place la distinction entre incorrect (codé zéro) et pas de réponse, étant donné qu'il ne sera pas possible d'analyser les tentatives de réponses à des items s'il n'y a pas de différentiation.

Outre les réponses aux items, les variables récapitulatives suivantes doivent être incluses dans les données brutes pour les tâches chronométrées :

1. **Subtask_time_remain.** C'est le temps restant dans une tâche si un élève a terminé la tâche avant l'expiration du temps alloué. Cette variable récapitulative servira à calculer la vitesse à la minute. Elle est consignée en secondes. On donne en général un maximum de 60 secondes pour réaliser une tâche. Le temps restant sera donc 60 secondes moins le temps que l'élève aura mis pour réaliser la tâche.
2. **Subtask_auto_stop.** Pour procéder efficacement à l'évaluation et éviter que les élèves ne marquent une pause prolongée pour répondre à des questions dont il est évident qu'ils ne connaissent pas la réponse, on met fin à l'évaluation quand un élève est incapable de répondre aux quelques premiers items—généralement les 10 premiers items (ou moins). On donne à un élève qui ne peut pas répondre avant l'arrêt automatique un code de 1 pour cette tâche, 1 signifiant oui, l'élève a fait l'objet d'un arrêt automatique. Ce score porte sur la tâche dans son ensemble et n'est pas consigné au niveau de l'item.

   Pour créer des variables récapitulatives, les réponses aux items individuels sont fixées à 1 pour les réponses correctes, 0 pour les réponses incorrectes et manquantes pour pas de réponse/n'a pas répondu.

On appelle souvent la vitesse à la minute le niveau de fluidité. Les tâches chronométrées sont généralement administrées sur une période de 60 secondes, si bien que seuls les élèves qui répondent aux items d'une tâche ou lisent le passage avant que s'écoule le délai feront état d'une valeur de fluidité différente de leur score brut. L'unité de mesure finale est soit le nombre de lettres correctes soit le nombre de mots corrects par minute.

Le taux per_minute [par minute] est calculé à l'aide de la formule suivante :

$$\text{tâche par minute} = \frac{\text{Score pour la tâche}}{\text{Temps alloué pour la tâche-temps restant pour la tâche}} \times 60$$

10.4 Tâches non chronométrées

Comme pour les tâches chronométrées, les réponses à ces items doivent être codées comme étant correctes, incorrectes ou pas de réponse/n'a pas répondu.

Pour créer des variables récapitulatives, les réponses aux items sont fixées à 1 pour les réponses correctes, 0 pour les réponses incorrectes et manquantes pour pas de réponse/n'a pas répondu.

Remarque sur l'activité de compréhension en lecture :

Comme il est de règle, si la compréhension en lecture est calculée à partir du même passage que celui employé pour évaluer la lecture à haute voix, les élèves sont évalués sur le nombre de questions de compréhension auxquelles ils ont répondu dans la section du passage qu'ils ont su lire.

Par exemple, si cinq questions de compréhension en lecture étaient fondées sur la lecture du passage jusqu'au 9^{e}, 17^{e}, 28^{e}, 42^{e} et 55^{e} mot respectivement et qu'un élève a lu jusqu'au 33^{e} mot, il sera évalué sur les trois premières questions de compréhension en lecture. Les tentatives de réponses sont marquées : correctes, incorrectes ou pas de réponse. Les deux dernières questions seront codées pas posées.

Bien que cet étalonnage puisse varier en fonction du contexte, les élèves sont en général considérés comme étant à même de lire couramment, avec compréhension, quand ils lisent un passage dans son intégralité et peuvent répondre correctement à 80 % ou plus des questions de compréhension en lecture. Pour calculer cette valeur, on crée une nouvelle variable récapitulative : **read_comp_score_pcnt80**, qui est correcte (codée 1) si le pourcentage du score de compréhension en lecture est de 80 % ou plus ; autrement, elle est déterminée comme étant incorrecte (codée 0).

10.5 Equivalence statistique

L'équivalence est une procédure statistique employée pour convertir les scores de plusieurs versions d'un test à une même échelle de mesure commune. Ce processus de conversion tient compte de tout problème résultant de différences entre plusieurs versions pour permettre d'ajuster le score d'une version à la valeur équivalente sur une autre. L'équivalence permet donc d'estimer le score que les enfants évalués avec une version auraient reçu s'ils avaient été évalués avec une différente version du test (Holland & Dorans, 2006 ; Kolen & Brennan, 2004).

Les recherches effectuées sur l'équivalence statistique portant sur un petit échantillon (ce qui convient pour presque toutes les procédures d'équivalence en rapport avec EGRA) ont montré que lorsque des différences réelles de scores entre plusieurs tâches sur deux versions de test sont inférieures à environ 1/10 d'un écart type, l'erreur peut en fait excéder le biais résultant d'une non équivalence (Hanson, Zeng, & Colton, 1994 ; Skaggs, 2005). L'équivalence n'est par conséquent pas recommandée pour les petits échantillons quand la différence entre les scores de plusieurs versions ne dépasse pas 1/10 d'un écart type.

Si l'équivalence est nécessaire, il conviendra de tenir compte de plusieurs points importants.

Le premier est que les réalisateurs de l'instrument doivent considérer et reconnaitre la convenance des tâches à l'équivalence. Quatre techniques sont employées pour l'équivalence statistique : *équivalence des items appartenant au même domaine, équivalence de personnes en commun, équivalence par théorie classique des tests (TCT) et équivalence par théorie de la réponse aux items (TRI).*

Equivalence des items appartenant au même domaine : employée quand les instruments ou les tâches comportent certains items que toutes les versions de test ont en commun.

Ces items communs (également appelés items fixes) devraient, dans des conditions idéales, représenter au moins 20 à 25 % du total des items contenus dans l'évaluation et constituent une mini-version de l'évaluation dans son ensemble (en termes de difficulté et de variation).[27] Il est également important de s'assurer que les items fixes restent à la même place sur toutes les versions du test (par ex. si un item fixe particulier est le cinquième sur la version A du test, c'est aussi le cinquième item sur la version B du test). Les items restants (c.-à-d. items non fixes) peuvent être soit des items tirés de l'instrument original et remaniés, soit des items entièrement nouveaux.

Le principe de base sur lequel repose l'équivalence d'items appartenant au même domaine est que la difficulté des items fixes est identique pour toutes les versions de l'évaluation. Les scores sont donc ajustés pour tenir compte de la difficulté générale du test en fonction du sous-score pour les items fixes. Il existe plusieurs méthodes pour procéder à l'équivalence d'items appartenant au même domaine (notamment équivalence enchaînée et post-stratification), mais leur complexité sort du champ du présent manuel.

En fin de compte, l'équivalence d'items appartenant au même domaine convient mieux aux t6aches comportant un nombre suffisant d'items (un minimum recommandé of 20 à 25 items), du fait qu'une erreur statistique est moins susceptible de se produire (en supposant une petite taille d'échantillon semblable).

[27] Il y a controverse quant à la proportion exacte d'items fixes requis mais 20 à 25 % est la recommandation souvent citée.

Equivalence de personnes en commun : également appelée modèle avec un seul groupe ou modèle de groupe aléatoirement équivalent, cette méthode est employée quand les instruments ou les tâches sont conçus pour mesurer des constructs identiques mais ne contiennent pas d'items fixes. C'est actuellement le type d'équivalence le plus communément employé pour EGRA parce qu'elle ne nécessite pas de connaissances en procédures d'équivalence au stade de conception de l'instrument. Pour cette démarche, plusieurs versions de l'évaluation EGRA sont pilotées avec un échantillon d'élèves (chacun prenant toutes les versions). Le principe de base est que les différences dans les scores du test pour toutes les versions de l'évaluation peuvent être considérées comme des différences dans la difficulté du test (plutôt que de différences dans les aptitudes des élèves), les mêmes élèves répondant à la même version du test. Cette démarche est nécessaire pour

l'exercice de facilité de lecture à haute voix de l'évaluation EGRA, étant donné qu'il n'est pas possible de créer des items fixes pour cette tâche (et que les informations au niveau de l'item ne sont pas pertinentes—prérequis pour l'équivalence TRI comme nous l'avons vu plus haut). Dans ce modèle, chaque passage figure dans chaque série (première, deuxième, troisième) et chaque passage figure avec six autres passages. L'ordre des passages

ETAPES REQUISES POUR L'EQUIVALENCE DE PERSONNES EN COMMUNAU COURS DE L'ETUDE PILOTE

Pour optimiser l'efficience et pour tirer parti au maximum du modèle d'équivalence de personnes en commun, il conviendra d'employer le scenario suivant au cours du stade pilote si l'on a le temps (et la prévoyance) de créer un grand nombre de passages parallèles et si l'on dispose du financement nécessaire pour procéder à un essai pilote auprès d'au moins 500 élèves.[28]

Dans ce scenario, il est suggéré que les réalisateurs de l'évaluation EGRA créent 10 passages de compréhension de la lecture comportant chacun cinq questions (10 séries), en ayant recours à des experts pour leur construction pour les rendre aussi parallèles que possible au début. Trois passages distincts (et les questions de compréhension les accompagnant) seraient ensuite administrés à chaque échantillon d'élèves. Le modèle pourrait (hypothétiquement) ressembler à ce qui est présenté à la **Figure 29** (avec 10 versions de 3 séries et 15 questions chacune).

[28] Ce seul essai pilote pourrait remplacer plusieurs pilotes portant sur 150 à 200 élèves (ce qui n'est pas rare dans un travail de développement). C'est simplement une question de coûts par rapport aux avantages et l'intérêt d'avoir 10 passages évalués.

Figure 29. Exemple de modèle contrebalancé

Nombre d'élèves	Première série	Deuxième série	Troisième série	Versions de test pilote, par lettre
50	1	2	4	A
50	2	3	5	B
50	3	4	6	C
50	4	5	7	D
50	5	6	8	E
50	6	7	9	F
50	7	8	10	G
50	8	9	1	H
50	9	10	2	I
50	10	1	3	J
500				

subit une rotation de manière à minimiser les effets de l'ordre. Cette démarcher nécessite un échantillon de 500 élèves (assignés au hasard en 10 sous-échantillons recevant chacun une des 10 versions du test).

Il est donc possible d'obtenir des mesures solides de la difficulté relative de chaque item et de chaque série. On fait ensuite correspondre les séries pour obtenir une comparabilité maximum pour les pré- et posttests, le degré de confiance changeant dans les scores au niveau de l'échantillon pouvant être utile.

Equivalence par théorie classique des tests : les modèles d'équivalence basés sur la TCT établissent des rapports entre le total des scores sur différentes versions du test. C'est une démarche plus « conventionnelle »

à l'étalonnage des tests et c'est celle la plus souvent employées pour l'équivalence portant sur de petits échantillons. Les modèles TCT sont notamment l'équivalence moyenne, linéaire, arc de cercle et calibration par percentiles égaux. Ce manuel n'explique pas en détail chacune de ces démarches mais on trouvera en **Annexe M** des recommandations supplémentaires sur celles-ci.

L'équivalence TCT est utile pour les données linéaires et pour de petits échantillons. Elle n'est pas recommandée pour les tâches comportant un nombre relativement limité d'items (moins de 10). Pour les tâches comportant 10 à 25 items, il peut être possible d'avoir recours à un modèle de pré-étalonnage TCT en pilotant plusieurs versions de test récemment mises au point ainsi que des versions de base et en comparant les statistiques au niveau des items sur toutes les versions. Dans le contexte des tâches EGRA, cette démarche est particulièrement utile pour étalonner la facilité de lecture à haute voix.

Equivalence par théorie de la réponse aux items : l'équivalence TRI repose sur le principe de l'établissement de rapports d'équivalence s'appuie par le biais de modèles qui relient les variables observables et les variables latentes. Cette démarche présente l'avantage d'employer le même modèle mathématique pour les caractéristiques des personnes et les caractéristiques des instruments. L'équivalence TRI a de plus l'avantage d'être mieux compatible avec la nature du test tout en permettant d'étalonner des tâches comportant peu d'items. L'équivalence TRI est cependant complexe tant du point de vue de son concept que celui de sa procédure et nécessite des échantillons considérablement plus grands que l'équivalence TCT (à l'exception su modèle de Rasch qui exige la même taille d'échantillon que la TCT— soit environ 100 à 150 participants).

L'équivalence TRI est en conséquence extrêmement utile pour le post-étalonnage (c.-à-d. d. étalonnage portant sur des données opérationnelles ou d'étude à grande échelle—par rapport à un post-étalonnage qui s'effectue à l'aide de données pilotes), quand on dispose de la capacité et de l'expertise technique nécessaires. Dans la majorité du travail EGRA, l'équivalence TRI (faisant intervenir des modèles Rasch) peut être particulièrement utile pour le pré-étalonnage portant sur des tâches comportant peu d'items (un désavantage des modèles d'équivalence TCT), tant que ces tâches comportent des données utiles au niveau des items. Il s'agira notamment de tâches de compréhension de la lecture, de compréhension à l'écoute, de dictée, de vocabulaire et de labyrinthe.

Une réflexion supplémentaire sur les procédures d'étalonnage des tests figure en **Annexe M.**

10.6 Mise de l'évaluation EGRA à la disposition du public

On s'attend à ce que l'USAID mette à la disposition du public des fichiers à usage public (FUP) contenant des données d'évaluations du niveau de lecture dans le primaire par le biais du portail d'Analyse secondaire pour le suivi des résultats en éducation [Secondary Analysis for Results Tracking Education (SART Ed)] et de la Bibliothèque de données de développement [Development Data Library (DDL)] et, de plus en plus, au travers de la plateforme EdStats de la Banque mondiale. D'autres donateurs pourront décider de suivre un processus similaire pour les collectes de données qu'ils financent.

Les fichiers à usage public sont des jeux de données nettoyées, finalisées et dépersonnalisées destinées au public. Ces jeux de données contiennent toutes variables pertinentes nécessaires à une bonne analyse de données mais toutes les informations identifiables qu'ils contiennent sont masquées pour protéger l'identité des individus et des établissements. La nouvelle version des directives en matière d'information (Update to Reporting Guidance) de l'USAID définit comme suit les données nettoyées, finalisées :

- **Données nettoyées.** L'exécutant a vérifié qu'il n'y a pas d'erreurs ou d'incohérences dans le contenu, qu'aucune information ne manque, etc. et y a éventuellement remédié.

- **Données finalisées.** Les jeux de données comportent tous indicateurs dérivés ou secondaires que l'exécutant a employés pour calculer les valeurs des indicateurs comprises dans les rapports. L'exécutant a terminé le traitement du jeu de données et aucune modification supplémentaire n'est anticipée.
- **Données dépersonnalisées.** Des mesures ont été entreprises pour protéger la vie privée et l'anonymat des individus et des écoles associés à l'évaluation. L'organisation chargée de la mise en œuvre a collaboré avec son Comité d'examen institutionnel pour s'assurer que les participants à l'évaluation sont correctement protégés.

On trouvera en **Annexe N** des recommandations et directives supplémentaires sur la manière de nettoyer, finaliser et dépersonnaliser les données de manière à pouvoir les distribuer au public. Une fois qu'un FUP d'un jeu de données a été créé, des informations bien documentées sont nécessaires pour permettre à l'utilisateur public de se familiariser avec les données. Pour les études EGRA menées avec le financement de l'USAID, les informations suivantes sont fournies aux utilisateurs :

- Informations contextuelles, définition de la population concernée par exemple—notamment la source de la base d'échantillonnage employée pour extraire l'échantillon, description du plan de sondage et date de la collecte des données.

- Toute documentation pertinente, notamment les questionnaires et les outils d''évaluation employés. (Si le sondage EGRA a été réalisé en vue d'évaluer l'impact du programme, les questionnaires ne sont publiés qu'après la fin du programme, de manière à ne pas compromettre les matériels pouvant être utilisés pour des études EGRA futures.)
- Le rapport écrit d'analyse des données soumis à l'USAID et approuvé.

Les exécutants doivent reconnaitre à quel point il est important de documenter les noms et les descriptions des variables clés, ainsi que les paramètres nécessaires à une bonne analyse des données. Des directives spécifiques sur la manière d'inclure les descriptions et les noms des variables à des fins d'analyse de données figurent en **Annexe N.**

A la rédaction de cette édition du manuel, plusieurs référentiels de données de l'USAID étaient en phase de développement. Il reste cependant important que les données d'évaluation du niveau de lecture dans le primaire soient accessibles au public.

Il est donc recommandé aux exécutants qui ont recueilli les données EGRA de :

- Afficher sur un site accessible en ligne les FUP accompagnés d'une documentation facile à localiser (par ex. tous les items se trouvent dans un fichier zippé ou le site Web comporte un lien permettant d'accéder à ces documents).
- Créer le FUP sous forme de fichier de données non exclusives et, si possible, un fichier de données exclusives prêtes à être analysées (c'est-à-dire le plan d'enquête complexe étant déjà précisé).

- Pour un fichier de données non exclusives, créer un fichier texte sous format CSV (valeurs séparées par des virgules)
- Pour un fichier de données exclusives, créer soit un fichier STATA .dta, soit un fichier SPSS .sav (accompagné du fichier SPSS .csaplan).

11 ANALYSE ET COMMUNICATION DES DONNEES

Cette section du manuel donne un bref aperçu des types d'analyses de données qui correspondent à divers modèles de recherche, ainsi que des éléments qu'il est nécessaire d'inclure dans les rapports EGRA.

Quand ils analysent les données EGRA, les chercheurs doivent employer des statistiques descriptives et / ou déductives pour décrire les données, examiner les tendances et tirer des conclusions. Il est cependant important de comprendre les différences entre ces deux types de statistiques, ainsi que leur objectif et leur valeur.

11.1 Statistiques descriptives (non déductives)

Les statistiques descriptives (ou non déductives) sont employées pour décrire et résumer les données—souvent dans le but de voir quelles tendances on peut constater. Les statistiques descriptives ne permettent pas de tirer des conclusions à partir des données ni de mettre à l'essai les hypothèses de recherche. Le principal objectif d'une analyse descriptive est de présenter les données de façon concrète pour en permettre l'interprétation (plutôt que de présenter simplement des données brutes). Les mesures les plus communes dont les analyses descriptives fendent compte sont les fréquences, les mesures de tendance centrale (moyennes et médianes, par ex.) et les mesures de dispersion (par ex. écarts-types et intervalles sommaires).

Comme leur nom l'indique, les statistiques descriptives servent de plus uniquement à décrire les données de l'échantillon. Dans une grande partie du travail EGRA, des échantillons sont sélectionnés qui sont représentatifs de populations plus importantes. Dans ces cas, les mesures signalées (fréquences, moyennes, etc.) sont basées sur des données pondérées et deviennent donc effectivement des statistiques déductives. On ne doit donc rendre compte de statistiques descriptives que pour des études qui sont conçues pour ne pas tirer de conclusions au-delà des échantillons ou comme fréquences non pondérées, moyennes non pondérées, etc., pour des données de sondage complexes.

Enfin, il est essentiel, avec des statistiques non déductives, de décrire entièrement l'échantillon en fonction du niveau de désagrégation à analyser et à communiquer. Par exemple, si les scores des élèves dans le rapport vont être désagrégés par langue et par classe, les statistiques descriptives de l'échantillon comprennent ces niveaux de désagrégation.

Des exemples de statistiques descriptives utiles dans les rapports EGRA seraient

notamment les fréquences et les moyennes de caractéristiques démographiques de base de l'échantillon, ainsi que les moyennes non pondérées des tâches pour tous les niveaux de désagrégation.

11.2 Statistiques déductives

Les statistiques déductives permettent aux évaluateurs de parvenir à des conclusions sur des populations entières en fonction d'un échantillon de cette population, pour tirer des conclusions sur des hypothèses concernant des paramètres de la population et pour comparer deux populations différentes (par ex. groupes expérimentaux et groupes témoins). Les statistiques déductives sont essentielles pour des évaluations d'impact et des rapports EGRA instantanés cherchant à faire des déclarations sur l'éducation pour tout un pays ou toute une région en fonction d'un échantillon d'élèves ou d'écoles dans ce pays ou cette région. Le type de statistiques déductives nécessaires pour un élève donné est fonction du modèle d'évaluation :

- **Modèles expérimentaux (ou essais contrôlés randomisés).** Deux groupes peuvent être comparés à l'aide de statistiques t appariées pour déterminer, par exemple, si les scores finaux des participants au groupe expérimental étaient supérieurs à ceux des participants au groupe témoin. Si la randomisation a été réussie et que l'échantillon est de taille suffisante, il n'est pas nécessaire de tenir compte également de différences entre les groupes expérimentaux et les groupes témoins en ternes de facteurs démographiques ou de scores initiaux, les deux groupes étant rendus identiques par le processus de randomisation.
- **Modèles quasi-expérimentaux mettant en œuvre un modèle longitudinal (qui suit chaque élève dans le temps) ou semi-longitudinal (qui suit les enseignants ou les écoles dans le temps).** Si toutes les conditions suivantes ont été remplies, les évaluateurs peuvent également employer des statistiques t et / ou des scores de progrès pour montrer le changement dans le temps : (1) les groupes expérimentaux et les groupes de comparaison étaient au départ uniformément répartis, (2) il n'existait pas de différences notables entre les deux groupes et (3) les écoles ou les élèves étaient suivis dans le temps dans les deux groupes au travers d'un modèle longitudinal. Sinon, des analyses quasi- expérimentales (telles que celles décrites plus bas) sont nécessaires pour mesurer avec précision le progrès et / ou l'efficacité du programme.
- **MQE ayant recours à un modèle transversal.** Pour comparer des groupes, les évaluateurs doivent employer une démarche d'analyse quasi-expérimentale, notamment différences de différences (DID), discontinuité de la régression, variables instrumentales ou analyse de la régression (avec de préférence une variable d'ajustement). Les DID soustraient les résultats initiaux des résultats finaux (créant ainsi des scores de progrès) pour le groupe expérimental et le groupe de comparaison, puis soustraient le score de progrès du groupe de comparaison du score de progrès du groupe expérimental pour obtenir l'effet de l'expérimentation. Cette démarche a pour but de tenir compte de différences initiales tout en reposant sur l'hypothèse que les trajectoires des résultats avant l'établissement d'une base de référence étaient cohérents d'un groupe à l'autre.

Il est souvent utile d'apparier cette démarche à une procédure correspondante (appariement des coefficients de propension par exemple) pour pouvoir s'assurer que les deux groupes sont aussi similaires que possible. Quand ils ont recours à une analyse de la régression, les évaluateurs doivent inclure des variables nominales de temps et d'expérimentation, ainsi qu'une variable nominale d'interaction pour temps × expérimentation pour déterminer l'effet (ce qui est essentiellement la démarche de régression pour une estimation DID). Des variables indépendantes additionnelles peuvent également être introduites pour tenir compte de différences entre les groupes (ce qui est important pour l'équilibre au niveau initial ainsi que pour les changements dans le temps). Des estimations de l'ampleur de l'effet sont inclues pour des analyses DID (voir **Annexe O, Sous-annexe O-1** où figure un exemple d'analyse DID)

Quel que soit le type de modèle employé, les évaluateurs doivent vérifier l'équilibre entre les groupes expérimentaux et les groupes témoins / de comparaison au niveau initial en examinant à la fois les variables des principaux résultats et les principaux prédicteurs pour assurer la compatibilité entre les deux groupes. Si les groupes ne sont pas comparables, les évaluateurs doivent envisager l'emploi de techniques d'appariement, appariement des coefficients de propension par exemple, pour améliorer la robustesse du modèle et des analyses. Si des données secondaires sont disponibles avant la collecte des données initiales, celles-ci peuvent servir à sélectionner des écoles, enseignants ou élèves de comparaison viables à partir desquels on pourra recueillir des données de base. Si ces données secondaires ne sont pas disponibles, l'équipe d'évaluation pourra cependant envisager d'augmenter la taille de l'échantillon pour au moins l'unité de comparaison pour s'assurer qu'il existe de bonnes correspondances pour chaque unité expérimentale, en supposant qu'elle dispose de ressources adéquates pour ce faire. Tous les rapports mettant en œuvre un MQE contiennent un tableau d'équilibrage et ce peut être le cas également de ceux qui emploient un modèle expérimental. Ce tableau est essential pour montrer l'équilibre entre les mesures et remédier à certains problèmes de biais de sélection.

La validité interne doit de plus être assurée pour tous types de modèles par le biais de l'examen des effets de l'attrition, de la mortalité, des retombées et des antécédents.

11.3 Types d'analyse de régression

Etant donné que la régression est la façon la plus courante d'analyser les rapports et les valeurs prévues des variables dans les données EGRA, il est important d'examiner brièvement les différents types d'analyses de régression pouvant être entreprises. L'analyse de régression par moindres carrés ordinaires (MCO) convient bien aux données EGRA qui ont normalement distribué des valeurs résiduelles, quand on emploie une variable continue telle que le score en fluence de lecture à haute voix.

Beaucoup de pays en voie de développement font cependant état de scores qui se regroupent autour de zéro, ce qui fait que la distribution des scores est très inégale. Confrontés à des données de ce type, les évaluateurs doivent envisager l'emploi d'une analyse de régression binomial, régression des probits ou régression logistique par exemple, qui permet aux évaluateurs d'examiner les résultats binomiaux (par exemple si un élève remplit les critères locaux en facilité de lecture ou si un élève obtient des scores de zéro pour une tâche de lecture particulière.

11.4 Rapport d'analyse des données

L'objectif de l'analyse des données EGRA est double : améliorer l'efficacité programmatique et fournir des résultats aux clients, organisations partenaires et représentants du gouvernement par le biais de brefs exposés et de rapports complets sur les programmes. Les principes d'orientation suivants sont nécessaires quand on reconnait que différents objectifs, ainsi que différentes audiences, détermineront la structure et le contenu de ces rapports :

1. **Objectifs et limitations.** Le rapport doit clairement faire état des objectifs de l'étude et de ses limitations.
2. **Simplicité du langage.** Les principaux résultats doivent être présentés en termes clairs, concis et non techniques.
3. **Visualisation des données.** Les données doivent être visualisées pour permettre au grand public d'interpréter les résultats. Les visualisations doivent être « indépendantes » de manière à ce que le visuel puisse être interprété sans qu'il soit nécessaire de lire un texte additionnel (voir **Annexe O, Sous-annexe O-2** où figure un exemple de graphique employé pour rendre compte visuellement de données.
4. **Analyses descriptives et déductives.** Le principal rapport présente un résumé des résultats de l'analyse descriptive des données, notamment distributions moyennes et distributions groupées. Les analyses statistiques déductives sont employées pour établir des pondérations, pondérations de post- stratification et les erreurs standard pour tenir compte (au besoin) de la complexité du modèle d'étude.
5. **Distributions des scores.** Pour chaque estimation de scores d'élèves dont il est rendu compte, il faudra représenter graphiquement une visualisation de la distribution des scores (voir **Annexe O, Sous-annexe O-3**) pour faciliter l'interprétation par le lecteur de l'estimation donnée ; par exemple, alors que le score moyen peut être produit, la distribution qui l'accompagne met en perspective la mesure dans laquelle l'estimation est « représentative » des scores des élèves. Ceci est particulièrement important si la distribution des scores des élèves est non-normale. Dans certains cas, il peut être utile de présenter les scores médians des élèves en plus des scores moyens et des distributions.
6. **Niveaux de désagrégation.** Les résultats de la désagrégation des données par sexe, classe, langue et autres variables dignes d'intérêt doivent être décrits de manière convenant au modèle de la recherche.

7. **Rapport de tous les résultats.** Toutes les fois où on mène des tests statistiques de comparaison des moyennes pour comparer divers groupes de sujets (sexe ou langue par exemple) ou des analyses statistiques bivariée / multivariée (corrélations par exemple) pour examiner les rapports entre différentes variables, les résultats doivent en être rapportés même s'ils ne sont pas statistiquement significatifs.
8. **Justification d'estimations déductives**. Les éléments suivants doivent accompagner toutes les estimations déductives rapportées (y compris, mais sans s'y limiter, moyennes, médianes, modes et proportions) :
    - Précision – soit comme intervalle de confiance (IC) de 95 % pour les estimations, soit un score t et une valeur p pour les comparaisons en plus des erreurs standard
    - Taille d'échantillon
9. **Ampleurs d'effet.** Toutes les fois que sont présentés les résultats de comparaison de données de plusieurs groupes (différences entre données initiales et données finales ou entre garçons et filles ou entre élèves d'écoles rurales ou urbaines, par exemple), il faudra rendre compte de l'ampleur d'effet de la différence.
10. **Equivalence.** Dans les modèles expérimentaux et quasi-expérimentaux, il faudra établir l'équivalence of des bases (What Works Clearinghouse, 2015).

ANNEXES REQUISES

Les chercheurs doivent présenter en annexe les détails de la méthodologie et les résultats de l'analyse. Ces annexes peuvent être longues et rédigées en langage technique. Les annexes suivantes doivent accompagner le rapport :

Détails de la méthodologie, des méthodes et de la collecte des données :

- Objectifs de l'étude
- Modèle
- Méthodes et processus de collecte des données
- Instruments de collecte de données
- Méthode et résultats d'équation si des différents outils ont été employés à différents points de l'étude
- Paramètres et attrition d'échantillonnage (pour les études longitudinales)
- Détails sur la pondération
- Limitations
- Résultats d'analyse de fiabilité de test (coefficient alpha de Cronbac, corrélations entre items et totaux)
- Coefficient de corrélation interne (CCI)

Détails d'analyses n'étant pas inclus dans le rapport principal :

- Description de l'échantillon
- Détails des analyses descriptives
- Détails des analyses bivariées / multivariées

12 EMPLOYER LES RESULTATS POUR GUIDER L'ACTION

12.1 Stratégie de diffusion

L'objectif ultime de l'EGRA est d'améliorer l'enseignement de la lecture et les résultats en lecture. Il est bien connu que la mise en œuvre d'une évaluation ne suffit pas à elle seule pour atteindre ce but. Les résultats doivent être employés de manière à informer la politique, la pratique de l'enseignement, le soutien pédagogique dans les classes et à l'extérieur de l'école et l'emploi de ressources pour combler les lacunes du système. Que la solution soit sous forme de formation des enseignants pour les orienter dans l'utilisation de meilleures méthodes d'enseignement, d'achat de livres à distribuer dans les écoles et les classes ou de mobilisation d'une communauté, le dialogue et les actions qui font suite à une évaluation EGRA sont tout aussi importants que la collecte de données.

« La motivation pour la création de l'évaluation EGRA était d'accéder rapidement à l'information pour informer l'amélioration de l'apprentissage dans les pays à faibles revenus »
– Dubeck & Gove (2015)

S'assurer que les résultats se traduisent par des mesures implique une démarche à plusieurs niveaux à commencer par un travail de planification et de mise en œuvre, dont on a parlé ailleurs dans ce manuel (par exemple, définir clairement l'objectif de l'évaluation, procéder soigneusement à l'échantillonnage de la population à évaluer, appliquer les instruments supplémentaires appropriés, comme observations de classe ou questionnaires, et impliquer les décideurs locaux dans la planification et la mise en œuvre). L'analyse des données après-mise en œuvre porte principalement sur des questions de recherche concrètes. Enfin, les résultats doivent être communiqués aux audiences concernées de manière culturellement et contextuellement appropriée pour être compris et donner suite à des actions.

Cette section porte principalement sur la mise au point d'une démarche de diffusion stratégique mais encourage vivement les lecteurs à ne pas oublier que la crédibilité—et en conséquence l'emploi efficace—des résultats vont dépendre en grande partie de l'exécution soigneuse des étapes précédentes dans la collecte de données.

Avant de planifier des activités de diffusion des résultats, les dirigeants de la mise en œuvre doivent se pencher au minimum sur les questions « qui ? », « quoi ? » et « comment ? »

1. **Qui** va utiliser les données ?

Et pour chaque type d'audience identifiée par la question ci-dessus :

2. **Quelles** données présenter ?

3. **Quel** type d'information (et sous quel format) l'audience comprend-elle mieux ?

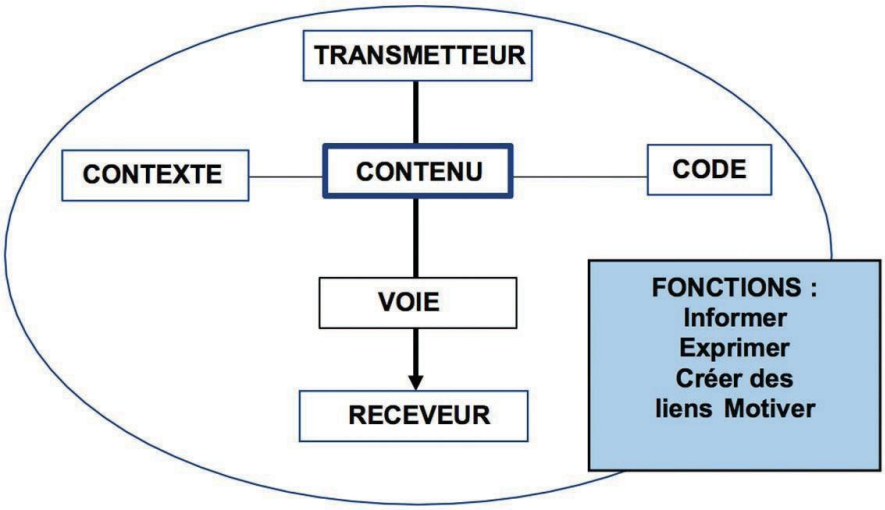
4. **Comment** ces données vont-elles être employées ?

Ces questions permettront de modeler la façon dont les résultats sont préparés et distribués aux différents intervenants (par ex. communautés, représentants au niveau des écoles, gouvernements, représentants du ministère, enseignants, parents). L'organisation mettant en œuvre l'EGRA n'a pas toujours besoin d'assurer la large diffusion des résultats si elle est plutôt à même de toucher les personnes concernées (au bon moment, avec le bon message). Quand le temps ou le budget est limité, on cherchera à toucher les individus qui sont influents et à même de contacter les décideurs et les enseignants.

12.1.1 Communication des résultats

La **Figure 30** rappelle brièvement les éléments impliqués dans tous types de communication. Il est facile de se concentrer uniquement sur le contenu du message à communiquer ; la clarté et l'impact de ce message sont cependant fonction du contexte dans lequel il se situe, du « code » (langage, ton) dans lequel il est rédigé et de la voie ou moyen de communication (imprimé, verbal, numérique) par lequel il est transmis. Cela signifie que le communicateur (« transmetteur ») doit connaître l'audience et être informé de la façon dont cette audience est habituée à évaluer et à traiter l'information—y compris compétences de base en lecture/écriture et familiarité avec la visualisation de données (sous forme de tableaux et graphiques par exemple). Pour certaines audiences, une information technique détaillée sera bien reçue, alors que pour d'autres la meilleure démarche peut être d'utiliser les résultats pour raconter une histoire qui en dépeint la signification dans le contexte. Le lieu d'accès à l'information et la personne qui la diffuse peuvent également changer la façon dont le message est reçu et interprété.

Figure 30. Cadre de communication



Source : adapté de Jakobsen (1960).

La **Figure 31** tirée des Notes d'orientation pour la planification et la mise en œuvre d'EGRA (RTI International & Comité international de secours, 2011, p. 82), donne un aperçu des audiences potentielles pouvant être ciblées pour la diffusion des résultats EGRA.

Figure 31. Aperçu des audiences potentielles

Niveau	Audience	Pertinence
International	Donateurs ou bailleurs ?	Les donateurs peuvent appuyer les efforts promotionnels, les interventions pilotes en matière de lecture, les évaluations futures et la mise à échelle des meilleures pratiques. Le soutien des donateurs est crucial, quand on considère les ressources limitées dont les ministères de l'éducation dans beaucoup de pays en voie de développement disposent pour l'innovation. Les donateurs peuvent faciliter l'exploitation des résultats EGRA, notamment quand les gouvernements ne peuvent ou ne veulent pas prendre des mesures.
	Universitaires ou professionnels	Les universitaires et autres professionnels s'intéressent souvent aux résultats des études EGRA parce qu'ils fournissent des informations précieuses sur les résultats d'apprentissage et les questions pédagogique dans de nombreux contextes peu étudiés et permettent de déterminer les meilleures pratiques
National	Ministère ou service de l'éducation	Le processus et les résultats du sondage EGRA peuvent encourager les représentants du gouvernement à faire des efforts supplémentaires pour mettre l'accent sur la lecture comme compétence fondamentale. Les résultats EGRA peuvent constituer un catalyseur pour diverses interventions du gouvernement portant sur la lecture dans les premières classes du primaire et peuvent également rehausser l'intérêt porté à l'intégration des sondages EGRA dans les évaluations nationales de l'éducation. Dans certains cas, le gouvernement (ou des éléments clés à l'intérieur de celui-ci) sont déjà convaincus qu'il est nécessaire d'agir et les évaluations EGRA ne font que confirmer et augmenter la précision des connaissances existantes. EGRA peut également permettre de formuler l'intervention qui s'ensuit.

Figure 31. Aperçu des audiences potentielles

Niveau	Audience	Pertinence
	Autorités budgétaires	Les résultats des évaluations EGRA peuvent aider à convaincre les autorités gouvernementales d'allouer des fonds pour diriger davantage de ressources publiques vers le service/ministre de l'éducation pour la lecture dans le primaire. Cela ne se produit généralement que lorsque les autorités financières sont convaincues que le gouvernement et ses partenaires disposent d'une stratégie viable pour remédier à la situation détectée par l'évaluation EGRA. L'idée qu'il existe des moyens d'améliorer les résultats peut donc être intégrée à la communication avec ce type d'autorités.
	Syndicats d'enseignants	Le soutien apporté aux enseignants pour la lecture dans le primaire est crucial et les interventions correspondantes sont essentielles. L'aide et la collaboration avec les syndicats peuvent influencer les perceptions des enseignants de l'évaluation EGRA comme plate-forme pour un changement positif plutôt que comme moyen de critiquer la performance des enseignants. Les syndicats d'enseignants peuvent mettre en lumière l'importance des attentes en matière de lecture à différents niveaux du primaire et communiquer aux enseignants ainsi qu'à d'autres audiences les principales conclusions de l'évaluation.
	Société civile et médias	La société civile et les médias peuvent susciter une prise de conscience, faire pression sur les décideurs (gouvernement) et, dans certains cas, promouvoir la durabilité.
Régional	Chefs de cabinets ou de bureaux	Les autorités en matière d'éducation au niveau de la province, de l'état ou du district constituent une importante audience et un partenaire clé, notamment dans les cas où les services de l'éducation sont décentralisés. Les résultats EGRA peuvent encourager tant la concurrence que la coopération.
Communautaire	Dirigeants de la communauté	Les dirigeants communautaires peuvent susciter une prise de conscience sur la lecture et les bonnes pratiques auprès des membres de la communauté—particulièrement les parents—et faire de plus pression sur les autorités locales et les écoles.
	Parents	Les parents ont une énorme influence sur les habitudes de lecture des enfants et des interventions de sensibilisation ciblant les parents peuvent susciter une prise de conscience sur les attentes en matière de lecture à différents niveaux scolaires, encourager la lecture à la maison et faire davantage pression sur les écoles et les décideurs pour qu'ils accordent une plus grande priorité à la lecture.
	Société civile	La société civile peut donner son appui aux activités de dissémination au niveau de la communauté et accroître la responsabilisation au niveau de base.
Ecole	Directeurs	Les directeurs des écoles doivent être conscients de l'importance de la lecture dans les premières années du primaire et des bonnes pratiques dans l'enseignement de la lecture pour mieux soutenir les élèves, les enseignants et les parents. Les directeurs des écoles peuvent également participer aux interventions.
	Enseignants	Les enseignants constituent une audience cruciale qui devra faire l'objet d'interventions et d'une sensibilisation aux attentes et aux bonnes pratiques d'enseignement en matière de lecture à différents niveaux du primaire.

Source : RTI International & Comité international de secours, 2011, p. 82 [Tableau 6.2].

12.1.2 Démarches de diffusion

Les résultats provenant des sondages EGRA ou instruments similaires d'évaluation dans les premières classes du primaire tendent à concerner les décideurs et les

représentants du gouvernement, notamment quand l'outil EGRA est employé comme diagnostic national au niveau du système ou quand les résultats sont associés à une évaluation de l'impact d'une innovation pédagogique évolutive (ou mise à échelle). S'il est vrai que les processus politiques peuvent « traduire la volonté du peuple en des politiques publiques et établir des règles qui apportent de manière efficiente et efficace des services à tous les membres de la société » (Crouch & Winkler, 2008, p. 3), aborder les questions ou les problèmes d'un système d'éducation (important service) commence alors avec la volonté politique. Une hypothèse clé est que les données faisant état des résultats d'apprentissage du système d'éducation peuvent servir à stimuler cette volonté politique.

« A long terme, augmenter la capacité de l'emploi de l'information pour informer l'instruction est critique à l'amélioration de l'apprentissage pour les quelque 250 millions d'enfants dans le monde entier qui n'acquièrent pas de compétences de base ».

ñ UNESCO, 2014

Niveau national

Il est important de faire valider les résultats auprès du gouvernement (ou autre client/intervenant) avant toute diffusion supplémentaire. Cela peut se faire par un évènement organisé sur un jour ou deux (souvent appelé « **dialogue politique** ») qui réunit soit certains représentants du gouvernement soit une représentation plus large de plusieurs groupes d'intervenants. Le format de présentation en personne permet aux intervenants de poser des questions et de fournir des informations sur le contexte local pouvant informer l'interprétation des résultats et améliorer le rapport final. Ce type de rencontre comprend la présentation de résultats par les chercheurs, des déclarations de politique et de pertinence par des représentants ou des agences du ministère de l'éducation mettant en œuvre des programmes d'amélioration de la lecture, des témoignages de collecteurs de données sur le terrain et l'établissement de groupes de travail pour débattre des résultats et des plans d'action. Des réunions régionales ou communautaires peuvent faire suite à cette rencontre nationale.

Naturellement, si l'application EGRA a été dès le début mise au point avec le gouvernement et qu'il existe d'étroits rapports entre les experts EGRA et les experts gouvernementaux, la validation aura tendance à se dérouler dans un climat positif. Ainsi donc, dans la plupart des cas (mais pas tous), une étroite collaboration dès le début réduira la possibilité d'un dialogue litigieux au cours de la validation.

Outre des ateliers nationaux et régionaux, d'autres stratégies de diffusion regroupent la préparation et la dissémination de rapports numériques ou imprimés, de prospectus, de banderoles et de matériels infographiques, des réunions communautaires ou des présentations multimédia (programmes de radio, clips vidéo, documentaires). L'emploi d'enregistrements audio ou vidéo au cours des séminaires ou comme stratégie de diffusion constitue un moyen évident et persuasif de voir les différences entre un niveau de lecture médiocre (par exemple un enfant qui lit sans comprendre 10 à 15 mots par minute) et un bon niveau de lecture (par exemple un enfant qui lit et comprend 60 mots par minute). Il est alors bien plus facile de placer les résultats quantitatifs dans ce cadre de référence.

ETUDE DE CAS : REFORME POLITIQUE AU NIVEAU NATIONAL AU YEMEN

En début 2012, le Ministère de l'éducation du Yémen a demandé au programme d'amélioration communautaire (CLP), initiative de développement financée par l'USAID et mise en œuvre par Creative Associates International d'appuyer la mise au point d'une nouvelle démarche d'enseignement de la lecture dans les classes du primaire. Les résultats d'une évaluation des compétences fondamentales en lecture administrée par EdData II en 2011 dans trois gouvernorats ont été présentés en mars 2012 lors d'un dialogue politique mettant en jeu divers intervenants, y compris le Ministère de l'éducation.

Le programme CLP de Creative Associates, basé à Sana'a, a facilité l'examen des résultats EGRA lors du développement du programme de lecture dans les premières classes du primaire au Yémen.

La médiocrité des résultats de l'évaluation EGRA 2011 (27 % des élèves de 3e année ne savaient pas lire un seul mot d'arabe) a renforcé la détermination du Ministère de l'éducation à réformer la façon dont la lecture de l'arabe est enseignée dans les premières classes du primaire. Le Ministère a donné la priorité à la réforme de la lecture dans les premières classes du primaire afin d'établir une fondation pour renverser des années de sous-développement dans le secteur de l'éducation. Durant cette période, le Ministre de l'Education a adopté un rôle central pour mobiliser le Ministère autour de la Stratégie de lecture au primaire au Yémen (YEGRA). Le programme YEGRA, qui bénéficie de l'appui de l'USAID, a été mis à l'essai dans 310 écoles au cours de l'année scolaire 2012–2013 et, suite à son succès, a été élargi à 800 écoles en 2013–2014 et à 1 200 écoles en 2014–2015.

Au cours du dialogue national en 2012, plusieurs factions souhaitaient mettre fin à tous changements du programme scolaire jusqu'à ce que la constitution soit finalisée. Il s'agissait d'une tentative visant à s'assurer que toutes les parties aient l'occasion de participer à tous nouveaux programmes académiques. La seule exception au moratorium sur la révision du programme scolaire, convenue par les délégués au dialogue nationale, était le programme scolaire pour les 1ère, 2e et 3e années, étant donné qu'on portait déjà une certaine attention au programme YEGRA qui avait commencé à faire état d'importantes améliorations dans la lecture au primaire, le perfectionnement et la motivation des enseignants et l'engagement parental.

Le Ministère de l'éducation a émis un certain nombre de décrets pour assurer la réussite du nouveau programme YEGRA. Un de ces décrets visait la transparence et la qualité dans la sélection des formateurs. Plutôt que de laisser le Ministère nommer des formateurs de districts et de gouvernorats, un processus de sélection rigoureux a été mis en place. Il portait notamment sur un processus de demande et de sélection basé sur divers critères pertinents pour l'enseignement de la lecture de l'arabe au primaire.

Un autre décret voulait que dans les écoles ressortissant du programme YEGRA le temps consacré à la lecture de l'arabe en 1ère, 2e et 3e années passe de 5 minutes à 70 minutes par jour, cinq jours par semaine. Le Ministère a également émis un décret visant à s'assurer que les enseignants du primaire participant à la formation YEGRA étaient effectivement les enseignants de ces classes. Le Ministère souhaitait éviter une situation courante dans laquelle des enseignants favoris étaient sélectionnés par le directeur de l'école pour participer à la formation, qu'ils soient ou non affectés aux classes et aux sujets faisant l'objet de la formation.

Une décision politique d'étendre initiative YEGRA à l'échelon national a été finalement prise au cours de la deuxième année d'essai du programme, quand le CLP financé par l'USAID a mis le programme

YEGRA en œuvre dans plus de 800 écoles ; l'agence d'aide allemande GIZ a participé à une mise en œuvre dans 72 écoles et, avec le financement de la Banque mondiale, le Ministère a pu le mettre en œuvre dans 200 écoles. La Banque mondiale a apporté son soutien au Ministère pour étendre le programme à 14 700 écoles additionnelles à l'échelon national en 2014–2015. Autrement dit, à l'année scolaire 2014–2015, après deux ans de mise à l'essai du programme YEGRA, toutes les 16 000 écoles du pays avaient mis en œuvre le programme de lecture dans le primaire—y compris les 70 minutes obligatoires par jour—dans toutes les classes de 1ère année.

Source : adapté de du Plessis, El-Ashry & Tietien (à paraître).

Niveau local

Les citoyens, notamment ceux qui administrent eux-mêmes le test EGRA (ou demandent simplement aux enfants de leur lire), se rendent rapidement compte que les enfants ne lisent pas et souhaitent intervenir pour participer au changement. Les membres de la communauté semblent souvent réaliser qu'il existe un grave problème de lecture parmi les enfants de leurs écoles. L'évaluation EGRA a dans certains cas facilité cette observation, mais dans d'autres cas il s'agit d'une réponse à des inquiétudes qui ont déjà été exprimées. Des évaluations du niveau de lecture dans le primaire ont récemment permis de recueillir un certain enthousiasme autour de l'éducation et d'autres mouvements populaires pour susciter une prise de conscience des compétences (ou incompétences) des élèves en matière de lecture. La mobilisation de la communauté et la sensibilisation au niveau local est une importante « étape sur la voie de l'accélération de réformes dans le domaine de l'éducation visant à améliorer l'alphabétisme » (Gove & Cvelich, 2011, p. 45).

A ce jour, les applications du test EGRA ont principalement servi à ouvrir une discussion au niveau national et à encourager les ministères à prendre des mesures. L'objectif est de rendre compte des résultats EGRA pour le niveau le plus bas ou les couches les plus basses de l'échantillon (souvent au niveau national mais parfois au niveau régional ou des districts). Les évaluations EGRA étant menées par échantillonnage, il est impossible d'en communiquer des résultats par école. Il n'est pas non plus généralement rentable de procéder à une évaluation EGRA dans chaque école (et auprès d'un nombre suffisant d'enfants dans chaque école) pour produire des résultats particuliers à une école. Du fait que la sensibilisation de la communauté et l'encouragement de son engagement en matière d'alphabétisation et de lecture dans le primaire présentent un double avantage, les professionnels peuvent cependant envisager plusieurs stratégies.

La première stratégie recommandée est la production des résultats EGRA sous forme de rapport bref à communiquer aux dirigeants des communautés et des écoles pour encourager le dialogue sur la situation en matière d'alphabétisme en général (pas particulièrement pour l'école mais pour le niveau où celle-ci se situe). Ce rapport s'accompagne d'explications sur la façon dont chaque tâche est en rapport avec l'instruction et sur ce que les enseignants peuvent faire pour améliorer les résultats des élèves. Des exemples de plans de cours et des activités suggérées peuvent

également être partagés avec les écoles pour indiquer comment les intervenants au niveau de la communauté peuvent eux-mêmes prendre des mesures locales.

Deuxièmement, pour obtenir et rendre compte des scores en alphabétisme au niveau des écoles, les professionnels peuvent avoir recours à d'autres outils d'évaluation du niveau d'alphabétisme, comme ceux utilisés par Pratham pour le Rapport annuel sur l'état de l'éducation (http://www.asercentre.org/p/141.html) ainsi qu'à des protocoles d'évaluation de l'alphabétisme administrés en groupe et mis au point par le biais de modèles d'échantillonnage par lots pour l'assurance de la qualité (LQAS) (voir l'étude de cas ci-dessous, ainsi que Batchelder, Betts, Mulcahy-Dunn & Stern, 2015, Mulcahy-Dunn, Valadez, Cummiskey & Hartwell, 2013 et Valadez, Mulcahy-Dunn & Sam-Bossman, 2014).

ETUDE DE CAS : PRATIQUE PILOTE ET CONTINUE DE CONTROLE LQAS AU GHANA

Le Ghana a établi son Conseil national d'inspection pour mettre au point des outils qui permettent de contrôler la qualité de l'éducation dans le cadre de son Programme national d'accélération de l'alphabétisation. Un programme LQAS pilote a été entrepris au Ghana de manière à tester un outil de ce type pour contrôler la qualité de l'éducation et déterminer les domaines pouvant bénéficier d'un soutien additionnel au niveau local. Ces activités pilotes, conçues pour améliorer les résultats scolaires, portaient notamment sur l'emploi de l'instrument EGRA pour mesurer les compétences des élèves en lecture.

Les résultats de l'étude LQAS pilote ont montré que les scores des élèves en lecture étaient faibles dans toutes les écoles sondées. Les compétences fondamentales en lecture, évaluées par le test EGRA, permettaient à la méthodologie LQAS de distinguer les résultats de l'évaluation d'une école à une autre. La démarche LQAS avait pour but de catégoriser les circonscriptions en districts « dont les résultats répondent aux attentes » et en districts « dont les résultats sont inférieurs aux attentes ». Ces classifications étaient « basées sur la détermination que 80 % des écoles répondent ou non aux critères d'intérêt spécifiés dans un ensemble d'indicateurs en rapport avec la performance des enseignants et les résultats des élèves » (Mulcahy-Dunn et al., 2013, p. 9).

Suite à un dialogue politique au Ghana sur les activités LQAS pilotes et le test EGRA de base, il a été procédé à une diffusion additionnelle des résultats sous forme de « forums de groupes de districts » orientés vers les collectivités. Ces forums ont permis de diffuser les résultats plus localement. C'est au cours de ces forums que les intervenants locaux ont déterminé qu'une surveillance continue était nécessaire. Du fait de cet intérêt et du succès de l'étude LQAS pilote au district d'évaluation et d'un rendement scolaire efficace et rentable, l'initiative de surveillance LQAS a reçu un financement additionnel pour s'étendre à plusieurs districts du Ghana.

12.2 Détermination de critères de référence propres aux pays

Une des vertus de l'évaluation EGRA est que les principes scientifiques sur lesquels elle repose correspondent relativement bien à l'idée que les personnes non initiées se font de ce que signifie « savoir lire » : la notion de « connaître ses lettres », d'être capable de lire sans hésitation et à une vitesse raisonnable et pouvoir répondre à quelques questions sur ce qu'on a lu.

Ainsi donc, pouvoir rendre compte que les enfants ne savent pas reconnaître les lettres ou ne peuvent les lire qu'extrêmement lentement est une mesure que la plupart des individus peuvent interpréter. Les données produites par les tests EGRA (ou d'autres types d'évaluations administrées oralement dans le primaire) permettent de relater la mesure dans laquelle les écoles desservent les besoins les plus fondamentaux des élèves.

Il est néanmoins utile, pour attirer l'attention des décideurs et représentants officiels sur la question de la façon dont les élèves apprennent à lire, de pouvoir comparer les résultats d'une manière ou d'une autre. Des critères de référence sont particulièrement utiles pour la lecture car ils établissent des attentes et des normes en matière de performance. Il est nécessaire d'avoir des critères de référence pour évaluer le progrès réalisé dans tout pays ou contexte donné. Un critère de référence stable peut permettre de traduire aisément un objectif établi en des mesures du progrès à des moments donnés. Par exemple, si l'objectif est que tous les enfants sachent bien lire à la fin de la 3e année, un critère de référence peut indiquer le pourcentage d'élèves parvenant à différents niveaux de facilité de lecture dans une classe et une année donnée—indiquant ainsi si un progrès est réalisé vers cet objectif général. De plus, les critères de référence sont utiles quand ils servent à communiquer publiquement le degré des progrès effectués (par ex. bulletins scolaires ou surveillance et rapport au niveau national).

Les normes permettent l'application d'attentes communes et mesurables à des populations régionales ou nationales mais permettent également une prise de décisions décentralisée sur la façon de guider les enfants vers la réalisation de ces objectifs. Ces mêmes mesures objectives servent de plus de mécanisme de responsabilisation, les écoles—et parfois les enseignants—étant tenus de rendre des comptes sur les résultats pédagogiques. Des études ont montré que les systèmes d'évaluation où les enjeux sont importants affectent le comportement des enseignants et des administrateurs, mais pas de manière cohérente ou prévisible. Il faut donc procéder prudemment à l'établissement de critères de référence pour veiller à ce que le système d'éducation puissant les employer pour mesurer le progrès et déterminer les domaines devant faire l'objet d'efforts supplémentaires, plutôt que pour prendre des sanctions sévères.

ETUDE DE CAS : ETABLISSEMENT DE POINTS DE REPERE NATIONAUX AU KENYA

En novembre 2015, l'USAID, au travers du projet EdData II, avait financé des ateliers d'établissement d'objectifs et d'analyse comparative dans 12 pays : Egypte, Ghana, Jordanie, Kenya, Liberia, Malawi, Mali, Pakistan, Philippines, Tanzanie, Cisjordanie et Zambie. Dans chacun de ces pays, les données portant sur les compétences fondamentales en lecture ont servi à l'établissement de critères de référence.

D'août 2011 à 2014, le Ministère kenyan de l'éducation, des sciences et de la technologie (MoEST) a mis en œuvre l'initiative Mathématiques et lecture dans le primaire pour améliorer les compétences de base en lecture des élèves (PRIMR). La conception de l'initiative PRIMR s'inspire d'un essai

expérimental d'amélioration de la lecture entrepris dans le district de Malindi au Kenya par la Fondation Aga Khan et RTI en 2007 (RTI International, 2008).

Au cours du programme PRIMR, les compétences en lecture d'élèves sélectionnés au hasard dans des écoles participant au programme et dans des écoles témoins ont été mesurées au travers d'un test EGRA. Le programme PRIMR étant conçu comme essai contrôlé randomisé, il a été faisable de déterminer son impact sur l'apprentissage. Les données recueillies par le test EGRA ont ensuite servi à informer les décisions politiques ministérielles d'investissement dans des méthodes d'enseignement particulières pouvant entraîner des améliorations tant pour les filles que pour les garçons et dans toutes les classifications socioéconomiques.

Le MoEST a de plus invité l'initiative PRIMR à « à mettre en œuvre des programmes de recherche standardisés en collaboration avec le Conseil national Kenyan des examens (KNEC) pour établir des critères de référence en lecture et en calcul » (RTI International, 2014b, p. 47). Les résultats du rapport PRIMR initial ont permis d'évaluer « des critères de référence appropriés sur la fluence et la compréhension en matière d'apprentissage des élèves » (RTI International, 2014b, p. 47). De plus, un outil pour la valorisation d'une éducation de qualité basé sur les outils EGRA employés pour l'initiative PRIMR a été mis au point pour le Kenya par le personnel chargé du suivi et de l'évaluation de ce programme pour mesurer l'apprentissage au primaire. Le MoEST a alors incorporé cet outil dans ses niveaux de classement de référence. Le personnel du programme PRIMR a présenté au comité de direction du KNEC le modèle de la recherche, les résultats initiaux et des premières recommandations, ainsi que des résultats pour l'établissement de critères de référence. Au cours de cette dernière présentation, le programme PRIMR a été à même de montrer son niveau de précision de mesure de l'apprentissage des élèves ayant influencé les critères de référence. Lors d'un exercice au cours de la réunion du KNEC, il a été demandé aux membres du comité directeur de déterminer les critères de référence appropriés pour ce qui est de la fluence et de la compréhension des élèves avec les données PRIMR initiales recueillies par les tests EGRA. L'outil de valorisation a été par la suite modifié pour incorporer les niveaux de classement de référence du MoEST.

12.2.1 Que sont des critères de référence ?

Les critères de référence sont définis comme étant « une norme ou un point de référence auquel des éléments peuvent être compares et en fonction desquels ils peuvent être évalués » (dictionnaire Oxford en ligne, http://www.oxforddictionaries.com) ; « un critère de performance à un moment donné (jalon) » et « des scores cibles dérivés empiriquement et critériés qui représentent un progrès adéquat en lecture » (Dynamic Measurement Group, Inc., 2010, p. 1).

Aux fins de ce manuel, un « critère de référence » est synonyme de « norme » en ce qu'il définit un niveau de performance désiré et pouvant être atteint à un moment donné. Une « évaluation de référence » est donc un diagnostic administré à intervalles réguliers et employé pour déterminer si les élèves continuent à progresser vers l'atteinte de normes désirées. Des « scores de référence » peuvent également être établis à de seuils limites qui permettent d'interpréter la signification d'un score particulier ; par exemple, l'établissement de seuils « de base », « intermédiaires » et « de bon niveau » peuvent permettre de déterminer les profils des élèves en fonction d'une définition de maîtrise partielle ou totale.

Les critères de référence peuvent également être associés à des « cibles » (buts, objectifs) qui définissent les attentes pour la population concernée ; par exemple, si la référence détermine à quelle hauteur il faut mettre la barre, l'objectif définit combien d'enfants vont franchir cette barre. Par exemple : « 60 % des élèves satisfont au critère de référence au cours de la première année ; 80 % des enfants satisfont au critère de référence au cours de la deuxième année ». L'établissement d'objectifs est particulièrement important quand les résultats sont médiocres. La cible définit une étape intermédiaire vers la réalisation de l'objectif.

Comme nous l'avons décrit plus haut dans la Section 12.1 portant sur la diffusion, les messages ne sont efficaces dans les activités de communication que si l'audience désirée peut les comprendre. Il n'est généralement pas efficace de fournir les résultats EGRA sans point de référence dans des environnements où les mesures de la fluidité (c.-à-d. 20 mots corrects par minute) sont mal connues ou où les évaluations tendent à être communiquées sous forme de pourcentage de réponses correctes. Un critère de référence est un point de repère qui permet d'interpréter la performance parce qu'il fournit un niveau de résultat attendu. Dans le cas de critères de référence pédagogiques, ceux-ci rendent les objectifs scolaires plus spécifiques (« devra pouvoir lire couramment » plutôt que « devra pouvoir lire à raison de 40 mots corrects par minute à la fin de la 2e année »). Ces attentes doivent cependant tenir compte de la réalité du pays plutôt que d'être adoptées d'autres pays ou d'autres langues. Les données EGRA peuvent servir à définir des critères de référence et les administrations ultérieures peuvent produire des données qui permettent d'évaluer la performance dans le temps en fonction de ces critères de référence. A des fins de comparaison, l'**Annexe P** présente des normes de fluidité de lecture à haute voix pour l'anglais.

Définitions

- Un objectif est une aspiration à long terme, peut-être sans valeur numérique

  Objectif : tous nos enfants doivent lire
- Une mesure est une unité de quantification valide et fiable

  Mesure : « mots corrects par minute dans la lecture d'un passage »
- Un critère de référence est une étape numérique vers la réalisation d'un objectif qui met en jeu la mesure

  Critère de référence : lit 45 mots corrects par minute et en comprend 80 %
- Une cible est une variable mettant en jeu le critère de référence

  Cible : % d'enfants répondant ou dépassant le critère de référence ou moyenne atteinte par les enfants mettant en jeu la mesure

Source : LaTowsky (2014)

12.2.2 Critères pour l'établissement de critères de référence

L'établissement de critères de référence peut mettre en œuvre un processus qui

allie l'analyse statistique de données portant sur les élèves dans le temps à des informations additionnelles comme la recherche sur la façon dont les enfants apprennent à lire, l'expérience acquise ailleurs, des connaissances tirées de sciences cognitives et la connaissance des contextes locaux.

Les critères de référence peuvent changer dans le temps pour s'aligner sur l'amélioration des résultats des élèves. Des normes et des critères de référence peuvent être mis au point de plusieurs façons mais doivent répondre aux conditions suivantes :

- Les critères de référence sont ambitieux mais réalistes et réalisables.
- Ils ne sont pas sujets à l'inflation des scores (l'augmentation des scores ne se généralise pas à d'autres mesures du même contenu parce qu'elle reflète principalement des activités étroites de préparation de test axées sur un test particulier) (Hamilton, Stechter & Yuan, 2008).
- Les critères de référence doivent pouvoir identifier les élèves susceptibles d'échouer dans l'atteinte d'un niveau de lecture indépendante. Les critères de référence sont particuliers à un moment donné (début d'année, fin d'année, classe, etc.) et les critères de référence subséquents sont dérivés en fonction de la probabilité que les enfants répondant au premier critère de référence vont également répondre au suivant (dans les conditions pédagogiques actuelles). (Dynamic Measurement Group, Inc., 2010).
- Les critères de référence sont basés sur une recherche qui examine la validité prédictive d'un score sur une mesure à un moment donné, comparé à des évaluations de résultats externes et des mesures ultérieures. Si un élève atteint un objectif de référence, il y a des chances que cet élève obtienne des résultats en lecture ultérieurs s'il/si elle reçoit un enseignement basé sur la recherche relevant d'un programme scolaire de base (Dynamic Measurement Group, Inc., 2010).
- Les meilleures données à employer sont les scores des tests de candidats réels dont la performance a été jugée de manière significative par des juges qualifiés (Zieky & Perie, 2006).
- Les critères de référence sont bien reliés à toutes les classes pour éviter des erreurs de classification des élèves ou la transmission de rapports trompeurs aux intervenants. Par exemple, s'il peut être approprié d'affecter une limite plus haute pour définir un élève de niveau avancé en 2e année plutôt que de définir un élève de niveau de base en 3e année, l'inverse n'est pas vrai (Zieky & Perie, 2006).

« Il n'existe pas de scores limites vrais ou corrects, mais seulement des scores plus ou moins justifiables. Cette justification est en grande partie fonction de la méthode employée pour établir les normes. Deuxièmement, il n'y a pas de méthode qui soit en elle-même meilleure ou correcte pour l'établissement de normes, mais plutôt diverses démarches pouvant être plus ou moins appropriées à une situation particulière ».

– Ferrara, Perie & Johnson, 2008

Tous les critères de référence sont en fin de compte basés sur des normes ou des jugements de ce qu'un enfant devrait pouvoir faire (Zieky & Perie, 2006). Un pays peut établir ses propres critères de référence en examinant les résultats dans les écoles dont on sait qu'elles sont performantes ou qui peuvent démontrer une bonne performance dans une évaluation de type EGRA mais qui n'ont pas d'avantage socioéconomique particulier ou un niveau non-durable d'utilisation de ressources.

Ces écoles vont généralement produire des critères de référence raisonnablement exigeants mais manifestement réalisables même par des enfants ne bénéficiant guère d'avantages socioéconomiques ou dans des écoles ne bénéficiant guère d'avantages en matière de ressources, tant qu'un enseignement de qualité est prodigué. L'Etude internationale 2001 du progrès en matière de lecture (PIRLS 2001), par exemple, a sélectionné quatre limites étiquetées « critères de référence internationaux » sur l'échelle combinée de maîtrise de la lecture. Ces critères de référence ont été sélectionnés pour correspondre aux scores obtenus par le quart inférieur, la médiane, le quart supérieur et les 10 % supérieurs des élèves de 4e année dans l'échantillon international de l'étude PIRLS 2001 (Institute of Education Sciences, s. d.).

12.2.3 Processus d'établissement de critères de référence

Comme il est mentionné plus haut dans une des études de cas, le projet EdData II de l'USAID avait, en novembre 2015, donné son appui à l'établissement de critères de référence dans une douzaine de pays. Un processus cohérent a permis de déterminer dans ces pays des niveaux acceptables de performance dans plusieurs domaines de développement des compétences en lecture et dans plusieurs classes. On trouvera ci-dessous des recommandations mises au point en fonction de certaines des leçons apprises suite au travail réalisé dans 12 pays.

Etape 1 : Commencer par déterminer le niveau de compréhension de lecture qui est acceptable (démonstration d'une compréhension totale d'un texte donné). La plupart des pays ont situé le niveau de compréhension acceptable à 80 % ou plus (au moins 4 réponses correctes sur 5 questions).

Etape 2 : En fonction d'un critère de référence en compréhension de lecture, les données EGRA sont employées pour montrer la plage des scores de facilité de lecture à haute voix (ORF)—mesurée en mots corrects par minute (mcpm)—obtenus par les élèves capables d'atteindre le niveau de compréhension désiré. Il conviendra ensuite de déterminer la valeur proposée comme critère de référence dans les limites de cette plage. Autrement, une plage peut indiquer les niveaux qui sont acceptables pour déterminer qu'un élève est « compétent » ou satisfait aux normes correspondant au niveau scolaire (par exemple, 40 à 50 mcpm).

Etape 3 : Une fois un critère de référence ORF défini, le rapport entre l'ORF et le décodage (lecture de non-mots) permet de déterminer la vitesse moyenne de lecture de non-mots correspondant au niveau d'ORF donné.

Etape 4 : Procéder ensuite de la même manière pour chaque domaine de compétence suivant.

12.3 Mises en garde et limites

Dans certains contextes, les réactions à une évaluation de type EGRA ne sont pas simples. Certains commentateurs, dans certains pays, s'interrogent sur l'utilité de la facilité de lecture à haute voix comme référence ou indicateur précurseur d'un apprentissage général ou même du niveau de lecture. Ils peuvent demander pourquoi l'évaluation comprend des items ou des formats qui ne reflètent pas directement l'enseignement scolaire (lecture de mots inventés, par exemple). C'est pourquoi il est important d'avoir accès à la documentation qui en explique les raisons ; cette documentation est mentionnée en partie dans ce manuel et également disponible sur le site Web de l'Association internationale pour l'alphabétisation (www.reading.org), sur les pages du National Reading Panel de l'Institut américain de la santé de l'enfant et du développement humain (www.nationalreadingpanel.org) et sur le site Web du Centre sur l'enseignement et l'apprentissage de l'université de l'Oregon pour les indicateurs dynamiques des compétences fondamentales précoces en alphabétisation (DIBELS, http://dibels.uoregon.edu/).

Dans d'autres cas, les audiences potentielles semblent percevoir que les initiatives EGRA s'efforcent à véhiculer la notion selon laquelle « la lecture est la seule chose qui compte ». Il est important de noter dans ces cas que la lecture est effectivement une importante compétence fondamentale qui influence la réussite académique tout au long du programme scolaire et que la lecture est également un bon indicateur de la qualité générale de la scolarisation. L'effort n'est cependant pas basé sur l'hypothèse que la lecture est la seule chose qui compte.

En général, toute tentative visant à mesurer la qualité de l'éducation, représentée par l'apprentissage, est sujette à ce type de débat. L'expérience accumulée avec l'application de l'instrument EGRA ou d'outils de type EGRA semble montrer que les enseignants, les personnes directement concernées par l'apport d'un soutien aux enseignants et les hauts fonctionnaires reconnaissent immédiatement la valeur d'une évaluation EGRA, alors que certains théoriciens dans le domaine pédagogique et de la lecture hésitent à trop simplifier peut-être la situation. Il est essentiel de comprendre que l'emploi pratique de l'évaluation EGRA et les stratégies d'amélioration qui en sont dérivées ne doivent être considérés que comme point de départ. Les données peuvent servir d'exemple de ce qui peut être réalisé en concentrant et en assurant le suivi de résultats particuliers. La leçon de base peut alors être appliquée à d'autres aspects de l'enseignement et de l'apprentissage.

La résistance à la méthodologie et aux résultats EGRA peut se faire sentir davantage là où ces résultats sont les plus faibles, et c'est pourquoi il est important de mettre en œuvre l'évaluation et d'analyser les résultats avec rigueur et objectivité. Des questionnaires contextuels additionnels (caractéristiques des élèves, des enseignants et des écoles, observations de classe, etc.) peuvent permettre d'expliquer les résultats mais leur administration et l'analyse de leurs résultats entraînent des coûts supplémentaires. Quand des instruments de sondage additionnels sont associés aux résultats de l'évaluation EGRA, les personnes qui en assurent la mise en œuvre doivent veiller soigneusement à la taille de l'échantillon

et à sa signification statistique et éviter d'associer une corrélation avec la causation jusqu'à la réalisation de recherches supplémentaires.

L'Initiative mondiale pour l'éducation avant tout lancée en septembre 2012 par le Secrétaire général des Nations Unies établit la nécessité d'une évaluation efficace des résultats scolaires pour améliorer les systèmes d'éducation. Une évaluation soigneuse et un suivi rapproché du degré d'efficacité opérationnelle d'un système peut influencer la politique en donnant aux représentants officiels et aux décideurs l'occasion « d'employer l'information pour orienter le soutien et les ressources vers des solutions efficaces » (Bureau du Secrétaire général des Nations Unies, 2012, p. 19).

BIBLIOGRAPHIE

Abadzi, H. (2006). *Efficient learning for the poor.* Washington, DC: The World Bank. https://openknowledge.worldbank.org/handle/10986/7023

Abadzi, H. (2012). *Developing cross-language metrics for reading fluency measurement: Some issues and options.* Global Partnership for Education working paper. Washington, DC: World Bank. Retrieved from https://www.academia.edu/3484052/Developing_Cross- Language_Metrics_for_Reading_Fluency_Measurement_Some_issues_and_opti ons._World_Bank_Global_Partnership_for_Education_working_paper

Abu-Rabia, S. (2000). Effects of exposure to literary Arabic on reading comprehension in a diglossic situation. *Reading and Writing: An Interdisciplinary Journal, 13,* 147–157.

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print.* Cambridge, MA: MIT Press.

Adolf, S. M., Catts, H. W., & Lee, J. (2010). Kindergarten predictors of second versus eighth grade reading comprehension impairments. *Journal of Learning Disabilities*, *43*(4), 332–345. http://dx.doi.org/10.1177/0022219410369067

Adolf, S.M., Perfetti, C. & Catts, H.W. (2011). Development changes in reading comprehension: Implications for assessments and instruction. In S.J. Samuels & A.E. Farstrup (Eds). *What research has to say about reading instruction?* Newark DE: International Reading Association, 186–214.

Arnaud, A. & Lancelot, C. (1660). *Grammaire générale et raisonnée – Grammaire de Port Royal*. Paris: Prault fils l'Ainé.

Arrington, C.N, Kulesz, P.A., Francis, D.J., Fletcher, J.M., Barnes, M.A. (2014). The contribution of attentional control and working memory to reading comprehension and decoding. *Scientific Studies of Reading*. 18(5), 325-346.

August, D., & Shanahan, T. (2006). *Developing literacy in second-language learners.* Prepared by the Center for Applied Linguistics and SRI International for the Institute of Education Sciences and the Office of English Language Acquisition, US Department of Education; and the US National Institute of Child Health and Human Development. Washington, DC: Lawrence Erlbaum Associates and the Center for Applied Linguistics.

Ayari, S. (1996). Diglossia and illiteracy in the Arab world. *Language, Culture and Curriculum, 9,* 243–253.

Backman, J., Bruck, M., Hebert, M. & Seidenberg, M. (1984). Acquisition and use of spelling-sound correspondence in reading. *Journal of Experimental Child Psychology*, 38, 114-133.

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual*

Review of Psychology, 63, 1-30

Bara, F., Gentaz, E., Colé, P. & Sprenger-Charolles, L.(2004). The visuo-haptic and haptic exploration of letters increases the kindergarten-children's understanding of the alphabetic principle. *Cognitive Development*, 19(3), 433-449.

Batchelder, K., Betts, K., Mulcahy-Dunn, A. & Stern, J. (2015). Lot quality assurance sampling (LQAS) pilot in Tanzania: Final report. Prepared for USAID under the EdData II project, Task Order No. AID-OAA-12-BC-00003 (RTI Task 20, Activity 5). Research Triangle Park, NC: RTI International.

Billard, C., Bricout, L., Ducot, B., Richard, G., Ziegler, J. & Fluss, J. (2010). Evolution des compétences en lecture, compréhension et orthographe en environnement socio-économique défavorisé et impact des facteurs cognitifs et comportementaux sur le devenir à deux ans. *Revue d'Epidémiologie et de Santé Publique*, 58, 101-110

Bradley, L. & Bryant, P. (1983). Categorizing sounds in learning to read: A causal connection. *Nature, 301*, 419-421

Braun, H., & Kanjee, A. (2006). Using assessment to improve education in developing nations. In H. Braun, A. Kanjee, E. Bettinger, & M. Kremer (Eds.), *Improving education through assessment, innovation, and evaluation* (pp. 1–46). Cambridge, MA: American Academy of Arts and Sciences. Retrieved from https://www.amacad.org/publications/braun.pdf

Bulat, J., Brombacher, A., Slade, T., Iriondo-Perez, J., Kelly, M., & Edwards, S. (2014). *Projet d'Amélioration de la Qualité de l'Education (PAQUED): 2014. Endline report of Early Grade Reading Assessment (EGRA) and Early Grade Mathematics Assessment (EGMA).* Prepared for USAID under Contract No. AID-623-A-09-00010. Washington, DC: Education Development Center and RTI International.

Casalis, S. & Louis-Alexandre, M. F. (2000). Morphological analysis, phonological analysis and learning to read French: A longitudinal study. *Reading and Writing: An Interdisciplinary Journal, 12*, 303–335.

Center for Global Development. (2006). *When will we ever learn? Improving lives through impact evaluation.* www.cgdev.org/files/7973_file_WillWeEverLearn.pdf

Chabbott, C. (2006). *Accelerating early grades reading in high priority EFA Countries: A desk review.* http://www.equip123.net/docs/E1- EGRinEFACountries-DeskStudy.pdf

Clay, M. M. (1993). *An observation survey of early literacy achievement.* Ortonville, MI.: Cornucopia Books.

Colé, P., Bouton, S., Leuwers, C., Casalis, S., Sprenger Charolles, L. (2012). Stem and derivational-suffix processing during reading by french second and third graders. *Applied Psycholinguistics, 33,* 97-120.

Collins, P., & Messaoud-Galusi, S. (2012). *Student performance on the Early Grade Reading Assessment (EGRA) in Yemen* [English version; also available in Arabic]. Report prepared for USAID under the EdData II project, Task Order

EHC-E-07-04-00004-00 (RTI Task 7). Research Triangle Park, NC: RTI International. http://pdf.usaid.gov/pdf_docs/PNADZ047.pdf

Coltheart M., Rastle K., Perry C., Langdon R., & Ziegler J. C. (2001). DRC: a dual-route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108,* 204–256.

Crouch, L., & Korda, M. (2008). *EGRA Liberia: Baseline assessment of reading levels and associated factors.* Report prepared for the World Bank under Contract No. 7147768. Research Triangle Park, NC: RTI International.

Crouch, L., & Winkler, D. (2008). Governance, management, and financing of Education for All: Basic frameworks and case studies. Background paper commissioned for the *Education for All global monitoring report 2009: Governance, management and financing of education for all.* Research Triangle Park, NC: RTI International. unesdoc.unesco.org/images/0017/001787/178719e.pdf

Cunningham, P.M., & Allington, R. L. (2015). *Classrooms that work: They can all read and write* (6th ed.). Boston, MA: Pearson.

Daniel, S. S., Walsh, A. K., Goldston, D. B., Arnold, E. M., Reboussin, B. A., & Wood, F. B. (2006). Suicidality, school dropout, and reading problems among adolescents. *Journal of Learning Disabilities*, *39*(6), 507–514. http://dx.doi.org/10.1177/00222194060390060301

Darney, D., Reinke, W. M., Herman, K. C., Stormont, M., & Ialongo, N. S. (2013). Children with co-occurring academic and behavior problems in first grade: Distal outcomes in twelfth grade. *Journal of School Psychology*, *51*(1), 117–128. http://dx.doi.org/10.1016/j.jsp.2012.09.005

Dehaene, S. (Ed), Dehaene-Lambertz G., Gentaz E., Huron C. & Sprenger-Charolles L. (2011). *Apprendre à lire. Des sciences cognitives à la salle de classe*. Paris: Odile Jacob; 2011.

Delattre, P. (1965). *Comparing the phonetic features of English, French, German and Spanish*. Heidelberg: Jumius Gross Verlag.

Desrochers, A. & Saint-Aubin, J. (2008). Sources de materiel en français pour l'élaboration d'épreuves de compétences en lecture et en écriture. *Revue Canadienne d'Education*, *31*(2), 305–326.

Doctor, E. & Coltheart, M. (1980). Phonological recoding in children's reading for meaning. *Memory and Cognition*, 80, 195-209. ;

Dubeck, M. M., & Gove, A. (2015). The early grade reading assessment (EGRA): Its theoretical foundation, purpose, and limitations. *International Journal of Educational Development, 40,* 315–322. http://dx.doi.org/10.1016/j.ijedudev.2014.11.004

Dunn, L. M., Thériault-Whalen, C. M., & Dunn, L. M. (1993). *Echelle de vocabulaire en images Peabody: Adaptation française du Peabody Picture Vocabulary Test.* Toronto, Canada: Psycan.

du Plessis, J., El-Ashry, F., & Tietjen, K. (Forthcoming). Oral reading assessments in Yemen: Turning bad news into a national reform. In *Understanding what works*

in oral reading assessments. Montreal: UNESCO Institute for Statistics (UIS)

Dynamic Measurement Group, Inc. (2010). *DIBELS® Next benchmark goals and composite score.* https://dibels.org/papers/DIBELSNextBenchmarkGoals.pdf

Ecalle, J. & Magnan, A. (2015). *L'apprentissage de la lecture et ses difficultés*. Paris, Dunod.

Ehri, L. C. (1998). Grapheme-phoneme knowledge is essential for learning to read words in English. In J. L. Metsala & L. C. Ehri (Eds.), *Word recognition in beginning literacy* (pp. 3–40). Mahwah, NJ: Lawrence Erlbaum Associates.

Ehri, L.C., Nunes, S.R., Stahl, S.A., & Willows, D.M. (2001). Systematic phonics instruction helps students learn to read: Evidence from the national reading panel's meta-analysis. *Review of Educational, 71(3),* 393-447.

Ehri, L.C., Nunes, S.R., Stahl, S.A., Willows, D.M. (2001a). Systematic Phonics Instruction Helps Students Learn to Read: Evidence from the National Reading Panel's Meta-Analysis, *Review of Educational Research*, 71: 393–447.

Ehri, L. C., Nunes, S. R., Willows, D. M., Schuster, B. V., Yaghoub-Zadeh, Z. & Shanahan, T. (2001b). Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel's meta-analysis. *Reading Research Quarterly, 36*(3), 250–287.

Expertise collective INSERM (2007). *Dyslexie, dysorthographie; dyscalculia: Bilan des données scientifiques.* Paris: INSERM https://www.google.fr/?client=firefox-b#q=dyslexie+dysorthographie+dyscalculie+bilan+des+donn%C3%A9es+scientifiques

Feldman, L. B., Rueckl, J., DiLiberto, K., Pastizzo, M. & Vellutino, F. (2002). Morphological analysis by child readers as revealed by the fragment completion task. *Psychonomic Bulletin, 9*(3), 529–535.

Ferguson, C. A. (1959). Diglossia. *Word, 15,* 325–340.

Ferrara, S, Perie, M., & Johnson, E. (2008). Matching the judgmental task with standard setting panelist expertise: The item-descriptor (ID) matching method. *Journal of Applied Testing Technology, 9*(1), 1–22.

Filmer, D., Hasan, A., & Pritchett, L. (2006). *A millennium learning goal: Measuring real progress in education.* Washington, DC: World Bank. Retrieved from http://dx.doi.org/10.2139/ssrn.982968

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.) New York: John Wiley.

Foulin, J. N. (2005). Why is letter-name knowledge such a good predictor of learning to read? *Reading and Writing: An Interdisciplinary Journal*, *18*, 129–155.

Fuchs, L., Fuchs, D., Hosp, M. K., & Jenkins, J. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*(3), 239–256.

Gambrell, L. B., & Morrow, L. M. (Eds). (2014). *Best practices in literacy instruction*

(5th ed.). New York, NY: Guildford.

Gaonac'h, D. & Fayol, M. [Ed] (2003). *Aider les élèves à comprendre: du texte au multimédia*. Paris, Hachette Education.

Gentaz, E., Sprenger-Charolles, L. & Theurel, A. (2015). Differences in the predictors of reading comprehension in first graders from low socio-economic status families with either good or poor decoding skills. *PLoS ONE*, March 2015 DOI: 10.1371/journal.pone.0119581

Glick, P., & Sahn, D. E. (2010). Early academic performance, grade repetition, and school attainment in Senegal: A panel data analysis. *The World Bank Economic Review, 24*(1), 93–120.

Good, R. H., Simmons, D. C., & Smith, S. (1998). Effective academic intervention in the United States: Evaluating and enhancing the acquisition of early reading skills. *School Psychology Review, 27,* 45–56.

Goswami, U. (2008). The development of reading across languages. *Annals of the New York Academy of Sciences, 1145,* 1–12.

Goswami, U., Gombert, J. E. & de Barreira, L. F. (1998). Children's orthographic representations and linguistic transparency: Nonsense word reading in English, French and Spanish. *Applied Psycholinguistics*, *19*, 19-52.

Gove, A., & Cvelich, P. (2011). *Early reading: Igniting education for all. A report by the Early Grade Learning Community of Practice* (rev. ed). Research Triangle Park, NC: RTI International. http://www.rti.org/publications/abstract.cfm?pubid=17099

Gove, A., & Wetterberg, A. (2011). The Early Grade Reading Assessment: An introduction. In A. Gove & A. Wetterberg (Eds.), *The Early Grade Reading Assessment: Applications and interventions to improve basic literacy* (pp. 1–37). Research Triangle Park, NC: RTI Press. http://www.rti.org/pubs/bk-0007-1109-wetterberg.pdf

Gove, A., & Wetterberg, A. (Eds.). (2011). *The Early Grade Reading Assessment: Applications and interventions to improve basic literacy.* Research Triangle Park, NC: RTI Press. http://www.rti.org/pubs/bk-0007-1109-wetterberg.pdf

Grainger, J., Bouttevin, S., Truc, C., Bastien, M. & Ziegler, J. (2003). Word superiority, pseudoword superiority, and learning to read: A comparison of dyslexic and normal readers. *Brain and Language*, *87(3)*, 432-440.

Hamilton, L. S., Stetcher, B. M., & Yuan, K. (2008). *Standards-based reform in the United States: history, research, and future directions.* Prepared under National Science Foundation Grant No. REC-0228295. Santa Monica, CA: RAND Corporation. http://www.rand.org/content/dam/rand/pubs/reprints/2009/RAND_RP1384.pdf

Hanson, B. A., Zeng, L., & Colton, D. (1994). *A comparison of presmoothing and postsmoothing methods in equipercentile equating* (ACT Research Report 94-4). Iowa City, IA: ACT.

Hanushek, E. A., & Woessman, L. (2009). *Do better schools lead to more growth?*

Cognitive skills, economic outcomes, and causation. Working Paper 14633. Cambridge, MA: National Bureau of Economic Research.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 187–220).Westport, CT: Praeger.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal, 2,* 127–160.

Hudson, R. F., Lane, H. B., & Pullen, P. C. (2005). Reading fluency assessment and instruction: What, why, and how? *The Reading Teacher, 58*(8), 702–714

Institute of Education Sciences, National Center for Education Statistics [US]. (n.d.). *International comparisons in fourth-grade reading literacy: Reading literacy by benchmarks* (Web page). http://nces.ed.gov/pubs2004/pirlspub/5.asp

Jakobsen, R. (1960). Closing statements: Linguistics and poetics. In T. A. Sebeok (Ed.), *Style in language* (pp. 350–377). Cambridge, MA: MIT Press.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology 80*(4), 437–447.

Kandel, S., Soler, O., Valdois, S. & Gros, L. (2006). Graphemes as motor units in the acquisition of writing skills *Reading and Writing: An Interdisciplinary Journal, 19*, 313–337.

Kanjee, A. (2009). *Assessment overview* [Presentation]. Prepared for the first READ Global Conference, "Developing a Vision for Assessment Systems," Moscow, October 1, 2009. http://www.worldbank.org/content/dam/Worldbank/document/Program/READ/Eve nts/READ-conference-2009/READ_GC_Presentation_5_AKanjee_Eng.pdf

Kleinman, L., Leidy, N. K., Crawley, J., Bonomi, A., & Schoenfeld, P. (2001). A comparative trial of paper-and-pencil versus computer administration of the quality of life in reflux and dyspepsia (QOLRAD) questionnaire. *Medical Care 39,* 181–189.

Kochetkova, E., & Dubeck, M. (In press). Assessment in schools. Chapter in *Understanding what works in oral reading assessments*. Montreal: UNESCO Institute for Statistics (UIS).

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking.* New York, NY: Springer-Verlag.

Kolinsky R., Morais J., Cohen L., Dehaene-Lambertz G. & Dehaene S. (2014). L'influence de l'apprentissage du langage écrit sur les aires du langage. *Revue de Neurospychologie*, 6(3), 173-181.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6,* 293–323.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

LaTowsky, R. (2014). *Towards possible early grade reading benchmarks for the West*

Bank (Presentation slides). Prepared for USAID under the Education Data for Decision Making (EdData II) project, Measurement and Research Support to Education Strategy Goal 1, Task Order No. AID-OAA-12-BC-00003 (RTI Task 20). Research Triangle Park, NC: RTI International. https://www.eddataglobal.org/countries/index.cfm?fuseaction=pubDetail&ID=778

Lété, B., Sprenger-Charolles, L. & Colé, P. (2004). MANULEX: A grade-level lexical database from French elementary school readers. *Behavior Research Methods, Instruments, and Computers, 36*(1), 156–166.

Linan-Thompson, S., & Vaughn, S. (2004). *Research-based methods of reading instruction: Grades K-3.* Alexandria, VA: Association for Supervision and Curriculum Development.

Linan-Thompson, S., & Vaughn, S. (2007). *Research-based methods of reading instruction for English-language learners: Grades K–4*. Alexandria, VA: Association for Supervision and Curriculum Development.

Lonigan, C., Wagner, R., Torgesen, J. K., & Rashotte, C. (2002). *Preschool Comprehensive Test of Phonological and Print Processing (Pre-CTOPPP).* Tallahassee: Department of Psychology, Florida State University.

Mann, V. & Singson, M. (2003). Linking morphological knowledge to English decoding ability: large effects of little suffixes. In E. M. H. Assink & D. Sandra (Eds.), *Reading complex words: Cross-language studies* (pp. 1-24). New York: Kluwer Academic/Plenum publishers.

Melby-Lervag, M., Lyster, S.A. & Hulme, C. (2012). Phonological skills and their role in learning to read: a meta-analytic review. *Psychological Bulletin*, 138(2), 322–352.

Metsala, J.L., Stanovich, K.E. & Brown, G.D.A. (1998). Regularity effects and the phonological deficit model of reading disabilities: A meta-analytic review. *Journal of Educational Psychology, 90(2),* 279-293.

Morais, J., Cary, L., Alegria, J. & Bertelson, P. (1979). Does awareness of speech as a sequence of phones arise spontaneously? *Cognition, 7*, 323–333.

Mulcahy-Dunn, A., Valadez, J. J., Cummiskey, C., & Hartwell, A. (2013). *Report on the pilot application of lot quality assurance sampling (LQAS) in Ghana to assess literacy and teaching in primary grade 3.* Prepared for USAID under the EdData II project, Task Order No. EHC-E-07-04-00004-00 (RTI Task 7). Research Triangle Park, NC: RTI International.

Muter, V., Hulme, C., Snowling, M. J., & Stevenson, J. (2004). Phonemes, rimes, vocabulary, and grammatical skills as foundation of early reading development: Evidence from a longitudinal study. *Developmental Psychology, 40,* 665–681.

Nag, S. (2007). Early reading in Kannada: The pace of acquisition of orthographic knowledge and phonemic awareness. *Journal of Research in Reading, 30*(1), 7– 22.

Nag, S. (2014). Akshara-phonology mappings: the common yet uncommon case of the consonant cluster. *Writing Systems Research, 6,* 105–119.

Nag, S., & Snowling, M. J. (2010). Cognitive profiles of poor readers of Kannada. Reading and Writing, 24(6), 657–676

Nagy, W. E., & Scott, J. (2000). Vocabulary processes. In M. E. A. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr, (Eds.), *Handbook of reading research* (Vol. III, pp. 269-284). Mahwah, NJ: Erlbaum.

National Center for Family Literacy (NCFL) [US]. (2008). *Developing early literacy: Report of the national early literacy panel. A scientific synthesis of early literacy development and implications for intervention.* Prepared under interagency agreement IAD-01-1701 and IAD-02-1790 between the Department of Health and Human Services and the National Institute for Literacy. Washington, DC: National Institute for Literacy. https://www.nichd.nih.gov/publications/Pages/pubs_details.aspx?pubs_id=5750

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, US Department of Health, Education and Welfare (DHEW). (1978). *Belmont Report: Ethical principles and guidelines for the protection of human subjects of research.* Report of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. DHEW Pub. No. (OS) 78-0012. Washington, DC: United States Government Printing Office. http://videocast.nih.gov/pdf/ohrp_belmont_report.pdf

National Institute of Child Health and Human Development (NICHD) [US]. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups* (NIH Publication No. 00-4754). Washington, DC: NICHD. https://www.nichd.nih.gov/publications/pubs/nrp/Pages/smallbook.aspx

Nielsen, D. (2014). *Early grade reading and math assessments in 10 countries: Dissemination and utilization of results—a review.* Report prepared for USAID under the Education Data for Decision Making (EdData II) project, Measurement and Research Support to Education Strategy Goal 1, Task Order No. AID-OAA- BC-12-00003 (RTI Task 20). Research Triangle Park, NC: RTI International. http://pdf.usaid.gov/pdf_docs/PA00K8RP.pdf

Office of the United Nations Secretary-General. (2012). *Global Education First Initiative: An initiative of the United Nations Secretary-General.* New York: United Nations. http://www.globaleducationfirst.org/files/GEFI_Brochure_ENG.pdf

Optimal Solutions Group, LLC. (2015). *Secondary Analysis for Results Tracking (SART) data sharing manual, USAID Ed Strategy 2011–2015, Goal 1.* Prepared for USAID under the Secondary Analysis for Results Tracking (SART) project, Contract AID-OAA-C-12-00069. Location: Optimal Solutions. Retrieved from https://sartdatacollection.org/images/SARTDataSharingManual-Feb2015.pdf

Orr, D. B., & Graham, W. R. (1968). Development of a listening comprehension test to identify educational potential among disadvantaged junior high school students. *American Educational Research Journal, 5*(2), 167–180.

Ouellette, G & Beers, A. (2010) A not-so-simple view of reading: How vocabulary and visual-word recognition complicates the story. *Reading and Writing, 23*, 189-208.

Pacton, S. (2008). L'apprentissage de l'orthographe du français. in A. Desrochers, F. Martineau et Y. C. Morin (Eds), *Normes et pratiques orthographiques* (pp. 331-354). Ottawa, Editions David

Pacton, S., Fayol, M. & Perruchet, P. (2005). Children's implicit learning of graphotactic and morphological regularities. *Child Development*, 76(2), 324-339.

Paris, S. G., & Paris, A. H. (2006). Chapter 2: Assessments of early reading. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology: Theoretical models of human development, 6th Edition* (Vol. 4: Child Psychology in Practice). Hoboken, New Jersey: John Wiley and Sons.

Parviainen T., Helenius P., Poskiparta E., Niemi P. & Salmelin R. (2006). Cortical sequence of word perception in beginning readers. *Journal of Neuroscience, 26*(22), 6052-6061.

Patrinos, H. A, & Velez, E. (2009). Costs and benefits of bilingual education in Guatemala: A partial analysis. *International Journal of Educational Development, 29*(6), 594–598.

Peereman R., Sprenger-Charolles L. & Messaoud-Galusi S. (2013). The contribution of morphology to the consistency of spelling-to-sound relations: A quantitative analysis based on French elementary school readers. *L'Année Psychologique*, 113(1), 3-33.

Perfetti, C. A. (2003). The universal grammar of reading. *Scientific Studies of Reading, 7*(1), 3–24.

Perfetti, C. A., Beck, I., Bell, L. & Hughes, C. (1987). Phonemic knowledge and learning to read are reciprocal: A longitudinal study of first-grade children. *Merrill-Palmer Quarterly, 33,* 283–319.

Perfetti, C. A., & Dunlap, S. (2008). Learning to read: General principles and writing system variations. In K. Koda & A. Zehler (Eds.), *Learning to read across languages* (pp. 13–38). Mahwah, NJ: Erlbaum.

Perfetti, C. A. & Zhang, S. (1995). The universal word identification reflex. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 33, pp. 159–189). San Diego, California: Academic Press.

Piper, B., & Mugenda, A. (2014). *USAID/Kenya Primary Math and Reading (PRIMR) Initiative: Endline impact evaluation.* Prepared under the USAID EdData II project, Task Order No. AID-623-M-11-00001 (RTI Task 13). Research Triangle Park, NC: RTI International. http://pdf.usaid.gov/pdf_docs/pa00k27s.pdf

Piquard-Kipffer A. & Sprenger-Charolles L. (2013). Early predictors of future reading skills: A follow-up of French-speaking children from the beginning of kindergarten to the end of the second grade (age 5 to 8). *L'Année Psychologique*, 113(4), 491-521.

Pothier, B. & Pothier, P. (2002). EOLE: Echelle d'acquisition en orthographe lexicale. Paris: Retz.

Prodigy Systems. (2011). *EGRA Yemen with iProSurveyor* [Presentation slides].

Sana'a: Prodigy Systems.

Quémard, Casalis, S. & Colé, P. (2011). The role of form and meaning in the processing of written morphology: A priming study in French developing readers. *Journal of Experimental Child Psychology, 109*, 478-496.

Rack, J., Snowling, M. J. & Olson, R. K. (1992). The nonword reading deficit in developmental dyslexia: A review. *Reading Research Quarterly, 27*, 29–53.

Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M.S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest, 2,* 31–74.

Rey, A., Ziegler, J. C. & Jacobs, A. (2000). Graphemes are perceptual reading units. *Cognition,*75, B1–12.

RTI International. (2008). *Early grade reading Kenya: Baseline assessment. Analyses and implications for teaching interventions design. Final report.* Prepared for USAID under the EdData II project, Task Order No. EHC-E-01-04- 00004-00 (RTI Task 4). Research Triangle Park, NC: RTI International. http://pdf.usaid.gov/pdf_docs/PNADL212.pdf

RTI International. (2011). *EGRA Plus: Liberia. Final report: October 2008–January 2011.* Prepared for USAID/Liberia under the EdData II Project, Task Order No. EHC--E-06-04-00004-00 (RTI Task 6). Research Triangle Park, NC: RTI International. http://pdf.usaid.gov/pdf_docs/PNADZ817.pdf

RTI International. (2014a). *Codebook for EGRA and EGMA* [Excel spreadsheet]. Research Triangle Park, NC: RTI. Retrieved from https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&id=389

RTI International. (2014b). *USAID/Kenya Primary Math and Reading (PRIMR) Initiative: Final report.* Prepared for USAID under the EdData II project, Task Order No. AID-623-M-11-00001. Research Triangle Park, NC: RTI. http://pdf.usaid.gov/pdf_docs/PA00K282.pdf

RTI International. (2015). *EGRA tracker.* Prepared for USAID under the EdData II project, Contract No. EHC-E-00-04-00004-00. Research Triangle Park, NC: RTI. https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&id=188

RTI International & International Rescue Committee (IRC). (2011). *Guidance notes for planning and implementing EGRA.* Research Triangle Park, NC: RTI and IRC. https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&id=318

Saiegh-Haddad, E. (2003). Linguistic distance and initial reading acquisition: the case of Arabic diglossia. *Applied Psycholinguistics, 24,* 115–135.

Scanlon, D. M., Gelzheiser, L. M., Vellutino, F. R., Schatschneider, C., & Sweeney, J. M. (2008). Reducing the incidence of early reading difficulties: Professional development for classroom teachers versus direct interventions for children. *Learning and Individual Differences*, *18*(3), 346–359. http://dx.doi.org/10.1016/j.lindif.2008.05.002

Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology, 94,* 143–174.

Share, D. L. (2008). On the Anglocentricities of current reading research and practice: The perils of overreliance on an "outlier" orthography. *Psychological Bulletin, 134*(4), 584–615.

Share, D. L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition, 55,* 151–218.

Share, D. L., Jorm, A., Maclearn, R., & Matthews, R. (1984). Sources of individual differences in reading acquisition. *Journal of Education Psychology, 76,* 1309–1324.

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement, 42,* 309–330.

Snow, C. E., Burns, M. S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children.* Prepared on behalf of the Committee on the Prevention of Reading Difficulties in Young Children under Grant No. H023S50001 of the National Academy of Sciences and the U.S. Department of Education. Washington, DC: National Academy Press.

Snow, C., & the RAND Reading Study Group. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Research prepared for the Office of Educational Research and Improvement (OERI), U.S. Department of Education. Santa Monica, CA: RAND Corporation.

Spencer, L. H., & Hanley, J. R. (2003). Effects of orthographic transparency on reading and phoneme awareness in children learning to read in Wales. *British Journal of Psychology, 94*(1), 1–28.

Spencer, M., Quinn, J., Wagner, R. (2014). Specific reading disability: Major problem myth or misnomer? *Learning Disabilities Research & Practice*, 2014; 29: 3-9.

Sprenger-Charolles, L. (2008). *Résultats d'enfants sénégalais des trois premiers grades du primaires ayant appris à lire en français et en wolof: EGRA (Early Grade Reading Assessment).* Report for the World Bank (version en anglais : http://pdf.usaid.gov/pdf_docs/Pnadl691.pdf ; version en français : http://pdf.usaid.gov/pdf_docs/Pnadq182.pdf).

Sprenger-Charolles, L. (2008). *Early grade reading assessment (EGRA).* Résultats d'élèves sénégalais des trois premiers grades ayant appris à lire en français *et en wolof*—Rapport pour la Banque Mondiale. Récupérée à partir du http://pdf.usaid.gov/pdf_docs/PNADL692.pdf

Sprenger-Charolles L. & Colé P. (2013). *Lecture et dyslexie: Approches cognitives*, Paris: Dunod (2nd édition, entièrement revue et actualisée, 1ère édition en 2003).

Sprenger-Charolles, L., Colé, P., Béchennec, D. & Kipffer-Piquard, A. (2005). French normative data on reading and related skills: From EVALEC, a new computarized battery of tests. *European Review of Applied Psychology, 55,* 157–186.

Sprenger-Charolles, L., Siegel, L., Béchennec, D. (1998a). Phonological mediation and semantic and orthographic factors in silent reading in French. *Scientific Studies of Reading, 2*, 3-29.

Sprenger-Charolles, L., Siegel, L., Béchennec, D. & Serniclaes, W. (2003). Development of phonological and orthographic processing in reading aloud, in silent reading and in spelling: A four year longitudinal study. *Journal of Experimental Child Psychology, 84*, 194–217.

Sprenger-Charolles, L., Siegel, L. S. & Bonnet, P. (1998b). Phonological mediation and orthographic factors in reading and spelling. *Journal of Experimental Child Psychology, 68,* 134–155.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21,* 360–406.

Stanovich, K. E. (2000). *Progress in understanding reading: Scientific foundations and new frontiers.* New York: Guilford Press.

Stanovich, K.E. (1980). Toward an interactive compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly, 16*, 32-71.

Stern, J. & Nordstrum, L. (2014). *Indonesia 2014: The National Early Grade Reading Assessment (EGRA) and Snapshot of School Management Effectiveness (SSME) survey.* Prepared for USAID/Indonesia under the Education Data for Decision Making (EdData II) project, Task Order No. AID-497-BC-13-00009 (RTI Task 23). Research Triangle Park, NC: RTI International. https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&id=680

Strigel, C. (2012). *Tangerine™—Electronic data collection tool for early reading and math assessments. January 2012 – Kenya field trial report: Summary.* Research Triangle Park, NC: RTI International. www.rti.org/files/tangerine_report_0112.pdf

Torgesen, J. K. (1998). Catch them before they fall: Identification and assessment to prevent reading failure in young children. *America Educator/American Federation of Teachers, 22*, 32–39.

Torgesen, J. K. (2002). The prevention of reading difficulties. *Journal of School Psychology*, *40*(1), 7–26. http://dx.doi.org/10.1016/s0022-4405(01)00092-9

Torgesen, J. K. & Davis, C. (1996). Individual difference variables that predict response to training in phonological awareness. *Journal of Experimental Child Psychology, 63*, 1–21.

Tunmer, W.E., & Chapman, J.W. (2012). The simple view of reading redux: Vocabulary knowledge and the independent components hypothesis. *Journal of Learning Disabilities,* 45, 453–466.

United Nations. (2015). *The Millennium Development Goals report 2015.* New York: United Nations. http://www.un.org/millenniumgoals/2015_MDG_Report/pdf/MDG%202015%20re v%20(July%201).pdf

United Nations Development Programme (UNDP). (2015). *Sustainable Development Goals (SDGs)* [Web page]. Retrieved from http://www.undp.org/content/undp/en/home/mdgoverview/post-2015- development-agenda.html

United Nations Educational, Scientific and Cultural Organization (UNESCO). (2014). *Education for All Global Monitoring Report 2013/4. Teaching and learning: Achieving quality for all.* Paris: UNESCO. http://en.unesco.org/gem- report/report/2014/teaching-and-learning-achieving-quality- all#sthash.n1q0vitl.dpbs

United States Agency for International Development (USAID). (2012). *How-to note: Preparing evaluation reports.* Monitoring and Evaluation Series, No. 1, Version 1.0. Washington, DC: USAID. Retrieved from https://www.usaid.gov/sites/default/files/documents/1870/How-to- Note_Preparing-Evaluation-Reports.pdf

Valadez, J. J., Mulcahy-Dunn, A., & Sam-Bossman, E. (2014). *Using lot quality assurance sampling to monitor impact of early grade reading programs* [87-slide training presentation plus handouts]. Prepared under the EdData II project, Task Order No. AID-OAA-12-BC-00003 (RTI Task 20), for a USAID-hosted webinar based in Washington, DC, July 9–10, 2014. Research Triangle Park, NC: RTI International. https://www.eddataglobal.org/reading/index.cfm?fuseaction=pubDetail&ID=602

Vaughn, S., & Linan-Thompson, S. (2004). *Research-based methods of reading instruction grades K-3.* Alexandria, VA: Association for Supervision and Curriculum Development.

Wagner, D.A. (2011). *Smaller, quicker, cheaper: Improving learning assessments for developing countries.* Paris: UNESCO International Institute of Educational Planning (IIEP) and Fast Track Initiative/World Bank. http://unesdoc.unesco.org/images/0021/002136/213663e.pdf

Wagner R. K., Torgesen J. K., & Rashotte C. A. (1994). Development of reading- related phonological processing abilities: New evidence of bi-directional causality from a latent variable longitudinal study. *Developmental Psychology, 30,* 73–87.

Walther, B., Hossin, S., Townend, J., Abernethy, N., Parker, D., & Jeffries, D. (2011). Comparison of electronic data capture (EDC) with the standard data capture method for clinical trial data. *PLoS One, 6*(9), e25348. http://dx.doi.org/10.1371/journal.pone.0025348

What Works Clearinghouse. (2015). *Procedures and standards handbook, version 3.0.* Washington, DC: Institute of Education Sciences, US Department of Education. http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_stan dards_handbook.pdf

World Bank. (2015a). *EdStats dashboards: Learning outcomes dashboard* [Web page]. Washington, DC: World Bank. http://datatopics.worldbank.org/education/wDashboard/tbl_index.aspx

World Bank. (2015b). *Learning outcomes* [Web page]. Washington, DC: World Bank. http://go.worldbank.org/GOBJ17VV90

World Bank: Independent Evaluation Group. (2006). *From schooling access to learning outcomes—An unfinished agenda: An evaluation of World Bank support*

to primary education. Washington, DC: World Bank. https://openknowledge.worldbank.org/handle/10986/7083

Yesil-Dağli, Ü. (2011). Predicting ELL students' beginning first grade English oral reading fluency from initial kindergarten vocabulary, letter naming, and phono- logical awareness skills. *Early Childhood Research Quarterly, 26*(1), 15–29.

Yovanoff, P., Duesbery, L., Alonzo, J., & Tindall, G. (2005). Grade-level invariance of a theoretical causal structure predicting reading comprehension with vocabulary and oral reading fluency. *Educational Measurement, Fall*, 4–12.

Ziegler, J. C., Bertrand, D., Toth, D., Csepe, V., Reis, A., Fasca, L., Saine, N., Lyytinen, H., Vaessen, A., & Blomert, L. (2010). Orthographic depth and its impact on universal predictors of reading: A cross-language investigation. Psychological Science, 21, 551–559.

Ziegler, J. C., Bertrand, D., Lété, B. & Grainger, J. (2014a). Orthographic and phonological contributions to reading
development: Tracking developmental trajectories using masked priming. *Developmental Psychology*, 50(4), 1026-1036

Ziegler, J.C. & Goswami, U. (2005). Reading acquisition, developmental dyslexia and skilled reading across languages: A psycholinguistic grain size theory. Psychological Bulletin, 13(1), 3–29.

Ziegler, J. C. & Goswami, U. (2006). Becoming literate in different languages: Similar problems, different solutions. Developmental Science, 9, 429–436.

Ziegler, J. C., Pech-Georgel, C., George, F. & Lorenzi, C. (2009). Speech-perception-in-noise deficits in dyslexia. *Developmental Science,*12(5), 732-745.

Ziegler, J. C., Perry, C. & Coltheart, M. (2003). Speed of lexical and nonlexical processing in French: The case of the regularity effect. *Psychonomic Bulletin & Review*, 10(4), 947-953.

Ziegler, J. C., Perry, C. & Zorzi, M. (2014b). Modelling reading development through phonological decoding and self-teaching: implications for dyslexia. *Philosophical Transactions of the Royal Society B-Biological Sciences, 369*(1634), 2012039, doi: 10.1098/rstb.2012.0397

Zieky, M., & Perie, M. (2006). *A primer on setting cut scores on tests of educational achievement.* Princeton, NJ: Educational Testing Service. https://www.ets.org/Media/Research/pdf/Cut_Scores_Primer.pdf

Zimmerman, R. (2008). *Digital data collection demonstration white paper. A comparison of two methodologies: Digital and paper-based.* Prepared for USAID under the Educational Quality Improvement Program 1 (EQUIP1), Cooperative Agreement No. GDG-A-00-03-00006-00. Washington, DC: American Institutes forResearch.http://www.equip123.net/docs/e1-DigitalDataCollection.pdf

ANNEXE A : INFORMATION SUR LES ATELIERS EGRA 2015

Source : site Web du projet EdData II, News and Events, https://www.eddataglobal.org/news/index.cfm

A1 Atelier sur la conception et la mise en œuvre de l'évaluation EGRA : comprendre les principes de base

24 mars 2015 – Kellie Betts

Le personnel technique de RTI a facilité du 2 au 4 mars 2015 l'atelier intitule « Conception et mise en œuvre d'évaluations des compétences fondamentales en lecture : comprendre les principes de base ». Cet atelier a été organisé par le Réseau mondial de lecture à University Research Co. (URC), LLC, à Bethesda, Maryland. Kate Batchelder, Alison Pflepsen et Sarah Pouezevara ont dirigé l'atelier EGRA, conçu pour enseigner à ses participants tant en personne qu'en ligne les principes de base relatifs à la conception et à la mise en œuvre d'évaluations des compétences fondamentales en lecture.

La fondation à la base du programme de formation était le Manuel EGRA et les Notes explicatives. Les connaissances et la pratique acquises sur le terrain ont été mêlées au contenu de plusieurs sessions. Richard Vormarwor du Centre de recherche et d'évaluation pédagogique (EARC) au Ghana et Eva Yusuf de Myriad Research en Indonésie—tous deux sous-traitants de RTI—ont fait part de récits et d'expériences particulières tirés de leurs contextes respectifs.

La session d'ouverture a été dirigée par Margaret (Peggy) Dubeck, expert RTI en alphabétisation et Amber Gove, directrice de la recherche chez RTI, qui ont donné un historique et un aperçu de l'instrument EGRA.

Tout au long de l'atelier, les participants ont reçu une orientation, ont été formés aux meilleures pratiques et ont participé à une pratique interactive portant sur plusieurs différents aspects de la conception et de la mise en œuvre de l'outil EGRA. La conception de la recherche et le cadre d'échantillonnage, l'adaptation, l'administration, la notation et la saisie de données, la collecte de données électroniques, l'évaluation et la sélection des collecteurs de données, la collecte de données, les évaluations pilotes et la diffusion des données étaient parmi les sujets traités.

Elena Vinogradova du Centre de développement de l'éducation (EDC) a présenté un bref exposé sur la saisie de données électroniques, après quoi les participants ont pu s'entraîner sur des outils comme « SurveyToGo » et Tangerine®. Ben Sylla et

Christie Vilsak de l'USAID ont également été invités à faire des présentations dans le cadre de l'atelier.

La session de conclusion, intitulée « Mise en œuvre : la pratique rend parfait », a permis aux participants de réunir tous les matériels et les informations tirés des sessions précédentes pour commencer la planification et la conception d'une évaluation des compétences fondamentales en lecture.

Cet atelier, qui s'est déroulé sur 3 jours, a été financé par l'ordre de mission EdData II USAID/Washington intitulé « Mesure et support de recherche à la stratégie en matière d'éducation – Objectif 1 ».

A2 Le personnel technique continue à améliorer la qualité des données EGRA

16 juin 2015 – Kellie Betts

Le Réseau mondial de lecture (Global Reading Network) a organisé deux journées de table ronde sur des sujets en rapport avec la conception, l'administration, l'analyse et le rapport de l'évaluation des compétences fondamentales en lecture (EGRA). Le personnel technique de RTI International, ainsi que d'autres groupes d'experts de diverses organisations, ont présenté des exposés sur chacun des sujets traités. La rencontre a eu lieu à University Research Co., LLC (URC) les 27 et 28 mai 2015.

La rencontre a été financée par l'Agence américaine pour le développement international (USAID), laquelle procède actuellement à la mise au point de directives pour la conception, l'administration et le rapport des tests EGRA. Cette série de présentations et de discussions a été conçue pour rassembler des organisations et des membres du Réseau mondial de lecture et leur permettre de faire part de leur expertise et de leurs expériences. Trente personnes étaient présentes à la rencontre en personne et des douzaines d'autres individus provenant de diverses organisations de mise en œuvre de l'instrument EGRA y ont participé en ligne par webinaire WebEx. Les participants disposaient d'une vaste expérience (actuelle ou passé) dans la planification, l'administration et/ou le rapport d'évaluations EGRA, ce qui leur a permis de contribuer aux présentations techniques et aux discussions fructueuses qui ont défini la rencontre.

Chaque session a duré deux heures et a compris des présentations de groupes d'experts ainsi que des discussions facilitées entre les panélistes et les participants. Chaque discussion portait sur le sujet de la présentation précédente et consistait généralement en des questions de clarification, des suggestions et autres commentaires généraux. A la conclusion de chaque discussion, le facilitateur terminé la session en résumant les domaines de consensus parmi les panélistes et les participants ou ceux méritant une discussion et une attention supplémentaires.

A l'ouverture de la rencontre mercredi 27 mai 2015, le personnel de l'URC et de l'USAID a accueilli les participants et présenté les panélistes, La première session

a porté sur la conception de la recherche et les cadres d'échantillonnage. Chris Cummiskey (RTI), ainsi que Matt Sloan (Mathematica Policy Research, Inc.) et Elena Vinogradova (Education Development Center [EDC]) faisaient partie des présentateurs. La discussion a été facilitée par Melissa Chiappetta de Social Impact.

La deuxième session, qui portait sur la création d'instruments EGRA comparables en plusieurs langues, a compris des présentations d'un groupe d'experts, dont Margaret (Peggy) Dubeck de RTI, Carol de Silva de FHI 360 et Fathi El Ashry de Creative Associates. Pooja Reddy Nakamura (Instituts américains de recherche [AIR]) a mené la discussion sur ce sujet.

La dernière session de la première journée de la rencontre a porté sur la concordance inter-évaluateurs. Simon King (RTI), Jeff Davis (Management Systems International [MSI]) et Abdullah Ferdous (AIR) figuraient parmi les présentateurs. Fathi El Ashry a facilité la discussion subséquente entre les panélistes et les participants.

La deuxième journée de la rencontre, jeudi 28 mai 2015, a commencé par une session sur la préparation et l'analyse des données EGRA / EGMA (évaluation des compétences fondamentales en mathématiques), facilitée par Agaia Zafeirakou (Partenariat mondial pour l'éducation [GPE]). Simon King (RTI), Elena Vinagradova (EDC) et Melissa Chiappetta (Social Impact) faisaient partie des panélistes pour cette session.

La deuxième session, qui portait sur l'équivalence de l'évaluation EGRA pour plusieurs applications dans la même langue, a été présentée par Jonathan Stern (RTI), Jeff Davis (MSI) et Zarko Vukmirovic (AIR). Alla Berezner du Conseil australien pour la recherche en éducation a dirigé la discussion qui s'en est ensuit.

Jeff Davis et Thomaz Alvares (MSI) ont présenté un bref exposé portant sur l'objectif no1 de la stratégie en matière d'éducation et les propositions actuelles sur l'amélioration de la méthodologie. Suite à cette présentation, Benjamin Sylla (USAID) a répondu aux questions et commentaires échangés entre l'audience et les panélistes.

La dernière session de la deuxième journée, facilitée par Jill Meekes (Chemonics), a porté sur trois présentations sur la création de fichiers à usage public pour les jeux de données. Chris Cummiskey et Kellie Betts (RTI), Thomaz Alvares (MSI) et Roger Stanton (Optimal Solutions) ont présenté des exposés sur ce sujet.

Ces deux journées de table ronde ont encouragé des discussions techniques et soulevé des questions devant faire l'objet d'un examen plus approfondi portant sur les divers aspects de l'administration de l'instrument EGRA. Les panélistes participant à cette rencontre prévoient de collaborer à l'élaboration de recommandations et de directives portant sur chacun des sujets présentés. La rencontre a été hautement appréciée, une grande majorité des participants sur place et en ligne convenant que le contenu des sessions était informatif, encourageait la participation et l'interaction et menait à des discussions intéressantes.

ANNEXE B : CONSIDERATIONS SUR LA TAILLE DE L'ECHANTILLON DANS LES EVALUATIONS DES COMPETENCES FONDAMENTALES EN LECTURE

B.1 Introduction

Cette Annexe établit des considérations portant sur la taille applicable aux échantillons destinés à l'évaluation de compétences fondamentales en lecture (EGRA). Elle a pour objectif d'informer le personnel du Ministère de l'éducation, les donateurs ou tous autres acteurs intéressés par la mise en place d'un instrument EGRA sur les exigences en matière de taille de l'échantillon et de calculs.

B.2 Méthode d'échantillonnage

La démarche d'échantillonnage appliquée aura une incidence sur les exigences en matière de taille de l'échantillon. Les autres paramètres étant les mêmes, la sélection aléatoire d'élèves sur une liste nationale nécessitera un échantillon plus petit, alors qu'un échantillonnage en grappes demanda des échantillons plus importants. Bien que cela semble contradictoire, des échantillons choisis purement au hasard sont plutôt coûteux lorsqu'on les compare avec d'autres méthodes d'échantillonnage. Si l'on essayait, par exemple, d'appliquer un échantillon simple choisi au hasard composé de 400 enfants, on serait confronté à une situation où il faudrait se rendre dans presque 400 écoles, puis ne tester qu'un seul enfant dans chaque école, ce qui augmenterait énormément les frais de transport et de main-d'œuvre.[29]

Il faudrait de plus en principe disposer d'une liste de tous les enfants scolarisés dans le pays et de leur adresse pour obtenir un échantillon d'enfants simple et aléatoire. De telles listes n'existent tout simplement pas dans la plupart des pays. Grâce au groupement par grappes des échantillons, les écoles sont d'abord sélectionnées, puis les élèves qui fréquentent ces mêmes écoles (groupes). Le fait de choisir au hasard les écoles en premier, puis les enfants ensuite, réduit les frais

[29] Il ne serait nécessaire d'aller que dans seulement presque 400 écoles, parce qu'en raison du jeu du hasard, et selon le nombre total d'écoles dans le pays, certaines écoles auraient plus d'un enfant sélectionné. Dans un pays avec, disons, seulement 500 écoles, sélectionner un échantillon de 400 enfants par simple échantillon aléatoire devrait probablement générer plusieurs cas où il y aura plus d'un enfant par école, alors que cela ne serait pas le cas dans un pays ayant, disons, 80 000 écoles.

de déplacement et la durée des trajets et élimine également le besoin de dépendre d'une liste nationale des élèves. Etant donné que la plus grande partie des frais encourus par les enquêtes sont en rapport avec le déplacement, on peut tester autant d'enfants que possible dans chaque école au cours d'une visite d'une journée afin d'augmenter la taille des échantillons à un prix relativement faible (voir l'Annexe C de ce manuel où figure des informations supplémentaires sur l'échantillonnage complexe et en grappes.)

Les applications précédentes d'EGRA ont montré qu'il est possible à un enquêteur d'interroger entre 4 et 10 enfants au cours d'une journée scolaire, selon le nombre de tâches et de questions posées à chaque élève.[30] Supposons, en guise d'exemple uniquement, un échantillon de 15 élèves par école, une taille d'échantillon de 400 élèves nécessiterait que l'on ne se rende que dans quelques 27 écoles— économie considérable comparé à des visites dans environ 400 écoles. (L'échantillon réel désiré d'enfants par école peut varier en fonction des caractéristiques du pays). Il est donc recommandé d'appliquer une méthode d'échantillonnage en grappes.

L'application de la méthode du groupement par grappes débouche toutefois sur une perte de réalisme, les enfants variant généralement moins au sein des mêmes écoles que l'« enfant représentatif » dans chaque école ne varie par rapport aux enfants d'autres écoles. Le coefficient de corrélation interne (ICC) entre en jeu, en ce que les enfants d'une école tendent à appartenir à la même classe sociale ou à posséder le même avantage ou désavantage linguistique ou bénéficient d'enseignants de même qualité et sont exposés à des pratiques de gestion similaires—à un degré supérieur à celui les enfants d'écoles différentes. Dans ce sens, la variabilité réelle ou de la population entre les enfants tend à être sous-estimée si l'on utilise une méthode d'échantillonnage par grappes—c'est-à-dire que l'efficacité des frais de transport et de main-d'œuvre est atteinte au prix d'une perte d'informations portant sur la variabilité et ainsi, en l'absence d'ajustement, une perte de précision en résultera. Un indicateur pourra heureusement nous dévoiler le degré auquel le groupement par grappes peut déboucher sur une sous-estimation de la variabilité. Cette mesure, connue sous le nom d'effet de sondage (DEFF), peut être utilisée en vue d'ajuster la taille de l'échantillon pour tenir compte de la perte de variabilité due au groupement.

Quatre éléments doivent figurer dans le calcul de la taille de notre échantillon :

1. *La variabilité* des scores en lecture des élèves (ou autre variable EGRA si on le souhaite) – tant la variabilité générale que la variabilité au sein des écoles et d'une école à une autre
2. *L'amplitude de l'intervalle de confiance (IC)* déterminé par le chercheur
3. *Le niveau de confiance* déterminé par le chercheur (généralement 95 %)
4. *L'effet de sondage (DEFF)* causé par application de l'échantillonnage en grappes

[30] Ce nombre spécifique d'enfants pouvant être interrogés dépend de la version de l'instrument EGRA qui est administrée, du nombre de langues dans lesquelles l'évaluation EGRA est réalisée et si EGRA fait partie d'autres études en cours dans l'école.

B.3 Calcul de la taille de l'échantillon pour un intervalle de confiance et un niveau de confiance donnés

La taille de l'échantillon requis peut être représentée par l'expression algébrique suivante :

$$n = 4\left(\frac{CLtvalue\ DEFT\ SD}{CI_{Width}}\right)^2,$$

dans laquelle :

n est la taille de l'échantillon requise,

CLtvalue est la valeur t associée au niveau de confiance retenu (généralement 1,96 pour 95 %),

DEFT est la racine carrée de l'effet de sondage (DEFF),

SD est l'écart type estimé, qui est une mesure de la variabilité dans la variable retenue,

CI_{Width} = l'amplitude de l'intervalle de confiance déterminée par le chercheur,

le chiffre 4 est dérivé de l'équation de base pour un intervalle de confiance[31]

Comme on peut l'observer à partir de cette équation, les augmentations du *niveau* de confiance, l'*effet* de sondage et la *variabilité* (telle que mesurée par le SD), œuvrent tous pour augmenter la taille de l'échantillon requise (*n*). Toute augmentation de l'*amplitude* de l'intervalle de confiance, inversement, réduit l'exigence de taille de l'échantillon mais, par définition, cette augmentation réduit également la précision.

Aux fins de la formulation de recommandations portant sur la taille de l'échantillon, la racine carrée de l'effet de sondage (DEFT étant la racine carrée de DEFF) et l'écart type (SD) sont calculés à l'aide de données provenant d'applications EGRA précédentes.

Le DEFF est calculé comme suit :

$$DEFF = 1 + (clustersize - 1)\ ICC,$$

[31] Cette équation est dérivée de la formule traditionnelle pour un intervalle de confiance: $\overline{X} \pm CLtvalue\frac{SD\ DEFT}{\sqrt{n}}$, dans laquelle l'expression à droite du signe ± représente l'amplitude unilatérale. L'amplitude bilatérale totale est alors $Width = 2\ CLtvalue\frac{SD\ DEFT}{\sqrt{n}}$. Une manipulation algébrique nous guidera alors jusqu'à l'équation utilisée dans le texte principal et montrera pourquoi le 2 devient un 4.

où :

clustersize est la taille du groupe moyen (le nombre d'enfants échantillonnés dans chaque école[32]) et

ICC est le coefficient de corrélation interne.

Des augmentations dans la taille du groupe ou de l'ICC auront pour effet d'augmenter l'effet de sondage. Si l'on sondait les élèves par échantillonnage aléatoire simple, la taille du groupe serait de 1 (un enfant par école dans l'échantillon), l'ICC serait de zéro parce qu'il n'y a aucun autre élève échantillonné dans l'école à comparer et le EFF serait de 1. Autrement dit, le groupement n'affecte pas la variation présumée si la taille du groupe n'est que de 1.

L'ICC indique dans quelle mesure la variabilité change en fonction des écoles et combien celle-ci varie au sein des écoles. Une manière intuitive d'analyser ceci est de considérer que l'ICC indique la probabilité de trouver deux observations similaires dans le groupe se rapportant à la découverte de deux observations identiques choisies au hasard. Par exemple, un ICC de 0,41 indiquerait qu'il y aurait 41 % plus de possibilités de trouver deux élèves ayant le même niveau de lecture sélectionnés au hasard dans deux écoles, quelles qu'elles soient.

Il existe plusieurs manières d'analyser l'ICC dans la documentation. L'ICC dans ce contexte précis suit l'usage qui en est fait dans le logiciel Stata et il est calculé de la manière suivante :

$$ICC = \frac{MSE_{between} - MSE_{within}}{MSE_{between} + (clustersize - 1)\,MSE_{within}},$$

où :

MSE est l'erreur quadratique moyenne et

clustersize est la taille moyenne des groupes (nombre d'enfants dans chacune des écoles sélectionnées).

$MSE_{between}$ mesure le degré de variation existant entre les écoles (groupes). Arithmétiquement, $MSE_{between}$ est la somme des écarts au carré entre chaque moyenne de groupes (écoles) et la moyenne globale, pondérée par la taille du groupe (le nombre d'enfants échantillonnés dans l'école).

MSE_{within} mesure le degré de variation existant entre les écoles (nos groupes). Arithmétiquement, MSE_{within} est la somme des écarts au carré entre enfant et la

[32] En supposant que les écoles soient le premier élément échantillonné. Dans certaines études, les régions géographiques sont échantillonnées en premier.

moyenne du groupe (école) divisé par le nombre total d'enfants moins le nombre de groupes. En symboles,

$$MSE_{between} = \frac{\sum_{j=1}^{cluster} n_j (\overline{X}_j - \widetilde{X})^2}{cluster - 1}$$

et

$$MSE_{within} = \frac{\sum_{j=1}^{cluster} \sum_{i \in j=1}^{n_j} (x_{ij} - \overline{X}_j)^2}{\sum_{j=1}^{cluster} n_j - cluster},$$

où :

$\widetilde{X}$ est la moyenne « globale » ou générale,

j est un indice pour les groupes,

$i \quad j$ est un indice pour le i^e enfant dans le groupe j,

$\overline{X}_j$ est la moyenne du j^e groupe (ou école),

cluster est le nombre de groupes ou l'indice du dernier groupe et

n_j est la taille du j^e groupe ou l'indice du dernier membre du j^e groupe.

La procédure d'analyse de la variance (ANOVA) dans Excel peut être utilisée pour calculer à la fois MSE_{within} et $MSE_{between}$.

La **Figure B-1** montre une série d'estimations de l'ICC et de la DEFT pour quelques cas particuliers et l'implication de ces variables pour le nombre d'écoles (groupes), ainsi que la taille totale de l'échantillon qui en résulte. Un écart-type (SD) de 29 est présupposé pour tous les cas, une amplitude totale de l'intervalle de confiance (amplitude bilatérale) de 10 est spécifiée et un niveau de confiance de 95 % est utilisé. L'ICC, la DEFT et clustersize sont des valeurs réelles tirées d'études EGRA.

Figure B-1. Estimation des valeurs *ICC* et *DEFT* dans divers pays et diverses classes montrant la taille moyenne des groupes dans chaque cas

Pays	ICC	DEFT	Clustersize	n
Pays A, 3e année	0,17	1,2	3,75	198
Pays B, 2e année	0,22	2,3	20	698
Pays C, 3e année	0,25	1,6	7,57	356
Pays D, 3e année	0,47	2,3	10,05	708
Pays E, 2e année	0,48	1,8	5,35	416

Source : calculé par les auteurs à partir de divers sondages EGRA

Les DEFT introduites à la Figure B-1 sont affectées par l'ICC et par la taille du groupe. Comme on peut l'observer dans l'équation correspondant à la DEFT, ces deux indices affectent la DEFT. Dans le Pays B, par exemple, la DEFT se révèle être légèrement élevée (2,3), même si l'ICC est faible (0,22), parce que la taille du groupe est de 20 ; on élimine donc une grande partie des variations en prenant un grand nombre d'enfants provenant d'écoles particulières. Dans le Pays D, un ICC élevé est à l'origine de l'augmentation de la DEFT. Dans le Pays A, la DEFT est la plus faible parce que la taille des groupes et l'ICC sont tous deux peu élevés. Les incidences sur la taille de l'échantillon requise sont importantes. Dans le Pays A, un échantillon de seulement 198 enfants (mais de quelque 53 écoles) serait nécessaire, alors que dans le Pays D, on aurait besoin d'un échantillon regroupant 708 enfants et 70 écoles

B.4 Recommandations sur les tailles d'échantillon pour les intervalles de confiance

Pour déterminer les tailles réelles recommandées pour les échantillons, il serait raisonnable d'exiger que les différences entre classes soient suffisamment « significatives » d'une manière ou d'une autre—c.-à-d. que les intervalles de confiance globaux soient suffisamment étroits pour que les intervalles de confiance pour les classes qui se suivent ne se chevauchent pas. Si l'on sait que la différence moyenne interclasse est de 14, une amplitude de 14 est raisonnable.

Si l'on suppose une amplitude de 14, un ICC de 0,45, une taille d'échantillon de 12 et un écart-type de 29, la « bonne » taille de l'échantillon est de 409 enfants. Noter que, au fur et à mesure que l'on passe à une classe supérieure, les scores des élèves augmentent ; l'écart-type augmente donc aussi presque certainement. Les classes supérieures exigent généralement en conséquence une taille d'échantillon plus importante pour obtenir le même niveau de précision.

Étant donné les différences généralement minimes entre la moyenne des résultats obtenus par les garçons et les filles dans les tâches EGRA (et/ou étant donné que les différences entre les sexes varient fortement d'un pays à l'autre, contrairement à la progression continue en fonction des années) et étant donné l'équation de la taille

de l'échantillon, il devrait être clair qu'une toute petite *amplitude* serait nécessaire pour détecter les différences entre sexes et donc un échantillon de très grande taille, environ 7 000. Il semblerait prudent d'accepter la notion que les tailles des échantillons les plus raisonnables ne devraient pas saisir *d'importantes différences* entre garçons et filles, ce qui souligne par ailleurs à quel point il est important de faire une distinction entre différence formelle et *différence statiquement importante*. En général, s'il existe une différence quelconque entre deux couches de population, même si celle-ci n'est pas de grande importance, les chercheurs doivent la « forcer » à revêtir une importance statistique en constituant un immense échantillon. En d'autres termes, de petites différences d'importance marginale peuvent être déterminées comme étant statistiquement importantes avec un échantillon de grande taille. Le jugement passé ici est que les différences entre sexes sont suffisamment minimes dans les évaluations EGRA pour que seuls de très grands échantillons puissent les détecter de manière à les rendre statistiquement importantes.

Fin 2015, des évaluations des compétences fondamentales en lecture avaient été menées dans de nombreux pays. Quand des évaluations supplémentaires sont entreprises dans des pays où des sondages ont déjà été menés, on peut employer les données des fichiers à usage public (FUP) provenant de la première évaluation pour procéder à des estimations plus précises pour le nouveau sondage. Les données FUP des missions EdData II (Données sur l'éducation pour la prise de décisions) peuvent être obtenues au travers du site Web EdData II (https://www.eddataglobal.org/datafiles/index.cfm?fuseaction=datafilesIndex). Par exemple, l'extraction de l'ICC et de l'écart-type à partir des données d'une évaluation EGRA menée en Zambie en 2012 a permis de procéder à une estimation plus précise de la taille de l'échantillon pour l'évaluation EGRA nationale 2014 en Zambie. Au sein de Stata, la commande *loneway* peut servir à déterminer l'ICC et la commande *summarize* peut permettre d'établir l'écart-type.

Il convient de tenir compte des différences potentielles entre les populations visées et les évaluations EGRA futures. Si différentes langues, différentes régions ou différentes classes sont évaluées, l'ICC et l'écart-type provenant de l'évaluation EGRA précédente peuvent ne pas être exactes pour des évaluations futures. Si ces obstacles peuvent être surmontés, l'emploi de données historiques présente cependant l'occasion de calculer des estimations raisonnables de taille d'échantillon. La **Figure B-2** montre comment varier le nombre d'élèves échantillonnés de 10 à 22 (en incréments de 2 points) et varier les amplitudes de l'intervalle de confiance entre 10, 12 et 14. Des valeurs fixes de 26 pour l'écart-type et de 0.45 pour l'ICC permettent alors de calculer la DEFT et d'estimer ainsi le nombre d'élèves et d'écoles nécessaires. Comme le montre la **Figure B-2**, avoir davantage d'élèves par école et moins d'écoles ne présente pas d'avantage, l'ICC étant élevé ; 10 élèves par école et 52 écoles permettent d'obtenir le même niveau de précision que 22 élèves par école et 49 écoles, une différence de 561 élèves (1 086 moins 525). En évaluant beaucoup plus d'élèves, on peut se rendre à juste trois écoles de moins—ce qui ne présente guère d'avantage financier.

Figure B-2. Estimation du nombre d'élèves et d'écoles nécessaires en fonction de la variation du nombre d'élèves par école et de l'amplitude de l'intervalle de confiance, l'ICC et l'écart-type restant les mêmes

Valeurs variables		Valeurs fixes			Résultats	
Nombre d'élèves échantillonnés par école	Amplitude d'intervalle de confiance de 95 %	Ecart-type	ICC	DEFT	Nobre total d'élèves	Nombre d'écoles
10	±5	26	0,45	2,25	525	52
12	±5	26	0,45	2,44	618	52
14	±5	26	0,45	2,62	712	51
16	±5	26	0,45	2,78	805	50
18	±5	26	0,45	2,94	899	50
20	±5	26	0,45	3,09	992	50
22	±5	26	0,45	3,23	1 086	49
10	±6	26	0,45	2,25	364	36
12	±6	26	0,45	2,44	429	36
14	±6	26	0,45	2,62	494	35
16	±6	26	0,45	2,78	559	35
18	±6	26	0,45	2,94	624	35
20	±6	26	0,45	3,09	689	34
22	±6	26	0,45	3,23	754	34
10	±7	26	0,45	2,25	268	27
12	±7	26	0,45	2,44	315	26
14	±7	26	0,45	2,62	363	26
16	±7	26	0,45	2,78	411	26
18	±7	26	0,45	2,94	458	25
20	±7	26	0,45	3,09	506	25
22	±7	26	0,45	3,23	554	25

B.5 Vérification d'hypothèse versus intervalles de confiance : implications pour l'échantillonnage

Pour décider de la taille des échantillons, un facteur à retenir est savoir si la base de comparaison entre groupes (c.-à-d., entre niveaux de fluence dans les différentes classes) doit être des intervalles de confiance qui ne se chevauchent pas ou des vérifications d'hypothèses unilatérales. Une pratique courante consiste à présenter les IC selon les variables clés et de déclarer ou de laisser supposer que des IC ne se chevauchant pas sont une première analyse utile en vue de déterminer si les différences entre groupes sont importantes. C'est une pratique courante, le chercheur ne sachant pas d'avance quel contraste ou quelle vérification d'hypothèse revêtira le plus grand intérêt. C'est pourquoi la présentation des IC pour des

variables clés dans EGRA semble être une pratique prudente. Généralement, les lecteurs s'intéressant de près à ce domaine se penchent de plus plutôt sur les paramètres réels qui sont évalués (les niveaux moyens de fluence, par exemple), ainsi que sur leur portée probable et s'intéressent moins à savoir si les différences entre les sous-populations de l'étude ont une importance statistique.

Essayer de réduire suffisamment les IC pour qu'ils ne se chevauchent pas et détecter ainsi une différence précise entre les moyennes requiert toutefois des échantillons plus grands. Il faudrait peut-être des échantillons plus grands pour effectuer des vérifications d'hypothèses. Par ailleurs, les vérifications d'hypothèses sont plus difficiles à interpréter, attirent peut-être trop l'attention sur l'« importance statistique » et s'éloignent des paramètres étudiés. De plus, certaines des économies réalisées par les vérifications d'hypothèses ne peuvent être enregistrées que si les vérifications d'hypothèses sont unilatérales.

Les avis sont partagés dans la littérature spécialisée quant aux conditions qui pourraient justifier une vérification unilatérale d'hypothèses. Le débat n'étant cependant pas concluant, il pourrait être utile de se rappeler les difficultés qui se présentent.

La vérification d'hypothèses postule en général une hypothèse « nulle » selon laquelle, par exemple (en prenant la fluence comme échantillon), la fluence pour une classe donnée est égale à celle d'une classe précédente ou que la fluence après une intervention est équivalente à celle avant cette intervention. On peut alors postuler d'autres hypothèses. Une forme d'hypothèse alternative est que le niveau de fluence dans une classe supérieure est tout simplement différent de celui d'une classe précédente ou que le niveau de lecture après intervention diffère du niveau de lecture avant intervention. Pour vérifier cette hypothèse, on exécute une vérification « bilatérale » d'hypothèses. Cette pratique est courante lorsque l'on s'intéresse aux analyses d'exploration, où un certain traitement ou une certaine variable (niveau de ruralité, expérience de l'enseignant, etc.) pourrait avoir un effet positif ou négatif sur quelque chose d'autre (les résultats des tests pourraient être influencés négativement ou positivement par le degré de ruralité et il n'existe pas a priori de raison valable de tester une hypothèse partant dans une direction plutôt que dans une autre).

Dans la plupart des applications EGRA, il semble raisonnable de penser que la majorité des hypothèses testées, ou la plupart des déclarations que l'on aimerait faire, sont unidirectionnelles. Il semblerait ainsi justifiable de postuler la vérification unilatérale d'hypothèses pour réaliser des économies sur la taille de l'échantillon. S'il existe de bonnes raisons de croire que l'analyse doit être de nature plus exploratoire et descriptive, il conviendra alors d'employer la technique de vérification bilatérale des hypothèses.

Les intervalles de confiance peuvent, si on le souhaite, être présentés avec les vérifications d'hypothèses. Le but d'une présentation des IC est d'encourager l'examen du paramètre en question, facilité de lecture à haute voix d'un texte simple,

par exemple. Il convient toutefois de noter que si les tailles

d'échantillon sont juste suffisamment importantes pour permettre la détection de différences dans des vérifications unilatérales d'hypothèses, les IC auront tendance à être relativement amples. La démarche EGRA décide donc d'abord si des vérifications unilatérales d'hypothèses sont acceptables, avec la condition que cela puisse entraîner des IC légèrement plus amples. La discussion suivante met en évidence les points difficiles.

Supposons que nous ayons deux moyennes d'échantillon, X_1 et X_2. Pour ne pas

compliquer les choses, disons que les écarts quadratiques moyens (SE) sont estimés comme étant les mêmes pour les deux, donc $SE_1 = SE_2 = SE$. Nous supposons également, sans grande perte due à la généralisation, que cela est dû à la similitude des écarts types ainsi que des tailles des échantillons.[33] Aux fins de cette discussion, nous retiendrons des tests de 5 % ou des IC de 95 %. On accepte que les ordonnées t correspondent aux degrés de liberté appropriés. Les IC de 95 % sont

$$\overline{X}_1 \pm t_{.025}\ SE$$

$$\overline{X}_2 \pm t_{.025}\ SE\ ,$$

où $t._{025}$ est l'ordonnée t requise pour un test bilatéral de 5% avec les degrés de liberté appropriés. L'exigence que les deux IC pour chaque moyenne ne se chevauchent pas équivaut à exiger que

$$\overline{X}_1 + t_{.025}SE < \overline{X}_2 - t_{.025}SE$$

ou

$$\overline{X}_2 - \overline{X}_1 > t_{.025}SE + t_{.025}SE = 2t_{.025}SE$$

si l'estimation de la première moyenne est inférieure à la seconde, et de manière similaire, mais avec des signes différents, si la seconde est inférieure, soit plus généralement :

$$\left|\overline{X}_1 - \overline{X}\right|_2 > 2t_{.025}\ SE\ ,$$

parce que les IC pour les moyennes sont symétriques aux abords de la moyenne et

[33] En fait, la plupart des ET et des SE différeront l'un de l'autre. L'égalité entre la taille de l'échantillon et l'ET est supposée pour cette exposition, uniquement à des fins de clarification.

ont la même amplitude, en supposant que les SE et les degrés de liberté (influencés par *n*) restent inchangés.

Mais l'exigence selon laquelle l'IC pour la *différence* ne se chevauche pas *avec 0* équivaut à exiger que

$$\left|\overline{X}_1 - \overline{X}_2\right| > 1.41\, t_{.025}\, SE,$$

en raison de l'équation de l'écart-type pour une différence entre moyennes, qui est comme suit, étant donné l'hypothèse retenue de l'égalité des écarts-type et de l'égalité des échantillons :

$$SD_{diff} = \sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}} = \sqrt{2\frac{SD^2}{n}} = 1.41\, SD.$$

Noter que le rapport de 2 à 1,41 est de 1,41, puisque tout nombre divisé par sa racine carrée est égal à sa racine carrée. Cela signifie que dans le premier cas, il faudrait un SE plus petit que dans le second cas, afin de ne pas créer de chevauchement des IC—1,41 fois moins. Etant donné que $SE = SD/\sqrt{n}$, une SE qui est 1,41 fois inférieure requiert un échantillon qui est 2 fois plus grand, puisque

$$\frac{SE}{1.41} = \frac{SD}{1.41\sqrt{n}} = \frac{SD}{\sqrt{2n}}.$$

Les tests instantanés suivants provenant de Stata (à l'aide de la commande « ttesti ») sont présentés à titre d'exemple. Ces tests se servent des valeurs qui ont déjà été utilisées dans les exemples ci-dessus. Pour illustrer le principe fondamental portant sur les différences entre intervalles de confiance et vérifications d'hypothèses, nous allons nous concentrer sur un cas dans lequel le DEFF est de 1. La *procédure* employée est celle qui est utilisée pour les variances inégales, bien qu'en pratique et pour faciliter l'exposé, les données d'entrée pour les écarts types figurant dans les exemples s'équivalent les uns les autres. Noter que la commande ttesti ne peut pas être utilisée pour la plupart de l'analyse EGRA parce qu'elle n'ajuste pas les erreurs types pour tenir compte du modèle d'échantillon en grappes. Nous sommes en premier lieu confrontés à un cas où l'intervalle de confiance pour la différence entre les deux moyennes ne chevauche pas zéro, mais le fait presque, comme on l'observe dans la zone surlignée d'en bas. Il convient de remarquer que le logiciel Stata présente les IC pour chaque variable, l'IC pour la *différence* entre les variables, ainsi que toutes les vérifications d'hypothèses pertinentes pour la différence entre les variables.

```
ttesti 34 20 29 34 34 29, unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      34          20    4.973459          29    9.881422    30.11858
       y |      34          34    4.973459          29    23.88142    44.11858
---------+--------------------------------------------------------------------
combined |      68          27    3.593661    29.63409    19.82702    34.17298
---------+--------------------------------------------------------------------
    diff |                 -14    7.033533               -28.0429     .042902
------------------------------------------------------------------------------
    diff = mean(x) - mean(y)                                      t =  -1.9905
Ho: diff = 0                     Satterthwaite's degrees of freedom =       66

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0253         Pr(|T| > |t|) = 0.0507          Pr(T > t) = 0.9747
```

Les IC pour les deux moyennes se chevauchent en grande partie, comme on peut le voir dans les deux zones surlignées du haut, mais l'IC pour la *différence* ne chevauche pas zéro (même s'il le fait presque, par effet voulu) comme cela est visible dans la zone surlignée du bas. Toutefois, cette manière d'interpréter l'exigence de détection de la différence entre les groupes est la bonne. Pour éviter tout chevauchement des IC pour les moyennes elles-mêmes, il faudrait doubler la taille des échantillons.

Le test suivant montre qu'en multipliant par deux la taille de l'échantillon, les CI pour les moyennes individuelles évitent de peu le chevauchement, comme on peut le voir dans les zones surlignées :

```
ttesti 69 20 29 69 34 29, unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      69          20     3.49119          29    13.03344    26.96656
       y |      69          34     3.49119          29    27.03344    40.96656
---------+--------------------------------------------------------------------
combined |     138          27    2.531281    29.73582    21.99457    32.00543
---------+--------------------------------------------------------------------
    diff |                 -14    4.937288               -23.76379   -4.236213
------------------------------------------------------------------------------
    diff = mean(x) - mean(y)                                      t =  -2.8356
Ho: diff = 0                     Satterthwaite's degrees of freedom =      136

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0026         Pr(|T| > |t|) = 0.0053          Pr(T > t) = 0.9974
```

Mais multiplier par deux la taille de l'échantillon est cher payé (et inutile) pour obtenir des IC ne se chevauchent pas pour les moyennes, plutôt qu'un IC ne chevauchant pas zéro pour la différence entre les moyennes. Cela peut être observé par le fait que l'IC pour la différence entre les moyennes est relativement éloignée du zéro (zone surlignée du milieu), ou par le fait qu'une vérification bilatérale d'hypothèses pour la différence entre les deux moyennes génère une valeur de probabilité en dessous du seuil de 5 % (zone surlignée du bas).

On a toutefois encore un peu plus de marge de manœuvre. La plus grande partie du gain en efficacité parmi les vérifications d'hypothèses concernant la notion d'« intervalles de confiance qui ne se chevauchent pas » est réalisé simplement en postulant le problème comme vérification d'hypothèses. Toutefois, si cela est désiré et si cela est justifié *a priori*, une petite amélioration de l'efficacité peut être obtenue en supposant un test unilatéral d'hypothèses. Noter que dans le premier imprimé Stata ci-dessus, malgré que l'IC de la différence touche presque zéro, une vérification unilatérale d'hypothèses est très forte—« excessivement » forte pour un test de 5 %. Du fait que l'IC de 95 % pour la différence se rapproche beaucoup de zéro, la valeur de probabilité pour une vérification *bilatérale* d'hypothèses est véritablement 0,05 (ou s'en rapproche) conformément à nos attentes, étant donné l'équivalence entre une vérification bilatérale d'hypothèses et un IC pour une différence entre moyennes n'incluant pas zéro. Mais la valeur de probabilité pour une vérification unilatérale d'hypothèses, lors du premier lancement ci-dessus, n'étant que de 0,025 (0,0249 en réalité), nous avons plus de degrés de liberté qu'il n'est nécessaire si nous ne désirons qu'un test de 5 %. Puisque la valeur t pour un test unilatéral d'hypothèses de 5 % est de 1,67 (plus ou moins, pour un n élevé), alors que ce qui était nécessaire pour un test bilatéral était d'environ 1,96, nous pourrions réduire l'échantillon d'un taux d'environ $\sqrt{1.67/1.96} = 0.73$.

Nous n'avons en fait seulement besoin que

$$\left|\overline{X}_1 - \overline{X}_2\right| > 1.41\, t_{.05}\, SE$$

pour un test unilatéral, où t ≈ 1,67 avec un *n* raisonnablement élevé.

Le test instantané Stata suivant démontre que lorsque l'on réduit la taille de l'échantillon, à partir de la première série de résultats, en fonction d'un taux de 0,73 sur 34, ou 25, le test unilatéral d'hypothèses a une valeur de probabilité juste en dessous de 0,05, comme cela est nécessaire (zone du bas surlignée). Les IC se superposent parfaitement à présent (zones du haut surlignées). L'IC de 95 % pour la différence se superpose même avec zéro, parce que le fait d'avoir besoin d'un IC ne se superposant pas avec zéro pour la différence serait équivalent à un test bilatéral d'hypothèses.

```
ttesti 25 20 29 25 34 29, unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      25          20         5.8          29    8.029388    31.97061
       y |      25          34         5.8          29    22.02939    45.97061
---------+--------------------------------------------------------------------
combined |      50          27    4.180518    29.56073    18.59893    35.40107
---------+--------------------------------------------------------------------
    diff |               -14    8.202439               -30.49211    2.492108
------------------------------------------------------------------------------
    diff = mean(x) - mean(y)                                      t =  -1.7068
Ho: diff = 0                     Satterthwaite's degrees of freedom =       48

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0472         Pr(|T| > |t|) = 0.0943          Pr(T > t) = 0.9528
```

En tenant compte des deux facteurs à la fois, la taille de l'échantillon requise pour un test unilatéral d'hypothèses est d'environ 0,36 de ce qui est requis pour créer des IC (bilatéraux) ne se superposant pas sur les deux moyennes.

Noter que si l'écart type est suffisamment augmenté par une DEFT de 2,44 (le résultat des mêmes hypothèses que celles utilisées pour établir la taille de l'échantillon de 409 pour un IC, c'est à dire un ICC de 0,45 et des groupes de 12), alors la taille de l'échantillon requise pour un test de 5% augmente, surtout jusqu'à $2{,}44^2$ fois 25, soit 148.

```
ttesti 148 20 70.7 148 34 70.7, unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |     148          20    5.811504        70.7    8.515112    31.48489
       y |     148          34    5.811504        70.7    22.51511    45.48489
---------+--------------------------------------------------------------------
combined |     296          27    4.122578    70.92751    18.88661    35.11339
---------+--------------------------------------------------------------------
    diff |               -14    8.218708               -30.17496    2.174957
------------------------------------------------------------------------------
    diff = mean(x) - mean(y)                                      t =  -1.7034
Ho: diff = 0                     Satterthwaite's degrees of freedom =      294

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0448         Pr(|T| > |t|) = 0.0895          Pr(T > t) = 0.9552
```

Ces facteurs permettent de réaliser quelques économies dans la taille d l'échantillon avec une vérification unilatérale d'hypothèses par opposition aux intervalles de confiance ne se superposant pas. Toutefois, il existe une pression inverse, celle de tenir compte de la puissance. En tenant compte de la puissance, en supposant une puissance de 0,8 et une vérification d'hypothèse de 5 % et en introduisant la notion selon laquelle les écarts-type *pourraient* être différents, une taille d'échantillon pour une vérification unilatérale d'hypothèses est

$$n = \frac{(SD_1^2\ DEFF + SD_2^2\ DEFF)(0.85 + 1.67)}{DIFF^2},$$

où :

0,85 est la valeur unilatérale t pour une puissance de 0,8

1,67 est la valeur unilatérale t pour un test de 5% (tous les deux avec 60 degrés de liberté, un nombre peu élevé choisi à dessein)

DIFF est la différence hypothétique, entre les classes par exemple.

En utilisant les mêmes paramètres que pour l'intervalle de confiance, c'est-à-dire un DEFF de 5,595 (DEFT de 2,44) (à cause d'un ICC de 0,45 et une taille des groupes fixée à 12) et des écarts types (ET) de 29 (ils sont similaires dans cet exemple mais l'équation permettant différents écarts types) et une DIFF de 14, la taille de l'échantillon requise est de 324. Dans le cas le plus pessimiste où les écarts types sont de 50, mais où on laisse la DIFF atteindre 20, la taille d'échantillon requise est de 472. Dans les deux cas, ceux- ci sont légèrement plus petits que nécessaire pour un intervalle de confiance de 95 %.

En guise de conclusion, et en se basant sur les discussions ci-dessus selon lesquelles les tests bilatéraux conviennent le mieux, l'équation correcte serait la suivante :

$$n = \frac{(SD_1^2\ DEFF + SD_2^2\ DEFF)(0.85 + 2)}{DIFF^2}.$$

Dans ce cas, et en utilisant les mêmes hypothèses que ci-dessus, la taille de l'échantillon avec un écart type de 29 est de 414 et, avec un écart type plus pessimiste de 50 mais une DIFF de 20, elle passerait à 603.

B.6 Résumé des tailles des échantillons sur la base des intervalles de confiance et des vérifications d'hypothèses

La **Figure B-3** résume une série de suggestions sur les tailles des échantillons. Si des données EGRA historiques sont cependant disponibles pour ce pays, ces données sont employées en priorité pour estimer des écarts types et des ICC qui conviennent mieux. Cette figure suppose un écart type de 29, un ICC de 0,45 (ce qui est proche de la limite supérieure de ce qui a été trouvé dans les études EGRA réalisées jusqu'ici), et une taille de groupes (nombre d'élèves échantillonnés par école) de 10. Dans le cas de la vérification d'hypothèses, on suppose une puissance de 0,8. Dans chaque cas, le nombre d'écoles nécessaire est dérivé en arrondissant le résultat de la division de l'échantillon par 10.

Figure B-3. Résumé des tailles des échantillons en fonction de divers critères

	Taille de l'échantillon	Nombre d'écoles
Niveau de confiance de 90 %		
Méthode pour l'intervalle de confiance :		
Amplitude bilatérale de l'intervalle : 10	475	48
Amplitude bilatérale de l'intervalle : 15	211	22
Méthode de vérification de l'hypothèse – unilatérale :		
Différence minimale détectable : 10	390	39
Différence minimale détectable : 15	173	18
Méthode de vérification de l'hypothèse – bilatérale ::		
Différence minimale détectable : 10	539	54
Différence minimale détectable : 15	239	24
Niveau de confiance de 95 %		
Méthode pour l'intervalle de confiance :		
Amplitude bilatérale de l'intervalle : 10	680	68
Amplitude bilatérale de l'intervalle : 15	303	31
Méthode de vérification de l'hypothèse – unilatérale :		
Différence minimale détectable : 10	539	54
Différence minimale détectable : 15	239	24
Méthode de vérification de l'hypothèse – bilatérale :		
Différence minimale détectable : 10	689	69
Différence minimale détectable : 15	306	31

Source : calculé par RTI International.

B.7 Echantillonnage et pondérations

En général, pour le sondage d'écoles, un échantillonnage par probabilité proportionnelle de la taille de la population (PPT) est la technique la plus fréquemment employée et recommandée. Avec cette technique, on sélectionne les écoles pour un échantillonnage de 1er niveau dans lequel la probabilité de sélection de c chaque école est proportionnelle au nombre d'élèves dans l'école (ou la classe) divisé par le nombre d'écoles dans la région ou le pays souhaité. La probabilité de sélection d'élèves dans l'école est le deuxième niveau d'échantillonnage, dans lequel la probabilité de sélection de chaque élève est le nombre d'élèves à sélectionner divisé par le nombre d'élèves dabs l'école ou dans la classe. La probabilité générale de sélection d'élèves est donc le produit de ces deux probabilités de sélection.

Si les niveaux individuels présentent différentes probabilités de sélection des unités d'échantillonnage pour ce niveau, les produits des deux niveaux présentent des probabilités générales de sélection égales quand le nombre d'élèves sélectionnés dans chaque école est le même.

La probabilité générale de sélection est égale à

Probabilité de 1er niveau × probabilité de 2e niveau

soit

$$\frac{\textit{Nombre d'élèves dans l'école x Nombre d'écoles sélectionnées}}{\textit{Nombre d'élèves dans la région / pays}} \times \frac{\textit{Nombre d'élèves sélectionnés dans l'école}}{\textit{Nbre d'élèves dans l'école}}$$

qui peut être simplifié comme suit :

$$\frac{\textit{~~Nombre d'élèves dans l'école~~ x Nombre d'écoles sélectionnées}}{\textit{Nombre d'élèves dans la région / pays}} \times \frac{\textit{Nombre d'élèves sélectionnés dans l'école}}{\textit{~~Nbre d'élèves dans l'école~~}}$$

ce qui donne

$$\frac{\textit{Nombre d'écoles sélectionnées x Nombre d'élèves sélectionnés dans l'école}}{\textit{Nombre d'élèves dans la région / pays}}$$

Les pondérations finales sont l'inverses des probabilités de sélection générales. Si les pondérations sont égales, il s'ensuit que la moyenne pondérée est la même que la moyenne non pondérée. Si c'est ce qui doit se produire en théorie, cela n'est généralement pas le cas du fait d'un nombre moins important d'élèves échantillonnés dans certaines écoles, de la stratification des écoles, de substitution d'écoles, etc. Un échantillonnage PPT donne cependant des pondérations proches les unes des autres, ce qui réduit le biais de l'échantillon.

ANNEXE C : ECHANTILLONNAGE COMPLEXE ET EN GRAPPES

Pour les importantes études portant sur l'éducation, il n'est ni rentable ni pratique de sélectionner au hasard des élèves dans l'intégralité de la population visée (c.-à-d. d'employer un échantillonnage aléatoire). Cela nécessiterait une liste courante de chaque élève dans la population concernée, et la plupart des ministères de l'éducation ne disposent pas de cette information. Même si c'était le cas, il ne serait pas rentable pour les équipes d'évaluation de se rendre dans une école donnée pour évaluer uniquement un ou deux élèves. Il est bien plus rentable et pratique d'échantillonner des écoles au hasard, puis d'échantillonner un groupe d'élèves au sein de chacune des écoles sélectionnées. On appelle cette méthodologie échantillon scolaire complexe. La plupart des échantillons EGRA sont des échantillons scolaires. Un échantillonnage scolaire implique souvent le sondage de l'ensemble des d'écoles, puis celui des élèves (ou parfois l'échantillonnage en premier de régions géographiques, districts par exemple, puis celui des écoles, suivi par celui des élèves). Quels que soient ces étapes d'échantillonnage, il y a une forme quelconque de regroupement d'élèves au sein des écoles échantillonnées.

Un grand nombre d'études EGRA impliquent une *inférence statistique* dans laquelle un échantillon aléatoire d'élèves est tiré d'une population visée explicite et les résultats des élèves échantillonnés sont employés comme estimations pour déduire les résultats pour la population. Appelée *statistique paramétrique déductive*, cette technique comprend généralement deux types d'estimations : (1) *estimations ponctuelles*, valeurs uniques calculées à partir des données pour représenter des paramètres inconnus et (2) *estimations de fidélité*, plage de valeurs probables. (On trouvera plus bas une définition plus complète des estimations ponctuelles et des estimations de fidélité).

En fonction de ces deux estimations et des degrés de liberté, on peut calculer un intervalle de confiance de 95 % et procéder à une analyse statistique formelle. Noter que ces deux types d'estimations sont directement affectées par la façon dont l'échantillon est prélevé.

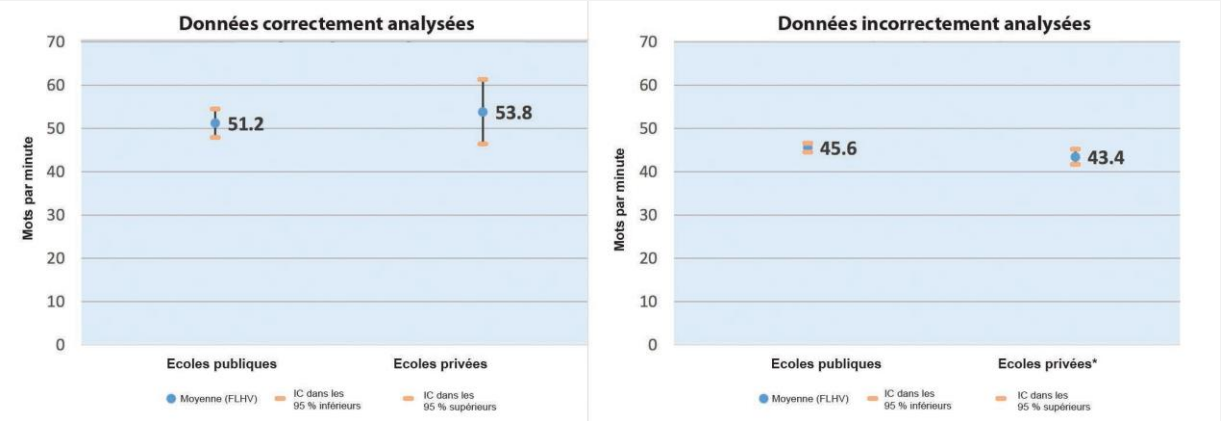
Si on ne tient pas compte de la méthodologie d'échantillonnage, le logiciel statistique supposera que les élèves ont été choisis par simple échantillonnage aléatoire. Cela entrainera un biais dans toutes les estimations ponctuelles des paramètres de la population. Toutes les estimations de fidélité seront de plus exagérément basses. Ces deux effets combinés peuvent entraîner une conclusion erronée du chercheur, à savoir qu'il e existe des différences statistiquement notables parmi les sous-groupes alors qu'en fait il n'y en a pas, comme on montre à la **Figure C-1.**

Figure C-1. Données correctement analysées et données incorrectement analysées

La facilité de lecture à haute voix (FLHV) a été analysée par type d'école. Le graphique de gauche montre des données qui ont été analysées avec le bon échantillonnage. Dans le graphique de droite, les données ont été analysées sans le bon échantillonnage. Autrement dit, les données n'ont pas été pondérées et le chercheur a laissé le logiciel d'analyse statistique passer par défaut à l'échantillonnage des élèves par simple échantillonnage aléatoire.

La moyenne des estimations de FLHV (mots par minute) dans la mauvaise analyse est biaisée vers la population, parce qu'elles supposent que seuls 25 % de la population des élèves provenaient de la région de Java–Bali d'Indonésie, alors qu'en fait 55 % de la population des élèves provenaient de cette région. L'échantillon sous-représente en conséquence la population des élèves provenant de la région de Java–Bali, ce qui entraîne in biais dans la moyenne des estimations.

De plus, comme on a laissé le programme d'analyse supposer qu'il s'agissait d'un simple échantillonnage aléatoire, il n'a tenu compte d'aucun des effets du plan de sondage qui accompagnent les échantillons complexes. Cela a entraîné une estimation d'erreur standard extrêmement basse, entraînant ainsi un degré de confiance exagéré dans l'estimation moyenne biaisée. Ce chercheur aura peut-être incorrectement conclu qu'il existait une différence statistiquement significative dans les compétences en lecture des élèves d'écoles publiques comparé aux écoles privées—alors que les données correctement analysées montrent que cela n'était pas le cas.



Source : évaluation EGRA nationale des élèves de 2e année en Indonésie (2014). * p < 0,05

ANNEXE D : ECHANTILLONNAGE POUR EVALUATIONS DE L'IMPACT

On trouvera dans cette annexe une description de la façon dont les calculs de puissance statistique peuvent permettre de déterminer correctement l'échantillon nécessaire pour estimer les impacts.

Passer d'estimations ponctuelles à une estimation des impacts affecte énormément les considérations de taille d'échantillons. Le calcul de puissance statistique démontre la mesure dans laquelle une évaluation peut distinguer des impacts réels de différences fortuites pour permettre de répondre à deux questions en corrélation l'une avec l'autre :

1. **Pour une taille d'échantillon donnée, de quelle dimension d'effet aurait-on besoin pour que l'étude puisse raisonnablement observer une différence statistiquement significative entre les groupes sondés et les groupes témoins ?** Cette question est pertinente dans une situation où l'échantillon disponible est fixé et ne peut être altéré. Une étude peut par exemple se limiter à un total de 200 écoles et des calculs de puissance statistique permettraient à un chercheur de déterminer la taille d'un changement qu'il serait nécessaire d'observer pour établir avec confiance que le changement est statistiquement significatif (autrement dit réel). Dans ce cas, le chercheur commence avec l'échantillon disponible, puis calcule la taille de l'effet qui serait nécessaire pour le détecter avec confiance.
2. **Pour une dimension d'effet donnée, de quelle taille d'échantillon aurait-on besoin pour s'assurer en toute probabilité que l'étude détecterait cette dimension d'effet s'il résultait de l'intervention ?** Ici, si l'échantillon est quelque peu flexible, les calculs de puissance statistique peuvent permettre de déterminer la taille d'échantillon nécessaire pour observer un changement statistiquement significatif dans un résultat d'une certaine magnitude observé. Par exemple, les chercheurs, l'USAID ou les partenaires de mise en œuvre peuvent savoir, soit par la théorie du changement, soit par des études d'interventions similaires, qu'ils s'attendent à un certain changement suite à l'intervention. Dans ce cas, le chercheur commencerait avec cet effet attendu, puis déterminerait la taille d'échantillon nécessaire pour pouvoir le détecter avec confiance.

Pour déterminer les tailles d'échantillon convenant à une évaluation, les évaluateurs calculent généralement les *impacts décelables minimaux* (MDI) qui représentent les plus petits impacts programmatiques réels—moyenne des différences entre le

groupe sujet de l'intervention et le groupe témoin—pouvant entraîner une estimation statistiquement significative avec forte probabilité pour une taille d'échantillon donnée.

Il est courant de normaliser les MDI en des unités d'ampleur d'effet—c'est-à-dire sous forme de pourcentages des écarts types des mesures du résultat. On appelle alors le MDI normalisé taille *d'effet décelable minimal* (MDES). La mise à échelle des estimations en des unités d'écart type facilite la comparaison de conclusions pour divers résultats mesurés sur différentes échelles.

La formule MDI peut être mathématiquement exprimée comme suit :

$$MDI = Facteur * SE(impact),$$

où SE(impact) est l'erreur standard de l'estimation de l'impact et Facteur est une constante qui est fonction des niveaux de puissance statistique et de signification. Le Facteur augmente au fur et à mesure que le niveau de signification baisse et que le niveau de puissance augmente. Le MDI augmente donc quand on cherche à réduire les chances d'erreurs de Type I et de Type II. La valeur SE(impact) varie en fonction du modèle d'évaluation de l'impact. En général, des échantillons plus grands réduisent la valeur SE(impact) et, en conséquence, le MDI, ce qui rend l'évaluation « plus puissante ». Une plus grande puissance est désirable, l'évaluation étant plus susceptible de détecter des impacts substantiellement significatifs, bien qu'une plus grande puissance coûte généralement plus cher.

La formule pour la MDES divise le MDI par SD(outcome), écart type de la mesure du résultat :

$$MDES = MDI / SD(outcome)$$

Une MDES est fonction de l'erreur standard de l'estimation de l'impact, du niveau de signification supposé et du niveau de puissance suppose. Le niveau de signification est la probabilité de commission d'une erreur de « Type I » ; cette erreur est un faux positif—qui conclut incorrectement qu'il y a un impact là où il n'y en a pas. Un niveau de signification conventionnel est de 5 pour cent. Le niveau de puissance est un moins la probabilité d'une erreur de « Type II » ; cette erreur est un faux négatif—la non détection d'un impact qui existe réellement.

Les évaluations tentent souvent de parvenir à une puissance de 80 pour cent. L'objectif est d'avoir une petite MDES pour que, si l'étude produit un effet plus grand que la MDES, on l'appelle statistiquement significatif et pense qu'il est réel. Toutes les études comportent un calcul de ce type, la formule étant correctement documentée pour guider les décisions concernant la composition et la taille de l'échantillon.

Autres facteurs. Il existe dans les calculs de puissance statistique pour les évaluations d'impact de nombreux facteurs dont il convient de tenir compte pour le développement d'échantillons pour une estimation ponctuelle. Un facteur

hautement pertinent pour les chercheurs tentants d'estimer les impacts à l'aide d'une évaluation EGRA est le regroupement de l'échantillon. De nombreuses études ayant recours à l'outil EGRA sont en grappes (soit les écoles soit les communautés sont sélectionnées en premier pour les groupes sujets et témoins, puis un certain nombre d'individus au sein de chaque groupe sont testés).

Dans cette situation, on suppose que les individus au sein du groupe ont certaines similitudes en commun. Par exemple, les enfants d'une classe, indépendamment de toute intervention, peuvent tous avoir le même enseignant. Cela réduit la variation individuelle au sein du groupe, et donc la contribution de chaque individu à l'estimation de l'impact. On appelle dette mesure du degré auquel les résultats des individus au sein des groupes sont en corrélation coefficient de corrélation interne (ICC). Dans une situation où l'ICC était plus proche de 1, l'addition d'individus supplémentaires aurait un effet limité ou nul sur la MDES. Le chercheur devra plutôt ajouter des groupes additionnels pour réduire la MDES. En fait, dans une étude en grappes présentant un ICC élevé, l'addition d'individus dans chaque groupe aura au mieux un effet positif minimal. Autres facteurs ayant une incidence sur les calculs de puissance statistique : nombre de contrastes (volets de l'intervention ou groupes d'étude) et si le test est unilatéral ou bilatéral (un test unilatéral chercherait à n'estimer l'impact que dans une direction).

Enfin, il est important de se rappeler que l'analyse de sous-groupes aura également une incidence sur la taille de l'échantillon. Par exemple, il peut être pertinent de comprendre comment une intervention particulière a une incidence différente pour les garçons et les filles. Dans ce cas, les calculs de puissance sont réalisés au nouveau du sous-groupe. Noter que dans ce cas chaque sous-groupe est suffisamment grand pour pouvoir détecter les impacts entre les membres de ce groupe seul dans les groupes concernés et les groupes témoins. Autrement dit, plus il y a de sous-groupes à analyser (désagrégations demandées par l'USAID ou d'autres intervenants), plus l'échantillon sera grand.

Résumé. Pour résumer, un chercheur emploierait les principes de calculs de puissance statistique suivants pour informer la taille de l'échantillon.

1. Plus l'échantillon est grand, plus la puissance est grande (pour une étude en grappes, plus le nombre de groupes est élevé, plus la puissance est grande).
2. La puissance est plus élevée quand l'écart type du résultat est petit que quand il est grand.
3. Plus l'amplitude d'effet est élevée, plus il y a de chances qu'une évaluation conclue à un effet significatif.
4. Il y a un compromise entre le niveau de signification et la puissance : plus le niveau de signification est rigoureux (plus il est bas), plus la puissance est faible.
5. La puissance est plus élevée avec un test unilatéral qu'avec un test bilatéral, tant que la direction supposée est correcte.

ANNEXE E : EVALUATION DE LA QUALITE TECHNIQUE DE L'INSTRUMENT EGRA

Il est important d'évaluer la qualité technique de tout instrument utilisé dans la mesure des résultats des élèves. L'instrument EGRA ne fait pas exception à cette règle. Les procédures employées pour ces contrôles ressortent du domaine de la psychométrie. Traditionnellement, ces procédures se sont concentrées sur deux principaux concepts : fiabilité et validité. Les équipes administrant l'évaluation EGRA doivent comporter un spécialiste en psychométrie qui puisse exécuter les contrôles nécessaires. La fiabilité et la validité sont expliquées à la Section 9.1.2. On trouvera ci-dessous des types d'analyses supplémentaires pouvant être envisagées pour établir la fiabilité et la validité des instruments.

E.1 Tests de fiabilité

La méthode tester-retester est une autre mesure de mise à l'épreuve de la fiabilité. Cette démarche, qui peut être entreprise dans le cadre du pilotage de l'instrument EGRA, implique principalement l'administration de l'instrument EGRA au même groupe d'élèves à deux moments différents (par exemple, avec un écart d'une semaine environ). Les élèves sélectionnés sont représentatifs de la population ciblée dans les domaines clés, tels que le sexe et l'âge, le statut socioéconomique/ historique familial, compétences cognitives, et ainsi de suite. Le coefficient de fiabilité pour la méthode tester-retester représente la corrélation entre les scores des élèves pour les deux administrations du test. Dans des conditions idéales, ces corrélations sont également établies sur les mesures sommaires des tâches (pourcentage correct, fluence, etc.) parce que si les corrélations sont calculées pour des items individuels au sein des tâches, le même biais à la hausse introduit par un teste EGRA chronométré sera présent dans un test–retest comme dans le coefficient alpha de Cronbach.

Le biais à la hausse peut être illustré à l'aide d'un exemple numérique. Supposons qu'on ait des résultats pour eux enfants—ou pour le même enfant après un certain délai—comme il est présenté à la **Figure E-1** dans laquelle les 0 et les 1 représentent respectivement des résultats « incorrects » ou « corrects ». Dans aucun des deux cas l'enfant a été capable d'aller au-delà du 5e mot. Si un analyste devait calculer la corrélation pour les mots du 1er au 5e, la corrélation serait très faible (0,17). Si l'analyste devait considérer comme incorrects les mots après le 5e, puis calculer la corrélation pour tous les 10 mots, la corrélation serait alors de 0,38— bien plus élevée mais incorrectement dérivée, tous les 0 après le 5e mot ayant gonfler artificiellement la corrélation.

Figure E-1. Exemple de résultats de tâche pour le calcul d'un biais à la hausse

Mot	Enfant 1	Enfant 2 (ou Enfant 1 plus tard)
1er	0	0
2e	1	1
3e	0	1
4e	1	0
5e	0	0
6e	0	0
7e	0	0
8e	0	0
9e	0	0
10e	0	0

Deux points dont il convient de tenir compte quand la méthode tester–retester est employée dans le domaine de l'éducation :

- Premièrement, si les tests sont trop éloignés l'un de l'autre, il est probable que les élèves auront acquis d'importantes connaissances et que le manque de fiabilité soit en fait une mesure de l'amélioration des résultats scolaires.
- Deuxièmement, limiter les délais entre les deux administrations de l'évaluation eut réduire l'impact de l'apprentissage mais augmenter la probabilité d'effets de report. Autrement dit, les scores pour la deuxième administration sont influencés par le fait que le test a été récemment administré aux mêmes élèves.

Une autre mesure de fiabilité du test est la **fidélité de versions parallèles**. Cette démarche emploie deux versions similaires de l'instrument EGRA. La procédure consiste dans ce cas à administrer la version 1 du test à chaque élève, puis à administrer la version 2 aux mêmes élèves. Il est recommandé d'inverser l'ordre d'administration des formulaires pour la moitié du groupe sélectionné. La corrélation entre les deux séries de scores offre une mesure du degré de fiabilité des scores EGRA pour les deux versions du test.

Cette méthode est particulièrement utile quand plusieurs versions d'évaluations sont créées pour mesurer les scores à plusieurs moments (résultats initiaux, à mi-parcours et finaux). Il est cependant important de se rappeler qu'une forte corrélation entre deux versions de test ne signifie pas nécessairement que les versions sont équivalentes ni qu'un étalonnage ne va pas être nécessaire. Comme pour l'application des corrélations uniquement aux mesures sommaires des tâches, il faudra ici faire preuve de prudence.

E.2 Tests de validité

Les **preuves liées au critère** font référence à l'étroitesse du rapport (corrélation) existant entre les résultats pour le test EGRA et autres indicateurs externes au test. Cela implique en général l'examen du rapport entre les scores EGRA et ceux portant sur les indicateurs de certains critères que le test est sensé prédire (par exemple, les résultats du test de compréhension dans les années supérieures), ainsi que les rapports aux autres tests postulés comme mesurant les mêmes constructs ou des constructs associés (par exemple, les scores des élèves dans d'autres évaluations des compétences fondamentales en lecture). Les données sur ces autres mesures peuvent être recueillies en même temps que les données EGRA ou plus tard (mais recueillies auprès des mêmes élèves). Ce type de preuve de validité sera difficile à collecter dans les pays disposant de peu d'indicateurs standardisés sur les résultats d'apprentissage des élèves. Il ne faut toutefois pas oublier que des travaux de recherche extensifs dans d'autres pays ont établi que des instruments de type EGRA révèlent des rapports étroits (0,7 et plus) avec les types de mesures externes fournis à titre d'exemples dans ce paragraphe.

Certains concepteurs recommandent la collecte d'un type supplémentaire de preuves dans le cadre de la validation du test, à savoir la **preuve des conséquences de l'utilisation des scores du test** sur les candidats au test et les autres intervenants. Cela implique la collecte de données pour déterminer si les effets bénéfiques désirés du test sont en cours de réalisation (dans le cas de l'évaluation EGRA, les bénéfices désirés portent sur la communication aux législateurs de résultats systémiques sur les compétences fondamentales en lecture pour leur permettre de cibler plus efficacement les ressources et la formation). Il faudra également rassembler des preuves sur toutes conséquences négatives involontaires de l'utilisation des scores du test (par exemple, sanction des écoles accusant une mauvaise performance avec EGRA en leur refusant des ressources qui leur sont destinées) et prendre des mesures afin de prévenir la récurrence de ces résultats négatifs.

ANNEXE F : RECOMMENDATIONS ET CONSIDERATIONS POUR DES COMPARAISONS INTER-LANGUES

F.1 Recommandations pour les caractéristiques des systèmes d'écriture

Pour faciliter les comparaisons entre les épreuves EGRA administrées dans des langues différentes, ceux qui adaptent l'EGRA doivent comprendre en profondeur les caractéristiques des systèmes d'écriture des langues en question.

Pour améliorer la qualité de ces comparaisons, il faut savoir si le système d'écriture de la langue en question est morphosyllabique, syllabique, alphasyllabique ou alphabétique (alphabet latin ou non latin).

Par la suite, on présente quelques recommandations selon le type d'orthographe.

F.1.1 Orthographes alphabétiques latines

Pour les langues avec une orthographe basée sur l'alphabet latin, il faut :

1. Savoir si l'orthographe de la langue en question est profonde (opaque) ou peu profonde (transparente).
    - Les enfants qui apprennent à lire avec les orthographes transparentes maitrisent le décodage plus vite que ceux qui apprennent à lire avec les orthographes profondes (Spence & Hanley, 2003). La profondeur de l'orthographe est aussi liée à la facilité et la rapidité du développement de la compréhension.
2. Connaitre la structure des syllabes de la langue en question.
    - Il faut accorder plus de temps pour apprendre à lire dans les langues qui permettent des syllabes complexes (c'est-à-dire, des syllabes avec des combinaisons des consonnes (C) et voyelles (V) telles que CCVCCC, comme dans le mot « starts » en anglais) que dans les langues dominées par une structure syllabique plus simple (telle que CV, comme dans le mot « sa »).

3. Savoir que la longueur moyenne des mots jouera un rôle dans la comparaison des résultats des épreuves administrées en différentes langues.
    - Les mots plus courts sont plus faciles à décoder que les mots plus longs. A comparer : les langues agglutinantes, qui relient plusieurs morphèmes l'un après l'autre dans un mot, ce qui n'est pas le cas dans les langues non-agglutinantes.
4. Savoir que la présence des signes diacritiques dans l'orthographe des langues tonales peut jouer un rôle sur la compréhension, tandis que ce n'est pas un facteur pour les langues non-tonales sans signes diacritiques.

F.1.2. Orthographes alphasyllabiques

Pour les langues avec une orthographe alphasyllabique (par exemple, le hindi, le thaï, le sinhala, le lao), il faut :

1. Savoir que le nombre de composantes voyelles ou consonnes (diacritiques phonémiques) dans chaque grappe syllabique (akshara) influencera la facilité de la lecture (Nag & Perfetti, 2014).
2. Savoir que le type de diacritique phonémique influencera la facilité de la lecture (Nag, 2014).
3. Savoir que la non-linéarité des composantes phonémiques dans une grappe de syllabes influencera la facilité de la lecture.
4. Savoir qu'en raison du grand ensemble orthographique à acquérir, il faut à peu près cinq ans d'enseignement pour atteindre un niveau de lecture fluide dans les orthographes alphasyllabaires des langues du Sud et Sud-asiatiques (par rapport à environ trois ans pour l'anglais) (Nag, 2007).

F.1.3. Orthographes alphabétiques non-latines

Pour les langues avec une orthographe alphabétique non latine (par exemple, l'arabe, le hébreu), il faut :

1. Savoir si l'orthographe de la langue en question est peu profonde (transparente) (par exemple, l'arabe écrit avec signes diacritiques représentant les voyelles) ou profond (opaque) (par exemple, l'arabe écrit sans signes diacritiques).
2. Savoir que l'arabe est un cas évident de diglossie (Ferguson, 1959).
    - La diglossie est un terme pour décrire une situation dans laquelle deux variétés d'une langue sont utilisées pour des fonctions socialement distinctes. La distinction fonctionnelle et sociolinguistique et, par conséquent, la distance linguistique (phonologique, syntaxique, morphosyntaxique, et lexicale) entre les deux variétés de l'arabe sont censé entraver, ou du moins ralentir, l'acquisition initiale de la lecture (Abu-Rabia, 2000; Ayari, 1996).

- La nature diglossique de l'arabe est étroitement liée à la profondeur orthographique et à la maitrise de la lecture.

3. Savoir que les voyelles sont perçues comme naturellement attachées aux consonnes.
4. Savoir que la recherche sur la lecture de scripts arabes et hébraïques écrits sans signes diacritiques a montré que la compréhension écrite dans ces langues n'est pas liée à l'exactitude de la lecture (Saiegh-Haddad, 2003).
5. Savoir que la forme des lettres est importante dans certaines orthographes (par exemple, l'arabe) comme les enfants ne voient pas beaucoup les lettres séparément.

F.2 Recommandations pour les évaluations du langage oral

Indépendamment du désir de faire des comparaisons inter-linguistiques, toutes les adaptations d'EGRA doivent tenir compte de plusieurs aspects du langage oral, tels que : les différences entre les dialectes ou la présence de la diglossie, la clarté des instructions, les niveaux de difficulté du contenu des épreuves de conscience phonologique, de la compréhension orale, et du vocabulaire. Pour ceux qui se concentrent sur les comparaisons inter-linguistiques, il est particulièrement important de :

1. Veiller à ce que les passages de lecture orale dans les différentes langues aient un niveau de difficulté comparable.
2. Veiller à ce que les mots de vocabulaire mesurent le même concept dans les deux langues.

F.3 Recommandations pour la connaissance de l'écrit et de l'orthographe

Le contenu des épreuves qui mesurent la connaissance de l'écrit et de l'orthographe peut être contrôlé de sorte qu'il y ait une certaine comparabilité entre les langues.

Les comparaisons inter-linguistiques permettraient de suivre le taux et la précision avec lesquels les élèves évalués dans des différentes langues reconnaissent les éléments appropriés pour leur niveau scolaire.

F.4 Recommandations pour la lecture des textes

Assurer l'adéquation technique et la comparabilité fondamentale des épreuves de la lecture administrées dans multiples langues exige plusieurs considérations:

1. Que le texte soit originel et préparé spécifiquement pour l'évaluation.
2. Que le texte traite d'un sujet approprié à l'âge dans une structure de texte

familière, pour minimiser l'influence des connaissances de base sur la compréhension.

3. Pour mieux faciliter la comparaison entre langues, que les textes dans les deux langues aient des éléments en commun et traitent des sujets familiers à tous les deux groupes linguistiques.
4. Que le passage évite l'utilisation de mots ambigus, tels que :
    - Un mot qui représente plus d'un sens toutefois étant écrit de la même manière (par exemple, les pluriels du fil et du fils sont tous deux les fils mais avec sens différent).
    - Un mot pour lequel plusieurs orthographes sont acceptables pour représenter le même sens.

F.5 Recommandations pour les apprenants d'une deuxième langue ou des apprenants multilingues

1. En comparant les résultats des épreuves dans des langues différentes, veiller à ce que les comparaisons se limitent aux apprenants de la même « classification linguistique ». Par exemple, pour une évaluation administrée en français à un groupe de francophones monolingues ou de locuteurs dont le français est leur langue maternelle, éviter de faire des comparaisons avec un groupe pour lequel le français est la langue seconde.
2. L'acquisition simultanée de la langue (ou l'apprentissage de deux ou plusieurs langues à partir de la naissance ou d'un âge précoce) est possible ; les enfants peuvent donc avoir deux langues « maternelles ».
3. Lorsque les enfants lisent dans une langue seconde, le transfert de compétences entre les langues est possible. Par exemple, la plupart des compétences de décodage se transfèrent d'une langue à une autre si les langues en question emploient des systèmes d'écriture semblables.
4. Le fait d'apprendre à lire dans une langue seconde sans instruction adéquate dans la première langue se reflètera dans les résultats d'une évaluation de la lecture. Il est probable que les enfants prendront plus de temps pour maitriser la lecture dans ces cas.

ANNEXE G : COMPARAISON DE LOGICIELS DE COLLECTE DE DONNEES

Caractéristiques	Tangerine	Magpi	SurveyToGo	DoForms	Droid Survey	ODK	Command Mobile
Prix		Licence gratuite à concurrence de 6 000 interviews/ formulaires remplis par an. Le prix varie pour un volume plus important.	0,10 à 0,15 USD par formulaire rempli, en fonction du volume. Des frais supplémentaires peuvent d'appliquer à la transmission/ stockage de fichiers photo/ vidéo/audio haute résolution.	9,95 USD par mois/99,95 USD pan par appareil pour la version professionnelle. 14,95 USD par mois/149,95 USD par an par appareil pour la version de distribution.	60 USD pour un mois, 280 USD pour six mois, 400 USD pour un an. Nombre illimité d'appareils et téléchargement de 3 000 résultats par mois	Gratuit	Version standard : 24,99 USD par mois, 64,99 USD par trimestre, 239,99 par an. Version évoluée : 69,99 USD par mois.
Compatibilité							
Android	Oui	Oui	Oui	Oui	Oui	Oui	Oui
iOS	Non	Oui	Non	Oui	Oui, avec une application distincte appelée iSurvey	Oui, prise en charge de tiers	Oui
Windows Mobile	Non	Non	Oui	Disponible prochainement	Non	Non	Oui
Symbian	Non	Oui	Non	Non	Non	Non	Non
Blackberry	Non	Non	Non	Non	Non	Non	Non
SMS	Non	Oui	Non	Non	Non	Non	Non
Développement de l'instrument							
Basé sur formulaire (aucune expertise de programmateur n'est nécessaire)	Oui	Oui	Oui	Oui	Oui	Oui	Oui
Modification d'instrument hors ligne	Oui	Oui, il est possible de modifier l'instrument hors ligne sous format Excel ou XForms, puis de le télécharger	Non	Oui	Non	Oui	Non
Compatible avec Unicode ; compatibilité avec beaucoup de langues/écritures/	Oui	Oui	Oui	Oui	Oui	Oui	Oui
Interface utilisateur pour localisation linguistique	Oui	Choix de cinq langues différentes	Oui	Anglais, espagnol, russe	Oui, choix d'environ 15 langues	Oui	N'est pas précisé
Modèles de principales tâches EGRA	Oui	Non	Non	Non	Non	Non	Non
Modèles de principales tâches EGMA	Oui	Non	Non	Non	Non	Non	Non
Possibilité de création d'instrument EGRA ?	Oui	Non, pas sans contrat de services de personnalisation	Oui, mais pas sans une certaine formation	Oui, avec toute version achetée	Oui (mais cela ne serait pas facile, ce tableau grille étant conçu pour plusieurs rangées de questions avec les mêmes choix multiples de réponses)—c.-à-d. qu'il n'est pas possible d'étiqueter les items de la grille	Oui, cela a été fait	Pas de démo disponible en ligne
Impression de formulaires ?	Oui	Non	Oui	Non	Non	Oui	Non

Caractéristiques	Tangerine	Magpi	SurveyToGo	DoForms	Droid Survey	ODK	Command Mobile
Collecte de données							
Collecte de données hors ligne	Oui	Oui	Oui	Oui	Oui	Oui	Oui
Données textuelles/numériques	Oui	Oui	Oui	Oui	Oui	Oui	Oui
Données d'étude chronométrée	Oui	Non	Oui	Oui	Non	Oui	Non
Tableaux grilles	Oui	Non	Oui	Oui	Oui	Oui	Oui
Réponse à choix unique/multiples	Oui	Oui	Oui	Oui	Oui	Oui	Oui
Emplacement GPS	Oui	Oui	Oui	Oui	Oui	Oui	Oui
Captures d'écran	Oui	Oui	Oui	Oui	Oui	Oui	Oui
Caméra	Non	Non	Oui	Oui	Oui	Oui	Oui
Vidéo	Non	Non	Oui	Oui	Non	Oui	Oui
Audio	Non	Non	Oui	Oui	Non	Oui	Oui
Code à barres	Non	Non	Oui	Oui	Oui	Oui	Oui
Signature	Non	Non	Oui	Oui	Oui	Oui	Oui
Fonctionnalités logiques							
Logique scellée	Oui	Oui	Oui	Oui	Oui	Oui	Oui
Validation personnalisée	Oui	Oui	Oui	Oui	Non	Oui	Oui
Affichage de forme conditionnelle	Oui	Oui	Oui	Oui	Non	Oui	Oui
Boucle	Non	Oui	Oui	Oui	Non	Oui	Non
Embranchement de questions	Oui	Oui	Oui	Oui	Non	Oui	Non
Téléchargement de données							
Wifi	Oui	Oui	Oui	Oui	Oui	Oui	Oui
Cellulaire	Oui	Oui	Oui	Oui	Oui	Oui	Oui
Câble	Non	Oui	Oui	Oui	Oui	Oui	Oui
Sauvegarde d'appareil à appareil	Oui	Oui, par carte mémoire	Non	Non	Non	Oui, avec prise en charge de tiers	Non
Stockage de données							
Excel	Oui	Oui	Oui	Oui	Oui	Oui	Oui
RSS	Non	Non	Non	Non	Non	Non	Oui
SPSS	Oui	Non	Oui	Non	Oui	Non	Non
MS Word	Non	Oui	Oui	Oui	Non	Non	Non
MS Access	Non	Oui	Oui	Oui	Non	Non	Non
XML	Non	Oui	Oui	Oui	Non	Oui	Non
HTML	Non	Non	Non	Oui	Non	Non	Non
PDF	Non	Oui	Non	Oui	Non	Non	Non
Google Docs	Non	Non	Non	Oui	Non	Oui	Non
Open Office	Non	Non	Non	Oui	Non	Non	Non
Stockage de données							
Volume de stockage inclus		Stockage en nuage illimité	20 Mo pour pièces jointes ; un volume supplémentaire peut être acheté. Stockage en nuage illimité	Stockage en nuage illimité	Stockage en nuage illimité	Illimité	250 Go inclus
Chiffrement pendant le transfert ?	Non	Chiffré à une Norme robuste de 256 bits	Oui, avec supplément	Oui—chiffrement SSL	Oui—chiffrement SSL	Oui	Oui
Source ouverte ?	Oui	Non, mais API disponible pour les clients Enterprise	Non	API disponible, mais pas source ouverte	API des résultats disponible sur demande	Oui	Non
Mode kiosque ?	Non	Oui	Oui	Non	Oui	Oui, avec prise en charge de tiers	Non

ANNEXE H: COMPARAISON DES INSTRUCTIONS EGRA POUR LES VERSIONS PAPIER ET TABLETTE

Instructions pour l'examinateur: PAPIER	Instructions pour l'examinateur: TABLETTE	Instructions pour l'élève (identique pour les versions papier/tablette)
Instructions Générales		
Il est important d'établir une relation détendue et enjouée avec les élèves qui vont être évalués, grâce à de simples conversations initiales (voir exemples ci-dessous). L'élève doit presque percevoir l'évaluation qui suit comme étant un jeu. Identifiez la langue que l'élève préfère utiliser pour communiquer. Lisez à haute voix clairement et lentement les sections encadrées UNIQUEMENT.	Il est important d'établir une relation détendue et enjouée avec les élèves qui vont être évalués, grâce à de simples conversations initiales (voir exemple ci-après). L'élève doit presque percevoir l'évaluation qui suit comme étant un jeu. Identifiez la langue que l'élève préfère utiliser pour communiquer. Lisez à haute voix clairement et lentement les sections encadrées UNIQUEMENT.	Bonjour! Je m'appelle____et j'habite____. Je souhaite te parler un peu de moi. J'ai des enfants qui, comme toi, aiment la lecture, le sport, et la musique. • **Et toi, comment t'appelles-tu? Qu'est-ce que tu aimes?** [Attendez la réponse de l'élève. Si l'élève semble à l'aise, passez directement au consentement verbal. S'il hésite ou a l'air peu à l'aise, posez la deuxième question avant de passer au consentement verbal]. • **Qu'est-ce que tu aimes faire lorsque tu n'es pas à l'école?**
Consentement verbal		
Si le consentement verbal n'est pas obtenu, remercier l'élève et passer au prochain élève. Réutilisez ce même formulaire.	Si le consentement verbal n'est pas obtenu, remercier l'élève et sélectionnez "non "sur l'écran. Sélectionnez ensuite "enregistrer "et "démarrer un nouveau test".	• **Laisse-moi t'expliquer pourquoi je suis là aujourd'hui. Le Ministère de l'Éducation nous a demandé d'étudier comment les élèves apprennent à lire. Tu as été sélectionné(e) pour participer à cette étude.** • **Ta participation est très importante, mais tu n'es pas obligé de participer si tel n'est pas ton désir.** • **Nous allons faire des jeux à l'oral, en lecture, et en écriture.** • **J'utiliserai ce(tte) chronomètre/tablette/gadget pour savoir à quelle vitesse tu lis.** • **Mais ce n'est pas un examen, et ce que tu fais avec moi ne changera pas tes notes à l'école.**

Instructions pour l'examinateur: PAPIER	Instructions pour l'examinateur: TABLETTE	Instructions pour l'élève (identique pour les versions papier/tablette)
		• **Je vais aussi te poser quelques questions sur ta famille et la langue que tu parles à la maison.** • **Je n'écris pas ton nom sur cette fiche, alors personne ne saura que ces réponses sont les tiennes.** • **Encore une fois, tu n'es pas obligé de participer si tu ne le veux pas. Si tu arrives à une question à laquelle tu préfères ne pas** • **répondre, ce n'est pas grave, on peut passer.** **As tu des questions? Peut-on commencer?**
Connaissance du son des graphèmes (lettres et groupes de lettres)		
Montrez à l'élève la feuille de lettres et groupes de lettres (graphèmes) dans le Cahier de Stimulus pendant que vous lui lisez les instructions suivantes: **[INSEREZ LES INSTRUCTIONS DE LA COLONNE DE DROITE ET LISEZ LES A L'ELEVE]** Faites démarrer le chronomètre quand l'élève lit la première lettre. • Suivez la lecture de l'élève sur votre page à l'aide du crayon. Barrez (/) les graphèmes incorrects ainsi que ceux que l'élève saute ou ne lit pas.	Montrez à l'élève la feuille de lettres et groupes de lettres (graphèmes) dans le Cahier de Stimulus pendant que vous lui lisez les instructions suivantes: Faites démarrer le chronomètre quand l'élève lit la première lettre. Suivez la lecture de l'élève sur votre écran et relevez les réponses incorrectes en touchant du doigt les graphèmes incorrects sur l'écran. Les graphèmes incorrects apparaitront alors en bleu. Si l'élève s'autocorrige (donne une réponse incorrecte puis la corrige), touchez à nouveau l'item, ce qui fera apparaitre le graphème en gris.	**Voici une page pleine de lettres et de groupes de lettres. Donne-moi le SON de ces lettres, pas le nom.** **Par exemple, cette lettre** [Montrez le "O" dans la ligne des exemples] **se lit "O" comme dans le mot "pot".** **Pratiquons maintenant: Lis-moi ce groupe de lettres** [Montrez le "ou" sur la rangée des exemples]: [Si l'élève répond correctement, dites:] **Très bien, ce groupe de lettres se lit "ou" comme dans le mot "cour".** [Si l'élève ne répond pas correctement, dites:] **Non, ce groupe de lettres se lit "ou" comme dans le mot "cour".**

Instructions pour l'examinateur: PAPIER	Instructions pour l'examinateur: TABLETTE	Instructions pour l'élève (identique pour les versions papier/tablette)
• Si l'élève donne une réponse incorrecte puis se corrige (autocorrection), entourez l'item si vous l'avez déjà barré (ø) et continuez. Comptez cette réponse comme étant correcte. • Si l'élève saute une ligne entière, tracez un trait au travers de la ligne pour indiquer que toutes les réponses sont incorrectes. • Ne dites rien sauf si l'élève ne répond pas et reste bloqué sur un graphème pendant plus de 3 secondes. Dans ce cas, dites-lui, "Continue", en lui montrant le prochain graphème. Notez le graphème sur lequel l'élève était bloqué comme incorrect. • Si l'élève prononce le NOM de la lettre au lieu du son, donnez la réponse correcte en prononçant le son, puis dites "Dis-moi le SON de la lettre". Attention, cette indication ne peut-être donnée qu'une seule fois pendant l'épreuve. Quand le chronomètre sonne et atteint 0, demandez à l'élève de s'arrêter, mettez un crochet ( ]) juste après le dernier graphème que l'élève a lu. Règle d'auto-stop: Si l'élève ne donne aucune réponse correcte parmi les dix premiers graphèmes (la première ligne), demandez-lui gentiment de s'arrêter, et cochez la case "auto-stop". [EN BAS DE LA PAGE, INCLURE LES INDICATIONS SUIVANTES] Nombre exact de secondes restantes indiquées sur le chronomètre ☐ Cochez ici si l'épreuve a été arrêté par manque de réponses correctes sur la première ligne (auto-stop): ☐	Ne dites rien sauf si l'élève ne répond pas et reste bloqué sur un graphème pendant plus de 3 secondes. Dans ce cas, dites-lui, "Continue", en lui montrant le prochain graphème. Notez le graphème sur lequel l'élève était bloqué comme incorrect. Si l'élève prononce le nom de la lettre au lieu du son, donnez la réponse correcte en prononçant le son, puis dites "Lis-moi le SON de la lettre". Attention, cette indication ne peut-être donnée qu'une seule fois pendant l'épreuve. Lorsque que la minute s'est écoulée et que le chronomètre s'arrête, l'écran s'affichera en rouge par flashs successifs. Cela signalera la fin du test. Demandez à l'élève de s'arrêter, indiquez le dernier graphème lu en touchant l'écran afin qu'un crochet rouge apparaisse. Appuyez ensuite sur la touche "suivant". Si l'élève atteint la fin du test avant que l'écran ne s'affiche en rouge, arrêtez le chronomètre vous même au moment où l'élève lit le dernier graphème du test. Puis touchez la dernière lettre sur l'écran afin qu'un crochet rouge apparaisse. Appuyez ensuite sur la touche "suivant". Règle d'auto-stop: Si l'élève ne donne aucune réponse correcte parmi les 10 premiers items (la première ligne), l'écran s'affichera en rouge par flashs successifs et le chronomètre s'arrêtera automatiquement. Demandez gentiment à l'éleve de s'arrêter, interrompez l'épreuve et passez à la suivante.	**Essayons un autre exemple. Lis-moi cette lettre:** [Montrez le "t" sur la rangée des exemples]: [Si l'élève répond correctement, dites:] **Très bien, cette lettre se lit "t"comme dans le mot "table".** [Si l'élève ne répond pas correctement, dites:] **Non, cette lettre se lit "t" comme dans le mot "table".** **Essayons encore un autre exemple. Lis-moi ce groupe de lettres:** [Montrez le "ch" sur la rangée des exemples]: [Si l'élève répond correctement, dites:] **Très bien, ce groupe de lettres se lit "ch" comme dans le mot "chat".** [Si l'élève ne répond pas correctement, dites:] **Non, ce groupe de lettres se lit "ch" comme dans le mot "chat".** **Lorsque je dis "Commence", commence à lire ici** [montrez lui le premier graphème] **et lis de gauche à droite** [faites glisser votre doigt vers la droite], **ligne par ligne en parlant fort pour que je puisse t'entendre. Montre chaque lettre du doigt quand tu la lis. Essaye de lire rapidement et correctement. Si tu n'arrives pas à lire une des lettres, continue et lis celle qui suit. Mets ton doigt sur la première lettre. Tu es prêt(e)? Commence.** [INSEREZ ICI LA LISTE D'ITEMS] **Merci bien! On peut passer à la prochaine activité!**

Instructions pour l'examinateur: PAPIER	Instructions pour l'examinateur: TABLETTE	Instructions pour l'élève (identique pour les versions papier/tablette)
Conscience Phonémique – Identification du son initial		
Cette épreuve n'est pas chronométrée. Retirez le Cahier de Stimulus de la vue de l'élève. Lisez les instructions à l'élève et donnez lui les exemples. **[INSEREZ LES INSTRUCTIONS DE LA COLONNE DE DROITE]** Lisez les instructions et prononcez les items deux fois. Prononcez chaque item lentement. Seul le son prononcé isolément est correct. Cochez la case correspondant à la réponse de l'élève. En cas de non-réponse, après 3 secondes cochez la case "Pas de réponse" et passez au prochain item. Règle d'auto-stop: Si l'élève ne donne aucune réponse correcte parmi les 5 premiers mots, demandez-lui gentiment de s'arrêter, et cochez la case "auto-stop". Passez à la prochaine épreuve.	L'épreuve n'est pas chronométrée. Retirez le Cahier de Stimulus de la vue de l'élève. Lisez les instructions et prononcez les items deux fois, lentement. Suivez les réponses de l'élève sur votre écran et relevez les réponses correctes et incorrectes. La réponse sélectionnée sur l'écran apparait en jaune. En cas de non-réponse, après 3 secondes cochez la case "Pas de réponse" et passez au prochain item. Règle d'auto-stop: Si l'élève ne donne aucune réponse correcte parmi les 5 premiers mots, l'écran s'affichera en rouge par flashs successifs.	**Cette épreuve est une épreuve orale. Je vais te dire un mot deux fois, puis je veux que tu me dises le tout premier son du mot que tu entends, d'accord? Par exemple: Le mot " soupe" commence àvec le son "sssss", n'est-ce pas? Je dirai chaque mot deux fois et tu me diras le tout premier son de chaque mot.** **Essayons encore quelques exemples:** **Quel est le tout premier son dans le mot "chic"? "chic"?** [Si l'élève répond correctement, dites-lui] **Très bien! Le premier son dans le mot "chic", c'est "ch"** [Si l'élève ne répond pas correctement, dites-lui] **Le premier son dans le mot "chic", c'est "ch"** **Quel est le tout premier son dans le mot "poule"? "poule"?** [Si l'élève répond correctement, dites-lui] **"Très bien! Le premier son dans le mot "poule", c'est "p**"

Instructions pour l'examinateur: PAPIER	Instructions pour l'examinateur: TABLETTE	Instructions pour l'élève (identique pour les versions papier/tablette)
[EN BAS DE LA PAGE, INCLURE LES INDICATIONS SUIVANTES] Cochez ici si l'épreuve a été arrêté par manque de réponses correctes parmi les 5 premiers items (auto-stop): ☐		[Si l'élève ne répond pas correctement, dites-lui] **"Le premier son dans le mot "poule", c'est "p".** **Tu es prêt(e)? Commence.** **Quel est le tout premier son dans le mot " "?** **" "?** [INSEREZ ICI LA LISTE D'ITEMS] **Merci bien! On peut passer à la prochaine activité!**
Lecture de mots familiers		
Montrez à l'élève la feuille de mots dans le Cahier de Stimulus pendant que vous lui lisez les instructions suivantes: [**INSEREZ LES INSTRUCTIONS DE LA COLONNE DE DROITE]** Faites démarrer le chronomètre quand l'élève lit le premier mot. • Suivez la lecture de l'élève sur votre page à l'aide du crayon. Barrez (/) les mots incorrects ainsi que ceux que l'élève saute ou ne lit pas. • Si l'élève donne une réponse incorrecte puis se corrige (auto-correction), entourez le mot si vous l'avez déjà barré (ø) et continuez. Comptez cette réponse comme étant correcte. • Si l'élève saute une ligne entière, tracez un trait au travers de la ligne. • Ne dites rien sauf si l'élève ne répond pas et reste bloqué sur un mot pendant plus de 3 secondes. Dans ce cas, dites-lui, "Continue", en lui montrant le prochain mot. Notez le mot sur lequel l'élève était bloqué comme incorrect.	Montrez à l'élève la feuille de mots dans le Cahier de Stimulus pendant que vous lui lisez les instructions. Faites démarrer le chronomètre quand l'élève lit le premier mot. Suivez la lecture de l'élève sur votre écran et relevez les réponses incorrectes en touchant du doigt les mots incorrects sur l'écran. Les mots incorrects apparaitront alors en bleu. Si l'élève s'autocorrige (donne une réponse incorrecte puis la corrige), touchez à nouveau l'item, ce qui fera apparaitre le mot en gris. Ne dites rien sauf si l'élève ne répond pas et reste bloqué sur un mot pendant plus de 3 secondes. Dans ce cas, dites-lui, "Continue", en lui montrant le prochain mot. Notez le mot sur lequel l'élève était bloqué comme incorrect. Lorsque que la minute s'est écoulée et que le chronomètre s'arrête, l'écran s'affichera en rouge par flashs successifs. Cela signalera la fin du test. Demandez à l'élève de s'arrêter, indiquez le dernier mot lu en touchant l'écran afin qu'un crochet rouge apparaisse. Appuyez ensuite sur la touche "suivant".	**Voici une page avec des mots en FRANCAIS. Essaye de lire autant de mots que tu peux. Il ne faut pas dire les lettres mais lire le mot. Par exemple, ce premier mot** [montrez le mot "ta"] **se lit "ta".** **Essayons. Peux-tu lire ce mot?** [montrez le mot "par" avec le doigt.] [Si l'élève lit correctement dites:] **Très bien, ce mot se lit "par".** [Si l'élève ne lit pas correctement, ou après 3 secondes de non-réponse, dites:] **Ce mot se lit "par"** **Essayons. Peux-tu lire ce mot?** [montrez le mot "lune" avec le doigt.] [Si l'élève lit correctement dites:] **Très bien, ce mot se lit "lune".** [Si l'élève ne lit pas correctement, ou après 3 secondes de non-réponse, dites:] **Ce mot se lit "lune"** **Lorsque je dis "Commence", commence à lire ici** [montrez lui le premier mot] **et lis de gauche à droite** [faites glisser votre doigt vers la droite], l**igne par ligne en parlant fort pour que je puisse t'entendre.**

Instructions pour l'examinateur: PAPIER	Instructions pour l'examinateur: TABLETTE	Instructions pour l'élève (identique pour les versions papier/tablette)
Quand le chronomètre sonne et atteint 0, demandez à l'élève de s'arrêter, mettez un crochet ( ]) juste après le dernier mot que l'élève a lu. Règle d'auto-stop: Si l'élève ne donne aucune réponse correcte sur la première ligne (5 premiers mots), demandez-lui gentiment de s'arrêter, et cochez la case "auto-stop". [**INSEREZ LA MENTION CI-DESSOUS EN BAS DE PAGE]** Nombre exact de secondes restantes indiquées sur le chronomètre ☐ Cochez ici si l'épreuve a été arrêtée par manque de réponses correctes sur la première ligne (auto-stop): ☐	Si l'élève atteint la fin du test avant que l'écran ne s'affiche en rouge, arrêtez le chronomètre vous même au moment où l'élève lit le dernier mot du test. Puis touchez le dernier mot sur l'écran afin qu'un crochet rouge apparaisse. Appuyez ensuite sur la touche "suivant". Règle d'auto-stop: Si l'élève ne donne aucune réponse correcte parmi les 5 premiers mots (la première ligne), l'écran s'affichera en rouge par flashs successifs et le chronomètre s'arrêtera automatiquement. Demandez gentiment à l'éleve de s'arrêter, interrompez l'épreuve et passez à la suivante.	**Montre chaque mot du doigt quand tu le lis. Essaye de lire rapidement et correctement. Si tu n'arrives pas à lire un des mots, continue et lis celui qui suit. Mets ton doigt sur le premier mot. Tu es prêt(e)? Commence.** [INSEREZ ICI LA LISTE D'ITEMS] **Merci bien! On peut passer à la prochaine activité!**
Lecture de mots inventés		
Montrez à l'élève la feuille de mots dans le Cahier de Stimulus pendant que vous lui lisez les instructions suivantes: **[INSEREZ LES INSTRUCTIONS DE LA COLONNE DE DROITE]** Faites démarrer le chronomètre quand l'élève lit le premier mot. • Suivez la lecture de l'élève sur votre page à l'aide du crayon. Barrez (/) les mots incorrects ainsi que ceux que l'élève saute ou ne lit pas. • Si l'élève donne une réponse incorrecte puis se corrige (autocorrection), entourez l'item si vous l'avez déjà barré (ø) et continuez. Comptez cette réponse comme étant correcte. • Si l'élève saute une ligne entière, tracez un trait au travers de la ligne.	Montrez à l'élève la feuille de mots dans le Cahier de Stimulus pendant que vous lui lisez les instructions. Faites démarrer le chronomètre quand l'élève lit le premier mot. Suivez la lecture de l'élève sur votre écran et relevez les réponses incorrectes en touchant du doigt les mots incorrects sur l'écran. Les mots incorrects apparaitront alors en bleu. Si l'élève s'autocorrige (donne une réponse incorrecte puis la corrige), touchez à nouveau l'item, ce qui fera apparaitre le mot en gris. Ne dites rien sauf si l'élève ne répond pas et reste bloqué sur un mot pendant plus de 3 secondes. Dans ce cas, dites-lui, "Continue", en lui montrant le prochain mot. Notez le mot sur lequel l'élève était bloqué comme incorrect.	**Voici une page avec des mots inventés qui ressemblent à des mots FRANCAIS. Essaye de lire autant de mots que tu peux. Il ne faut pas dire les lettres mais lire le mot. Par exemple, ce premier mot** [montrez le mot "bi"] **se lit "bi".** **Essayons. Peux-tu lire ce mot?** [montrez le mot "tok" avec le doigt.] [[Si l'élève lit correctement dites:] **Très bien, ce mot se lit "tok".** [[Si l'élève ne lit pas correctement, ou après 3 secondes de non-réponse, dites:] **Ce mot se lit "tok".** **Essayons. Peux-tu lire ce mot?** [montrez le mot "sar" avec le doigt.] [Si l'élève lit correctement dites:] **Très bien, ce mot se lit "sar".**

Instructions pour l'examinateur: PAPIER	Instructions pour l'examinateur: TABLETTE	Instructions pour l'élève (identique pour les versions papier/tablette)
• Ne dites rien sauf si l'élève ne répond pas et reste bloqué sur un mot pendant plus de 3 secondes. Dans ce cas, dites-lui, "Continue", en lui montrant le prochain mot. Notez le mot sur lequel l'élève était bloqué comme incorrect. Quand le chronomètre sonne et atteint 0, demandez à l'élève de s'arrêter, mettez un crochet ( ]) juste après le dernier mot que l'élève a lu. Règle d'auto-stop: Si l'élève ne donne aucune réponse correcte sur la première ligne (5 premiers mots), demandez-lui gentiment de s'arrêter, et cochez la case "auto-stop". **[INSEREZ LA MENTION CI-DESSOUS EN BAS DE PAGE]** Nombre exact de secondes restantes indiquées sur le chronomètre ☐ Cochez ici si l'épreuve a été arrêtée par manque de réponses correctes sur la première ligne (auto-stop): ☐	Lorsque que la minute s'est écoulée et que le chronomètre s'arrête, l'écran s'affichera en rouge par flashs successifs. Cela signalera la fin du test. Demandez à l'élève de s'arrêter, indiquez le dernier mot lu en touchant l'écran afin qu'un crochet rouge apparaisse. Appuyez ensuite sur la touche "suivant". Si l'élève atteint la fin du test avant que l'écran ne s'affiche en rouge, arrêtez le chronomètre vous même au moment où l'élève lit le dernier mot du test. Puis touchez le dernier mot sur l'écran afin qu'un crochet rouge apparaisse. Appuyez ensuite sur la touche "suivant". Règle d'auto-stop: Si l'élève ne donne aucune réponse correcte parmi les 5 premiers mots (la première ligne), l'écran s'affichera en rouge par flashs successifs et le chronomètre s'arrêtera automatiquement. Demandez gentiment à l'éleve de s'arrêter, interrompez l'épreuve et passez à la suivante.	[Si l'élève ne lit pas correctement, ou après 3 secondes de non-réponse, dites:] **Ce mot se lit "sar".** **Lorsque je dis "Commence", commence à lire ici** [montrez lui le premier mot] **et lis de gauche à droite** [faites glisser votre doigt vers la droite], **ligne par ligne en parlant fort pour que je puisse t'entendre.** **Montre chaque mot du doigt quand tu le lis. Essaye de lire rapidement et correctement. Si tu n'arrives pas à lire un des mots, continue et lis celui qui suit. Mets ton doigt sur le premier mot. Tu es prêt(e)? Commence.** [INSEREZ ICI LA LISTE D'ITEMS] **Merci bien! On peut passer à la prochaine activité!**
Lecture du texte (petite histoire)		
Montrez à l'élève la petite histoire dans le Cahier de Stimulus. Lisez à l'élève les instructions suivantes: **[INSEREZ LES INSTRUCTIONS DE LA COLONNE DE DROITE]** Faites démarrer le chronomètre quand l'élève lit le premier mot.	Montrez à l'élève la petite histoire dans le Cahier de Stimulus. Lisez à l'élève les instructions. Faites démarrer le chronomètre quand l'élève lit le premier mot. Suivez la lecture de l'élève sur votre écran et relevez les réponses incorrectes en touchant du doigt les mots incorrects sur l'écran.	**Voici une petite histoire. Lis la à haute voix en essayant de lire rapidement et correctement; après, je vais te poser quelques questions sur l'histoire. Lorsque je dis "Commence", tu commenceras à lire. Si tu vois un mot que tu ne sais pas lire, essaye le prochain. Mets ton doigt sur le premier mot. Tu es prêt(e)? Commence.** [INSEREZ ICI LE TEXTE]

Instructions pour l'examinateur: PAPIER	Instructions pour l'examinateur: TABLETTE	Instructions pour l'élève (identique pour les versions papier/tablette)
• Suivez la lecture de l'élève sur votre page à l'aide du crayon. Barrez (/) les mots incorrects ainsi que ceux que l'élève saute ou ne lit pas. • Si l'élève donne une réponse incorrecte puis se corrige (autocorrection), entourez l'item si vous l'avez déjà barré (ø) et continuez. Comptez cette réponse comme étant correcte. • Si l'élève saute une ligne entière, tracez un trait au travers de la ligne. • Ne dites rien sauf si l'élève ne répond pas et reste bloqué sur un mot pendant plus de 3 secondes. Dans ce cas, dites-lui, "Continue", en lui montrant le prochain mot. Notez le mot sur lequel l'élève était bloqué comme incorrect. Quand le chronomètre sonne et atteint 0, demandez à l'élève de s'arrêter, mettez un crochet ( ]) juste après le dernier mot que l'élève a lu. Quand l'élève a fini de lire, RETIREZ le passage de sa vue. Règle d'auto-stop: Si l'élève ne donne aucune réponse correcte sur la première ligne, demandez-lui gentiment de s'arrêter, et cochez la case "auto-stop". [**INSEREZ LA MENTION CI-DESSOUS EN BAS DE PAGE]** Nombre exact de secondes restantes indiquées sur le chronomètre ☐ Cochez ici si l'épreuve a été arrêtée par manque de réponses correctes sur la première ligne (auto-stop): ☐	Les mots incorrects apparaitront alors en bleu. Si l'élève s'autocorrige (donne une réponse incorrecte puis la corrige), touchez à nouveau l'item, ce qui fera apparaitre le mot en gris. Ne dites rien sauf si l'élève ne répond pas et reste bloqué sur un mot pendant plus de 3 secondes. Dans ce cas, dites-lui, "Continue", en lui montrant le prochain mot. Notez le mot sur lequel l'élève était bloqué comme incorrect. Lorsque que la minute s'est écoulée et que le chronomètre s'arrête, l'écran s'affichera en rouge par flashs successifs. Cela signalera la fin du test. Demandez à l'élève de s'arrêter, indiquez le dernier mot lu en touchant l'écran afin qu'un crochet rouge apparaisse. Appuyez ensuite sur la touche "suivant". Si l'élève atteint la fin du test avant que l'écran ne s'affiche en rouge, arrêtez le chronomètre vous même au moment où l'élève lit le dernier mot du test. Puis touchez le dernier mot sur l'écran afin qu'un crochet rouge apparaisse. Appuyez ensuite sur la touche "suivant". Quand l'élève a fini de lire, RETIREZ le passage de sa vue Règle d'auto-stop: Si l'élève ne réussit pas à lire correctement les mots nécessaires à la passation de la première question de compréhension, l'écran s'affichera en rouge par flashs successifs et le chronomètre s'arrêtera automatiquement. Demandez gentiment à l'éleve de s'arrêter, interrompez l'épreuve et passez à la suivante.	

Instructions pour l'examinateur: PAPIER	Instructions pour l'examinateur: TABLETTE	Instructions pour l'élève (identique pour les versions papier/tablette)
Compréhension écrite		
Avant de poser les questions, pensez à RETIRER la petite histoire de la vue de l'élève. Posez les questions qui correspondent aux lignes du texte jusqu'à la ligne dans laquelle se trouve le crochet ( ]), c'est-à-dire, jusqu'à l'endroit où l'élève a cessé de lire. Notez les réponses de l'élève. Seules les réponses identiques ou similaires aux réponses fournies dans le protocole (listées à côté de chaque question) peuvent être considérées comme correctes. Si l'élève ne donne aucune réponse après 10 secondes, passez à la question suivante et sélectionnez "Pas de réponse". Si l'élève dit "je ne sais pas", cochez la réponse "incorrecte". Ne répétez pas les questions. Lisez à l'élève les instructions suivantes: [**INSEREZ LES INSTRUCTIONS DE LA COLONNE DE DROITE]**	Avant de poser les questions, pensez à RETIRER la petite histoire de la vue de l'élève. Posez toutes les questions qui sont présentées à l'écran. Elles correspondent automatiquement à la portion du texte que l'élève a lu. Sélectionnez les réponses de l'élève comme "Correcte" ou "Incorrecte". La réponse sélectionnée apparaitra en jaune. Si l'élève ne répond pas dans les 10 secondes, sélectionnez "Pas de réponse" et passez à la question suivante. Ne répétez pas les questions. Seules les réponses identiques ou similaires aux réponses fournies dans ce protocole (listées à côté de chaque question) peuvent être considérées comme correctes. Si l'élève dit "je ne sais pas", cochez la réponse "incorrecte".	**Maintenant, je vais te poser quelques questions sur l'histoire. Essaye de répondre aux questions du mieux possible. Tu peux répondre dans la langue que tu préfères.** [INSEREZ ICI LES QUESTIONS]
Compréhension Orale		
L'épreuve n'est pas chronométrée et aucun mot écrit n'est montré à l'élève. Retirez le Cahier de Stimulus de la vue de l'élève. Lisez les instructions à l'élève **[INSEREZ LES INSTRUCTIONS DE LA COLONNE DE DROITE]** Vous allez lire le passage à haute voix UNE SEULE FOIS, lentement (à peu près un mot par seconde). Posez toutes les questions à l'élève. Ne laissez pas l'élève voir l'histoire ou les questions.	L'épreuve n'est pas chronométrée et aucun mot écrit n'est montré à l'élève. Retirez le Cahier de Stimulus de la vue de l'élève. Lisez les instructions à l'élève. Vous allez lire le passage à haute voix UNE SEULE FOIS, lentement (à peu près un mot par seconde). Posez toutes les questions à l'élève. Ne laissez pas l'élève voir l'histoire ou les questions. Sélectionnez les réponses de l'élève comme "Correcte" ou "Incorrecte". La réponse sélectionnée apparaitra en jaune.	**Maintenant, je vais te lire une histoire UNE fois. Après cela, je vais te poser quelques questions sur cette histoire. Tu vas bien écouter, et ensuite tu répondras aux questions le mieux que tu peux. Tu peux répondre dans la langue que tu préfères. D'accord? Commençons! Ecoute bien:** [INSEREZ ICI L'HISTOIRE ET LES QUESTIONS] **Merci bien! On peut passer à la prochaine activité!**

Instructions pour l'examinateur: PAPIER	Instructions pour l'examinateur: TABLETTE	Instructions pour l'élève (identique pour les versions papier/tablette)
Notez les réponses de l'élève. Seules les réponses identiques ou similaires aux réponses fournies dans le protocole (listées à côté de chaque question) peuvent être considérées comme correctes. Si l'élève ne donne aucune réponse après 10 secondes, passez à la question suivante et sélectionnez "Pas de réponse". Si l'élève dit "je ne sais pas", notez la réponse comme incorrecte. Ne répétez pas les questions.	Si l'élève ne répond pas dans les 10 secondes, sélectionnez "Pas de réponse" et passez à la question suivante. Ne posez chaque question qu'une seule fois. Seules les réponses identiques ou similaires aux réponses fournies dans le protocole (listées à côté de chaque question) peuvent être considérées comme correctes. Si l'élève dit "je ne sais pas", notez la réponse comme incorrecte.	
Dictée		
Prenez la dernière page des tests pour les élèves où vous trouverez une page de lignes pour que l'élève puisse écrire. Placez cette page devant l'élève qui y écrira la phrase dictée. Prenez le Cahier de Stimulus à la dernière page. Vous y trouverez les mêmes instructions que celles présentées ci-dessous. Donnez un crayon à l'élève et dites lui: [**INSEREZ LES INSTRUCTIONS DE LA COLONNE DE DROITE]** Lisez la phrase suivante UNE SEULE FOIS (à peu près un mot par seconde). [**INSEREZ LA PHRASE A DICTER]** Lisez la phrase une deuxième fois. Faites une pause de 10 secondes entre chaque groupe de mots. **[INSEREZ LA PHRASE A DICTER EN ESPACANT LES MOTS POUR MARQUER LES PAUSES]** Après 15 secondes, lisez toute la phrase à nouveau. **[INSEREZ LA PHRASE A DICTER]** Attendez 15 secondes que l'élève ait finit d'écrire, puis interrompez l'épreuve.	Prenez la dernière page des tests pour les élèves où vous trouverez une page de lignes pour que l'élève puisse écrire. Placez cette page devant l'élève qui y écrira la phrase dictée. Prenez le Cahier de Stimulus à la dernière page. Vous y trouverez les mêmes instructions que celles présentées ci-dessous. Donnez un crayon à l'élève. Lisez la phrase suivante UNE SEULE FOIS (à peu près un mot par seconde). Puis lisez la phrase une deuxième fois et faites une pause de 10 secondes entre chaque groupe de mots. Après 15 secondes, lisez toute la phrase à nouveau. Attendez 15 secondes que l'élève ait finit d'écrire, puis interrompez l'épreuve.	**Je vais te dicter une petite phrase. Ecoute bien. Je vais lire la phrase une fois. Ensuite, je vais lire des petites parties pour que tu puisses écrire ce que tu entends. Je vais ensuite lire une dernière fois pour que tu puisses vérifier. Tu es prêt(e)?** **Merci bien! On peut passer à la prochaine activité!**

ANNEXE I : EXEMPLE DE PROGRAMME DE FORMATION DES EVALUATEURS

Formation des collecteurs de données EGRA

Jour et heure	Jour 1	Jour 2	Jour 3	Jour 4	Jour 5	Jour 6
Objectifs quotidiens:	• Comprendre l'objectif de l'évaluation EGRA • Savoir appliquer les règles d'administration et d'attribution de score sur papier	• Comprendre les fonctions et l'administration de la tablette • Savoir télécharger les données	• Améliorer les compétences d'administration de test • Se familiariser avec l'administration du questionnaire	• Perfectionner les compétences d'administration du test EGRA et l'exactitude de l'attribution de scores	• Perfectionner les compétences d'administration du test EGRA et l'exactitude de l'attribution de scores	• Formation des superviseurs • Préparation des équipes
8 h 30-9 h 00	• Accueil/présentations	• Passer en revue le Jour 1	**Visite d'école 1 : pratique du test EGRA**	**Visite d'école 2 : EGRA + questionnaires**	**Visite d'école 3 : EGRA + questionnaires**	• Formation des superviseurs • Préparation des équipes à la collecte de données
9 h 00-10 h 30	• Aperçu de l'objectif EGRA et du contenu de l'instrument • Objectif d'EGRA sans ce contexte	• Aperçu des fonctions de base de la tablette				
10 h 30-11 h 00	***Pause***	***Pause***				
11 h 00-13 h 00	• Aperçu de l'instrument • Démonstration et pratique de tâches	• Pratique de l'évaluation EGRA sur tablettes (petits groupes)				
13 h 00-14 h 00	***Déjeuner***					
14 h 00-15 h 30	• Poursuite de la démonstration et de la pratique de tâches • Questionnaire des élèves	• Problèmes de fonctionnalité de la tablette • Téléchargement de données	• Compte-rendu de la visite d'école • *Instruments d'enquête additionnels si elle est administrée*	• Compte-rendu de la visite d'école • Discussion sur les résultats de la mesure de précision des évaluateurs 2	• Compte-rendu de la visite d'école • Discussion sur les résultats de la mesure de précision	
				Pratique du test EGRA sur tablettes en groupes de 2 (principales tâches/ problèmes)	des évaluateurs 2 ï **Logistique de collecte de données**	
15 h 30-15 h 45	***Pause***					
15 h 45-17 h 30	• Poursuite de la pratique et de la correction en petits groupes	• **Procédures d'échantillonnage EGRA** • Logistique des visites d'écoles	• Pratique du test EGRA sur tablettes en groupes de 2 (principales tâches/ problèmes) • **Mesure de la précision des évaluateurs** • Passage en revue de la logistique des visites d'écoles	• **Mesure de la précision des évaluateurs 2** *Instruments d'enquête additionnels si elle est administrée*	• **Mesure de la précision des évaluateurs 3**	

Le nombre de jours de formation et le contenu des sessions dépendent en grande partie du nombre d'instruments qui vont être administrés (EGRA plus autres questionnaires ou en plusieurs langues), du nombre d'évaluateurs à former et de leur niveau d'expérience. Si les évaluateurs vont apprendre à administrer le test EGRA en deux langues, une durée de formation additionnelle sera nécessaire. Il est donc recommandé de réduire à 2 le nombre de visites d'écoles pour consacrer davantage de temps à l'apprentissage de l'instrument.

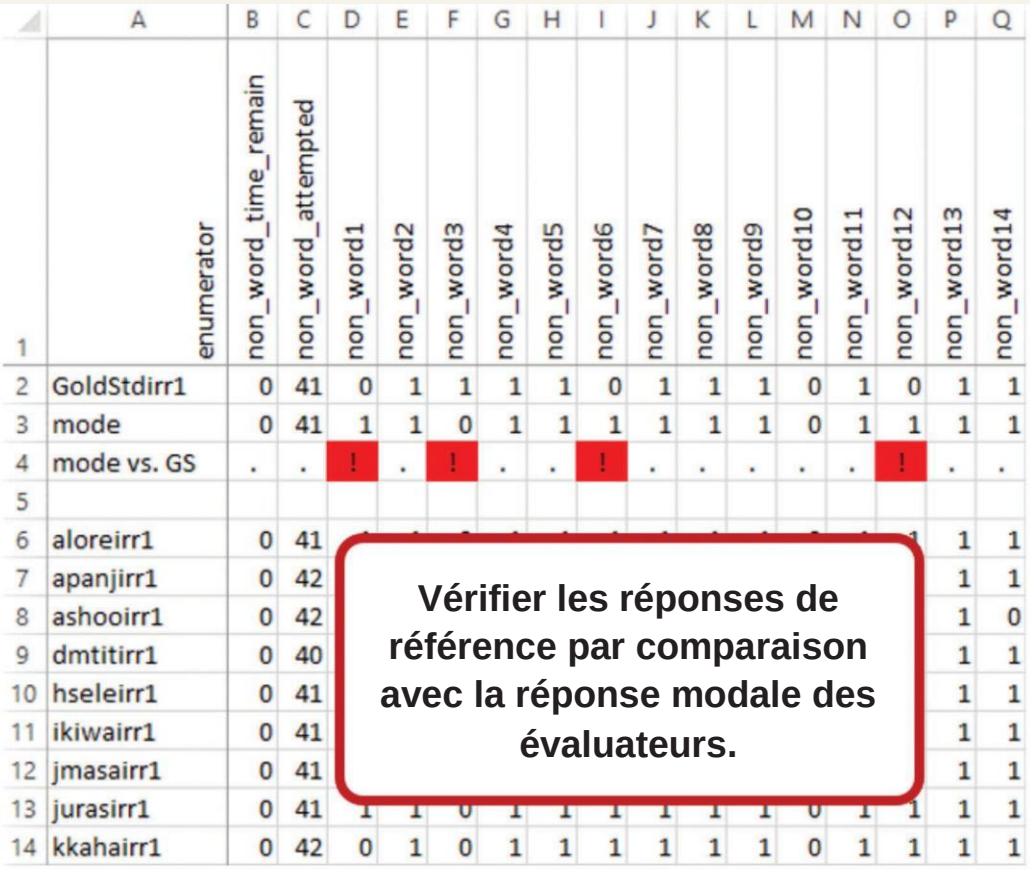
ANNEXE J : ANALYSE DES DONNEES ET DIRECTIVES STATISTIQUES POUR LA MESURE DE LA PRECISION DES EVALUATEURS

On trouvera dans cette annexe des détails sur la gestion des données recueillies pour juger de la précision des évaluateurs, notamment certains termes et conseils statistiques.

J.1 Préparation des données

La **Figure J-1** est un exemple qui montre (indiqué par les cellules grisées) au niveau de l'item où la norme de référence et le mode diffèrent. L'équipe de formation devra en établir la raison. Une explication possible peut être que la norme de référence était inexacte, qu'il y a un problème au niveau de l'instrument ou que les stagiaires ont mal interpréter cet item, auquel cas une formation additionnelle sera nécessaire.

Figure J-1 : Exemple de tableau Microsoft Excel comparant la norme de référence à la réponse modale de l'évaluateur



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	enumerator	non_word_time_remain	non_word_attempted	non_word1	non_word2	non_word3	non_word4	non_word5	non_word6	non_word7	non_word8	non_word9	non_word10	non_word11	non_word12	non_word13	non_word14
2	GoldStdirr1	0	41	0	1	1	1	1	0	1	1	1	0	1	0	1	1
3	mode	0	41	1	1	0	1	1	1	1	1	1	0	1	1	1	1
4	mode vs. GS	.	.	!	.	!	.	.	!	.	.	.	.	.	!	.	.
5																	
6	aloreirr1	0	41													1	1
7	apanjirr1	0	42													1	1
8	ashooirr1	0	42													1	0
9	dmtitirr1	0	40													1	1
10	hseleirr1	0	41													1	1
11	ikiwairr1	0	41													1	1
12	jmasairr1	0	41													1	1
13	jurasirr1	0	41	1	1	0	1	1	1	1	1	1	0	1	1	1	1
14	kkahairr1	0	42	0	1	0	1	1	1	1	1	1	0	1	1	1	1

J.2 Analyse des données

Le pourcentage de concordance par évaluateur est alors calculé par tâche. Cette mesure est la concordance entre l'évaluation de l'enfant par l'évaluateur et l'évaluation correcte de l'enfant. Pour calculer le score de chaque évaluateur (pour chaque tâche et pour l'ensemble de l'évaluation), le dirigeant de la formation fait le total du nombre de concordances avec la norme de référence et exprime ce chiffre sous forme de pourcentage du nombre d'items dans l'évaluation de la tâche, comme l'illustre la **Figure J-2.**

Figure J-2 : Exemple de tableau Microsoft Excel calculant le pourcentage de concordance avec la norme de référence par tâche

enumerator	Non word	non_word_time_remain	non_word_attempted	non_word1	non_word2	non_word3	non_word4	non_word5	non_word6	non_word7	non_word8	non_word9	non_word10	non_word11	non_word12	non_word13	non_word14
Average	88%	95%	59%	32%	100%	14%	100%	100%	0%	100%	100%	100%	73%	100%	0%	100%	95%
aloreirr1	91%	1	1	0	1	0	1	1	0	1	1	1	1	1	0	1	1
apanjirr1	81%	1	0	0	1	0	1	1	0	1	1	1	1	1	0	1	1
ashooirr1	75%	1	0	0	1	0	1	1	0	1	1	1	1	1	0	1	0
dmtitirr1	89%	1	0	1	1	0	1	1	0	1	1	1	1	1	0	1	1
hseleirr1	91%	1	1	0	1	0	1	1	0	1	1	1	1	1	0	1	1
ikiwairr1	91%	1	1	0	1	1	1	1	0	1	1	1	1	1	0	1	1
jmasairr1	89%	1	1	0	1	0	1	1	0	1	1	1	1	1	0	1	1
jurasirr1	91%	1	1	0	1	0	1	1	0	1	1	1	1	1	0	1	1
kkahairr1	89%	1	0	1	1	0	1	1	0	1	1	1	1	1	0	1	1
lkayoirr1	85%	1	0	1	1	0	1	1	0	1	1	1	0	1	0	1	1
mkyejirr1	79%	1	0	0	1	1	1	1	0	1	1	1	0	1	0	1	1
mndolirr1	93%	1	1	1	1	0	1	1	0	1	1	1	1	1	0	1	1
mpaziirr1	91%	1	1	1	1	0	1	1	0	1	1	1	1	1	0	1	1
mramairr1	91%	1	1	0	1	0	1	1	0	1	1	1	1	1	0	1	1
nkibonairr1	79%	0	0	1	1	0	1	1	0	1	1	1	0	1	0	1	1

Le calcul est effectué à l'aide de la formule suivante :

$$\text{Score de l'évaluateur pour la tâche (\%)} = \frac{\text{nombre de concordances avec la norme de référence}}{\text{nombre d'items dans la tâche}}$$

La concordance moyenne au niveau de l'item peut également être calculée pour tous les évaluateurs à l'aide de la formule suivante :

$$\text{Concordance au niveau de l'item (\%)} = \frac{\text{nombre de concordances avec la norme de référence pour l'item}}{\text{nombre de réponses (évaluateurs) pour l'item}}$$

Si la norme de référence fait état d'items manquants parce que l'enfant n'a pas répondu à tous les items d'une tâche, les résultats de concordance par évaluateur comprennent également la concordance avec les items manquants.

Pour les tâches chronométrées comme la fluidité de lecture à haute voix et les sons de lettres corrects par minute, si un enfant termine la tâche dans les délais voulus, il est important que l'évaluateur vérifie le temps que l'enfant a mis pour cette tâche. Si le temps enregistré par l'évaluateur se situe à plus ou moins 2 secondes du temps de référence restant, on considère que l'évaluateur est en accord avec la norme de référence. Un pourcentage de concordance général est alors calculé pour toutes les variables de temps restantes.

Un pourcentage de concordance générale par l'évaluateur est une moyenne des concordances de pourcentages de tâche et de temps restants. Un pourcentage de concordance d'évaluation générale est calculé comme étant une moyenne du pourcentage général de l'évaluateur.

Le résumé des résultats est ainsi rapport pour chaque évaluation et comporte les éléments suivants :

- Par évaluateur : pourcentage de concordance par tâche et générale
- Moyenne du pourcentage de concordance générale
- Pourcentage de concordance générale par tâche

J.3 Glossaire et définitions statistiques

% de concordance brut

Mesure le degré auquel les évaluateurs parviennent à la même conclusion
Coefficient Kappa
Mesure le degré auquel deux évaluations différentes du même sujet auraient pu se produire au hasard. Les valeurs Kappa vont de -1,0 à 1,0. Plus la valeur est élevée, moins il y a probabilité de concordance due au hasard.

Coefficient de corrélation interne (ICC)

Décrit la cohérence des scores attribués aux élèves par différentes évaluateurs. Les valeurs ICC vont de 0,0 à 1,0. Plus la valeur est élevée, plus il y a concordance parmi les évaluateurs.

J.4 Références pour la concordance des évaluateurs

% de concordance brut

Etant donné qu'il n'y a pas de détails provenant uniquement de cette statistique, aucune référence n'est possible. Les évaluateurs s'efforcent à obtenir un pourcentage de concordance aussi élevé (aussi proche de 100 %) que possible quand ils évaluent les élèves. Cependant, quel que soit le pourcentage de concordance, les évaluateurs doivent se reporter aux statistiques Kappa pour interpréter la qualité de la statistique « pourcentage de concordance ».

Coefficient Kappa

OPTION 1

Source : Landis & Koch (1977)

Statistique Kappa	Degré de concordance
moins de 0,0	Faible
de 0,0 à 0,20	Léger
de 0,21 à 0,40	Assez bon
de 0,41 à 0,60	Modéré
de 0,61 à 0,80	Considérable
de 0,81 à 1,00	Presque parfait

OPTION 2

Source : Fleiss (1981)

Statistique Kappa	Degré de concordance
moins de 0,40	Faible
de 0,40 à 0,75	D'intermédiaire à bon
plus de 0,75	Excellent

Coefficient de corrélation interne

Source : Fleiss (1981)

Statistique Kappa	Degré de concordance
moins de 0,40	Faible
de 0,40 à 0,75	D'intermédiaire à bon
plus de 0,75	Excellent

ANNEXE K : EXEMPLES DE PLANS POUR LE CONTROLE DE LA FIABILITE INTER-EVALUATEURS

Cette annexe accompagne le protocole de Save the Children à la Section 8.7 qui décrit comment évaluer en continu la fiabilité inter-évaluateurs (IRR) au cours d'une étude EGRA. Les tableaux ci-dessous présentent une manière systématique de varier la composition d'équipes de trois, quatre ou cinq évaluateurs pour la première évaluation de la journée à chaque école. Bien que la taille totale d'échantillon nécessaire pour une IRR puisse varier en fonction du modèle de l'étude (nombre d'écoles et d'élèves évalués dans l'ensemble), il est recommandé d'évaluer deux fois un minimum de 150 élèves. Une taille d'échantillon de moins de 100 pour une IRR ne produira probablement pas d'information utile.

Tous tableaux et texte à l'appui :

© 2015 Save the Children. Reproduit avec autorisation. Tous droits réservés.

Scénarios de fiabilité inter-évaluateurs en fonction du nombre d'évaluateurs

Equipe de 3 évaluateurs

Ecole	Jumelage des évaluateurs	Evalue & enregistre	Ecoute & enregistre
Ecole 1	A & B C	Evaluateur A Evaluateur C	Evaluateur B -----
Ecole 2	B & C A	Evaluateur B Evaluateur A	Evaluateur C -----
Ecole l 3	C & A B	Evaluateur C Evaluateur B	Evaluateur A -----
Ecole 4	A & C B	Evaluateur A Evaluateur B	Evaluateur C -----
Ecole 5	B & A C	Evaluateur B Evaluateur C	Evaluateur A -----
Ecole 6	C & B A	Evaluateur A Evaluateur C	Evaluateur B -----
Etc...			

Equipe de 4 évaluateurs

Ecole	Jumelage des évaluateurs	Evalue & enregistre	Ecoute & enregistre
Ecole 1	A & B C & D	Evaluateur A Evaluateur C	Evaluateur B Evaluateur D
Ecole 2	A & C B & D	Evaluateur A Evaluateur B	Evaluateur C Evaluateur D
Ecole 3	A & D B & C	Evaluateur A Evaluateur B	Evaluateur D Evaluateur C
Ecole 4	B & A D & C	Evaluateur B Evaluateur D	Evaluateur A Evaluateur C
Ecole 5	C & A D & B	Evaluateur C Evaluateur D	Evaluateur A Evaluateur B
Ecole 6	D & A C & B	Evaluateur D Evaluateur C	Evaluateur A Evaluateur B
Etc...			

Equipe de 5 évaluateurs

Ecole	Jumelage des évaluateurs	Evalue & enregistre	Ecoute & enregistre
Ecole 1	A & B	Evaluateur A	Evaluateur B
	C & D	Evaluateur C	Evaluateur D
	E	Evaluateur E	-----
Ecole 2	E & A	Evaluateur E	Evaluateur A
	B & C	Evaluateur B	Evaluateur C
	D	Evaluateur D	-----
Ecole 3	D & E	Evaluateur D	Evaluateur E
	A & C	Evaluateur A	Evaluateur C
	B	Evaluateur B	-----
Ecole 4	C & E	Evaluateur C	Evaluateur E
	B & D	Evaluateur B	Evaluateur D
	A	Evaluateur A	-----
Ecole 5	A & D	Evaluateur A	Evaluateur D
	E & B	Evaluateur E	Evaluateur B
	C	Evaluateur C	-----
Ecole 6	B & A	Evaluateur B	Evaluateur A
	D & C	Evaluateur D	Evaluateur C
	E	Evaluateur E	-----
Ecole 7	A & E	Evaluateur A	Evaluateur E
	C & B	Evaluateur C	Evaluateur B
	D	Evaluateur D	-----
Ecole 8	E & D	Evaluateur E	Evaluateur D
	C & A	Evaluateur C	Evaluateur A
	B	Evaluateur B	-----
Ecole 9	E & C	Evaluateur E	Evaluateur C
	D & B	Evaluateur D	Evaluateur B
	A	Evaluateur A	-----
Ecole 10	D & A	Evaluateur D	Evaluateur A
	B & E	Evaluateur B	Evaluateur E
	C	Evaluateur C	----
Etc…			

Si l'équipe d'évaluation comporte un nombre impair d'évaluateurs, assurer la rotation des équipes en excusant une personne de l'évaluation de la fiabilité inter-évaluateurs ce jour-là. Il est impératif de créer un calendrier d'évaluation comme ci-dessus pour éviter toute confusion de la part des évaluateurs.

Adresser toutes questions ou demandes de clarification supplémentaires à l'équipe de recherche de la Direction de l'éducation et de la protection de l'enfance à learningassessment@savechildren.org.

ANNEXE L : EXEMPLE DE CODE

Section : démographique	Format	l'étiquette de l'étiquette	Valeurs de l'étiquette	Etiquette de la variable
Country	Chaîne	—	(plus grande variable géographique)	Dans quel pays l'évaluation a-t-elle été administrée ?
Project	Chaîne			Quel projet dans le pays ?
Year	Nombre entier (2000-2020)	—	—	Quelle année l'évaluation a-t-elle été menée ?
Month	Ordinal (1-12)	month	1 janvier 2 février . . .12 décembre	Quel mois l'évaluation a-t-elle été menée ?
Date	Format de date	—	—	Quelle date l'évaluation a-t-elle été menée ?
State	Nominal	state	liste propre au pays (deuxième plus grande variable géographique, après Country)	Dans quel état est située l'école de l'élève ?
Region	Nominal	region	liste propre au pays (troisième plus grande variable géographique, après State)	Dans quelle région est située l'école de l'élève ?
District	Nominal	district	liste propre au pays (plus petite variable géographique, après Region)	Dans quel district est située l'école de l'élève ?
School_name	Chaîne	school	liste propre au pays	Quel est le nom de l'école de l'élève ?
School_code	Nombre entien	—	liste propre au pays	Code national de l'école
EMIS	Nombre entier	—	—	Code du Système d'information pour la gestion de l'éducation
School_type	Nominal	school_type	Fixer les étiquettes de valeur en fonction du projet	Quel type d'école l'élève fréquente-t-il ?
Treatment	Dichotomique	treatment	0 « Témoin » 1 « Intervention partielle » 2 « Intervention complète »	Quel niveau d'intervention l'école reçoit-elle ?
Treat_year	Ordinal (0-12)	—	—	Depuis combien d'années l'école fait-elle l'objet de l'intervention ?
Treat_phase	Ordinal (1-6)	treat_phase	Fixer les étiquettes de valeur en fonction du projet	A quelle phase de l'étude se situe cet élève de l'école faisant l'objet de l'intervention ?

Section : donnée démographique	Format	Intitulé de l'étiquette	Valeurs de l'étiquette	Etiquette de la variable
Urban	Dichotomique	urban	0 Zone rurale 1 Zone urbaine	L'école est-elle située en zone urbaine ?
Shift	Ordinal (0-2)	shift	0 « Pas de roulement » (journée complète) 1 Matin 2 Après-midi 3 En alternance	L'élève fréquente-t-il l'école par roulement ?
Dbl_shift	Dichotomique	yes/no	0 Non 1 Oui	Est-ce que l'école applique le système du double horaire ?
Admin	Nominal	admin	liste propre au pays	Qui a administré le test ? (numéro de code)
Admin_name	String	—	—	Qui a administré le test ?
ID	String	—	Doit être unique !!!!	Numéro unique d'identification de l'élève
Grade	Nombre entier (1-8)	grade	1 première, 2 deuxième, 3 troisième, 4 quatrième, 5 cinquième, 6 sixième, 7 septième, 8 huitième	Quel est le niveau scolaire de l'élève ?
Level	Nombre entier	—	*Comme pour la classe mais pour les élèves ne relevant pas d'une tranche d'âge traditionnelle*	*Pour les élèves de tranche d'âge non traditionnelle, à quel « niveau solaire » apprennent-ils ?*
Section	Nombre entien	—	liste propre au pays	Dans quelle section de classe est l'élève ?
Female	Dichotomique	female	0 Masculin 1 Féminin	S'agit-il d'une élève de sexe féminin ?
Multigrade	Dichotomique	yes/no	0 Non 1 Oui	L''élève est-il dans une salle de classe à années multiples ?
Teacher	Nombre entier	teacher	liste propre au pays	Quel est le nom de l'enseignant de l'élève ?
Age	Nombre entier (5-18)	—	—	Quel âge a l'élève ?
Start_time	Time (hh:mm)	—	—	Heure de commencement de l'évaluation ?
End_time	Time (hh:mm)	—	—	Heure de fin de l'évaluation ?
Assess_time	Time (m)	—	—	Temps mis pour réaliser l'évaluation (en minutes) ?
Language	Nombre entier	language	employer les codes ISO 639-3	Langue d'évaluation
Consent	Dichotomique	yes/no	0 Non 1 Oui	Est-ce que le participant a donné son consentement/ assentiment à l'évaluation ?

ANNEXE M : RECOMMENDATIONS POUR L'ETALONNAGE

Un panel d'experts et des participants à l'atelier USAID 2015 intitulé « Améliorer la qualité des données EGRA : consultation pour contribuer à l'orientation de l'USAID sur l'administration d'évaluations des compétences fondamentales en lecture » ont traité du sujet de l'équivalence de tâches de même langue pour plusieurs versions d'un instrument. On trouvera ci-dessous les recommandations techniques détaillées du panel à ce sujet, ainsi que des questions devant faire l'objet de débats supplémentaires.

M.1 Recommandations

1. **Pour les tâches comportant seulement quelques items (par ex. 10 à 25), piloter plusieurs versions de test récemment mises au point parallèlement aux versions initiales.** Comparer ensuite les statistiques au niveau des items pour toutes les versions et se servir de cette information (valeurs p et bisériales ponctuelles) pour élaborer des versions de mi-parcours et finales se rapprochant le plus des statistiques de la version initiale. C'est une démarche simplifiée de pré-équivalence de personnes en commun mettant en jeu la théorie classique des tests (TCT).
2. **Ne pas appliquer les démarches d'équivalence TCT aux tâches comportant seulement quelques items.** Les raisons en sont claires pour les tâches comportant 3 à 5 items (écoute et compréhension de lecture par exemple), mais ce point doit faire l'objet de débats supplémentaires quand le nombre d'items reste entre 10 et 25. S'il peut être possible de procéder à une équivalence à l'aide de démarches TRI, celles-ci exigent des tailles d'échantillon d'au moins 500 à 1 000 élèves pour des modèles plus complexes (modèles à deux ou trois paramètres). Les modèles Rasch et les méthodes TCT requièrent des tailles d'échantillon similaires—le choix devient donc une question de satisfaction aux hypothèses (et de convenance de l'analyse des données au niveau des items).
3. **Pour les données linéaires et les petits échantillons, employer des méthodes d'équivalence TCT.** Un pilotage avec personnes en commun peut dans ces cas-là être employé pour les passages de fluidité de lecture à haute voix (FLHV) et des démarches d'équivalence moyenne ou linéaire peuvent être appliquées (et choisies en fonction de la convenance visuelle, du biais et de l'erreur). Pour des données non linéaires, le processus se complique. Cette méthode convient mieux que les procédures d'équivalence par théorie de réponse aux items (TIR), étant donné que les mesures de la FLHV donnent un score total (sans données utiles au niveau des items).

4. **S'assurer que les échantillons pilotes et opérationnels sont aussi similaires que possible.** Etant donné que beaucoup de démarches d'étalonnage pour EGRA reposent sur le pilotage d'échantillons aléatoirement équivalents ou de personnes en commun (notamment pour la FLHV), l'échantillon pilote doit être aussi représentatif de l'échantillon opérationnel que possible pour que les ajustements d'équivalence appliqués à l'échantillon pilote conviennent aux données opérationnelles.
5. Quand on a recours à une calibration par percentiles égaux pour étalonner des données FLHV non linéaires, **s'assurer que l'échantillon comporte des élèves à tous les points de score possibles**—ce qui nécessite souvent une taille d'échantillon plus importante qu'il n'est faisable dans un pilotage avec personnes en commun pour des études basées sur le test EGRA.

La **Figure M-1** représente des recommandations portant sur les variables EGRA sommaires pouvant et ne pouvant pas être étalonnées à l'aide des méthodes d'équivalence traditionnelles pour de petits échantillons. Des questions restent à débattre sur ce tableau, notamment en ce qui concerne les scores de zéro, les tâches comportant entre 10 et 25 items et le pourcentage correct de tentatives.

Figure M-1. Résumé des variables EGRA à étalonner (recommandations)

Tâches EGRA	d'items	Score chronométré	Score Zéro	Score	% de score	de tentatives	% correct de tentatives
Phonétique / sons de syllabes	-20	TRI - Rasch	Non	Non	Non	Non	Non
Vocabulaire	5-10	TRI - Rasch	Non	Non	Non	Non	Non
Noms de lettres	100	Item fixe	Non	Item fixe	Item fixe	Non	Non
Sons de lettres	100	Item fixe	Non	Item fixe	Item fixe	Non	Non
Mots familiers	50	Item fixe	Non	Item fixe	Item fixe	Non	Non
Non-mots	50	Item fixe	Non	Item fixe	Item fixe	Non	Non
Fluence de lecture à haute voix	-50	Personnes en commun (percentiles égaux)	Non	Personnes en commun (percentiles égaux)	Personnes en commun (percentiles égaux)	Non	Non
Compréhension de lecture	-5	TRI - Rasch*	Non	Non	Non	Non	Non
Compréhension à l'écoute	-5	TRI - Rasch*	Non	Non	Non	Non	Non
Dictée	10-15	TRI - Rasch	Non	Non	Non	Non	Non
Labyrinthe	10-15	TRI - Rasch	Non	Non	Non	Non	Non

* = une enquête plus approfondie est nécessaire.

Remarque : ces méthodes ne sont recommandées que si on pense que le pilote suit une distribution similaire à celle de l'étude à grande échelle (tel que déterminé par échantillonnage aléatoire).

M.2 Questions devant faite l'objet de débats supplémentaires

1. **Quelles méthodes peut-on employer pour étalonner les tâches de compréhension de lecture (ou autres tâches comportant seulement 5 items) ?** Les méthodes d'équivalence TCT traditionnelles ne conviennent pas dans ces circonstances, mais on pourra envisager des démarches progressives ou autres méthodes non linéaires. Un étalonnage TRI devra être examiné de près pour en déterminer la faisabilité avec les tâches les plus brèves.
2. **Quelles méthodes peut-on employer pour étalonner les passages FLHV quand on a déterminé l'existence de rapports non linéaires dans toutes les versions ?** Il semblerait que l'étalonnage par arc de cercle et la calibration par percentiles égaux conviennent à ces situations, mais ces méthodes s'accompagnent toutes deux de limitations et doivent être examinées de plus près.
3. **Quels sont les compromise entre des démarches d'étalonnage TCT et TRI ?** Ces questions sont en rapport avec l'expertise technique, la taille de l'échantillon, les procédures de pilotage, etc. Quand les données au niveau des items sont enregistrées, il faudra en définitive accorder la préférence à des analyses Rasch pour petits échantillons.
4. **Comment traiter les scores de zéro au cours de l'étalonnage ?** Faut-il exclure des calculs d'étalonnage les élèves obtenant des scores de zéro sur tous les formulaires d'évaluation (ou seulement ceux obtenant des scores de zéro sur n'importe lequel) ? Est-il possible d'obtenir un score de zéro sur une version de test donnée mais d'avoir un ajustement d'équivalence qui produise un score de non zéro pour cet élève ? Dans quelle mesure le traitement des scores de zéro est-il fonction de la méthode d'étalonnage à appliquer ?
5. **Quelles sont les implications de l'emploi de données pilotes plutôt que de données opérationnelles pour l'étalonnage ?** Dans la majorité des cas, nous sommes limités à l'emploi de données pilotes pour étalonner la FLHV, mais quels sont les compromis dans les circonstances où l'une ou l'autre méthode est possible ? Le post-étalonnage (donnée opérationnelles) étant susceptible de fournir des rapports d'équivalence plus fiables, est-ce qu'il y a une raison de reposer sur un pré-étalonnage (données pilotes) quand les deux options sont disponibles ?
6. **Comment peut-on analyser les scores étalonnés ?** Si les classes sont étalonnées séparément, faut-il aussi les analyser séparément (réfutant ainsi une analyse générale/combinée) ? Si les scores bruts sont étalonnés, peut-on procéder à des analyses sur le pourcentage de tentatives réussies ?
7. **Les effets de l'ordre devront être examinés pour les tests dans lesquels les items sont agencés au hasard dans les rangées.** Le regroupement par inadvertance d'items difficiles pourrait avoir une incidence sur les scores du test ; cette question mérite un examen plus approfondi.
8. **Il faut explorer la dimensionnalité entre les tâches.** Si on peut prouver un niveau raisonnable d'unidimensionnalité, il est possible que l'étalonnage provenant de certaines tâches puisse être extrapolé pour d'autres. Sinon, tout l'étalonnage est restreint au niveau du composant (tâches), ce qui peut limiter la généralisabilité en ce qui concerne les résultats d'ensemble en lecture.

ANNEXE N : RECOMMANDATIONS TECHNIQUES DETAILLEES SUR LES FICHIERS A USAGE PUBLIC

La Section 10.6 de ce manuel présente les mesures qu'il conviendra de prendre avant de mettre les données EGRA à la disposition du public. On trouvera dans cette annexe les recommandations techniques additionnelles du panel sur les fichiers à usage public, tel que convenu à l'atelier USAID 2015 intitulé « Améliorer la qualité des données EGRA : consultation pour contribuer à l'orientation de l'USAID sur l'administration d'évaluations des compétences fondamentales en lecture ».

N.1 Recommandations spécifiques pour le nettoyage, la finalisation et l'anonymisation des données

N.1.1 Nettoyage

1. L'USAID renforce l'utilité de l'adoption d'un code maître pour les partenaires d'évaluation/mise en œuvre.
2. Le code maître est fondé sur le code élaboré par le projet EdData II (Données sur l'éducation pour la prise de décisions)[34] de l'USAID. Le panel recommande la mise au point d'un code pour les instruments complémentaires (données démographiques dans le questionnaire de l'élève par exemple).
3. Dans la mesure du possible, les noms des variables sont définis avec un maximum de 12 caractères et les étiquettes des variables avec un maximum de 80 caractères.
4. Les FUP sont auto-descriptifs, comportent des données catégoriques et emploient les catégories comme valeurs plutôt qu'un code numérique.
5. Pour éviter les généralisations non fondées, les données doivent être éliminées pour toutes les zones géographiques qui n'ont pas été utilisées à des fins d'échantillonnage et qui comportent trop peu d'écoles dans l'échantillon pour obtenir des estimations suffisamment précises (district, zone de dénombrement, localité et quartier).

[34] Le code (RTI International, 2014a), sous format de fichier Excel, est disponible sur le site Web EdData II : https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&id=389. Voir aussi l'Annexe L de ce manuel où figure un exemple de code.

N.1.2 Finalisation

Dans les cas où on a eu recours à une méthodologie d'échantillonnage complexe, les spécifications de l'étude devront dans la mesure du possible être établies dans le fichier de données FUP pour minimiser les erreurs de spécification par les utilisateurs publics.

Pour atténuer les problèmes d'erreurs de spécification au cours de l'analyse des données par les utilisateurs publics, le jeu de données d'analyse sert de base au FUP. Dans la mesure du possible et s'il y a lieu, les chercheurs fusionnent (données relatives aux enseignants et aux élèves par exemple) ou ajoutent (données initiales et finales par exemple) des fichiers de données pour éviter que les utilisateurs publics n'aient à manipuler plusieurs fichiers de données.

N.1.3 Anonymisation

Tous les éléments d'identification personnelle doivent être éliminés des jeux de données avant qu'ils ne soient mis à la disposition du public. On trouvera ci-dessous des recommandations générales sur l'élimination et l'anonymisation des renseignements personnels.[35]

1. Eliminer les renseignements personnels tels que nom, domicile, numéro de téléphone et numéro national d'identification.
2. Eliminer le nom des écoles et le nom de tous autres établissements et individus pouvant avoir été recueillis au cours du processus d'acquisition de données.
3. Eliminer toute information employée pour contacter et trouver les écoles ou établissements (adresse, numéro de téléphone, nom du principal, coordonnées GPS, etc.).
4. Les données employées à des fins d'échantillonnage peuvent comprendre des éléments d'identification personnelle : anonymiser les données mais ne pas les détruire. Il est important de conserver un jeu de données restreintes pour faire correspondre les valeurs des variables anonymisées aux valeurs des variables qui ne le sont pas.
5. Anonymiser toutes les variables qui contiennent les codes officiels du pays (codes des écoles ou établissements, codes des enseignants, par exemple).

N.2 Diffusion des données FUP

On s'attend à ce que l'USAID rende publics les fichiers FUP contenant les données des évaluations des compétences fondamentales en lecture au travers de l'analyse secondaire pour projet de suivi des résultats, portail de l'éducation (SART Ed) et de la bibliothèque des donnés de développement (DDL).

[35] Se reporter à l'Annexe A d'Optimal Solutions Group (2015) où figure une discussion détaillée.

Pour faciliter l'exploration des données par le public, le panel recommande de fournir une documentation d'accompagnement en plus des fichiers de données.

Il faudra également afficher des informations bien documentées accompagnant le FUP pour permettre aux utilisateurs publics de se familiariser avec les données. Les informations suivantes devront être fournies aux utilisateurs :

1. Rapport écrit d'analyse des données soumis à l'USAID et approuvé par celle-ci.
2. Questionnaires et outils d'évaluation employés pour la collecte de données. (Pour ne pas compromettre les matériels pouvant être utilisés pour des études futures dans le cadre du même projet, ces éléments ne peuvent être fournis qu'une fois le projet terminé.)
3. Information contextuelle et toute documentation pertinente

Outre l'exigence des rapports et instruments de collecte des données, l'USAID souligne de nouveau à l'intention des partenaires d'évaluation/mise en œuvre à quel point il est important de documenter les noms et descriptions des variables et des paramètres clés nécessaires à une bonne analyse des données, notamment :

1. Une définition explicite de la population visée, notamment la source de la liste employée pour le prélèvement de l'échantillon. La documentation indique le nombre total d'écoles et une estimation du nombre d'élèves que l'échantillon est censé représenté. Ces chiffres correspondent également aux estimations de données pondérées. Si l'étude comporte une intervention/contrôle, le chiffres sont communiqués par intervention/contrôle.
2. Les variables nécessaires pour analyser correctement les données complexes en fonction de la méthodologie d'échantillonnage (par ex., pour chaque stade de l'échantillonnage : tous les éléments échantillonnés, la variable de stratification et la variable de correction de la population finie, ainsi que la variable de pondération finale).
3. Les variables pour le modèle de recherche (par ex. intervention, année et cohorte si le modèle de recherche est un rapport d'évaluation de l'impact échelonné).
4. Une explication des critères de la méthodologie d'échantillonnage (par ex. prise en charge de la variance) pour que les caractéristiques du modèle d'étude puissent être employées indépendamment du logiciel propriétaire.
5. Un code complet contenant :
    a. La liste de toutes les variables du jeu de données.
    b. L'étiquette et le format de chaque variable, ainsi que (le cas échéant) les étiquettes des valeurs.
    c. Une description formelle du calcul employé pour produire les variables calculées (par ex. fluidité de lecture à haute voix).
    d. Le nombre total d'observations dans le jeu de données.

Pendant la construction des référentiels de données prévus de l'USAID, les partenaires de mise en œuvre et les évaluateurs mettent leurs FUP contenant des données d'évaluation des compétences fondamentales en lecture à la disposition du public.

1. Afficher le FUP sur un site accessible en ligne, accompagné de sa documentation tel que précisé plus haut (c.-à-d. tous les éléments sont placés dans un seul fichier comprimé ou le site Web comporte un lien vers ces documents).
2. Créer le FUP à l'aide d'un fichier de données non propriétaire et, dans la mesure du possible, un fichier de données propriétaire.
3. Pour le fichier propriétaire, créer un fichier texte csv (valeurs séparées par des virgules).
4. Pour le fichier non propriétaire, créer soit un fichier Stata .dta et / ou un fichier SPSS .sav (ainsi que le fichier SPSS.csaplan).

Le panel encourage également l'USAID à envisager la mise au point de directives pour des rapports d'évaluation basés sur les rapports des compétences fondamentales en lecture, similaires aux recommandations générales de l'USAID sur la préparation de rapports d'évaluation (USAID, 2012), ceci permettant au grand public de trouver plus facilement les informations dans les rapports. Cela garantirait de plus de pouvoir trouver les mêmes informations de base dans tous les rapports.

ANNEXE O : ANALYSE DES DONNEES EGRA

Pour chaque estimation de score d'élève communiquée, une représentation visuelle telle que celles figurant aux **Figures O-1 à O-3** doit être fournir sous forme de graphique à l'appui de l'interprétation par le lecteur de l'estimation donnée.

Figure O-1. Exemple d'analyse de différence parmi les différences (DID)

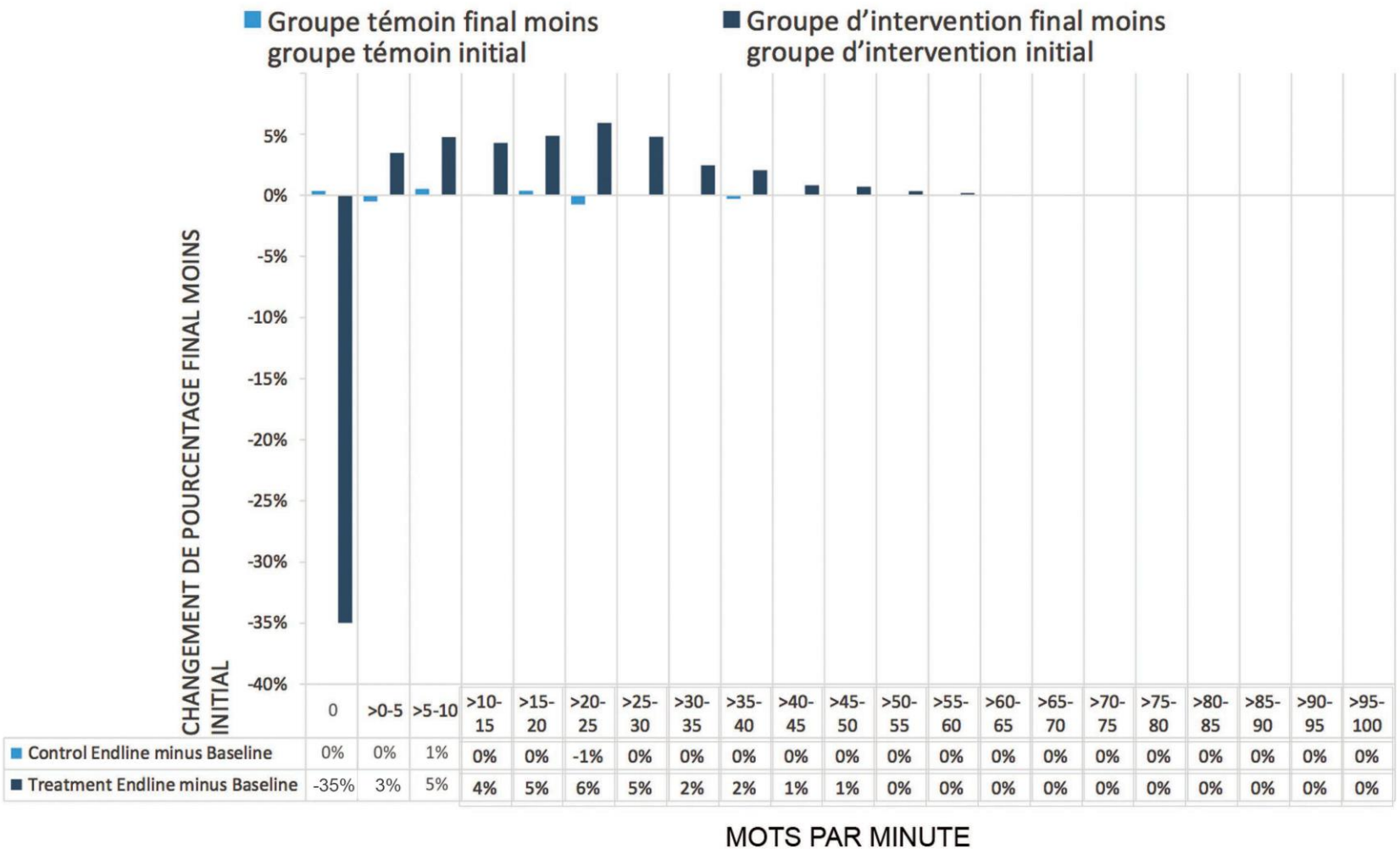
	Ini					F							
Cible	**Fluence moyenne (mpm)**	**Erreur type**	**Nombre d'élèves sondés**	**Stat t**	**Valeur p**	**Fluence moyenne (mpm)**	**Erreur type**	**Nombre d'élèves sondés**	**Stat t**	**Valeur p**	**Différence parmi les différences**	**Valeur p (DID)**	**Ampleur d'effet**
Témoin	4,5	0,6	656	–	–	9,5	1,6	475	–	–	–	–	–
Intervention	5.2	1,2	349	0,510	0,611	11,7	1,1	480	1,189	0,236	1,5	0,490	0,12

Calculs :

Différence parmi les differences : (cible finale moyenne – cible initiale moyenne) – (témoin final moyen – témoin initial moyen)

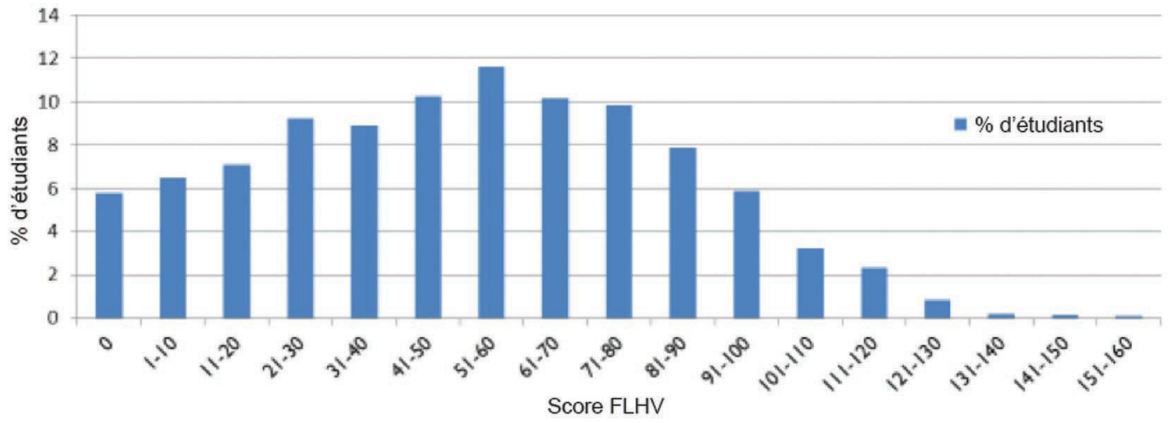
Ampleur d'effet (d de Cohen) : différence parmi les différences / écart-type cumulé

Figure O-2. Exemple de comparaison de répartition des différences entre le groupe témoin et le groupe d'intervention



	0	>0-5	>5-10	>10-15	>15-20	>20-25	>25-30	>30-35	>35-40	>40-45	>45-50	>50-55	>55-60	>60-65	>65-70	>70-75	>75-80	>80-85	>85-90	>90-95	>95-100
Control Endline minus Baseline	0%	0%	1%	0%	0%	-1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Treatment Endline minus Baseline	-35%	3%	5%	4%	5%	6%	5%	2%	2%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Figure O-3. Distribution de la fluidité de lecture à haute voix (FLHV) – Indonésie, 2013



Source : Stern, J. & Nordstrum, L. (2014). Indonésie 2014 : évaluation nationale des compétences fondamentales en lecture (EGRA) et étude Aperçu de l'efficacité de la gestion scolaire (SSME). Préparé pour USAID/Indonésie dans le cadre du projet EdData II (Données sur l'éducation pour la prise de décisions). Ordre de mission no AID-497-BC-13-00009 (Mission RTI 23). Research Triangle Park, NC : RTI International. https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&id=680

ANNEXE P : NORMES DE FLUIDITE DE LECTURE EN ANGLAIS

Le tableau ci-dessous est un exemple de normes de fluidité de lecture à haute voix basé sur des recherches effectuées en anglais. Ce tableau est inclus dans la version française du manuel à titre d'exemple de ce qui peut être établi dans d'autres langues.

ReadNaturally®

Hasbrouck & Tindal Oral Reading Fluency Data

This table shows the oral reading fluency rates of students in grades 1 through 8, based on an extensive study conducted by Jan Hasbrouck and Gerald Tindal. The results of their study are published in a technical report entitled, "Oral Reading Fluency: 90 Years of Measurement," which is available on these websites:

- **ERIC website:** eric.ed.gov/?id=ED531458
- **BRT website:** www.brtprojects.org/publications/technical-reports

This table can help you assess the oral reading fluency of your students relative to their peers. Students scoring 10 or more words below the 50th percentile using the average score of two unpracticed readings from grade-level materials need a fluency-building program. Teachers can also use the table to set long-term fluency goals for struggling readers.

For more information:

- **Essential Components of Reading:** readnaturally.com/components
- **Correlation Between Oral Reading Fluency and Overall Reading Achievement:** readnaturally.com/correlation
- **Read Naturally Tools for Assessing Fluency:** readnaturally.com/assessment-tools
- **Read Naturally Intervention Programs That Develop Fluency:** readnaturally.com/fluency-interventions

Grade	Percentile	Fall WCPM*	Winter WCPM*	Spring WCPM*	Avg. Weekly Improvement**
1	90		81	111	1.9
	75		47	82	2.2
	50		23	53	1.9
	25		12	28	1.0
	10		6	15	0.6
2	90	106	125	142	1.1
	75	79	100	117	1.2
	50	51	72	89	1.2
	25	25	42	61	1.1
	10	11	18	31	0.6
3	90	128	146	162	1.1
	75	99	120	137	1.2
	50	71	92	107	1.1
	25	44	62	78	1.1
	10	21	36	48	0.8
4	90	145	166	180	1.1
	75	119	139	152	1.0
	50	94	112	123	0.9
	25	68	87	98	0.9
	10	45	61	72	0.8
5	90	166	182	194	0.9
	75	139	156	168	0.9
	50	110	127	139	0.9
	25	85	99	109	0.8
	10	61	74	83	0.7
6	90	177	195	204	0.8
	75	153	167	177	0.8
	50	127	140	150	0.7
	25	98	111	122	0.8
	10	68	82	93	0.8
7	90	180	192	202	0.7
	75	156	165	177	0.7
	50	128	136	150	0.7
	25	102	109	123	0.7
	10	79	88	98	0.6
8	90	185	199	199	0.4
	75	161	173	177	0.5
	50	133	146	151	0.6
	25	106	115	124	0.6
	10	77	84	97	0.6

*WCPM = Words Correct Per Minute

www.readnaturally.com

**Average words per week growth

United States Agency for International Development
Office of Education
Bureau for Economic Growth, Education, and Environment (E3)
1300 Pennsylvania Avenue, N.W.
Washington, DC 20523, USA
www.USAID.gov