EDK | CDIP | CDPE | CDEP |
Schweizerische Konferenz der kantonalen Erziehungsdirektoren
Conférence suisse des directeurs cantonaux de l'instruction publique
Conferenza svizzera dei direttori cantonali della pubblica educazione
Conferenza svizra dals directurs chantunals da l'educaziun publica

VÉRIFICATION DE L'ATTEINTE DES COMPÉTENCES FONDAMENTALES: VISION D'ENSEMBLE

Rapport du Bureau de coordination pour la mise en œuvre de l'accord intercantonal sur l'harmonisation de la scolarité obligatoire (Bureau HarmoS)

23 novembre 2022

222.0-4.5 pl/cvb

Generalsekretariat | Secrétariat général
Haus der Kantone, Speichergasse 6, Postfach, CH-3001 Bern | T: +41 31 309 51 11, F: +41 31 309 51 50, www.edk.ch, edk@edk.ch

IDES Informationszentrum | Centre d'information | T: +41 31 309 51 00, F: +41 31 309 51 10, ides@edk.ch

Table des matières

Introduction 2

1. État des lieux 3
1.1 Bases et travaux préliminaires réalisés en amont de la vérification de l'atteinte des compétences fondamentales 3
1.2 Mise en œuvre de la vérification de l'atteinte des compétences fondamentales 9
1.3 Enquêtes menées dans le cadre de la vérification de l'atteinte des compétences fondamentales 14
1.4 Futures enquêtes de vérification ayant fait l'objet d'une décision 21

2 Analyse et discussion 25
2.1 Approche, objectifs et paramètres liés à la vérification de l'atteinte des compétences fondamentales 25
2.2 *Design* des enquêtes 29
2.3 Autres aspects liés à la mise en œuvre de la vérification 32
2.4 Gouvernance de la vérification 33

3 Étapes en vue du développement de la vérification 36
3.1 Nouveau *design* d'enquête 36
3.2 Adaptations en matière de gouvernance 45
3.3 Conséquences financières 48

Bibliographie 49

Annexe 1: Autres aspects liés à la mise en œuvre de la vérification 53

Annexe 2: Expertise scientifique sur le projet Vérification (en allemand) 61

Introduction

Le 25 octobre 2019, l'Assemblée plénière de la CDIP a chargé le Bureau de coordination pour la mise en œuvre de l'accord intercantonal sur l'harmonisation de la scolarité obligatoire (Bureau HarmoS) d'établir une vision d'ensemble du programme «afin de jeter les bases nécessaires [...] pour se prononcer sur la poursuite de la vérification de l'atteinte des compétences fondamentales après 2022». Il était prévu de réaliser quatre enquêtes principales jusqu'en 2022 dans le cadre de la vérification de l'atteinte des compétences fondamentales (à savoir en 2016, auprès d'élèves de 11e année en mathématiques; en 2017, auprès d'élèves de 8e année dans la langue de scolarisation et dans la première langue étrangère; en 2020, auprès d'élèves de 11e année, dans la langue de scolarisation, mais aussi dans la première et la deuxième langues étrangères; en 2022, auprès d'élèves de 4e année dans la langue de scolarisation et en mathématiques). En raison de la pandémie de coronavirus et de la nécessité de coordonner la planification avec celle de l'étude PISA, il a fallu reporter la troisième et la quatrième enquêtes à respectivement 2023 et 2024.

La présente vision d'ensemble s'articule en trois parties principales, à savoir un volet «état des lieux», un autre proposant une «analyse et discussion» et un dernier portant sur les «étapes en vue du développement de la vérification». L'état des lieux donne un aperçu des bases et de la manière dont la vérification a été mise en œuvre jusqu'à l'enquête prévue auprès d'élèves de 4e année. Le deuxième volet propose une analyse critique de différents aspects nécessitant d'être clarifiés ou consolidés pour permettre le développement du projet. Le Bureau HarmoS a fait appel à une équipe d'experts externes, provenant de l'étranger comme de la Suisse, et leur a confié le mandat de réaliser une expertise scientifique étendue pouvant servir de base à l'élaboration de cette deuxième partie; il a également invité à s'exprimer d'autres experts, en leur qualité de personnes ayant contribué ou contribuant encore, à différents degrés, au projet Vérification de l'atteinte des compétences fondamentales. Le troisième volet décrit les principaux aspects de la poursuite du projet tels que les soumet le Bureau HarmoS aux organes de la CDIP afin qu'ils puissent y réfléchir puis prendre une décision. Ces aspects concernent au premier chef le choix du futur *design* d'enquête défini à long terme ainsi que diverses modifications en rapport avec la gouvernance du projet.

La vision d'ensemble s'adresse aux membres de la CDIP. Elle a pour but de leur servir de base de réflexion et d'argumentation en vue des décisions à prendre en rapport avec la poursuite du projet.

1. État des lieux

Le volet «**état des lieux**» présente les bases conceptuelles, politiques et juridiques du projet, mais aussi les structures et acteurs impliqués dans sa mise en œuvre, les enquêtes qui ont été menées jusqu'à présent tout comme celles qui sont en cours de réalisation. Il y est aussi question des adaptations conceptuelles apportées à la suite des expériences engrangées, des comptes rendus relatifs aux résultats des enquêtes, de l'utilisation qui est faite des données des enquêtes de même que des aspects financiers.

1.1 Bases et travaux préliminaires réalisés en amont de la vérification de l'atteinte des compétences fondamentales

1.1.1 Renforcement de la coopération en matière d'éducation en Suisse par des niveaux de compétence communs

Après de premières discussions menées en l'an 2000, l'Assemblée plénière de la CDIP adopte le 7 juin 2001, sous la conduite de son nouveau secrétaire général Hans Ambühl (entré en fonction en janvier 2000), un modèle d'action (CDIP, 2001a) de même que son premier programme de travail (CDIP, 2001b). Les lignes directrices visaient principalement à renforcer la coopération en matière d'éducation et de formation en Suisse, dans le but de maintenir la compétitivité, d'encourager les échanges entre les cultures et de faciliter la mobilité dans le pays. Afin de renforcer la coopération et la coordination à l'échelle intercantonale, un élargissement du concordat scolaire de 1970 (CDIP, 1970) est envisagé, de même qu'une série de mesures de coordination. Il s'agit notamment d'harmoniser de manière plus contraignante les transitions vers les degrés secondaires I et II, en définissant, à l'échelle du pays, des niveaux de compétence uniforme pour la fin de la 6^{e} et de la 9^{e} années (c'est-à-dire la 8^{e} et 11^{e} années selon la numérotation HarmoS), dans l'idée de fixer des «points de rencontre». Il est également question, afin d'assurer la qualité de l'ensemble, d'introduire des évaluations de compétences à effectuer de manière continue, notamment à la fin de la scolarité obligatoire. À partir des années 2000, la nécessité de disposer d'objectifs communs en matière d'éducation et de vérifier de manière constante la réalisation de ces objectifs est considérée comme un élément essentiel des efforts d'harmonisation du système éducatif suisse. L'idée est de définir des niveaux de compétences, applicables de manière contraignante et à appréhender comme des «standards», afin de les inclure dans le concordat scolaire de 1970.

Durant ces années-là, d'autres pays travaillent également à introduire des objectifs de formation, sous la forme de standards, et des systèmes permettant de vérifier leur atteinte. Parmi les éléments déclencheurs à l'origine de ce développement, on peut notamment citer l'étude PISA, qui a été menée pour la première fois en l'an 2000 et qui, en de nombreux endroits, a créé la surprise par des résultats moins bons que prévu. En Suisse aussi, le groupe de pilotage chargé de tirer les enseignements des résultats de PISA 2000 recommande d'évaluer, à intervalles de plusieurs années, les performances disciplinaires et les compétences transversales sur la base de standards de formation définis (Buschor et al., 2003 p. 31).

1.1.2 Le projet HarmoS

Le 6 juin 2002, l'Assemblée plénière de la CDIP approuve l'organisation et le financement du projet appelé HarmoS. Une première phase est planifiée pour les années 2003 à 2006 (CDIP, 2004b p. 1), avec un mandat décliné dans un premier temps en trois étapes:

1. développement de modèles de compétences pour quatre disciplines (niveaux de compétence, descriptions de niveaux);
2. vérification empirique de l'atteinte des exigences à l'aide d'échantillons d'élèves;
3. détermination de standards contraignants, confiée à la politique de l'éducation.

S'ensuit une deuxième phase, qui a pour but le transfert des standards dans les plans d'études, les moyens d'enseignement ainsi que dans des outils d'évaluation et de test à différents niveaux du système éducatif.

Début 2003, les travaux démarrent sous la conduite d'une commission de coordination composée de représentants du Secrétariat général de la CDIP (SG CDIP) et des secrétaires des quatre conférences régionales de la CDIP. Un conseil consultatif, réunissant des représentantes et représentants des quatre régions de la CDIP, du corps enseignant ainsi que des milieux scientifiques, est nommé pour assumer une fonction de conseil.

Les travaux menés dans le cadre de ces étapes sont précédés de discussions de fond sur la manière dont il convient de décrire les attentes en matière de performances à destination des élèves, donc sur la question de la définition des standards de formation. La réponse à cette question est fondamentale pour les trois étapes. Enfin, on prend aussi pour référence l'expertise intitulée *Le développement de standards nationaux de formation: une expertise*, menée sous la direction du chercheur en éducation E. Klieme (CDIP, 2004b pp. 7-11; Klieme et al., 2004) et publiée pour la première fois en Allemagne en 2003. Les standards de formation suisses sont ainsi conçus sous la forme de standards de performance centrés sur les résultats ou les *outputs* (appelés aussi *performance standards* ou *outcome standards*), qui se concentrent sur les domaines noyaux des disciplines que l'on retrouve également dans les plans d'études cantonaux et régionaux. Dans le cadre du projet HarmoS, il est question d'élaborer pour ces domaines noyaux des modèles relatifs à l'acquisition de compétences, qui montrent les progressions au fil des niveaux. Il s'agit ensuite de définir les points sur les échelles ainsi définies qui correspondent aux standards minimaux. Ces standards minimaux, qui revêtent un caractère obligatoire, décrivent les compétences minimales pouvant être attendues de (pratiquement) tous les élèves à certains moments. L'expertise de Klieme et al. propose d'adopter, en plus de ces standards minimaux contraignants, une différenciation supplémentaire des exigences sous la forme de niveaux de compétences, ce qui doit permettre de mettre en évidence les évolutions dans les apprentissages et de déterminer des niveaux et des profils. En Suisse, il a été décidé de ne pas établir de descriptions de niveaux à l'échelle nationale (par ex. pour les modèles impliquant une différenciation selon les performances au degré secondaire I) et de laisser plutôt cette possibilité aux cantons. Le fait de définir des modèles de compétences et des standards de performances minimaux (devenus par la suite les «compétences fondamentales») qui sont contraignants à l'échelle nationale permet de garantir une harmonisation des disciplines concernées quant à leurs domaines noyaux, mais certains aspects essentiels ne sont pas réglementés, ce qui laisse une marge de manœuvre suffisante aux niveaux inférieurs, et même jusqu'à celui de l'enseignement. Lors de l'élaboration des plans d'études régionaux, mandat a été donné de reprendre, aussi littéralement que possible, les objectifs nationaux de formation dans les descriptions des exigences de base du plan d'études. Les contenus qui vont au-delà du domaine noyau, les processus d'enseignement et d'apprentissage de même que les conditions d'apprentissage peuvent, du point de vue national, être organisés librement au niveau du plan d'études, des supports pédagogiques et de l'enseignement, pour autant que l'atteinte des standards minimaux soit garantie. Les standards de performance ne laissent pas seulement une marge de manœuvre pour l'élaboration des plans d'études et l'aménagement de l'enseignement, ils peuvent également servir de base pour la définition d'objectifs de performance et le développement de tests, par ex. de tests portant sur la vérification de l'atteinte de standards minimaux.

Dès ses débuts, le projet HarmoS s'est limité à l'élaboration de standards pour quatre domaines disciplinaires jugés fondamentaux et particulièrement importants en vue de contribuer à faciliter la mobilité entre les cantons; il s'agit de la langue de scolarisation, des langues étrangères, des mathématiques et des sciences naturelles. En ce qui concerne les langues, des travaux avaient déjà été lancés dans les années 1990 en vue de coordonner leur enseignement. Le 25 mars 2004, l'Assemblée plénière de la CDIP adopte, au terme de ce processus, sa stratégie pour l'enseignement des langues (CDIP, 2004a), qui prévoit notamment, parmi les différents instruments de coordination envisagés, que soient définis des niveaux de compétence devant obligatoirement être atteints et dont l'atteinte peut être vérifiée, dans la langue de scolarisation comme dans les langues étrangères. Constat est d'ailleurs fait qu'il est possible de mettre à profit l'expérience acquise dans le cadre de l'étude PISA pour la description et la vérification de compétences dans les domaines d'enseignement que sont la langue de scolarisation, les mathématiques et les sciences naturelles. Quant à savoir s'il faut ou non intégrer d'autres domaines disciplinaires – comme une matière artistique ou le sport, en discussion au début – la question est dans un premier temps reportée, afin notamment de ne pas charger excessivement le projet (cf. point 1.1.4: décision de mise en œuvre relative au concordat HarmoS).

Pour la mise en œuvre de la première et de la deuxième étapes du projet HarmoS, la CDIP travaille en collaboration avec des consortiums de didactique disciplinaire pour les quatre domaines choisis; durant la deuxième étape, un groupe spécialisé en méthodologie s'ajoute pour partager les compétences acquises lors de la réalisation des études PISA. Composés de représentantes et représentants des différentes régions, ces consortiums préparent des modèles de compétences sur ce qui est unanimement considéré comme faisant partie du domaine noyau de leur discipline respective, et traduisent des parties de ces modèles en tâches. Au printemps 2007, des mesures de performances sont réalisées en Suisse alémanique et en Suisse romande auprès d'échantillons, représentatifs au niveau régional, d'élèves arrivant au terme du degré primaire et à la fin du degré secondaire I. Le but de ce relevé est de valider les modèles de compétences et les tâches élaborées. Par manque de temps et de moyens financiers, il a fallu renoncer à des analyses approfondies des résultats et à un processus de détermination de la valeur seuil régi par des normes scientifiques. Du fait que l'étude de validation n'a pas spécifiquement porté sur les compétences minimales, il ne s'est finalement pas non plus avéré possible d'obtenir, au travers des tâches et des résultats des tests, autant d'informations qu'espéré sur les performances des élèves possédant les compétences minimales définies (cf. rapport final du groupe Méthodologie: Ramseier et al., 2008). Les consortiums proposent néanmoins, à la suite d'un processus de consolidation, des standards minimaux pour la 2e année (sauf en langues étrangères), la 6e et la 9e années (selon l'ancienne numérotation) (cf. rapports scientifiques des consortiums: Consortium HarmoS Langue de scolarisation, 2010; Consortium HarmoS Langues étrangères et al., 2009; Consortium HarmoS Mathématiques et al., 2009; Consortium Sciences naturelles+, 2009).

Quant à la troisième étape du projet HarmoS, à savoir la détermination de standards contraignants par la politique en matière d'éducation, elle ne peut être achevée qu'à un stade ultérieur. En 2010, les standards de performance élaborés par les consortiums font dans un premier temps l'objet d'une procédure d'audition (cf. art. 7 du concordat HarmoS). S'ensuit un remaniement en collaboration avec les consortiums disciplinaires, puis les standards de formation nationaux sont finalement adoptés par l'Assemblée plénière, le 16 juin 2011, sous l'appellation de «compétences fondamentales». Les quatre domaines disciplinaires font chacun l'objet d'une brochure, qui décrit de manière générale les compétences fondamentales devant être acquises au terme des degrés d'enseignement (CDIP, 2011b, 2011a, 2011c, 2011d). Les objectifs de formation adoptés s'adressent en premier lieu aux professionnels qui élaborent les plans d'études, les moyens d'enseignement et les outils d'évaluation. Il est prévu de les intégrer dans les plans d'études régionaux déjà disponibles (plan d'études romand) ou en préparation (*Lehrplan 21* pour les cantons alémaniques[1] et *piano di studio* pour le canton du Tessin), ce qui effectivement se concrétise par la suite.

Les *compétences fondamentales* décrivent les compétences que doivent avoir acquises (pratiquement) tous les élèves au terme des trois degrés d'enseignement de la scolarité obligatoire (4e, 8e et 11e années, langues étrangères en 8e et en 11e années). Pour chacun des domaines disciplinaires, la brochure sur les objectifs de formation décrit un modèle de compétences qui aide à articuler la structure du domaine et ainsi à établir la planification des tests. Dans sa dimension verticale, la progression des compétences n'est visible qu'au fil des degrés scolaires. La description donnée pour chaque degré présente uniquement ce que cela signifie que d'avoir acquis les compétences fondamentales. La différenciation ne va pas au-delà[2]. C'est la raison pour laquelle les tests basés sur ces standards doivent (uniquement) servir à vérifier l'atteinte du seuil correspondant aux «compétences fondamentales».

1.1.3 Articles constitutionnels sur la formation

Parallèlement aux travaux menés en vue de compléter le concordat scolaire au niveau intercantonal, des efforts sont aussi déployés, dès le début, au niveau fédéral afin d'harmoniser les systèmes scolaires en Suisse et conduisent finalement à une modification de la Constitution. Le 21 mai 2006, le peuple suisse

[1] Une version spécifique du *Lehrplan 21* est élaborée pour la Suisse romanchophone.

[2] Initialement, il était prévu de décrire également, en plus des compétences fondamentales, des niveaux de compétences plus avancés. Les consortiums disciplinaires avaient même présenté un certain nombre de propositions allant dans ce sens. Ces travaux n'ont toutefois pas pu, par manque de ressources principalement, être menés à terme. Et il n'a pas non plus été possible de concevoir les descriptions des objectifs de formation de la même manière pour les quatre domaines disciplinaires.

accepte, par votation, les articles révisés de la Constitution fédérale sur l'éducation. Depuis lors, les cantons sont notamment tenus d'harmoniser à l'échelon national certains éléments importants concernant la scolarité obligatoire, à savoir l'âge de l'entrée à l'école, la durée et les objectifs des niveaux d'enseignement et le passage de l'un à l'autre, ainsi que la reconnaissance des diplômes (art. 62, al. 4, de la Constitution fédérale). Cette disposition confère également à la Confédération une compétence subsidiaire pour légiférer dans l'hypothèse où les efforts de coordination intercantonale n'aboutiraient pas à des solutions satisfaisantes. Le mandat consistant à harmoniser les objectifs des niveaux d'enseignement est d'une importance primordiale pour le développement et la mise en œuvre des objectifs de formation et, partant, de la vérification de l'atteinte des compétences fondamentales.

1.1.4 Concordat HarmoS

Près d'un an plus tard, le 14 juin 2007, l'Assemblée plénière de la CDIP adopte l'accord intercantonal sur l'harmonisation de la scolarité obligatoire (concordat HarmoS; CDIP, 2007a)[3], qui prévoit à son art. 7 la définition de standards nationaux de formation.

Concordat HarmoS

Art. 7 Standards de formation

[1] Aux fins d'harmoniser les objectifs de l'enseignement dans l'ensemble du pays sont établis des standards nationaux de formation.

[2] Ces standards de formation peuvent être de deux ordres, à savoir:

a. des standards de performance fondés, par domaine disciplinaire, sur un cadre de référence incluant des niveaux de compétence;
b. des standards qui déterminent des contenus de formation ou des conditions de mise en œuvre dans l'enseignement.

[3] Les standards nationaux de formation sont construits et validés scientifiquement sous la responsabilité de la CDIP. Ils doivent faire l'objet d'une consultation au sens de l'art. 3 du concordat sur la coordination scolaire du 29 octobre 1970.

[4] Ils sont adoptés par l'Assemblée plénière de la CDIP à la majorité des deux tiers de ses membres, parmi lesquels doivent figurer les représentantes ou représentants d'au moins trois cantons à majorité linguistique non germanophone. Ils sont révisés par les cantons concordataires selon une procédure analogue.

Art. 8 Plans d'études, moyens d'enseignement et instruments d'évaluation

[1] L'harmonisation des plans d'études et la coordination des moyens d'enseignement sont assurées au niveau des régions linguistiques.

[2] Plans d'études, moyens d'enseignement et instruments d'évaluation, ainsi que standards de formation sont coordonnés entre eux.

[3] Les cantons collaborent au sein des régions linguistiques à la mise en œuvre du présent accord. Ils peuvent prendre les dispositions d'organisation nécessaires à cet effet.

[4] La CDIP et les régions linguistiques se concertent au cas par cas pour développer des tests de référence sur la base des standards de formation.

Art. 10 Monitorage du système d'éducation

[3] Pour obtenir des informations sur le contexte et la genèse du concordat HarmoS, il est possible de se référer au commentaire rédigé alors par le Secrétariat général de la CDIP (Ambühl, 2011).

[1] En application de l'art. 4 du concordat sur la coordination scolaire du 29 octobre 1970, les cantons concordataires participent avec la Confédération à un monitorage systématique, continu et scientifiquement étayé de l'ensemble du système suisse d'éducation.

[2] Les développements et les performances de l'école obligatoire sont régulièrement évalués dans le cadre de ce monitorage. La vérification de l'atteinte des standards nationaux de formation, notamment au moyen de tests de référence au sens de l'art. 8, al. 4, fait partie intégrante de cette évaluation.

L'art. 7, al. 2, du concordat fait mention de deux types de standards: ceux qui ont trait aux performances, tels que les objectifs nationaux de formation, et ceux qui se rapportent aux contenus de formation et à l'aménagement de l'enseignement. Les standards correspondant à cette deuxième catégorie sont prévus pour les domaines disciplinaires tels que les arts visuels ou le sport, dans lesquels il n'est guère envisageable d'appliquer des standards de performance.

L'art. 7, al. 4, du concordat prévoit une procédure à suivre pour la révision des objectifs de formation. De manière générale, cette possibilité s'avère nécessaire, compte tenu de la constante évolution que connaît le système éducatif, et elle peut aussi être mise à profit si l'on vient par exemple à constater, lors de la vérification de l'atteinte des compétences fondamentales, que les objectifs de formation, tels qu'adoptés, ne donnent pas entière satisfaction. À cet égard, il importe de rappeler que les objectifs de formation en langues étrangères ont été définis avant que soit appliquée la décision d'avancer le début de l'apprentissage des langues étrangères, et que donc les effets ne pouvaient que difficilement être anticipés.

L'art. 8 crée la base nécessaire pour le développement, ultérieur ou déjà mené à terme (dans le cas de la Suisse romande), de plans d'études régionaux communs et en confie la responsabilité aux conférences régionales ou sous-régionales de la CDIP, qui sont également chargées de la coordination des moyens d'enseignement. Les standards de formation communs sont ainsi intégrés dans les plans d'études.

L'art. 10 établit, en relation avec l'art. 8, al. 4, du concordat, la base sur laquelle reposent le développement et la réalisation des enquêtes de vérification de l'atteinte des compétences fondamentales, le but étant que celle-ci participe à un monitorage systématique, continu et scientifiquement étayé de la scolarité obligatoire.

Le concordat HarmoS entre en vigueur le 1er août 2009. Une fois que le concordat a été ratifié par une majorité des cantons, où vit environ 80 % de la population résidente, l'harmonisation des éléments structurels de la scolarité obligatoire et des objectifs nationaux de formation voulue par la Constitution s'effectue de manière générale en fonction des solutions prévues dans le concordat. En l'état actuel, le nombre de cantons ayant adhéré au concordat est de quinze (CDIP, 2021a).

Les objectifs de formation et les tests de référence sont davantage précisés dans la décision de mise en œuvre relative au concordat HarmoS prise lors de l'assemblée plénière des 25 et 26 octobre 2007 (CDIP, 2007b).

- Décision est prise de définir des objectifs nationaux de formation qui fixent des standards pour les domaines disciplinaires que sont la langue de scolarisation, les langues étrangères, les mathématiques et les sciences naturelles, mais aussi, à un stade ultérieur en mettant à profit l'expérience acquise, pour de nouveaux domaines en donnant la priorité aux TIC et à l'éducation au mouvement, à la musique et aux arts visuels (points 5.1, 3.1).
- Les épreuves de référence (points 5.3 à 5.7) qu'il s'agit de développer doivent remplir les deux fonctions suivantes:
  1) évaluation du système sur le plan national et le plan régional ou cantonal dans le cadre du monitorage de l'éducation; l'élaboration de ces instruments est confiée à la CDIP (point 5.3.1).
  2) évaluation individuelle des élèves pour prévoir un encouragement adapté, notamment en vue du passage au degré secondaire II; l'élaboration de ces instruments est coordonnée par la CDIP mais s'effectue au niveau des régions linguistiques.
- Il est décidé que l'évaluation du système par la vérification de l'atteinte des compétences fondamentales ainsi que par l'étude PISA doivent dès lors fonctionner de manière complémentaire; en d'autres termes,

une fois mise en place la vérification de l'atteinte des compétences fondamentales, PISA doit uniquement servir à des fins de comparaison internationale, et non plus intercantonale[4]. Le texte même de la décision ne précise pas explicitement que le critère de comparaison change aussi dans les comparaisons intercantonales réalisées à travers la vérification: en effet, si les élèves sont classés sur des échelles de compétences dans l'étude PISA, la comparaison dans le cadre des enquêtes de vérification porte uniquement sur la proportion d'élèves qui atteignent ou non les compétences fondamentales.

- Il est également précisé que les épreuves de référence ne doivent pas pouvoir être utilisées à des fins de classement des écoles (*ranking*) ni d'évaluation des performances du corps enseignant.

Complément d'information sur les bilans de compétences individuels: les bilans de compétences individuels à des fins formatives, mentionnés dans la décision relative à la mise en œuvre et dont l'élaboration incombe aux régions, ne sont pour l'instant pas proposés partout ni réalisés de la même manière dans les différentes régions et les cantons. En outre, les tests cantonaux ou soutenus par les cantons ne visent pas toujours premièrement la planification individuelle des apprentissages, mais peuvent aussi porter par exemple sur le monitorage du système scolaire. Au moment où la banque de données de tâches (BDT) de la CDIP est mise en place, la CDIP partait encore du principe que les tests aujourd'hui appelés *Checks* effectués dans l'espace de formation Suisse du Nord-Ouest (*Bildungsraum Nordwestschweiz*), la palette de tests *Stellwerk* des éditions scolaires saint-galloises et la BDT utiliseraient des formats compatibles, qui permettraient des échanges de tâches. Cette idée ne s'est toutefois pas imposée. À ce jour, la BDT n'a pas pu intervenir dans l'établissement de bilans de compétences. La CDIP alémanique a essayé, aux alentours de 2012, de mettre en place une coordination des mesures de performances pour la Suisse alémanique, mais ce projet n'a pas abouti. La coordination s'est faite à plus petite échelle. L'espace de formation Suisse du Nord-Ouest (c'est-à-dire les cantons d'Argovie, de Bâle-Campagne, de Bâle-Ville et de Soleure) introduit, en utilisant les *Checks*, un système de tests complet pour l'établissement de bilans individuels répétés. Le canton de Zoug utilise lui aussi les *Checks* de manière ponctuelle. Des séries de tâches appelées *Mindsteps* servent de complément facultatif aux *Checks* à des fins d'utilisation formative par les écoles. Dans la même ligne que ce qu'offrent *Checks* et *Mindsteps*, les cantons de Saint-Gall et de Zurich ont développé *Stellwerk* pour disposer de systèmes d'apprentissage et de test nommés *Lernlupe* (degré primaire) et *Lernpass Plus* (degré secondaire I). Ces deux cantons, de même que de nombreux autres cantons alémaniques et la principauté du Liechtenstein[5], appliquent *Stellwerk* (intégré à présent dans *Lernpass Plus*) pour établir des bilans de compétences, souvent en lien avec le choix de la profession. L'utilisation de *Stellwerk* varie dans sa portée selon les cantons, et les compétences en la matière tout comme le caractère contraignant qui y est associé ne sont pas non plus toujours les mêmes. Au Valais, des épreuves cantonales sont menées dans les deux parties linguistiques du canton. Dans le reste de la Suisse romande, la pratique varie fortement d'un canton à l'autre. Des épreuves certificatives et formatives sont organisées, de même que des examens menés à des fins de monitorage cantonal. Les cantons de Genève, du Jura, de Neuchâtel, de Vaud et du Valais organisent des examens répétés dans différentes disciplines. Depuis 2016, l'Institut de recherche et documentation pédagogique (IRDP) travaille, sur mandat de la Conférence intercantonale de l'instruction publique de la Suisse romande et du Tessin (CIIP), à développer, dans le cadre du projet EpRoCom, une banque romande d'items fondés sur les objectifs et progressions du Plan d'études romand (PER), afin de les mettre à la disposition des cantons et du corps enseignant. La première étape a consisté à sélectionner des exercices de mathématiques et de français de 8e année, parmi les épreuves élaborées par les cantons romands, et à les soumettre à un processus de validation. Un dispositif complet révisé existe depuis 2020 pour le projet *EpRoCom – Banque d'items*. En 2021, une sélection de tâches a été mise à la disposition du corps enseignant sur la plateforme du PER, à des fins d'illustration principalement. Quant au canton du

[4] Dès lors que la participation à PISA sert uniquement à la comparaison à l'échelle internationale, la Suisse ne sélectionne plus qu'un échantillon national d'élèves de 15 ans pour prendre part à l'étude, comme c'est le cas habituellement dans les autres pays; il n'est donc plus question d'avoir recours à un échantillon élargi, qui serait en même temps représentatif pour les élèves de 11e année de scolarité dans chaque canton.

[5] La principauté du Liechtenstein procède à des vérifications obligatoires de ses standards de formation en allemand et en mathématiques (5e, 7e et 10e années, plus l'anglais en 7e et en 10e) (CDIP/IDES, 2021).

Tessin, il dispose dans un certain nombre de branches d'épreuves cantonales et travaille à mettre en place un système de test et de monitorage intégrant des enquêtes nationales et cantonales. Pour l'instant, l'idée d'établir des bilans de compétences individuels à des fins formatives n'a pas été réalisée. Des épreuves cantonales standardisées sont organisées à la fin du degré primaire et du degré secondaire I (l'accent étant mis sur les mathématiques et sur l'italien), et les résultats obtenus servent de retours d'information au corps enseignant. D'autres examens cantonaux liés au plan d'études tessinois sont organisés, mais uniquement au secondaire I. Le choix des années et des matières à tester tout comme l'échantillonnage ne s'effectuent pas, pour l'heure, selon un concept fixe.

Dans le cadre de l'enquête qu'il effectue chaque année auprès des cantons, le centre d'information et de documentation IDES de la CDIP fait un relevé des épreuves externes obligatoires qui sont utilisées pour tester l'ensemble des élèves dans les cantons suisses et la principauté du Liechtenstein (CDIP/IDES, 2021).

1.1.5 Bureau de coordination HarmoS

La décision relative à la mise en œuvre du concordat HarmoS marque le passage, sous l'angle de l'harmonisation, d'une structure de projet (c'est-à-dire du projet HarmoS) à des structures permanentes au sein de la CDIP. Le Bureau de coordination pour la mise en œuvre de l'accord intercantonal sur l'harmonisation de la scolarité obligatoire (Bureau HarmoS) débute son activité dans le courant de 2008. Il s'agit d'une commission permanente de la CDIP au sens de l'art. 21 des statuts de la CDIP. Le Bureau HarmoS a pour mission d'assurer une mise en œuvre coordonnée du concordat HarmoS et remplit cette fonction en s'occupant de la question elle-même et en faisant le lien entre les différents niveaux d'intervention intercantonaux. Sa composition, ses tâches et son mode de fonctionnement sont définis dans le règlement du 24 janvier 2008 relatif au Bureau HarmoS (CDIP, 2013b) ainsi que dans le règlement d'organisation concernant la réalisation de la vérification de l'atteinte des compétences fondamentales (CDIP, 2014b).

1.2 Mise en œuvre de la vérification de l'atteinte des compétences fondamentales

1.2.1 Bases légales de la mise en œuvre

Le 20 juin 2013, l'Assemblée plénière de la CDIP adopte le projet relatif à la réalisation des deux premières enquêtes à mener dans le cadre de la vérification de l'atteinte des compétences fondamentales et décide du lancement des premières enquêtes (CDIP, 2013a). Le règlement d'organisation sera ensuite édicté le 8 mai 2014.

La décision adoptée le 20 juin 2013 prévoit que la vérification de l'atteinte des compétences fondamentales porte sur les mathématiques en 11^{e} année (selon numérotation HarmoS) en 2016 ainsi que sur la langue de scolarisation et la première langue étrangère dans des classes de 8^{e} année en 2017. De plus, les considérations exposées en rapport avec cette décision prévoient également un cycle d'enquête sur les mathématiques et les sciences naturelles en 2019 auprès d'élèves de 4^{e} année. Ce projet n'est toutefois pas mis à exécution.

Le projet élaboré pour la vérification détaille en particulier la structure organisationnelle, un calendrier pour les enquêtes, la forme et l'étendue de l'échantillonnage parmi les élèves, et présente également une planification financière. Le SG CDIP se voit ainsi attribuer le rôle de service de coordination, tandis que la BDT et un consortium scientifique (encore à mandater) se répartissent l'exécution des travaux (pour plus d'informations à ce sujet, se référer au point 1.2.3). Le calendrier établi pour les futures enquêtes montre qu'à partir de l'enquête 2017, il deviendrait possible d'obtenir des coupes longitudinales au niveau des individus, de manière à suivre la progression des élèves durant leur temps de scolarité. Lors de la mise en œuvre de la vérification, il a été décidé de ne pas prélever systématiquement de données longitudinales au niveau de l'individu mais de se limiter plutôt à des études longitudinales par cohortes, donc en procédant à des relevés auprès d'échantillons de la même volée. Ainsi, si la pandémie de coronavirus n'avait pas, au dernier moment, empêché ce projet, les compétences linguistiques auraient fait l'objet de tests en 8^{e} année en 2017, puis en 11^{e} année en 2020.

D'après le règlement d'organisation, le but de la vérification est d'évaluer les performances du système de la scolarité obligatoire et de fournir aux cantons des informations au niveau systémique leur indiquant dans quelle mesure les compétences fondamentales sont atteintes à certains degrés de l'enseignement. De ce fait, l'accent est mis sur les résultats des tests mesurés aux objectifs de formation. Le règlement d'organisation ne fait pas mention de la saisie de données contextuelles (par ex. relatives au milieu social des élèves), qui peuvent servir à l'interprétation des résultats.

Le règlement d'organisation définit en particulier l'organisation, les compétences et le pilotage de la vérification par la CDIP, et notamment

- la participation des cantons aux enquêtes;
- les compétences du ressort de l'Assemblée plénière (dont l'aval des résultats de la vérification en vue de l'établissement des rapports et l'approbation de la définition de la valeur seuil correspondant aux compétences fondamentales);
- les tâches du Bureau de coordination HarmoS dans le contexte de la vérification;
- les qualifications et compétences du consortium scientifique chargé de la réalisation de la vérification;
- la sécurité des données et partant la gestion des données collectées.

À la suite de la décision relative aux premières enquêtes, prise en 2013, l'Assemblée plénière de la CDIP approuve, le 22 juin 2017 (CDIP, 2017), la réalisation des deux enquêtes suivantes à mener dans le cadre de la vérification. Il est ainsi décidé que l'atteinte des compétences fondamentales serait vérifiée en 2020 auprès d'élèves de 11e année dans la langue de scolarisation mais aussi dans la première et la deuxième langues étrangères, et en 2022, auprès d'élèves de 4e année dans deux autres domaines disciplinaires. Le 25 octobre 2019 (CDIP, 2019c), l'Assemblée plénière de la CDIP décide qu'il s'agira de la langue de scolarisation et des mathématiques. Cependant, l'enquête 2020 doit être interrompue à cause de la pandémie, et l'Assemblée plénière de la CDIP fixe un nouveau calendrier le 25 juin 2020: celui-ci prévoit que l'enquête en 11e année aura lieu en 2023, et celle à mener dans des classes de 4e année, en 2024 (CDIP, 2020b).

1.2.2 Banque de données de tâches (BDT)

Le 25 octobre 2012, l'Assemblée plénière de la CDIP décide de créer la Banque de données de tâches (BDT) de la CDIP (CDIP, 2012b), et le 23 janvier 2014, le Comité de la CDIP édicte le règlement d'organisation de la Banque de données de tâches de la CDIP (règlement BDT) (CDIP, 2014a). Selon ce règlement, la BDT se voit confier une série de tâches qui revêtent une importance primordiale pour la mise en œuvre des art. 8 et 10 du concordat HarmoS (tests de référence réguliers):

a) mettre à disposition une structure technique multifonctions et plurilingue;
b) garantir une échelle édumétrique applicable dans toute la Suisse;
c) instaurer un processus pour le développement des tâches et des tests et pour l'étalonnage de tâches par rapport aux compétences fondamentales;
d) coordonner l'échange des tâches servant à la vérification de l'atteinte des objectifs nationaux de formation (compétences fondamentales) ainsi que l'échange des tâches servant aux projets cantonaux.

En tant que «gardienne des compétences fondamentales» (ayant pour mission d'étalonner les tâches sur les compétences fondamentales et de garantir une échelle édumétrique applicable dans toute la Suisse), la BDT occupe une position différente de celle notamment des institutions qui participent par la suite, sur la base de conventions conclues avec la CDIP, à la réalisation des enquêtes. La BDT est placée sous la responsabilité de la CDIP, qui confie les tâches stratégiques et opérationnelles au Bureau de coordination HarmoS et au secrétariat de la BDT. Celui-ci est rattaché au Centre suisse de coordination pour la recherche en éducation (CSRE). Décision a été prise le 26 octobre 2018 de poursuivre pour le moment l'exploitation de la BDT jusqu'à 2025.

Dans la mise en œuvre, la BDT assume finalement un ensemble de tâches moins étendu que prévu dans le règlement. Des travaux ont certes été lancés en vue de constituer une banque de données de tâches

multifonctions au sens de la let. a, mais sa maintenance est suspendue, principalement parce que les possibilités qu'elle offre ne semblent pas rencontrer de demande du côté des régions et des cantons. La plateforme technique qui est utilisée dans un premier temps est le système d'évaluation en ligne TAO, qui est aussi employé lors de l'enquête 2016 (en 11e année). Pour la réalisation de cette dernière, TAO est utilisé via internet au moyen des infrastructures des écoles elles-mêmes; une solution équivalente consisterait à employer TAO à partir de clés USB. Dans le cadre de l'enquête 2017 (en 8e année), on passe à l'utilisation d'appareils (tablettes dans des mallettes) qui sont emmenées dans les écoles pour le relevé des données. La raison principale en est que les infrastructures informatiques sont moins développées dans les écoles primaires. Cette nouvelle solution consistant à emmener le matériel sur place s'accompagne aussi d'un changement de logiciel pour les tests; le choix se porte sur TBA ItemBuilder de l'institut DIPF (*Leibniz-Institut für Bildungsinformation und Bildungsforschung*). Les collaborateurs et collaboratrices de la HTW Chur (intégrée aujourd'hui dans la haute école spécialisée des Grisons, la FHGR), chargés de la mise en œuvre technique de la vérification, connaissent déjà le logiciel pour l'avoir employé dans d'autres contextes, tout comme les personnes chargées du développement des tâches pour les tests en langues étrangères. L'échange des tâches prévu à la let. d ne s'est développé que de manière tout à fait rudimentaire, et ce via la validation d'une partie des tâches des cycles de vérification menés en 2016 et 2017. Ces tâches peuvent être obtenues, comme les données de la vérification, auprès du Centre de compétences suisse en sciences sociales FORS de l'Université de Lausanne, ce que *Stellwerk* a aussi fait.

Les travaux accomplis jusqu'ici par la BDT se concentrent sur les domaines de tâches b) et c). Il s'agit des travaux réalisés dans le contexte des premiers cycles de vérification menés en 2016 et 2017 de même que des travaux déjà menés à terme pour préparer l'enquête 2023 (initialement prévue pour 2020) et ceux qui sont encore en cours pour l'enquête 2024 (initialement prévue pour 2022). Pour les quatre enquêtes, la BDT a développé, en collaboration avec des groupes de didacticiennes et didacticiens des disciplines, des tâches et des tests en vue de la vérification de l'atteinte des compétences fondamentales, analysé les données issues des tests conduits durant les enquêtes pilotes et principales de même que dirigé le processus de détermination des valeurs seuils dans les précédents cycles de vérification et contribué au compte rendu des résultats.

1.2.3 Consortium scientifique, direction du projet et commission Vérification

La CDIP travaille très tôt dans la perspective d'un transfert de la direction scientifique du projet à un consortium scientifique conduit par une haute école universitaire, lequel serait chargé de coordonner la réalisation des enquêtes et d'assurer l'ancrage de la vérification dans le monde scientifique. Cette perspective a été inscrite à l'art. 7 du règlement du 8 mai 2014 concernant la réalisation de la vérification.

Dans le but de constituer un consortium scientifique, plusieurs tables rondes sont organisées à partir de 2012 pour les personnes intéressées. Le but de ces rencontres est d'élaborer, avec la contribution de spécialistes de l'édumétrie, des critères pour une institutionnalisation durable et réussie des structures du projet dans une ou plusieurs hautes écoles en tenant compte des besoins spécifiques aux régions linguistiques. Durant cette phase une difficulté est identifiée dans le fait que l'on ne dispose pas d'une expérience suffisante en matière d'évaluations à large échelle dans le contexte helvétique. Par ailleurs, le financement serré ne permet de loin pas de mettre sur pied une structure scientifique fonctionnelle qui puisse porter le projet de la vérification.

En 2014, la mise en place d'un consortium scientifique n'est pas suffisamment avancée pour assumer la responsabilité scientifique de la réalisation des deux premiers cycles de vérification. C'est pourquoi le Secrétariat général de la CDIP nomme une direction de projet pour la vérification, laquelle se charge des activités opérationnelles dans l'intervalle précédant la mise sur pied du consortium scientifique. Outre le SG CDIP, y sont représentés le secrétariat de la BDT, les centres régionaux s'occupant de la mise en œuvre et qui se chargent déjà de l'étude PISA (Service de la recherche en éducation SRED, Haute école pédagogique de Saint-Gall [PHSG], *Scuola universitaria professionale della Svizzera Italiana* [SUPSI]), la *Hochschule für Technik und Wirtschaft* HTW Chur, qui est responsable des aspects techniques (et qui est aujourd'hui intégrée dans la FHGR), l'*Institut für Bildungsevaluation* IBE, responsable de l'échantillonnage

(institut associé de l'Université de Zurich) ainsi que le Centre de compétences suisse en sciences sociales FORS de l'Université de Lausanne, qui a pour tâche le traitement et la mise à disposition des données. Fait également partie de la direction de projet la direction du groupe chargé de constituer le futur consortium scientifique.

Afin de compléter ces structures opérationnelles de manière appropriée au niveau stratégique, le Bureau de coordination HarmoS décide, lors de sa séance du 21 novembre 2014, de tirer de ses rangs une commission pour la vérification de l'atteinte des compétences fondamentales (commission Vérification), qui agit en tant qu'élément représentant la coordination et la gouvernance vis-à-vis des institutions scientifiques. Toutes les régions linguistiques de même que l'espace de formation Suisse du Nord-Ouest (*Bildungsraum Nordwestschweiz*) y sont représentés.

Le groupe chargé de constituer le futur consortium scientifique, composé de représentantes et représentants de l'Université de Berne (direction), de l'Université de Genève et de la *Scuola universitaria professionale della Svizzera italiana* (SUPSI), élabore jusqu'à fin 2016 et en accord avec d'autres institutions un *règlement d'organisation du Consortium scientifique pour la recherche et l'évaluation en éducation* (WIKO BFE). Le Bureau de coordination HarmoS en prend connaissance et l'approuve le 27 janvier 2017. Le projet de règlement d'organisation du 7 novembre 2016 prévoit que le consortium se charge de la coordination et de la mise en œuvre de projets d'intérêt national dans le domaine de la recherche et de l'évaluation en éducation, notamment PISA et la vérification de l'atteinte des compétences fondamentales; il précise également que le consortium veille à favoriser l'intégration scientifique durable de projets de ce type, qu'il vise à ancrer ces projets dans les réseaux scientifiques et qu'il agit en tant qu'interlocuteur et partenaire de discussion pour les autorités responsables de l'éducation. Les organes du WIKO BFE sont les suivants:

a) un conseil scientifique assumant des fonctions de conseil, de direction et de décision (Universités de Berne et de Genève, SUPSI; év. d'autres membres dans un rôle consultatif),
b) une chaire de professeur/e pour la recherche et l'évaluation en éducation de l'Université de Berne,
c) une direction de projet,
d) une Assemblée des délégués, remplissant un rôle consultatif et constituée de représentantes et représentants de hautes écoles suisses, intéressés par le domaine, de même que
e) un conseil d'expert/e/s internationaux, constitué de trois spécialistes issus de régions de langues différentes.

En septembre 2018, la CDIP et l'Université de Berne concluent un accord de coopération en vue du transfert de la direction scientifique des enquêtes 2020 et 2022, et donc de la coordination du consortium scientifique. Le SG CDIP se retire ainsi de la direction de projet. C'est l'*Interfaculty Centre for Educational Research* (ICER) qui reprend la direction de la vérification (comme, ultérieurement, celle de PISA 2021, respectivement 2022). L'ICER de l'Université de Berne travaille, dans le cadre de ses fonctions de direction, avec les mêmes partenaires que précédemment le SG CDIP, à l'exception du centre FORS. Si FORS ne fait plus partie du projet, il continue toutefois, dans le cadre de sa mission de service public, de mettre à disposition les données issues de la vérification sur ses plateformes FORSbase, puis SWISSUbase (depuis 2022). Sont représentés dans la direction de projet: l'ICER, la BDT, la FHGR, l'IBE, la PHSG, le SRED et la SUPSI. L'ICER (ou plus précisément l'Université de Berne) conclut directement des conventions avec tous les partenaires, sauf avec la BDT et en partie la FHGR. La BDT fait partie de la CDIP; avec la FHGR, la CDIP a signé un accord de coopération à part, à titre de solution transitoire, portant sur les prestations informatiques à fournir en lien avec l'enquête 2020 (ou plutôt 2023). L'accord de coopération conclu entre la CDIP et l'Université de Berne prévoit que cette dernière (et partant l'ICER) assume, dans le cadre de la vérification, des tâches relevant d'une part de la coordination et, de l'autre, de la direction. Conformément à ce que prévoit un manuel élaboré par le SG CDIP sur la réalisation d'enquêtes à large échelle, une série de groupes de travail thématiques (GT) sont constitués pour se charger des travaux à effectuer: GT Développement des tests, GT Échantillonnage, GT Questionnaire contextuel, GT Livraison des tests, GT Gestion de l'enquête, GT Traitement des données, GT Archivage et diffusion, GT Détermination de la valeur seuil, GT Premier rapport.

La nouvelle structure définie conformément à la convention de coopération conclue entre la CDIP et l'Université de Berne ne correspond pas à une mise en application directe du projet de règlement d'organisation esquissé pour le consortium. On relèvera notamment que la responsabilité scientifique incombe bien plus fortement que prévu dans le projet de règlement à une seule et même haute école, à savoir l'Université de Berne. Celle-ci fournit également le comité directeur de l'ICER, qui assume les fonctions équivalant au conseil scientifique, envisagé de manière interinstitutionnelle, de même qu'à la chaire de professeur/e mentionnée à la let. b du projet de règlement d'organisation. Une chaire spécialisée n'a à ce jour pas été créée à l'Université de Berne. Les tâches relevant de la direction de projet sont essentiellement assumées par la direction de l'ICER, qui fait partie de la direction du projet Vérification et qui se charge de sa coordination et d'une partie de sa direction. L'assemblée de délégués des hautes écoles intéressées dont il est fait mention à la let. d n'existe pas en tant que telle. Cependant, l'ICER s'est engagé à réunir à intervalles réguliers ce groupe cible en organisant des tables rondes. Par ailleurs, les chercheurs et chercheuses qui sont intéressés ont la possibilité, à certaines conditions, d'intégrer leurs intérêts de recherche dans le questionnaire contextuel utilisé dans le cadre de la vérification. Quant au conseil d'expert/e/s internationaux, il a été composé selon d'autres principes que ceux décrits à la let. e: il s'agit d'un organe de l'ICER, ce qui signifie qu'il traite également de projets de l'ICER qui ne concernent pas spécifiquement des intérêts nationaux relevant du domaine de la recherche et de l'évaluation en éducation.

1.2.4 Collaboration avec la didactique des disciplines

Comme mentionné au point 1.1.2, des représentantes et représentants des didactiques des quatre domaines disciplinaires pour lesquels des standards de formation ont été développés sont activement impliqués dans les travaux de la CDIP dès le début. Cependant, les didactiques des disciplines ne sont pas représentées en permanence dans les structures de la vérification. La CDIP procède plutôt de cette manière: lorsqu'il s'agit de vérifier l'atteinte des compétences fondamentales dans un domaine disciplinaire, elle conclut avec les institutions des didacticiennes et didacticiens intéressés une convention de coopération portant sur le développement des tâches et des tests ainsi que sur la contribution à la détermination des valeurs seuils. Les trois plus grandes régions linguistiques sont généralement chacune représentées dans les consortiums de didactique disciplinaire. Seule exception à cette règle: les langues étrangères lors de l'enquête 2017. Le développement des tâches est alors entrepris uniquement par l'Institut de plurilinguisme de l'Université de Fribourg et de la Haute école pédagogique de Fribourg. Les consortiums de didactique disciplinaire organisent leur collaboration en concertation avec le secrétariat de la BDT et le SG CDIP directement. Les didacticiennes et didacticiens ont aussi la possibilité de participer à l'élaboration du questionnaire contextuel. Cet apport est fourni au travers des tables rondes du consortium scientifique, hors convention avec la CDIP. Certains représentants et représentantes des didactiques disciplinaires concernées contribuent également au premier rapport établi sur les résultats de la vérification. Des conventions séparées sont conclues à cet effet.

1.2.5 Le réseau des référentes et référents cantonaux

Afin de coordonner les enquêtes dans les cantons et notamment les contacts avec les écoles prenant part aux enquêtes, le Bureau de coordination HarmoS décide le 10 avril 2014 d'établir un réseau de référentes et référents cantonaux pour la vérification de l'atteinte des compétences fondamentales. Les cantons y sont généralement représentés par une personne ou, dans le cas des cantons plurilingues, par une personne par région linguistique. Sur mandat de leur département cantonal, ces personnes amorcent les enquêtes auprès des écoles en établissant le contact entre elles et les centres de réalisation régionaux. Pendant toute la durée du processus, elles se tiennent à la disposition tant des écoles que des centres de réalisation pour répondre aux questions liées à la mise en œuvre des enquêtes. Depuis la séance de lancement du 18 novembre 2014, les référentes et référents cantonaux se retrouvent généralement deux à trois fois par année pour recevoir des informations sur l'état d'avancement des travaux et discuter de questions de planification, de communication et de réalisation. Le 24 janvier 2019, le Comité de la CDIP approuve un mandat pour le Groupe de travail Référentes et référents cantonaux pour les évaluations à large échelle en Suisse.

1.3 Enquêtes menées dans le cadre de la vérification de l'atteinte des compétences fondamentales

Les sections suivantes décrivent les principaux éléments constitutifs des enquêtes 2016 et 2017. Ces deux premiers cycles de vérification sont menés alors que les nouveaux plans d'études, qui se rapportent aux objectifs de formation, commencent seulement à être appliqués. Il s'agit donc de mesures initiales, qui permettent d'obtenir une première référence.

Le descriptif proposé ci-après reprend largement ce qui est exposé dans le rapport plus complet qui a été rédigé sur les enquêtes 2016 et 2017 (CDIP, 2019a). En lien plus spécifiquement avec l'enquête 2016, il est fait référence au rapport d'audit (Fischbach & Ugen, 2018), qui a été commandé à ce sujet, et un aperçu est donné, pour les deux premiers cycles de vérification, sur la manière dont les résultats ont été présentés, la façon dont les données sont mises à disposition et les moyens qui ont été engagés pour ces enquêtes.

1.3.1 Enquête 2016: mathématiques en 11e année

La première enquête visant à vérifier l'atteinte des compétences fondamentales se déroule au printemps 2016. Elle porte sur les compétences fondamentales en mathématiques et est conduite auprès d'élèves arrivant au terme de leur scolarité obligatoire (donc en 11e année). Au total, 22 423 élèves issus de tous les cantons suisses y participent. Est responsable de la réalisation de l'enquête 2016 la direction de projet intérimaire décrite au point 1.2.3.

Développement des tâches et des tests

La base sur laquelle repose le développement des tâches et des tests de l'enquête 2016 est formée par les objectifs nationaux de formation pour les mathématiques adoptés en 2011 par la CDIP. Pour l'enquête 2016, des tâches sont développées en lien avec les cinq domaines de compétence de même qu'avec cinq des huit aspects de compétence contenus dans le modèle de compétence. Les tâches sont élaborées par des didacticiennes et didacticiens rattachés à des hautes écoles pédagogiques et des universités situées dans trois régions linguistiques de Suisse (d, f, i), en collaboration avec le secrétariat de la BDT. Les tâches sont soumises à plusieurs processus d'évaluation et de validation, et font l'objet d'une enquête pilote au printemps 2015 auprès d'un échantillon représentatif au niveau national. Durant la phase pilote, on a constaté que de nombreux items étaient plutôt difficiles à résoudre pour les élèves. Et certains des items ne présentaient pas la même difficulté pour les élèves des différentes régions linguistiques (*differential item functioning*, DIF), ce qui s'explique essentiellement par un certain nombre d'incohérences présentes dans les traductions. À la suite de cette phase pilote, les traductions sont donc remaniées, et des tâches supplémentaires développées pour le domaine inférieur des compétences. L'enquête principale est ainsi menée avec des tâches qui n'ont, pour certaines, pas été vérifiées dans le cadre de la phase pilote, ou qui ne l'ont pas été dans leur forme finale (cf. à ce propos Angelone & Keller, 2019a).

Questionnaire contextuel

Afin de pouvoir principalement identifier les raisons et les causes à l'origine des performances ou, plus précisément, des différences de performances, certaines informations contextuelles sont collectées au moyen d'un questionnaire destiné aux élèves. Le questionnaire contextuel comporte en particulier des questions sur l'origine sociale, la ou les langues parlées à la maison, le statut migratoire, les émotions ressenties vis-à-vis de l'école, les attitudes faces aux mathématiques et la motivation pour cette branche, de même que des aspects liés à la qualité, telle que perçue, de l'enseignement des mathématiques. Pour l'enquête 2016, deux versions de questionnaire sont développées, lesquelles présentent un chevauchement quant au milieu social mais couvrent sinon des aspects différents. Les deux versions sont réparties de manière aléatoire entre les différents élèves.

Réalisation de l'enquête

L'enquête principale de la vérification en 2016 est menée entre le 2 mai et le 10 juin; elle se déroule selon une procédure standardisée. La passation des tests est supervisée par des responsables formés à cet effet, qui se déplacent dans les écoles. Les élèves passent les tests de compétence et répondent au questionnaire en ligne, sur les ordinateurs de leur école. L'enquête dure en tout 3 heures et 15 minutes.

Détermination de la valeur seuil

La valeur seuil, qui est celle à partir de laquelle les compétences fondamentales sont considérées comme atteintes, est déterminée selon une procédure standardisée, par un groupe composé d'experts en didactique disciplinaire et d'enseignantes et enseignants issus de trois régions linguistiques de Suisse (d, f, i), à l'aide d'une méthode *Bookmark* adaptée. Parmi les didacticiennes et didacticiens du groupe d'experts, il en est qui ont aussi contribué au développement des tâches. La sélection des membres du groupe d'experts est effectuée par les secrétariats des conférences régionales de la CDIP (cf. Angelone & Keller, 2019a).

Résultats

L'enquête 2016 a principalement montré que près de deux tiers des élèves ont atteint les compétences fondamentales en mathématiques au terme de leur scolarité obligatoire. Cependant, on observe des différences relativement importantes entre les cantons, et celles-ci ne s'expliquent pas uniquement par la composition des classes. Ainsi, dans les régions francophones des cantons de Fribourg et du Valais, plus de 80 pour cent des élèves ont atteint les compétences fondamentales attendues en mathématiques, tandis que dans le canton de Bâle-Ville, cette proportion ne s'élève qu'à 44 pour cent, et dans les cantons de Berne (partie germanophone), Bâle-Campagne, Lucerne et Soleure, le pourcentage d'élèves atteignant les compétences fondamentales ne se monte qu'à environ 55 pour cent (cf. CDIP, 2019b; Nidegger, Roos, Petrucci, Verner, Helbling, Arnold, Erzinger, Pham, Schönenberger, et al., 2019).

1.3.2 Enquête 2017: langue de scolarisation et première langue étrangère en 8e année

La deuxième enquête réalisée dans le cadre de la vérification de l'atteinte des compétences fondamentales se déroule au printemps 2017. Elle porte sur les compétences fondamentales dans la langue de scolarisation (français, allemand ou italien) ainsi que dans la première langue étrangère (deuxième langue nationale ou anglais), et est conduite auprès d'élèves de 8e année. En tout, 20 177 élèves, issus de tous les cantons suisses, prennent part à l'enquête 2017. En raison de sa situation particulière, le canton des Grisons ne participe pas aux tests dans le domaine disciplinaire des langues étrangères.

Développement des tâches et des tests

Les tâches destinées à la langue de scolarisation sont développées par des didacticiennes et didacticiens de hautes écoles pédagogiques représentant trois régions linguistiques (PH FHNW, HEP Vaud, SUPSI), en collaboration avec le secrétariat de la BDT. Elles visent à tester l'atteinte des compétences fondamentales dans les domaines de compétence de la lecture et de l'orthographe. Les tâches sont soumises à plusieurs cycles d'évaluation et de validation, et font l'objet d'une enquête pilote au printemps 2016 auprès d'un échantillon représentatif au niveau national (cf. Angelone & Keller, 2019b).

Pour la première langue étrangère, les tâches sont développées par l'Institut de plurilinguisme de l'Université de Fribourg et de la Haute école pédagogique de Fribourg, en collaboration avec le secrétariat de la BDT. Les domaines de compétence testés sont la compréhension écrite et la compréhension orale. En Suisse romande, la première langue étrangère est l'allemand; au Tessin et dans les cantons alémaniques voisins de la Suisse romande, c'est le français (région de l'ancien plan d'études *Passepartout*). Dans les autres cantons de Suisse alémanique, la première langue étrangère enseignée est l'anglais. Les tâches sont soumises à plusieurs processus d'évaluation, et font d'abord l'objet d'une enquête pré-pilote dans trois régions linguistiques (d, f, i), avant d'être soumises au printemps 2016 à une enquête pilote auprès d'un échantillon représentatif au niveau national (cf. Angelone & Keller, 2019b).

Questionnaire contextuel

Pour l'enquête 2017, le questionnaire contextuel est élaboré par des experts issus d'institutions impliquées dans la direction de projet ou dans le développement des tâches, en collaboration avec des chercheurs et chercheuses ayant un intérêt spécifique et n'étant pas autrement impliqués dans le projet Vérification. Il est testé en 2016, dans le cadre de l'enquête pilote menée à l'échelle nationale. Il permet de collecter, en plus des informations sur l'origine sociale, le statut migratoire et la ou les langues parlées à la maison, une série d'aspects complémentaires, essentiellement ceux qui sont considérés comme particulièrement pertinents pour les compétences linguistiques acquises dans le cadre scolaire, par exemple ceux qui concernent l'apprentissage des langues, les contacts linguistiques, la motivation à apprendre les langues, les émotions en lien avec les langues et la perception de la manière dont l'enseignement est donné (cf. Erzinger et al., 2019).

Réalisation de l'enquête

L'enquête principale menée en 2017 se déroule entre le 24 avril et 2 juin. Les séances de tests sont là aussi supervisées par des responsables formés à cet effet, selon une procédure standardisée. Contrairement à ce qui s'est fait durant l'enquête 2016, les élèves passent les tests de compétence et répondent au questionnaire sur des tablettes, que les responsables des tests emmènent avec eux dans les écoles. L'enquête dure 2 heures et 55 minutes.

Détermination de la valeur seuil

Le processus de détermination de la valeur seuil pour la langue de scolarisation et les langues étrangères correspond à celui appliqué pour la valeur seuil en mathématiques. Il est à nouveau procédé selon une méthode *Bookmark* adaptée, sous la direction de la BDT. Les groupes d'experts sont constitués de didacticiennes et didacticiens ayant contribué au développement des tâches et d'autres n'y ayant pas participé, mais aussi d'enseignantes et enseignants exerçant au niveau primaire et d'enseignantes et enseignants spécialisés (cf. Angelone & Keller, 2019b).

Résultats

Les principaux résultats de l'enquête 2017 peuvent être synthétisés de la manière suivante (cf. CDIP, 2019b; Nidegger, Roos, Petrucci, Verner, Helbling, Arnold, Erzinger, Pham, Ambrosetti, et al., 2019):

- Les compétences fondamentales de lecture en langue de scolarisation sont atteintes par environ 88 % des élèves. Dans la partie francophone du canton de Fribourg, le résultat (91,3 %) se situe au-dessus de la moyenne suisse, et la différence est statistiquement significative. De manière statistiquement significative également, les taux enregistrés dans les cantons de Bâle-Ville (80,9 %) et de Glaris (80,5 %) se situent nettement au-dessous de la moyenne suisse.
- En orthographe, les élèves dont l'allemand est la langue de scolarisation atteignent en moyenne les compétences fondamentales à 84 % environ; ceux qui ont le français comme langue de scolarisation obtiennent un taux d'environ 89 % et ceux dont la langue de scolarisation est l'italien y parviennent à près de 80 %.
- Pour les cantons ayant l'allemand comme première langue étrangère enseignée (Suisse romande), le taux d'atteinte des compétences fondamentales en compréhension de l'oral est de 88 % en moyenne; il est d'environ 72 % pour la compréhension de l'écrit. Les différences entre les cantons sont minimes. Seule la partie francophone du canton de Fribourg se situe nettement au-dessus de la proportion moyenne d'atteinte des compétences fondamentales en compréhension de l'écrit (différence statistiquement significative).
- Dans les cantons ayant le français comme première langue étrangère enseignée (région *Passepartout* de la Suisse alémanique et Tessin), la proportion d'élèves ayant atteint les compétences fondamentales en compréhension de l'oral se situe aux alentours de 89 % en moyenne; en compréhension de l'écrit, le taux avoisine les 65 %. On observe toutefois des différences considérables entre les cantons. Dans le canton du Tessin comme dans la partie germanophone du canton de Fribourg, le pourcentage d'élèves ayant atteint les compétences fondamentales est plus élevé (différence statistiquement significative),

alors que dans les cantons de Bâle-Ville, de Bâle-Campagne et de Soleure, il se situe nettement au-dessous de la moyenne (différence statistiquement significative).
- Pour ce qui est de la compréhension de l'oral en anglais langue étrangère, la moyenne des élèves ayant atteint les compétences fondamentales est d'environ 95 %; pour la compréhension de l'écrit, elle est d'à peu près 86 %. Les différences entre les cantons sont minimes. Seuls les cantons de Schaffhouse (compréhension de l'écrit et de l'oral) et d'Appenzell Rhodes-Intérieures (compréhension de l'écrit) se démarquent de façon statistiquement significative par des pourcentages plus élevés.

Les analyses montrent également que les différences de pourcentages obtenus dans les cantons ne s'expliquent pas – ou que dans une très faible mesure – par les différences de composition des populations cantonales d'élèves.

1.3.3 Rapport d'audit sur l'enquête 2016

Le 27 octobre 2017, l'Assemblée plénière prend connaissance des premiers résultats de la vérification de l'atteinte des compétences fondamentales en mathématiques. Le SG CDIP est alors chargé de procéder à des analyses supplémentaires étudiant en particulier l'adéquation scientifique des processus de développement des tâches et de définition des valeurs seuils. En novembre 2017, Antoine Fischbach et Sonja Ugen sont chargés, en leur qualité d'experts du *Luxembourg Centre for Educational Testing* (LUCET) de l'Université du Luxembourg, de rédiger un rapport d'audit sur ces aspects. Les principales questions auxquelles il s'agit de répondre dans le rapport sont les suivantes:

- Les processus de développement des tâches et des tests ainsi que de définition des valeurs seuils répondent-ils aux critères scientifiques usuels?
- Les tâches utilisées dans le cadre de l'enquête permettent-elles d'évaluer les compétences fondamentales adoptées par la CDIP?
  a) en ce qui concerne les domaines et les aspects de compétences?
  b) en ce qui concerne le degré de difficulté?

Le rapport d'audit livre les résultats suivants (Fischbach & Ugen, 2018):

- Le processus de développement des tâches devrait être amélioré. Il n'a pas suffisamment été tenu compte de ce qui sert de référence dans le monde scientifique au cours de ce processus, qui est décrit comme étant le maillon faible de l'enquête 2016. Cependant, la qualité du test qui a finalement été utilisé, après certains remaniements, est jugée acceptable, et les données qui en résultent sont plausibles et peuvent faire l'objet d'interprétations.
- La détermination des valeurs seuils, la conception du test de même que la collecte et l'analyse des données ont été effectuées avec un degré de professionnalisme élevé et sont conformes aux normes de qualité scientifique internationales.
- Les tâches en mathématiques correspondent aux compétences fondamentales telles qu'elles ont été définies. Dans les tâches qui ont été développées, les descripteurs théoriques sont opérationnalisés le plus simplement possible. L'élément décisif selon les experts provient du fait que les objectifs de formation suisses en mathématiques sont très ambitieux en comparaison internationale et qu'ils n'ont pas fait l'objet d'une validation empirique. Les compétences fondamentales définies pour les mathématiques se comportent plutôt, d'un point de vue empirique, comme des standards «réguliers».

Quatre mesures sont proposées pour améliorer et renforcer la vérification sur le long terme (Fischbach & Ugen, 2018, p. 30 ss), à savoir:

1. *Le développement des tâches devrait être renforcé et ainsi la qualité des tâches améliorée.* Ce sont elles qui représentent, à côté des standards minimaux formulés, la partie la plus importante des enquêtes de vérification. Il convient donc de revoir à la hausse les ressources allouées au développement des tâches et de leur donner un soutien plus large en faisant davantage appel à des experts en tests ainsi qu'à des enseignantes et enseignants. Il s'agit également d'améliorer la qualité des tâches en réalisant davantage de prétests et en appliquant des processus de traduction plus solides.

2. *Il s'agit de reconsidérer les compétences fondamentales en mathématiques.* Les documents de référence actuels sont très ambitieux pour des standards minimaux et, de plus, la manière dont ils ont été formulés est loin d'être optimale pour le développement des tests. Il manque également une validation empirique des compétences fondamentales en mathématiques.
3. *L'organisation de projet devrait être reconsidérée.* Lors de l'enquête 2016, c'est le SG CDIP qui, tout en ayant le rôle d'autorité contractante, assurait en même temps la direction opérationnelle. Cette double casquette n'est pas idéale. La direction opérationnelle doit passer en d'autres mains et être professionnalisée.
4. *Le consortium scientifique prévu devrait être mis en place et entrer en fonction; il devrait être complété par un conseil scientifique composé d'experts (inter)nationaux indépendant.* Des organes consultatifs de ce type pourraient largement contribuer à assurer la qualité scientifique au fil du processus de réalisation des enquêtes ainsi qu'à conseiller la direction opérationnelle du projet sur les plans politique et stratégique.

Les mesures suggérées dans le rapport d'audit ont entre-temps été mises en œuvre de la manière suivante:

1. En ce qui concerne le développement des tâches, de premières mesures, qui correspondent en fait aux recommandations faites ultérieurement dans le rapport d'audit, ont été appliquées en amont déjà de l'enquête 2017: renforcement du mandat de coordination du secrétariat de la BDT; intégration d'enseignantes et enseignants de même que d'élèves à un stade précoce; amélioration des processus de traduction (mais sans traductions doubles exécutées séparément, comme c'est le cas par exemple avec PISA) et phase pilote approfondie pour les tâches. En vue du développement des tâches pour l'enquête 2024 en 4e année, le Bureau de coordination HarmoS a attribué au secrétariat de la BDT un mandat visant à renforcer le développement des tâches. Ce mandat comporte les éléments suivants:
   - Tâches soumises à des essais qualitatifs, accompagnés d'entretiens rétrospectifs avec des élèves pour la validation cognitive des tâches (suivi des processus de résolution); cette manière de procéder est déjà appliquée de manière complète lors du développement des tâches pour l'enquête 2020, reportée à 2023;
   - Évaluation des tâches par des enseignantes et enseignants, des didacticiennes et didacticiens n'ayant pas contribué au projet ainsi que des enseignantes et enseignants spécialisés de différentes régions linguistiques;
   - Phase pré-pilote sur les tâches, avec à chaque fois sept classes par région linguistique selon une procédure standardisée, qui anticipe autant que possible la procédure à suivre lors de l'enquête principale; puis évaluation du processus et analyse des données;
   - Enquête pilote avec des échantillons d'élèves représentatifs des régions linguistiques. La détermination de la valeur seuil s'effectue, dans la mesure du possible, sur la base des résultats de cette enquête.

   La direction de projet dispose des moyens nécessaires pour réaliser des phases pilotes plus complètes.
2. Parallèlement au développement des tâches et aux enquêtes, les objectifs de formation adoptés en mathématiques pour la 4e année de scolarité sont affinés et concrétisés. En février 2025, l'équipe chargée du développement des tâches présentera un rapport final à ce sujet. Pour les enquêtes suivantes, il est prévu d'adapter de manière analogue les objectifs de formation des deux autres degrés scolaires.
3. Depuis le début de la planification des enquêtes de vérification, les travaux sont menés dans la perspective d'un transfert de la direction scientifique du projet à une haute école universitaire. Pour les deux enquêtes prévues en 2023 et 2024 (cf. point 1.4), il a été possible de conclure une convention de coopération correspondante avec l'Université de Berne (cf. point 1.2.3). L'*Interfaculty Centre for Educational Research* (ICER) reprend donc les tâches centrales de coordination au niveau opérationnel et scientifique. Les organes de la CDIP restent impliqués dans les tâches opérationnelles, mais ils le sont beaucoup moins désormais (cf. à ce sujet le point 2.4.1).
4. Le point 1.2.3 donne des précisions sur la manière dont le projet d'organisation pour le consortium scientifique pour la recherche et l'évaluation en éducation (WIKO BFE) s'est traduit dans la structure actuelle.

1.3.4 Compte rendu des résultats et documentation

La CDIP a pour mandat d'établir les premiers rapports sur les résultats obtenus dans le cadre de la vérification de l'atteinte des compétences fondamentales. Ainsi, les rapports sur les deux premiers cycles sont publiés simultanément, le 24 mai 2019 (Nidegger, Roos, Petrucci, Verner, Helbling, Arnold, Erzinger, Pham, Ambrosetti, et al., 2019; Nidegger, Roos, Petrucci, Verner, Helbling, Arnold, Erzinger, Pham, Schönenberger, et al., 2019). Ils présentent des informations de base sur la vérification de même que sur la manière de procéder pour le cycle en cours, sur les plans d'études et sur les moyens d'enseignement. Ils détaillent également – et c'est ce qui en constitue l'élément central – les résultats obtenus en lien avec l'atteinte des compétences fondamentales ainsi que des comparaisons entre les cantons établies sur la base de ces résultats. Suivent un certain nombre de comparaisons «ajustées», pour ainsi dire plus équitables, dans lesquelles on a neutralisé (contrôlé) du point de vue statistique les effets engendrés par les différences de composition sociale entre les populations d'élèves des différents cantons. La deuxième partie des rapports comporte pour chaque canton (ou pour chaque région linguistique dans le cas des cantons bilingues) un «portrait cantonal», qui présente sur quelques pages les principales informations obtenues en lien avec l'atteinte des compétences fondamentales dans chaque domaine disciplinaire et domaine de compétence: d'abord en rapport avec l'ensemble des élèves, puis avec répartition en fonction des groupes sociaux définis par les caractéristiques que sont le genre, l'origine sociale, la langue parlée à la maison ou encore le statut migratoire, ainsi que pour la 11e année en fonction également du programme scolaire suivi. Dans le cadre de ce compte rendu initial, la CDIP publie aussi, à destination d'un plus large public, une synthèse sur le projet Vérification et sur les deux cycles 2016 et 2017. Cette synthèse comporte notamment des exemples de tâches (CDIP, 2019b).

Les rapports principaux sont complétés par une série de rapports portant sur certains aspects scientifiques et publiés sur le site web COFO Suisse (rubrique «Rapports nationaux»): développement des tâches et des tests, échantillonnage, questionnaire contextuel (contenu, traitement des données, analyses).

Par rapport aux possibilités existantes, ces rapports initiaux ne proposent que peu de pistes explicatives sur les différences de résultats observées. Des analyses en profondeur sur les données issues de la vérification ne sont réalisées qu'en fonction des besoins et des intérêts manifestés. À cet effet, des jeux de données sont mis à la disposition des acteurs scientifiques (y compris les centres de recherche cantonaux) sur FORSbase et SWISSUbase. La direction de projet propose en outre aux cantons d'accepter des mandats portant sur la réalisation d'analyses approfondies à partir des données de la vérification.

Une fois les résultats publiés, l'ICER met à la disposition des cantons qui le demandent une documentation complète avec des représentations, sous forme de graphiques et de tableaux, portant sur les différentes variables des jeux de données. À ce jour, six cantons (AG, FR, GL, SG, TG, ZH) ont demandé à bénéficier de cette offre de base, tous en lien avec l'enquête 2016. En tout, sept cantons ont demandé à des membres de la direction de projet de réaliser des analyses approfondies basées sur des questions de recherche: six cantons en lien avec l'enquête 2016 (BL, GE, SG, TI, TG, ZH) et quatre en lien avec l'enquête 2017 (BL, GL, GE, TI). Dans le cas des cantons de Genève et du Tessin, ces analyses ont été établies en interne au canton.

Pour l'instant, les résultats de la vérification ont principalement conduit à la prise de mesures décisives dans le canton de Bâle-Campagne. En réaction aux résultats obtenus lors de l'enquête 2016, qui ont été perçus comme décevants, le canton de Bâle-Campagne a lancé en 2019 un programme qui s'appelle *Zukunft Volksschule*. Il s'agit d'un vaste ensemble de mesures qui doivent être mises en œuvre de 2022 à 2028 dans les degrés primaire et secondaire I. L'accent est mis en particulier sur l'encouragement des élèves du secondaire dans la filière A (programme à exigences de base du degré secondaire I). Le but est de leur permettre de réussir la transition vers le degré secondaire II.

Les auteurs des rapports initiaux font remarquer à différents endroits que les résultats de la vérification ont aussi pour fonction de permettre d'identifier les points problématiques qui ne peuvent être suffisamment étudiés avec les seules données issues de la vérification, puisqu'il s'agit de données d'observation qui, par

nature, sont limitées dans les possibilités qu'elles offrent de démontrer des liens de causalité. Pour pouvoir étudier de telles relations, il est nécessaire de lancer des projets de recherche qui traitent ces questions avec des méthodes scientifiques appropriées.

1.3.5 Mise à disposition et utilisation des données issues de la vérification pour des analyses secondaires

Une fois les résultats publiés, en mai 2019, les données issues des enquêtes 2016 et 2017 sont anonymisées et remises à FORS sous la forme d'un fichier à usage scientifique (*scientific use file*). Les données ainsi traitées ne permettent pas de tirer des conclusions sur les écoles, les classes et les enseignants ou élèves impliqués. Ces jeux de données peuvent en particulier être utilisés dans le cadre du monitorage national de l'éducation (SKBF CSRE, 2018) et des processus cantonaux de développement de la qualité. Les résultats de la vérification peuvent ainsi être intégrés dans le développement continu du système éducatif. Ils sont également mis à la disposition de la communauté scientifique à des fins de recherche.

Afin de réglementer la gestion des données collectées, notamment la protection et la sécurité des données ainsi que les conditions d'utilisation avant et après leur publication, l'Assemblée plénière de la CDIP adopte, le 22 mars 2018, un dispositif concernant l'utilisation des données dans le cadre de la vérification de l'atteinte des compétences fondamentales. Ce dispositif est ensuite remanié sur la base des expériences faites et adopté le 7 mai 2021 par le Comité de la CDIP (CDIP, 2021b). Outre le fichier à usage scientifique, certaines données brutes, qui doivent être tenues secrètes et qui ne sont pas anonymisées ou pas traitées, sont aussi stockées chez FORS, tout comme les tâches utilisées dans les tests. Pour y accéder, il est nécessaire de déposer une demande d'utilisation des données. Les demandes d'utilisation du fichier à usage scientifique qui sont déposées par des chercheurs ou chercheuses associés à des hautes écoles ou d'autres établissements de droit public sont traitées par le SG CDIP et l'ICER selon une procédure simplifiée. L'utilisation des autres données doit faire l'objet d'une décision du Bureau de coordination HarmoS. Ce principe s'applique en particulier aux demandes qui visent l'appariement de données issues de la vérification à d'autres jeux de données. De tels appariements ne peuvent être réalisés que par l'Office fédéral de la statistique (OFS). Avant toute utilisation des données issues de la vérification, un contrat d'utilisation doit être signé, dans lequel il est notamment fait mention des critères relatifs à la sécurité des données tels qu'énoncés dans le règlement d'organisation.

Durant les années 2020-2022, le stock de données chez FORS sur les deux cycles de vérification achevés a encore été complété. Il est en particulier possible, pour des projets de recherche portant sur les tâches ou les items, d'accéder aux réponses brutes données par les élèves ainsi qu'à toutes les tâches qui ont été utilisées (celles qui sont publiques comme celles qui sont gardées sous clé). De plus, les deux fichiers à usage scientifique peuvent à présent être utilisés en deux versions: l'une sous la forme d'un jeu comprenant les données manquantes (par ex. refus de répondre) et l'autre sous la forme d'ensembles de 20 jeux de données complets comprenant des valeurs substituées statistiquement.

Au cours de l'hiver 2021-2022, la FHGR a créé, pour les deux cycles de vérification terminés, des archives contenant les tâches numériques en état de fonctionnement. S'y ajoutent les fichiers log du logiciel de test ainsi que les questionnaires électroniques incluant les métadonnées. Les archives sont en outre documentées en vue de leur utilisation. Des demandes peuvent également être faites pour l'utilisation de ces données à des fins de recherche. Pour l'instant, le contrat signé pour l'archivage de ces données court jusqu'à fin 2025.

La documentation des données chez FORS est certes complète, mais elle peut manquer de clarté pour les chercheurs et chercheuses n'ayant pas pris part aux cycles de vérification. Les membres de la direction de projet ont à plusieurs reprises été amenés à apporter leur soutien pour l'utilisation des données.

Si FORS recueille les demandes d'utilisation qui sont déposées, et qui contiennent de brefs descriptifs des projets, il ne tient cependant pas de registre sur l'obtention et l'utilisation effective des données. Les demandes sont pour la plupart déposées par des chercheurs et chercheuses ayant un lien avec la

vérification. L'édition 2/2020 de la *Revue suisse des sciences de l'éducation* contient par exemple six contributions de chercheurs et chercheuses proches du projet Vérification qui, en se basant sur les données de l'enquête 2016, ont étudié différents liens entre les compétences mesurées et les facteurs d'influence jouant un rôle dans les performances: contexte scolaire, qualité de l'enseignement, notes, motivation, bien-être et auto-évaluation des compétences. Une série de demandes d'utilisation des données de l'enquête 2016 ont été déposées par des personnes impliquées dans le projet TREE (*Transitions de l'École à l'Emploi*), avec lequel une collaboration s'était établie lors de l'élaboration du questionnaire contextuel. Le Centre suisse de coordination pour la recherche en éducation (CSRE) utilise les données des deux cycles de vérification en combinaison avec d'autres ensembles de données pour les analyses en cours en vue du *rapport 2023 sur l'éducation en Suisse*. D'autres projets de recherche ont une orientation sociologique ou relèvent du domaine de la formation professionnelle. Trois projets utilisent les données des enquêtes 2016 et 2017 pour des études portant spécifiquement sur les tâches employées dans les tests. Plusieurs demandes d'utilisation des données sont aussi liées à des travaux de qualification réalisés dans le cadre de hautes écoles.

1.4 Futures enquêtes de vérification ayant fait l'objet d'une décision

Le 22 juin 2017, l'Assemblée plénière de la CDIP décide de mener, en 2020 et en 2022, deux nouvelles enquêtes de vérification de l'atteinte des compétences fondamentales (CDIP, 2017). Les domaines disciplinaires sur lesquels doivent porter les tests sont définis lors des assemblées plénières du 26 octobre 2018 et du 25 octobre 2019 (CDIP, 2018, 2019c). Il est prévu de faire porter l'enquête de 2020 sur l'atteinte des compétences fondamentales par des élèves en fin de 11e année dans la langue de scolarisation ainsi que dans la première et la deuxième langues étrangères. Il s'agit ainsi de tester pour la deuxième fois la même cohorte qu'en 2017 dans la langue de scolarisation et, pour la deuxième fois également, dans la première langue étrangère. Pour l'enquête 2022, il est prévu de vérifier, pour la première fois, l'atteinte des compétences fondamentales auprès d'élèves de 4e année, et ce dans les domaines disciplinaires que sont la langue de scolarisation et les mathématiques. Comme mentionné au point 1.2.3, c'est la première fois que la nouvelle structure de projet est amenée à fonctionner, avec l'Université de Berne et plus précisément l'ICER à la tête de la mise en œuvre de ces deux nouveaux cycles de vérification.

En raison de la pandémie de coronavirus, l'Assemblée plénière de la CDIP décide le 1er avril 2020 de renoncer à réaliser l'enquête principale 2020, qui devait débuter peu après (CDIP, 2020a). Le 25 juin 2020, l'Assemblée plénière de la CDIP définit un nouveau calendrier, qui tient notamment compte de l'étude PISA, également reportée à 2022 (CDIP, 2020b). L'enquête 2020 est donc reportée à l'année 2023 et l'enquête 2022 à 2024.

Les sections suivantes décrivent ce qui est déjà connu des principales composantes des deux enquêtes de vérification ayant fait l'objet de décisions.

1.4.1 Enquête 2023: langue de scolarisation ainsi que première et deuxième langues étrangères en 11e année

Développement des tâches et des tests

Les tests prévus dans le cadre de l'enquête 2023 en 11e année de scolarité portent sur la langue de scolarisation ainsi que sur la première et la deuxième langues étrangères. Selon la région, les tests en langue de scolarisation portent sur le français, l'allemand ou l'italien, et ce dans les domaines de compétence que sont la lecture et l'orthographe. Pour ce qui est des langues étrangères, les compétences en compréhension orale et écrite sont évaluées dans chacune des deux langues formant la paire de langues étrangères enseignées. Ces paires de langues sont les suivantes: en Suisse romande, il s'agit de l'allemand et de l'anglais, en Suisse alémanique du français et de l'anglais et au Tessin du français et de l'allemand.

Pour le domaine de la langue de scolarisation, le secrétariat de la BDT développe les tâches en collaboration avec le même groupe que pour l'enquête 2017, à savoir des didacticiennes et didacticiens rattachés à des

hautes écoles pédagogiques (PH FHNW, HEP Vaud, SUPSI) ainsi que des collaborateurs et collaboratrices du *Dipartimento dell'educazione, della cultura e dello sport* (DECS) du canton du Tessin, ce qui leur permet de se référer à des travaux et des connaissances préalables. Pour les langues étrangères, il est fait appel à davantage d'institutions qu'en 2017, de façon à donner un appui plus large au développement des tâches. Le secrétariat de la BDT travaille en plus avec des didacticiennes et didacticiens et des spécialistes du *testing* rattachés aux institutions suivantes: l'Institut de plurilinguisme de l'Université de Fribourg et de la HEP Fribourg, la PH FHNW, la PH Luzern et la PH Zürich. Le développement des tâches est de plus suivi par un groupe d'experts composé de didacticiennes et didacticiens de Suisse romande et de Suisse italienne.

Afin d'obtenir suffisamment tôt des indications sur la qualité des tâches, celles-ci sont toutes soumises à une série d'essais qualitatifs menés auprès d'un certain nombre d'élèves et accompagnés d'entretiens rétrospectifs sur le processus de résolution des tâches, ce qui permet de dégager des indications sur les processus de résolution (validation cognitive). Au printemps 2019, les tâches font enfin l'objet d'un test auprès d'un échantillon national de plus de 2 000 élèves. Certaines tâches sont alors adaptées, sur la base d'analyses psychométriques des données, avant d'être validées pour l'enquête principale.

En tout, le nombre de tâches pouvant être utilisées dans le domaine de la langue de scolarisation est le suivant: domaine de compétence de la lecture: 44 items; domaine de compétence de l'orthographe (tests spécifiques pour chaque langue): 54 items en français, 50 en allemand et 49 en italien.

Dans les langues étrangères, les items pouvant être utilisés dans le domaine de compétence de la compréhension de l'écrit est de 47 et, pour la compréhension de l'oral, de 52. Ces items se déclinent, selon les langues étrangères testées, dans les versions suivantes: allemand comme langue testée (langue cible), instructions et consignes en français ou en italien; français comme langue cible, instructions et consignes en allemand ou en italien; anglais comme langue cible, instructions et consignes en français ou en allemand. Comme les exigences formulées dans les compétences fondamentales ne diffèrent pas entre la première et la deuxième langues étrangères, les mêmes tâches sont utilisées dans toutes les langues cibles. Le choix des textes pour les tâches résulte d'une analyse approfondie des manuels employés pour les trois langues cibles.

Contextualisation

La contextualisation de l'enquête 2020 s'effectue à nouveau au travers d'un questionnaire contextuel destiné aux élèves. En termes de contenu, ce questionnaire reprend, après examen approfondi, de nombreux construits et variables de celui développé pour l'enquête 2017. La direction de projet travaille à appliquer un «module de base» fixe dans (si possible) toutes les enquêtes de vérification, de façon à permettre des analyses entre les différents cycles également. Ce questionnaire permet de collecter des informations telles que l'origine sociale ou la situation familiale et scolaire des élèves. Il comporte également des questions sur les conditions et les paramètres d'apprentissage individuels, la motivation, l'intérêt personnel ainsi que sur différentes caractéristiques de l'enseignement (par ex. activités en classe).

Réalisation de l'enquête

Il est prévu que l'enquête 2023 se déroule à nouveau selon une procédure standardisée, assurée par des responsables formés à cet effet. Comme durant l'enquête 2017, les élèves passent les tests sur des tablettes amenées sur place. Les tests individuels se composent de trois parties de 40 minutes chacune (langue de scolarisation, première et deuxième langues étrangères). Les élèves remplissent ensuite le questionnaire contextuel.

Tous les cantons suisses, à l'exception du canton de Zoug, prennent part à l'enquête 2023. Sur une population totale d'environ 80 000 élèves (scolarisés en 11^{e} année dans les cantons prenant part à l'enquête), on tire des échantillons représentatifs au niveau cantonal du même ordre que pour les premiers cycles de vérification (à savoir au moins 20 000 personnes). Les élèves ne suivant plus d'enseignement en langues étrangères en 11^{e} année ne sont en principe pas dispensés de participer à l'enquête.

1.4.2 Enquête 2024: langue de scolarisation et mathématiques en 4e année

Développement des tâches et des tests

Le développement des tâches pour l'enquête 2024 s'appuie sur les travaux préliminaires déjà réalisés dans le cadre de l'étude de faisabilité sur la vérification de l'atteinte des compétences fondamentales en 4e année (Angelone et al., 2019). Cette étude de faisabilité a été commandée en 2018 par le SG CDIP en vue de montrer, comme base pour la planification ultérieure, les possibilités et les défis à escompter avec une enquête standardisée, menée au moyen de tablettes, sur les compétences fondamentales d'enfants âgés d'environ huit ans. Dans le cadre de cette étude, des didacticiennes et didacticiens ont élaboré de premiers cadres théoriques pour les tests et des tâches à titre d'exemples pour les domaines disciplinaires des mathématiques, des sciences naturelles et de la langue de scolarisation. Ils montrent quelles sont les compétences fondamentales, parmi celles formulées, qui peuvent faire l'objet de tests dans le cadre d'une enquête auprès d'élèves de 4e année. Par ailleurs, le secrétariat de la BDT a aussi testé dans quatre classes une passation de test standardisée, et la HEP de Saint-Gall l'utilisation d'un questionnaire pour élèves à effectuer sur tablette.

Les tâches pour les tests sont élaborées, en 2020 et 2021, par des didacticiennes et didacticiens des deux domaines disciplinaires, en collaboration avec le secrétariat de la BDT. Pour la langue de scolarisation, une coopération s'établit entre la PH FHNW, la HEP Vaud, la PH Zug, la SUPSI et le DECS du canton du Tessin. Pour les mathématiques, le travail se fait avec l'IRDP, la PH Thurgau et la SUPSI. Ces établissements travaillent à leur tour en collaboration avec des experts d'autres institutions (Université de Genève, HEP Vaud, PHLU, PHZH).

Comme exposé au point 1.3.3 dans les conclusions émanant du rapport d'audit sur l'enquête 2016, le développement des tâches pour l'enquête 2024 est substantiellement renforcé via un mandat de la CDIP que le Bureau de coordination a adopté le 13 novembre 2020 à l'attention de la BDT. Outre les mesures ayant pour but d'améliorer la qualité des tâches, le mandat vise aussi à affiner et concrétiser les descriptions existantes pour les objectifs de formation en mathématiques sous la forme d'un cadre de test (*test framework*) comprenant des tâches à titre d'exemples. Ce cadre forme, pour l'enquête 2024, un maillon intermédiaire entre les objectifs de formation adoptés d'un côté et, de l'autre, le développement des tâches et des tests.

Contextualisation

S'agissant de la contextualisation de l'enquête en 4e année, la direction de projet teste de nouvelles possibilités en vue de collecter des données contextuelles suffisamment fiables en dépit du jeune âge des élèves. D'une part, on cherche, en faisant participer des enfants du groupe cible, à trouver une forme qui soit adaptée aux enfants pour le questionnaire à remplir. On prévoit par exemple d'inclure une chouette parlante qui guide les élèves tout au long du questionnaire. Et d'autre part, on essaie de trouver des sources d'information complémentaires. Dans cette perspective, une démarche est en cours pour savoir dans quelle mesure les données des registres de l'OFS peuvent être utilisées. En s'appuyant sur ses propres ressources de projet, l'ICER met également à profit l'enquête pilote menée au printemps 2022 en vue de l'enquête 2024 pour tirer des éléments de conclusion quant à l'utilisation possible et à l'utilité d'un questionnaire qui serait destiné aux parents et d'un questionnaire qui serait adressé aux directions d'école. Des résultats fiables découlant de ces démarches, notamment concernant les taux de réponse aux questionnaires, devraient être disponibles dans le courant de l'année 2022.

Réalisation de l'enquête

Si l'enquête pilote se déroule bien au printemps 2022, il est prévu que l'enquête 2024 soit menée dans le même cadre que les enquêtes 2017 et 2023.

1.4.3 Financement de la vérification et de la BDT

Depuis 2014, la vérification de l'atteinte des compétences fondamentales est financée par des contributions cantonales annuelles d'un montant quasiment identique à chaque fois. Comme des enquêtes ont dû être reportées et des travaux répétés à cause de la pandémie de coronavirus, l'Assemblée plénière de la CDIP a décidé le 24 juin 2021 de revoir à la hausse les contributions cantonales prévues pour les années 2022 à 2025.

Financement passé et futur de la vérification pour la période allant de 2014 à 2025 (quatre enquêtes):

	Contribution par an	Montant total
Contributions cantonales de 2014 à 2016 (3 ans)	CHF 1 123 333	CHF 3 370 000
Contributions cantonales de 2017 à 2025 (9 ans)	CHF 1 130 000	CHF 10 170 000
Augmentation des contributions cantonales en raison de la pandémie de coronavirus pour les années 2022 à 2025	CHF 232 500	CHF 930 000
Total des contributions cantonales fixées au titre de la vérification pour la période allant de 2014 à 2025		**CHF 14 470 000**

Comme mentionné au point 1.2.2, la Banque de données de tâches assume des tâches importantes pour la vérification de l'atteinte des compétences fondamentales. Mais elle accomplit également toute une série de tâches, conformément au règlement, qui ne sont pas liées à la vérification et qui sont financées séparément. À ce jour, la BDT a reçu les montants suivants pour les années 2012 à 2025:

	Contribution par an	Montant total
Crédit de démarrage provenant des fonds du monitorage de l'éducation (2012-2013)		CHF 765 000
Contributions cantonales de 2014 à 2025 (12 ans)	CHF 370 000	CHF 4 440 000
Total pour le crédit de démarrage et les contributions cantonales allouées entre 2014 et 2025 à la BDT		**CHF 5 205 000**

La réalisation des cycles de vérification de 2023 et 2024 (initialement planifiés pour 2020 et 2022) est garantie par les contributions allouées pour les années 2020 à 2025 au projet Vérification et à la Banque de données de tâches.

Dans l'ensemble, les dépenses de la Suisse pour le monitorage de l'éducation sont faibles par rapport à celles de l'Allemagne et de l'Autriche.

2 Analyse et discussion

Dans ce chapitre, le Bureau de coordination HarmoS se fonde en particulier sur une expertise scientifique dont le Secrétariat général de la CDIP a confié la réalisation à des spécialistes possédant d'excellentes connaissances en matière de gouvernance du système éducatif et/ou d'évaluations à large échelle. L'équipe chargée de l'expertise était composée des personnes suivantes:

- Prof. Dr. Katharina Maag Merki, Université de Zurich (coordinatrice)
- Dr. Anja Giudici, University of Oxford
- Prof. i.R. Dr. Eckhard Klieme, DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation, Frankfurt/Main
- Prof. Dr. Hans Anand Pant, Humboldt-Universität zu Berlin
- Prof. Dr. Claudia Schreiner, Universität Innsbruck

Kai Schudel, collaborateur scientifique de l'Université de Zurich, a soutenu le groupe dans son travail.

Ce chapitre a également été élaboré sur la base des avis transmis par plusieurs expertes et experts externes. Il s'agit notamment d'Olivier Maradan, qui a été responsable de l'Unité de coordination Scolarité obligatoire au Secrétariat général de la CDIP, de Christian Nidegger qui a, entre autres fonctions, assumé le rôle de coordinateur national de l'étude PISA et de principal auteur des rapports sur les enquêtes de vérification de 2016 et 2017, mais aussi de Stefan Wolter, qui dirige le Centre suisse de coordination pour la recherche en éducation (CSRE) et qui contribue, en tant que responsable de l'équipe de rédaction, à l'élaboration du rapport sur l'éducation en Suisse (SKBF CSRE, 2018), outil central pour les rapports établis sur l'éducation et la formation. De plus, bon nombre d'observations fournies par l'ICER et, plus largement, par la direction du projet Vérification ont également été intégrées dans cette partie.

2.1 Approche, objectifs et paramètres liés à la vérification de l'atteinte des compétences fondamentales

Les sections suivantes apportent un éclairage critique sur différents éléments constitutifs de la vérification de l'atteinte des compétences fondamentales: l'approche du pilotage du système éducatif par les résultats (*output*); les principaux objectifs et fonctions; l'interprétation des résultats et leur traitement; l'accent mis sur les compétences fondamentales de même que le fait de se limiter à quatre domaines disciplinaires.

2.1.1 Pilotage du système par le biais des résultats

La vérification de l'atteinte des compétences fondamentales doit permettre d'assurer la qualité et d'harmoniser sur le fond, à l'échelle nationale, les systèmes éducatifs cantonaux au moyen de standards nationaux de formation s'exprimant sous la forme de standards de performance (cf. point 1.1).

Les résultats (*output*) des systèmes scolaires sont mesurés et comparés au travers des performances des élèves. La vérification de l'atteinte des compétences fondamentales s'inscrit ainsi dans un ensemble de systèmes comparables de pilotage des résultats qui se sont mis en place dans différents pays européens depuis les années 2000.

L'expertise menée a permis d'examiner la question de savoir si l'approche choisie est encore adaptée et d'actualité. La conclusion en a été que la mise en place d'un monitorage de l'éducation s'accompagne de certains défis et qu'elle ne garantit pas d'emblée un système éducatif dans lequel la totalité des élèves puisse atteindre les objectifs fixés. Cependant, le groupe chargé de l'expertise s'exprime clairement en faveur du maintien du programme de vérification, moyennant certains ajustements substantiels à apporter au *design* en général, à l'aménagement et à l'ancrage institutionnel de même qu'à l'attribution des ressources humaines et financières (voir sections suivantes).

Le Bureau de coordination HarmoS estime indispensable de veiller à un pilotage de certains aspects des systèmes éducatifs cantonaux par le biais d'évaluations des performances en se référant au mandat d'harmonisation défini dans la Constitution. Renoncer à ces évaluations de performances risquerait de nuire

au processus d'harmonisation et de remettre en question l'objectif, poursuivi à l'échelle suisse, du développement de la qualité dans la scolarité obligatoire. Toutefois, pour que l'évaluation des performances puisse déployer pleinement ses effets dans le pilotage du système, il importe d'apporter une série de changements sur la base des propositions qui ont été faites à la suite de l'expertise menée et qui ont été approuvées voire, pour certaines, encore accentuées par les autres spécialistes consultés.

2.1.2 Principales fonctions associées à la vérification

Comme précisé au point 1.2.1, le but de la vérification de l'atteinte des compétences fondamentales consiste, conformément au règlement y relatif, à évaluer les performances du système de la scolarité obligatoire ainsi qu'à fournir aux cantons des informations relatives au système et portant sur le degré d'atteinte des objectifs de formation. La collecte de données contextuelles permet d'identifier les liens entre les différences de performance et les caractéristiques contextuelles.

Durant les entretiens menés avec des acteurs de la politique de l'éducation et de la formation, le groupe chargé de l'expertise a constaté que les fonctions associées à la vérification sont les trois suivantes:

1. évaluation du système,
2. mise à disposition d'éléments de réponse permettant de comprendre les différences de performances observées et
3. mise à disposition de connaissances opérationnelles pour la pratique scolaire.

La *fonction d'évaluation du système* est largement reconnue, même si des réserves parfois importantes ont été constatées en ce qui concerne la manière dont la vérification a été réalisée jusqu'à présent.

Quant à la *volonté de dégager des éléments de réponse concernant les différences de performances* entre les régions, les cantons et les groupes sociaux qui composent la population d'élèves, si elle est jugée centrale par le groupe chargé de l'expertise comme par les autres spécialistes consultés, elle ne semble que partiellement atteinte avec la manière dont la vérification a été mise en œuvre jusqu'à présent. Pour que la vérification puisse réellement contribuer au développement de la qualité, il faut que sa fonction explicative soit renforcée, ce qui aurait pour effet d'augmenter l'utilité réelle et perçue de la vérification, et donc d'améliorer son acceptation. Les mesures proposées à cet effet sont présentées dans les points développés ci-dessous.

Pour le Bureau de coordination HarmoS, il est essentiel que les évaluations de performances présentent une qualité élevée et soient crédible, et que leur utilité soit plus grande pour l'évaluation du système, le monitorage de l'éducation et le pilotage de la formation. Elles doivent fournir une base solide pour permettre une politique fondée sur des données probantes.

Pour ce qui est de la fonction *Mise à disposition de connaissances opérationnelles pour la pratique scolaire,* qui selon les personnes interrogées est également associée à la vérification, elle n'est pas confirmée par les travaux d'expertise et ne reçoit pas non plus le soutien des spécialistes consultés ni du Bureau de coordination HarmoS. Une enquête telle que prévue dans le cadre de la vérification, qui ne vise pas l'individu, ne peut pas fournir de tels résultats. Il faudrait pour ce faire qu'elle soit conçue autrement et qu'elle se charge de tâches qui, aux termes du concordat HarmoS, reviennent aux régions et aux cantons.

2.1.3 Interprétation des résultats

Les cantons sont tenus par la Constitution fédérale d'harmoniser les objectifs des degrés d'enseignement. C'est là le mandat constitutionnel qui est concrétisé par la vérification de l'atteinte de ces objectifs.

Comme l'a relevé l'expertise, il n'a pas été déterminé, en lien avec les résultats des précédentes enquêtes, à partir de quel degré d'atteinte les compétences fondamentales sont jugées suffisamment acquises et jusqu'à quel écart entre les cantons les systèmes éducatifs cantonaux peuvent être considérés comme harmonisés. Proposition est donc faite de prévoir un mécanisme aux fins de cet examen.

Compte tenu du fait que les compétences fondamentales doivent, par définition, être acquises par (pratiquement) tous les élèves, il revient, de l'avis du Bureau Harmos, à la politique en matière d'éducation et de formation de décider, une fois les résultats d'une enquête obtenus,

- si des mesures doivent être prises selon la part d'élèves n'ayant pas atteint les compétences minimales et, le cas échéant, à quel(s) niveau(x);
- dans quelle mesure le mandat constitutionnel visant l'harmonisation est rempli.

2.1.4 Accent mis sur les compétences fondamentales et réflexion sur l'ensemble du spectre de compétences

Si la vérification permet actuellement de dégager des indications sur le degré d'atteinte des compétences fondamentales en mettant l'accent sur ces dernières, elle ne livre que peu voire pas de conclusions sur les performances de la totalité des élèves, puisque les tests se concentrent sur les compétences fondamentales, conformément à ce qui avait été demandé. Mais dans les tests menés jusqu'ici, il était plutôt rare que l'accent soit mis clairement sur la tranche inférieure, de sorte que les tests ont aussi permis, sans que cela soit prévu, d'opérer une différenciation entre les élèves les plus performants, et ce dans plusieurs domaines de compétence.

Depuis les débuts de la vérification, divers acteurs ont suggéré de ne pas limiter le programme à la vérification de l'atteinte des compétences fondamentales, mais d'avoir recours à des tests qui permettent de saisir avec fiabilité l'ensemble du spectre de compétences. Cela impliquerait de décrire également l'intégralité du spectre de compétences pour la construction et l'interprétation des tests, et ce de préférence à l'aide d'échelles de compétences comprenant plusieurs niveaux, comme c'est le cas pour les échelles PISA. Quant aux personnes qui ont été interrogées pour l'expertise, elles ont des avis partagés sur la question de savoir si l'accent mis dans le cadre de la vérification doit porter sur les compétences fondamentales (c'est-à-dire sur un seuil de compétences) ou sur l'ensemble du spectre de compétences.

L'expertise conclut qu'il vaut mieux s'en tenir à la vérification des compétences fondamentales et continuer, lors de la première publication des résultats (publication primaire), de traiter uniquement des pourcentages d'élèves ayant atteint les compétences fondamentales. Cette recommandation se justifie ainsi:

- Cette manière de procéder est généralement bien acceptée.
- La vérification peut ainsi contribuer à ce que pratiquement tous les élèves atteignent effectivement des compétences minimales.
- Si l'on adoptait des échelles de compétences, il faudrait redéfinir les objectifs de formation, ce qui s'avérerait probablement assez complexe compte tenu des différences qui existent entre les plans d'études des régions linguistiques.
- Les tests utilisés jusqu'à présent dans le cadre de la vérification offrent déjà une dispersion relativement importante, permettant des analyses secondaires dans un spectre de performances plus étendu.
- Concernant la mise à l'échelle des résultats des tests, on établit déjà à l'heure actuelle des échelles continues, lesquelles sont également mises à disposition pour des analyses secondaires.

Le Bureau de coordination HarmoS est aussi d'avis que les enquêtes doivent continuer d'avoir pour objectif central de vérifier l'atteinte des compétences fondamentales dans le but de garantir l'acquisition de celles-ci par les élèves. En même temps, il conclut également, en sa fonction d'organe de pilotage du programme, que les enquêtes devraient, pour les raisons présentées ci-dessous, permettre à l'avenir de générer davantage de données utiles pour le monitorage de l'éducation et le pilotage de l'éducation, de façon que ce dernier puisse se fonder sur des données probantes:

- Le rapport entre les ressources consacrées aux enquêtes et l'utilité des données n'est pas satisfaisant si les tests sont uniquement axés sur les compétences fondamentales.
- Les connaissances utiles pour le pilotage que l'on peut en retirer se limitent à un groupe trop restreint.
- On n'est pas en mesure de répondre aux attentes des milieux politiques et du public vis-à-vis des enquêtes: la vérification a notamment été comprise comme apportant une suite aux anciennes compa-

raisons cantonales, qui étaient habituelles avec l'étude PISA (cf. CDIP, 2009, 2012a). Grâce à des échelles comprenant plusieurs niveaux, PISA parvient à représenter la répartition des élèves en fonction de leurs niveaux et propose également une analyse critique de l'amplitude de la dispersion.

À l'avenir, les tests devraient relever avec fiabilité les compétences sur l'ensemble du spectre et devront permettre d'appliquer aussi une différenciation aux élèves se situant dans les parties supérieure et inférieure de l'échelle. Puisqu'il est nécessaire de disposer d'une description d'échelles adaptée pour pouvoir mener des enquêtes testant les compétences sur l'intégralité du spectre, il s'agirait de compléter les objectifs de formation déjà adoptés en vue de la construction des tests et de l'interprétation des résultats (pour plus de précisions, cf. annexe 1: *Autres aspects liés à la mise en œuvre de la vérification*, point *Bases concrètes pour le développement des tests*). Ces descriptions d'échelles complémentaires doivent découler d'analyses des plans d'études et éventuellement de moyens d'enseignement utilisés dans les différentes régions linguistiques, pour permettre de formuler, au niveau des compétences, des attentes qui soient justes pour toutes les parties concernées. Il ne s'agit toutefois pas d'adapter les objectifs de formation adoptés par la CDIP en 2011, qui ne perdent rien de leur validité pour obtenir des indications en lien avec l'atteinte des compétences fondamentales et donc en lien avec l'objectif politique d'harmonisation, du moins pour ce qui a trait aux compétences fondamentales, et comme base pour les plans d'études des régions linguistiques.

L'expertise a relevé le fait que les enquêtes menées jusqu'ici présentaient déjà une large dispersion des performances des élèves, mais ce constat ne semble pas pertinent au Bureau HarmoS car il s'agit d'un effet jugé indésirable du point de vue du développement des tests. Les tests en question portaient moins que souhaité sur les élèves de la partie inférieure du spectre de performances. Ce que l'on peut dire, de manière générale, c'est que les exigences des tâches composant les tests doivent correspondre à l'objectif fixé. Ainsi, si l'on ne teste systématiquement que l'atteinte des compétences fondamentales qu'il semble raisonnable d'attendre de la part des élèves, le test ne permet pas de saisir de manière différenciée les performances des élèves ayant plus de facilité. De fait, les élèves ayant un niveau moyen, bon ou très bon pourraient tous produire des résultats (presque) parfaits. Les échelles de test qui en résultent ne sont que partiellement adaptées pour mettre en évidence des liens avec d'autres variables, comme le contexte social, ou pour analyser l'évolution d'un groupe donné au sein de la cohorte entre une passation et une autre (par ex. entre la 4e et la 8e année). Or ce sont précisément les analyses de ce type qui sont intéressantes pour le monitorage de la formation et pour la recherche quantitative en éducation.

2.1.5 Restriction aux compétences fondamentales dans les quatre domaines disciplinaires

L'expertise a examiné le degré d'acceptation des descriptions des objectifs de formation adoptés en 2011 (compétences fondamentales) et l'a trouvé globalement positif. Le groupe chargé de l'expertise recommande cependant de poursuivre, pour les autres années également, les travaux de remaniement des descriptions des compétences fondamentales en mathématiques qui ont été entamés en 2021 pour la 4e année de scolarité. Cette suggestion est approuvée par le Bureau HarmoS. Les compétences fondamentales attendues dans le domaine des langues étrangères ont été définies avant que l'on puisse estimer avec certitude de quelle manière la mise en œuvre de la stratégie des langues allait se répercuter sur les compétences, et notamment la question de l'enseignement de deux langues étrangères au degré primaire déjà. Les résultats des cycles 2017 et 2023 devraient permettre de déterminer si les objectifs de formation décrivent effectivement des compétences fondamentales.

L'expertise s'est aussi penchée sur la question de savoir si les domaines disciplinaires qui font l'objet des tests ont été bien choisis et s'il faudrait également formuler des objectifs de formation pour d'autres domaines, en plus de la langue de scolarisation, des mathématiques, des langues étrangères et des sciences naturelles, ainsi que cela était prévu en option dans le projet décrivant initialement la vérification (cf. 1.1.2, 1.1.4).

Les réponses reçues révèlent une très forte approbation des domaines disciplinaires que sont la langue de scolarisation et les mathématiques. L'approbation est un peu moins forte pour les langues étrangères, et les sciences naturelles présentent un écart supplémentaire. Par contre, il n'y a guère de propositions faites concernant d'autres domaines disciplinaires pour lesquels des objectifs de formation nationaux devraient être élaborés. Le groupe chargé de l'expertise ne voit pas non plus de raison d'étendre le programme à

d'autres disciplines. Il souligne en revanche que, pour les sciences naturelles, il reste à démontrer si et comment les tests fonctionnent, surtout si – comme c'était le cas dans l'étude de validation réalisée en 2007 – il est aussi question d'utiliser des tâches faisant intervenir l'expérimentation dans le large échantillon.

Dans la recherche en éducation, les compétences acquises en langue de scolarisation et en mathématiques sont considérées comme des facteurs suffisamment prédictifs de la réussite ultérieure dans le parcours de formation. Si les capacités à disposition dans le cadre de la vérification ne suffisent pas pour tester régulièrement (au sens du point 2.1.1) davantage de domaines disciplinaires, il convient, d'après le directeur du CSRE, de donner la priorité à ces domaines-là. Du point de vue de l'école, cet ordre de priorités se justifie en outre par la pertinence des branches pour la promotion des élèves.

Le Bureau HarmoS ne voit pas de raison, du point de vue de la politique de l'éducation, de modifier le catalogue des domaines disciplinaires qui doivent faire l'objet des enquêtes. L'importance des langues étrangères est jugée élevée. Lors de l'enquête 2023, il sera possible de tester la langue de scolarisation et deux langues étrangères dans le même cycle. Il est souhaitable que des combinaisons de domaines disciplinaires puissent également être mises en œuvre à l'avenir. Le groupe chargé de l'expertise propose de tester à chaque fois, dans le cadre d'une même enquête, soit la langue de scolarisation et les langues étrangères (comme jusqu'ici), soit les mathématiques et les sciences naturelles, sous réserve toutefois que la faisabilité des tests soit démontrée pour les sciences naturelles.

2.2 *Design* des enquêtes

Les cycles déjà menés et prévus jusqu'en 2024 revêtent un caractère de mesures ponctuelles (études transversales) initiales. Pour les langues, il était prévu d'adopter une seule fois (selon ce qui était envisagé initialement) un *design* longitudinal par cohortes, lequel n'a ensuite pas pu être mis en œuvre en raison de la pandémie. Aucune cadence fixe n'avait été définie pour des mesures de tendances similaires à celles obtenues dans le cadre de l'étude PISA[6].

2.2.1 Planification à long terme permettant d'obtenir des mesures de tendances

Déplorant l'absence de planification fiable pour les enquêtes, l'expertise souligne la grande utilité que peut apporter, en matière de politique d'éducation et de formation, un programme qui se déploie sur le long terme:

- Un tel programme permet de générer des données sur les évolutions survenant dans le système éducatif et de dégager ainsi des tendances.
- Il a pour effet de rendre les résultats systématiquement utilisables et donc intéressants pour la politique et l'administration de l'éducation, les partenaires scientifiques, le monitorage de l'éducation ainsi que pour la recherche à mener ultérieurement.
- Il va permettre une coordination avec les systèmes régionaux et cantonaux de monitorage.
- Il apporte plus de continuité et probablement une meilleure qualité dans le développement des tâches, des tests et des questionnaires contextuels.
- Il assure une continuité du côté des spécialistes qui mettent en œuvre la vérification.
- Il donne de la pertinence à la vérification au sein du débat public.

Les études de tendances présentent des avantages importants, pour le pilotage de l'éducation, par rapport aux études transversales décidées au cas par cas. Si elles ne sont pas intégrées dans un *design* de ten-

[6] L'étude PISA repose sur un *design* de tendance: depuis l'an 2000, des mesures sont effectuées tous les trois ans auprès d'un échantillon représentatif d'élèves de 15 ans pour tester leurs compétences dans les trois domaines suivants: mathématiques, langue de scolarisation et sciences naturelles. En utilisant à chaque fois un même noyau de tâches PISA, il est possible de déterminer de manière fiable, à l'aide de méthodes statistiques, l'évolution des compétences dans la population au fil du temps. S'y ajoute une contextualisation assurée au moyen d'une série fixe d'items dans le questionnaire qui permettent également l'observation de tendances.

dances, les enquêtes de monitorage ne permettent pas d'obtenir autant d'observations utiles pour le pilotage. Les comparaisons entre cantons de même que celles qui se fondent sur des caractéristiques contextuelles d'enquêtes individuelles permettent tout au plus de faire des suppositions quant aux causes éventuelles et aux pistes nécessaires pour définir des mesures; de plus, les effets des mesures prises ne peuvent être suivis que de manière approximative. Les études de tendance offrent des bases déterminantes pour une recherche plus approfondie des causes. Elles révèlent l'évolution qui s'est produite, pendant un certain temps, dans un domaine donné (par ex. au niveau de la compréhension de l'écrit chez les jeunes de 15 ans, comme dans le cas de PISA). Le fait de réaliser des tests et des collectes de données contextuelles comparables de manière répétée sur une période assez étendue permet d'établir des liens entre les changements mesurés et de les interpréter.

L'expertise recommande un intervalle de maximum six ans entre les mesures de tendances (cadence de collecte). Cela permet d'évaluer au moins tous les six ans l'atteinte des compétences fondamentales dans la même combinaison de domaine et d'année de scolarité. Ce n'est que lorsque des tendances ont été observées sur plusieurs points de mesure que des écarts entre les mesures peuvent être interprétés à la lumière d'une tendance. Le directeur du CSRE estime qu'avec une cadence de trois ans pour les enquêtes, il faut prévoir déjà au moins douze ans pour pouvoir créer une base solide nécessaire à l'interprétation des tendances comme des écarts constatés par rapport à ces tendances. Pour la politique de l'éducation et de la formation, il est important que les nouvelles tendances et les écarts importants par rapport à celles-ci soient identifiés suffisamment tôt pour que le système éducatif puisse réagir à temps aux changements indésirables. Une cadence élevée de mesures de tendances permettrait par exemple d'identifier relativement rapidement les conséquences de la pandémie.

2.2.2 Combinaison du *design* de tendances et de l'échantillonnage longitudinal par cohortes

Comme variante ou complément à un design orienté uniquement sur les tendances, l'expertise suggère de combiner celui-ci avec un *design* visant l'analyse longitudinale (par cohortes). Les échantillons longitudinaux suivent ainsi la même cadence que le *design* de tendances (par ex. tous les quatre ans). Dans le cas de la vérification, cela signifie que la volée d'élèves sélectionnée (cohorte) participe trois fois à une enquête dans le même domaine disciplinaire au cours de sa scolarité obligatoire, à savoir en 4^{e}, en 8^{e} puis en 11^{e} année (langues étrangères uniquement en 8^{e} et en 11^{e} année). Des échantillons indépendants sont tirés de la cohorte pour chacun des trois cycles de tests. Le groupe chargé de l'expertise déconseille d'opter pour un *design* visant des analyses longitudinales systématiquement axées sur l'individu en raison du rapport coût/utilité défavorable que cela représenterait. Dans les petits cantons, où la totalité ou presque des élèves font partie de chacun des échantillons participant à la vérification, le *design* longitudinal par cohortes offre d'office de nombreux échantillons longitudinaux individuels, et donc la possibilité d'étudier divers aspects liés au développement individuel. En outre, certaines méthodes statistiques permettent d'analyser l'évolution de groupes au sein d'une cohorte donnée même lorsque l'on en tire, pour deux prises de mesure, des échantillons aléatoires qui ne présentent pas ou que peu de chevauchement au niveau des élèves.

Selon le concordat HarmoS (CDIP, 2007a, art. 10), l'objectif du monitorage de l'éducation, qui se doit d'être «systématique», «continu» et «scientifiquement étayé», est d'évaluer «les développements et les performances de l'école obligatoire». Pour ce qui est de suivre les développements de l'école, il convient de recourir à des études de tendances. Pour évaluer les performances de l'école, on analyse généralement l'évolution des élèves, ou de différents types ou groupes d'élèves, entre deux prises de mesure, par exemple dans l'intervalle d'un cycle de l'école obligatoire. Pour procéder à une telle évaluation, il est essentiel de disposer d'une mesure d'entrée et d'une mesure de sortie. C'est là la seule façon d'évaluer efficacement la performance de la partie du système considérée. Si donc la vérification est mise en œuvre sous la forme d'études longitudinales par cohorte, elle pourra mettre à la disposition du monitorage un potentiel nettement plus important que des études de tendances (comme PISA) prises isolément.

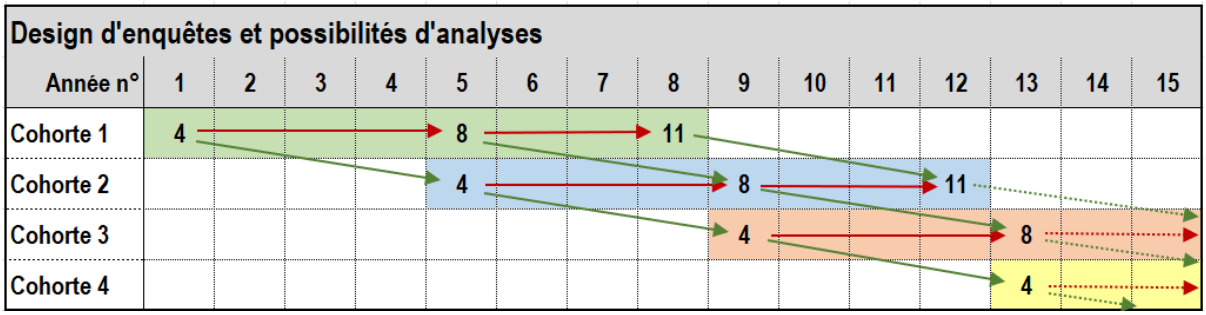
Le Bureau de coordination HarmoS adhère majoritairement aux réflexions présentées sur la nécessité d'appliquer un *design* de tendances défini à plus long terme ainsi que sur les avantages supplémentaires que peuvent apporter des études longitudinales par cohorte du fait qu'elles offrent une excellente base pour

obtenir des analyses étayées et des indications fiables concernant le développement de la qualité et le processus d'harmonisation du système scolaire. Ainsi, un *design* de tendances combiné à des études longitudinales par cohorte se prête particulièrement bien à l'évaluation du système et à la formulation de pistes explicatives. De plus, une telle approche offre également une base de données extrêmement intéressante pour la recherche en éducation en dehors du processus de monitorage officiel. Cet avantage sera alors immanquablement profitable également à la politique et au pilotage en matière d'éducation et de formation puisque, comme on le sait, le rapport sur l'éducation ne se fonde pas uniquement sur ses propres analyses mais aussi, dans une large mesure, sur des études réalisées par d'autres chercheuses et chercheurs.

Parmi les arguments en faveur d'un *design* de tendances combiné à des études longitudinales par cohorte, on relève également le fait que cette variante de *design* ne requiert pas plus d'efforts de mise en œuvre qu'un *design* purement orienté sur les tendances. Ce qu'elle demande, par contre, c'est une plus grande discipline dans la planification.

Après considération du rapport coût/utilité de différents rythmes d'enquêtes, le Bureau HarmoS estime qu'un intervalle de quatre ans entre les mesures de tendances est approprié. Ce rythme s'intègre également bien dans le reste du monitorage national de l'éducation: le rapport sur l'éducation paraît tous les quatre ans, tout comme les objectifs politiques communs concernant l'espace suisse de formation (p.ex. DEFR & CDIP, 2019). Par ailleurs, l'étude PISA suivra aussi un rythme quadriennal à partir de 2029. Une telle cadence permet d'obtenir suffisamment de données dans un laps de temps raisonnable pour disposer d'une base solide pour l'interprétation des mesures de tendances. En général, des intervalles de quatre ans permettent aussi de ne pas constater trop tard des évolutions d'importance.

L'illustration suivante donne un aperçu d'un *design* longitudinal (par cohortes) suivant un intervalle de tendances (cadence) de quatre ans sur une période de 15 ans à partir de la première enquête principale.



Ill. 1: design d'enquêtes pour la vérification permettant des analyses par cohortes de la 4e à la 11e année avec une cadence de quatre ans. Possibilités d'analyse: les flèches vertes indiquent les mesures de tendances entre différentes cohortes qui se suivent directement, les flèches rouges les changements au sein d'une même cohorte au cours d'un cycle de la scolarité obligatoire.

Un nouvel échantillonnage longitudinal est lancé tous les quatre ans pour chacune des deux combinaisons de disciplines. Une telle analyse comprend à chaque fois trois périodes d'enquête (4e, 8e et 11e années de scolarité) se succédant à intervalles réguliers. Avec une cadence de quatre ans, des enquêtes principales sont réalisées la même année en 8e année de scolarité pour une cohorte donnée et en 4e année pour la cohorte suivante. La planification des travaux préliminaires et des enquêtes pilotes prévoit toutefois une certaine flexibilité. Avec le *design* d'enquête présenté, neuf enquêtes principales seront réalisées au total sur une période de quinze ans. S'y ajouteront normalement encore neuf enquêtes pilotes, qui évalueront des échantillons des différentes régions linguistiques, et près du double d'enquêtes pré-pilotes, qui porteront chacune sur quelques classes de chaque région linguistique.

Les flèches vertes continues de la première illustration montrent qu'il est possible d'établir six observations de tendances entre les périodes successives de relevé au cours de ces quinze années. Ce qui n'apparaît pas explicitement, c'est le fait que les tendances s'étendant sur de plus longues périodes peuvent également être suivies, par ex. en se basant sur les quatre mesures effectuées en 4e année. Plus les séries de mesures

sont menées sur de longues périodes, meilleure est la fiabilité des études de tendances. Ce n'est qu'après avoir obtenu un certain nombre de mesures qu'il est possible d'identifier les valeurs aberrantes et les changements de tendance comme tels.

Les cinq flèches rouges continues représentent cinq possibilités d'étudier l'évolution de chacune des cohortes d'un degré à l'autre. La possibilité de comparer les évolutions de différentes cohortes, par ex. des cohortes 1 et 2, n'apparaît pas, tout comme celle d'analyser les différences au sein des différences afin d'étudier efficacement les causes de ces dernières. Le potentiel d'une base de données pour mener des études comparatives augmente dès lors qu'on réalise des enquêtes de manière systématique et persistante.

Avec un *design* longitudinal, le cadre conceptuel (*framework*) des tests et de la contextualisation, sur lequel se basent les enquêtes, doit satisfaire à des exigences plus élevées. En effet, celui-ci devrait dès le départ prévoir les trois moments de l'enquête dans des domaines clés, de manière à ne pas manquer inutilement des occasions d'analyses verticales entre les années de scolarité. Cela concerne en particulier les modèles de compétences des domaines disciplinaires, qui devraient si possible présenter une dimension évolutive entre les années de scolarité.

Si l'on adopte un *design* longitudinal, il faudra presque nécessairement élargir l'éventail des tests pour couvrir l'ensemble des compétences des élèves (voir point 2.1.4). Si le spectre de compétences des élèves ne peut être saisi que de manière insuffisante aux différents moments, il ne sera en effet pas possible d'établir des comparaisons verticales probantes. L'évolution des groupes au sein d'une cohorte ne pourra pas, par exemple, être suivie de manière fiable en raison du manque de précision des mesures.

2.2.3 Degré de standardisation des enquêtes

L'expertise s'est penchée sur la question de savoir s'il fallait que les cantons puissent mener eux-mêmes des enquêtes de vérification aux moments qu'ils choisiraient. Tant les acteurs interrogés que le groupe chargé de l'expertise ont répondu par un avis défavorable. L'expertise souligne que des mesures et des comparaisons telles que celles effectuées dans le cadre de la vérification sont sensibles à l'âge, à la période, à la cohorte et à la mise en œuvre. Pour pouvoir générer des données comparables, il faut donc contrôler bon nombre d'aspects différents et les traiter de la même manière dans chaque canton. Qui plus est, il faudrait déployer des efforts considérables pour garantir la sécurité des tests si ceux-ci étaient administrés de manière décentralisée et décalée dans le temps. Tous les supports devraient être conservés de manière sécurisée et accessibles par voie sécurisée également, tout comme pour une procédure centralisée, mais cela est nettement plus facile à faire quand il y a centralisation.

Selon le Bureau de coordination HarmoS, on peut y ajouter certaines objections de fond: en effet, la vérification n'entend pas permettre des formes d'auto-évaluation aux cantons individuellement; il s'agit au contraire d'un projet d'évaluation commun des cantons impliqués, permettant de saisir de manière standardisée et de comparer certaines prestations de formation fournies par les cantons. Pour toutes les raisons exposées ci-dessus, il semble indiqué de poursuivre la pratique actuelle.

La direction du projet et d'autres acteurs interrogés ont avancé l'idée que les enquêtes pourraient, pour des raisons de coûts, être conduites dorénavant sous la direction du corps enseignant en utilisant les ordinateurs à la disposition des écoles. Le groupe chargé de l'expertise conseille la prudence à cet égard car, dans d'autres contextes, on a pu observer certains effets liés au mode de réalisation: le fait de passer d'une administration externe assurée par des responsables de tests à une administration par le corps enseignant a eu un impact sur les résultats dans certains types de tâches. Il est notamment déconseillé de changer les conditions de réalisation entre l'enquête pilote et l'enquête principale.

2.3 Autres aspects liés à la mise en œuvre de la vérification

L'expertise se prononce sur toute une série d'aspects techniques et disciplinaires de la vérification et donne des avis et des recommandations sur ces éléments. Il s'agit des points suivants:

- Développement des tests et détermination des valeurs seuils
  - Ancrage didactique
  - Bases conceptuelles et concrètes pour le développement des tests
  - Adéquation de la difficulté des tâches
  - Traduction des tâches
  - Études de validation
  - Calendrier appliqué au processus de développement des tests
  - Création de *pools* de tâches
  - détermination des valeurs seuils
- Planification et mise en œuvre de la contextualisation (principalement des questionnaires)
- Utilisation des données pour des analyses secondaires
- Documentation relative aux cycles de vérification
- Rapports et travail de communication.

Le Bureau de coordination HarmoS a examiné ces différents points. Comme il s'agit essentiellement d'une discussion technique concernant principalement le niveau de la mise en œuvre, elle ne sera pas reproduite ici. Elle peut toutefois être consultée à l'annexe 1.

2.4 Gouvernance de la vérification

Comme mentionné au point 1.2.3, la vérification n'a reçu la structure prévue dans les bases légales qu'en vue des cycles 2020 et 2022 (respectivement 2023 et 2024). Elle consiste à laisser l'Université de Berne et son centre interfacultaire, l'*Interfaculty Centre for Educational Research* (ICER), assumer des fonctions essentielles de coordination et de direction au niveau opérationnel. L'expertise salue le réaménagement des structures comme constituant un progrès, mais identifie aussi d'autres améliorations à apporter dans le domaine de la gouvernance.

2.4.1 Délimitation entre politique et administration de l'éducation d'un côté et science de l'autre

Le groupe chargé de l'expertise rappelle que la science joue deux rôles différents en ce qui concerne la vérification: 1) prestations scientifiques, 2) participation à la vérification par intérêt propre (notamment pour la contextualisation ou via des analyses secondaires des données issues de la vérification). Il est à noter que, dans le cas de la vérification, l'intérêt de la science est en principe subordonné à celui de la politique. Cependant, la réalisation des travaux scientifiques doit s'effectuer sans influence politique, dans le seul respect des normes scientifiques.

Le groupe chargé de l'expertise considère que les tâches suivantes, qui sont actuellement assumées par la politique et l'administration de l'éducation, incombent aux acteurs scientifiques:

- Détermination des valeurs seuils: le règlement relatif à la vérification prévoit que c'est l'Assemblée plénière de la CDIP qui décide des valeurs seuils et peut donc influer sur la proportion d'élèves considérés comme ayant acquis les compétences fondamentales. De plus, le processus technique de détermination de ces valeurs, dont découle les propositions de valeurs seuils, relève de la compétence de la Banque de données de tâches (BDT), qui fait également partie de la CDIP. Le groupe chargé de l'expertise estime que la détermination des valeurs seuils constitue une tâche scientifique, qui ne devrait pas être exécutée par un organisme administratif.
- Approbation du questionnaire contextuel: le questionnaire contextuel pour les élèves utilisé dans le cadre de la vérification doit, selon les bases légales actuellement en vigueur, être présenté au Bureau de coordination HarmoS pour consultation et au Comité de la CDIP pour approbation. Cette pratique s'est établie parce que certaines questions, en particulier celles qui portent sur le contexte social de l'élève, sont jugées délicates et pas toujours adaptées aux circonstances actuelles. L'expertise considère le développement des questionnaires comme constituant une tâche purement scientifique et propose qu'au lieu d'une approbation définitive du questionnaire, un cadre conceptuel (*framework*) soit

négocié et adopté de manière contraignante entre les mandants et les mandataires au début des travaux de contextualisation.

- Rattachement de la BDT au Secrétariat général de la CDIP (SG CDIP): si la BDT est bien responsable des domaines (scientifiques) essentiels du développement des tâches et des tests, de même que de la détermination des valeurs seuils, elle n'est guère impliquée dans la vérification sur le plan scientifique. Selon l'expertise, ce positionnement non seulement affaiblit la crédibilité de la détermination des valeurs seuils, mais limite également les possibilités d'échanges pour le personnel scientifique, et ne lui offre que peu de perspectives. Par ailleurs, la faible dotation en personnel (actuellement, une seule personne; auparavant deux) s'accompagne de risques importants de perte de savoir-faire. L'expertise recommande donc que la BDT soit détachée du SG CDIP et que ses fonctions soient intégrées dans les structures scientifiques.

Le Bureau de coordination HarmoS est favorable à une séparation fonctionnelle claire entre le niveau de la politique et de l'administration de l'éducation et celui de la réalisation scientifique de la vérification. Comme outil pour y parvenir, il semble judicieux de convenir d'un cadre conceptuel (*framework*) qui définisse de manière contraignante les principaux paramètres sur la base des objectifs de la vérification, et plus précisément d'une enquête, surtout pour ce qui est des domaines de l'échantillonnage, du développement des tests et de la contextualisation. Une fois l'engagement mutuel pris vis-à-vis de ce cadre, l'institution scientifique mandataire réalise les travaux de développement avec ses partenaires contractuels, selon les normes scientifiques admises, et assume la responsabilité du résultat vis-à-vis de la CDIP.

Le Bureau de coordination HarmoS trouve également judicieux d'intégrer dans les structures scientifiques les fonctions de la BDT qui sont liées à la vérification.

En même temps, il propose de renoncer à certaines fonctions qui sont attribuées à la BDT dans le règlement, mais qu'elle ne doit pas remplir à l'heure actuelle. Sont concernées les fonctions suivantes:

a) mettre à disposition une structure technique multifonctions et plurilingue,
b) garantir une échelle édumétrique applicable dans toute la Suisse (et pas uniquement pour la vérification),
d) coordonner l'échange des tâches servant à la vérification de l'atteinte des objectifs nationaux de formation (compétences fondamentales) ainsi que l'échange des tâches servant aux projets cantonaux.

Le Bureau HarmoS fait remarquer que la nécessité de garantir une échelle édumétrique commune ainsi que l'échange des tâches dans toute la Suisse n'a pas disparu, même si les prestations de la BDT en question n'ont pas été sollicitées au cours des dernières années.

2.4.2 «Diffusion» dans la fonction de direction de l'ICER et risques d'interface

Selon la convention de coopération conclue entre la CDIP et l'Université de Berne en ce qui concerne la réalisation des deux cycles prévus, les fonctions que doit assumer l'ICER dans le projet Vérification relèvent tantôt de la coordination, tantôt du pilotage et de la direction. Parallèlement, l'Université de Berne assume également la responsabilité du résultat final. L'expertise y voit une forme de «diffusion» de la fonction de direction. À cet égard, il convient de distinguer deux aspects. Le premier concerne les relations contractuelles entre les institutions qui sont responsables de la mise en œuvre de la vérification: si l'Université de Berne conclut de son côté des conventions de coopération avec la plupart des institutions impliquées (IBE, PHSG, SRED, SUPSI et, pour l'enquête 2024, avec la FHGR également), elle a aussi particulièrement besoin, pour mener à bien le projet, des contributions essentielles pour le contenu que fournit la BDT, qui elle fait partie du SG CDIP. Le groupe chargé de l'expertise estime qu'il en résulte certains risques d'interface, principalement pour l'Université de Berne, et que ces risques devraient être écartés. Il recommande donc, pour ces raisons notamment, d'intégrer les fonctions de la BDT (et de la FHGR) dans la structure de projet placée sous la conduite de l'ICER. Le second aspect à mentionner concerne la fonction de conduite de l'ICER au sein de l'organe chargé de diriger le projet et qui comprend des éléments de coopération, de coordination, de pilotage et de direction à proprement parler. L'expertise recommande d'adopter, pour la direction du projet Vérification, des structures de compétence, de décision et de responsabilité qui soient claires et simplifiées, et qui puissent aussi résister aux conflits ou à la pression du temps. L'expertise indique

que la mise en place d'une structure organisationnelle efficace dépend aussi du fait qu'une planification à long terme définisse les paramètres principaux, par exemple le rythme des enquêtes. Selon le groupe chargé de l'expertise, le but serait de disposer d'un centre de compétences stable, doté d'un savoir-faire constant.

Le Bureau de coordination HarmoS partage le constat selon lequel des compétences définies de manière imprécise ou partielle comportent des risques d'interface et salue donc la proposition d'intégrer les fonctions de la BDT dans les structures scientifiques. La question d'une plus grande hiérarchisation des relations entre l'ICER et les autres institutions actuellement représentées au sein de l'organe de direction doit être examinée avec soin. Il s'agit notamment de tenir compte des éléments suivants:

- La CDIP tient à ce que la vérification soit menée au sein de son propre système éducatif, notamment parce qu'il lui semble important de développer et maintenir les compétences nécessaires à ce projet.
- Il existe une volonté politique de faire collaborer les trois régions linguistiques dans le cadre du projet Vérification. La convention passée entre la CDIP et l'Université de Berne prévoit par exemple, au point 3.2, que l'ICER travaille en collaboration avec les régions linguistiques.
- La collaboration entre les institutions de toutes les régions linguistiques semble non seulement souhaitable, mais aussi indispensable. La vérification, avec ses grands échantillons, doit réussir à suivre et à motiver durablement les acteurs d'un domaine pratique étendu, ce qui n'est guère envisageable sans un ancrage régional.
- La réponse que l'on peut apporter lorsqu'il est fait mention de craintes face à d'éventuels risques d'interface, c'est le fait que la direction du projet a manifestement trouvé jusqu'ici des formes de collaboration qui sont tout à fait fructueuses.

Le Bureau de coordination HarmoS a examiné, dans le contexte de la gouvernance, la question de la mise en œuvre du Consortium scientifique pour la recherche et l'évaluation en éducation (cf. point 1.2.3), qui est un sujet de discussion depuis le début du projet, tout comme la question d'une prise en compte équivalente de toutes les régions linguistiques dans le cadre de la vérification. Il soumettra un certain nombre de propositions à ce sujet.

2.4.3 Place des didactiques des disciplines

Jusqu'à présent, les didactiques des disciplines n'étaient intégrées qu'indirectement dans les structures de projet; autrement dit, elles ne font pas partie des groupes de travail thématiques de la direction de projet. Elles fournissent une partie essentielle du développement des tâches, en collaboration avec la BDT, contribuent aux ateliers menés en vue de déterminer les valeurs seuils et apportent aussi des éléments importants pour le travail de compte rendu des résultats. À l'instar d'autres scientifiques, elles ont en outre la possibilité de faire valoir leurs propres intérêts (construits) dans la contextualisation. Le groupe chargé de l'expertise recommande d'intégrer davantage les didactiques des disciplines dans le projet et, en même temps, de les professionnaliser en vue de leurs tâches dans le cadre de la vérification. L'intégration des didactiques des disciplines doit notamment permettre d'ajouter une composante didactique dans l'utilisation des données et des résultats, ce qui peut avoir un effet positif sur la formation et le perfectionnement du corps enseignant et sur l'enseignement, et contribuer à l'efficacité des objectifs de formation.

Pour le Bureau de coordination HarmoS, une meilleure intégration, avec un rôle plus central, de la didactique des disciplines dans les structures scientifiques de la vérification est une étape attendue depuis longtemps, qui constitue d'ailleurs un complément logique à la redistribution des fonctions de la BDT associées à la vérification. Le développement des tâches et des tests doit rester le principal domaine d'activité des didactiques des disciplines. À cet égard, ces dernières doivent encore voir leur assise renforcée (cf. annexe 1, point *Développement des tests, détermination des valeurs seuils*). Elles devront par ailleurs jouer à l'avenir un rôle important dans l'élaboration et le renouvellement périodique du cadre conceptuel (*framework*) de chaque enquête.

3 Étapes en vue du développement de la vérification

Le Bureau de coordination HarmoS accompagne les travaux de la vérification depuis le début. Il a été institué par les organes de la CDIP en vue de créer les bases nécessaires, d'assister la mise en œuvre et d'élaborer des bases décisionnelles adéquates. Sur mandat des organes de la CDIP, il s'est penché de manière approfondie sur l'évaluation de la mise en œuvre de la vérification, telle qu'elle a été pratiquée jusqu'à présent, en se basant sur l'expertise réalisée et en maintenant un dialogue avec d'autres spécialistes. En vue de la poursuite du projet, il a pris connaissance des constats et des recommandations formulés à cet égard et a évalué plusieurs options.

La grande majorité des membres du Bureau HarmoS sont favorables à la poursuite et au développement de la vérification. Pour certains membres cependant, on pourrait renoncer complètement à réaliser ces enquêtes au profit de l'étude PISA, qui porterait alors aussi sur des échantillons représentatifs au niveau cantonal d'élèves de 11^{e} année, même si le concordat HarmoS de 2007 prévoit explicitement une vérification de l'atteinte des compétences fondamentales comme instrument de pilotage et d'harmonisation.

En vue des décisions à prendre s'agissant de la poursuite de la vérification, le Bureau HarmoS présente aux organes de la CDIP, dans un premier temps, les principaux aspects d'un monitorage des compétences fondamentales. Ces aspects concernent le futur *design* d'enquête à adopter ainsi que diverses composantes de la gouvernance (institutions, structures, compétences). Il est proposé d'apporter plusieurs améliorations à la pratique actuelle. Le but des aspects principaux présentés ici en lien avec la poursuite du programme est d'accroître la valeur et l'utilité des enquêtes pour le monitorage de la formation et ainsi de permettre un pilotage des systèmes éducatifs en Suisse qui soit fondé sur des données probantes.

Dès lors que ces principaux aspects auront été définis, il conviendra d'adapter les bases légales pour la mise en œuvre du monitorage des compétences fondamentales puis de les soumettre aux organes de la CDIP.

3.1 Nouveau *design* d'enquête

Les quatre enquêtes réalisées jusqu'à présent et planifiées jusqu'en 2024 dans le cadre de la vérification de l'atteinte des compétences fondamentales revêtent un caractère de mesures individuelles uniques. Un *design* d'enquête défini à plus long terme, suivant un rythme fixé de manière ciblée et établissant des liens entre les différentes enquêtes, présente des avantages considérables, par rapport aux études transversales: cela permet de dégager des tendances et d'évaluer de manière plus pertinente les performances de l'école obligatoire. Par ailleurs, la contribution des enquêtes au monitorage national et cantonal de la formation gagne en prévisibilité et s'en trouve donc plus pertinent pour toute une série d'acteurs.

Après avoir mené, lors de quatre séances qui se sont tenues entre le 24 janvier et le 28 septembre 2022, une réflexion incluant différentes variantes pour le *design* d'enquête, le Bureau de coordination HarmoS présente finalement quatre variantes au Comité. Toutes les variantes prévoient un *design* d'enquête basé sur la régularité et une planification à long terme.

Les variantes se déclinent selon les quatre caractéristiques suivantes:

- les **domaines disciplinaires** testés: seuls peuvent être testés, dans le cadre de la vérification, les quatre domaines disciplinaires dans lesquels des compétences fondamentales ont été définies, à savoir la langue de scolarisation, les mathématiques, les langues étrangères et les sciences. Pour une question de coûts, il a été décidé de renoncer, dans toutes les variantes, à la vérification des compétences fondamentales en sciences.

  Dans l'étude PISA, trois domaines disciplinaires doivent obligatoirement être testés, à savoir la langue de scolarisation (lecture), les mathématiques et les sciences (selon la définition donnée par l'OCDE), et ce, auprès des jeunes de 15 ans.

Deux variantes proposent que tous les cantons participent à PISA avec à chaque fois un échantillon représentatif de la 11e année. Depuis 2015, la Suisse participe à PISA avec un échantillon national de jeunes de 15 ans. Cette participation sera maintenue indépendamment de la variante qui sera choisie.

- les **degrés scolaires** auxquels les tests sont menés (4e, 8e et/ou 11e années).
- la **cadence** des enquêtes: le point de départ, dans toutes les variantes, correspond au cycle quadriennal du rapport sur l'éducation et des études PISA à partir de 2029. La variante B prévoit une cadence de huit ans pour les langues étrangères.
- des **études par cohortes**: toutes les variantes (A à D) permettent de dégager des tendances grâce à la répétition des différentes mesures à un rythme quadriennal. Toutefois, les variantes se distinguent par la possibilité de suivre l'évolution des différentes volées au cours de la scolarité obligatoire, et par la durée de ce suivi (études longitudinales par cohortes). La variante D ne permet pas de suivre l'évolution de cohortes. Il n'est alors plus possible d'évaluer les performances des différents cycles scolaires.

Tableau 1: Variantes pour le design d'enquête

	Programme actuel	**Variante A**	**Variante B**	**Variante C**	**Variante D**
Domaines disciplinaires HarmoS	Langue de scolarisation Mathématiques Langues étrangères	Langue de scol. Mathématiques Langues étrangères	Langue de scol. (4e/8e) Mathématiques (4e/8e) Langues étr. (8e /11e)	Langue de scol. Mathématiques Langues étrangères	-
Domaines disciplinaires PISA (échantillon supplémentaire en 11e année)	-	-	Langue de scol. Mathématiques Sciences	-	Langue de scol. Mathématiques Sciences
Degrés scol.	4e/8e /11e	4e/8e/11e	4e/8e/11e	8e/11e	11e
Cadence	non défini	tous les 4 ans	tous les 4 ans / pour les langues étr.: tous les 8 ans	tous les 4 ans	tous les 4 ans
Études par cohortes	-	4e→8e→11e	4e→8e→11e	8e→11e	-
Nb d'enquêtes (vérification) en 15 ans	non défini	9	10	7	-
Nb d'enquêtes PISA auprès d'él. de 11e année en 15 ans	-	-	4	-	4
Coûts du programme par an Ø 2025-2041*	1,5 million de francs	2,25 millions de francs	2,3 millions de francs	1,95 million de francs	-
Coûts additionnels PISA en 11e année par an	-	-	0,5 million de francs	-	0,5 million de francs
Coûts de PISA pour les jeunes de 15 ans par an	0,6 million de francs	0,6 million de francs	0,6 million de francs	0,6 million de francs	0,6 million de francs
Total des coûts par an	2,1 millions de francs	2,85 millions de francs	3,4 millions de francs	2,55 millions de francs	1,1 million de francs

* La base de planification utilisée est celle couvrant les enquêtes principales qui seront menées durant la période de 15 ans allant de 2027 à 2041. Si une enquête principale doit être réalisée en 2027, il faut que les travaux préparatoires commencent au plus tard en 2025.

3.1.1 Les variantes proposées

Les sections suivantes présentent les variantes A à D par une brève description et les évaluent lorsque c'est possible à l'aide de différents critères particulièrement pertinents pour l'analyse. Les membres du Bureau de coordination HarmoS ne sont pas tous d'accord sur l'ampleur des charges pour chaque *design* d'enquête, c'est pourquoi aucune évaluation de ces dernières ne sera réalisée.

La préférence des membres du Bureau HarmoS va plutôt aux variantes A et C, qui ont obtenu une adhésion nettement plus grande que les variantes B et D. Dans l'ensemble, le Bureau recommande de mettre en œuvre la variante C.

- **Variante A**

La variante A prévoit une collecte de données sur les performances et le contexte vers la fin de la 4e, de la 8e et de la 11e année. Elle permet notamment de vérifier l'atteinte des compétences fondamentales dans trois degrés scolaires et dans les disciplines que sont la langue de scolarisation, les mathématiques et les langues étrangères. Les sciences n'entrent pas en ligne de compte.

Variante A: vérification en 4e, 8e et 11e années

Année n°	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Cohorte 1	4				8			11							
Cohorte 2					4				8			11			
Cohorte 3									4				8		
Cohorte 4													4		

Les mesures faites à trois moments différents pour une même cohorte permettent, de manière profitable pour le monitorage de l'éducation, d'évaluer les performances du 2e et du 3e cycle de la scolarité obligatoire, grâce aux mesures d'entrée et de sortie qui sont effectuées pour ces deux cycles.

Tous les quatre ans (cadence), les enquêtes commencent en 4e année avec une nouvelle cohorte. On peut ainsi identifier et suivre au fil des ans les tendances qui se dessinent en 4e, 8e et 11e année.

Indépendamment de cela, la Suisse ne participe à l'étude PISA qu'avec un échantillon représentatif à l'échelle nationale de jeunes de 15 ans, comme cela se passe depuis 2015. Y sont testés les domaines dits «principaux» de la langue de scolarisation (lecture), des mathématiques et des sciences.

Évaluation globale: en comparaison avec les autres variantes, c'est la variante A qui permet le mieux de mettre en œuvre les objectifs fixés dans le concordat HarmoS. Les résultats sont d'une grande pertinence du fait qu'on teste tous les domaines disciplinaires – hormis les sciences – à tous les niveaux entrant en ligne de compte.

Tableau 2: Évaluation de la variante A sur la base de différents critères

Critère	**Évaluation**
Concordat HarmoS de 2007: vérification de l'atteinte des objectifs de formation qui servent à l'harmonisation	+ vérification des objectifs de formation en langue de scolarisation + vérification des objectifs de formation en mathématiques + vérification des objectifs de formation en langues étrangères – pas de vérification des objectifs de formation en sciences
Concordat HarmoS de 2007: contribution à une évaluation systématique et continue des développements survenant dans le système et de ses performances	+ mesures d'entrée et de sortie mises en œuvre pour tous les cycles de la scolarité obligatoire qui sont concernés: • mesures de tendances sur les développements en 4e, 8e et 11e années • évaluation des performances des 2e et 3e cycles
Utilité pour: - le monitorage de l'éducation - le pilotage de l'éducation - la recherche en éducation	+ création d'une base de données exhaustive pour les 2e et 3e cycles de la scolarité obligatoire (compétences et contexte) • mesures d'entrée et de sortie complètes pour les études de tendances et les études par cohortes • livraison régulière de données • possibilité d'intégrer des intérêts de recherche propres dans la contextualisation • possibilités d'appariement avec d'autres ensembles de données des pouvoirs publics
Développement et maintien des compétences dans le domaine des évaluations à large échelle	+ des structures professionnelles permanentes sont indispensables pour une mise en œuvre réussie – exclusion de la didactique des sciences
Charge pour les écoles et les administrations occasionnée par les enquêtes	En 15 ans auraient lieu: • 9 enquêtes principales pour la vérification, 9 enquêtes pilotes, au max. 18 enquêtes pré-pilotes • 4 enquêtes principales pour PISA (avec échantillons nationaux), plus 4 enquêtes pilotes (*field trials*)
Coûts moyens par an sur 15 ans à partir de la première enquête principale (pour la vérification, compter en plus 2 ans de préparation)	• pour la vérification: env. 2,25 millions de francs par an (coûts du projet) • PISA: pas de coûts supplémentaires, car seulement échantillon national composé de jeunes de 15 ans (coûts pris en charge à 50 % par la Confédération; part de la CDIP équivalant à env. 0,6 million par an)

- **Variante B**

La variante B se caractérise par une imbrication des enquêtes menées au titre de la vérification et des enquêtes PISA, avec des échantillons représentatifs à l'échelle cantonale d'élèves de 11e année (comme pour la vérification en 11e année).

Variante B: imbrication des études de vérification et de l'étude PISA en 11e année															
Année n°	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Cohorte 1			PISA												
Cohorte 2				8			PISA								
Cohorte 3				4				8			PISA				
Cohorte 4								4				8			PISA
Cohorte 5												4			
Cohorte 1 LE		8			11										
Cohorte 2 LE										8			11		

En 4e et en 8e années sont menées, tous les quatre ans, des enquêtes visant à évaluer l'atteinte des compétences fondamentales dans les domaines disciplinaires que sont la langue de scolarisation et les mathématiques. En 11e année, la même cohorte participe à chaque fois à l'étude PISA, avec des échantillons représentatifs à l'échelle cantonale. Indépendamment de cela, dans d'autres cohortes, des mesures sont effectuées tous les huit ans en 8e et en 11e années pour tester les compétences en langues étrangères.

Ce *design* permet, comme dans la variante A, d'évaluer les performances du 2e et du 3e cycle dans le cadre du monitorage de l'éducation. Pour ce qui est de la vérification de l'atteinte des compétences fondamentales, cette variante implique de renoncer non seulement aux sciences, mais aussi aux domaines disciplinaires langue de scolarisation et mathématiques en fin de scolarité obligatoire. Le fait d'intégrer l'étude PISA dans la série de mesures entraîne également des conséquences pour la contextualisation: les questionnaires contextuels doivent être adaptés à PISA pour remplir leur objectif (à savoir essentiellement le suivi des changements). La question de la contextualisation en 8e et 11e années pour les langues étrangères n'a pas encore été réglée.

La variante B permet aussi, tout comme la variante A, de dégager des mesures de tendances pour les 4e, 8e et 11e années. Mais, pour le domaine des langues étrangères, l'intervalle de huit ans entre les mesures constitue un laps de temps très long pour ce but.

La participation à l'étude PISA se fait, dans la variante B, avec d'une part des échantillons représentatifs à l'échelle cantonale d'élèves de 11e année (ce qui constitue une nouveauté) et, de l'autre, avec un échantillon représentatif à l'échelle nationale de jeunes de 15 ans, comme cela s'est fait ces dernières années.

Évaluation globale: la variante B permet d'obtenir des constatations dans les quatre domaines disciplinaires. Mais les exigences correspondent en partie à la conception de l'OCDE. Au terme de la scolarité obligatoire, il n'est pas possible de dégager des indications quant à l'atteinte des objectifs de formation définis, sauf pour les langues étrangères. Du fait de l'imbrication qu'elle prévoit entre la vérification et PISA, la variante B s'avère compliquée et ambitieuse en termes de mise en œuvre et de communication.

Tableau 3: Évaluation de la variante B sur la base de différents critères

Critère	**Évaluation**
Concordat HarmoS de 2007: vérification de l'atteinte des objectifs de formation qui servent à l'harmonisation	± vérification des objectifs de formation en langue de scolarisation en 4e et en 8e années; pas de vérification en 11e année; PISA à la place ± vérification des objectifs de formation en mathématiques en 4e et en 8e années; pas de vérification en 11e année; PISA à la place

Critère	Évaluation
	– pas de vérification des objectifs de formation en sciences; PISA en 11e année + vérification des objectifs de formation en langues étrangères, mais à une cadence moins élevée (tous les 8 ans)
Concordat HarmoS de 2007: contribution à une évaluation systématique et continue des développements survenant dans le système et de ses performances	+ mesures d'entrée et de sortie mises en œuvre pour tous les cycles de la scolarité obligatoire qui sont concernés: • mesures de tendances sur les développements en 4e, 8e et 11e années • évaluation des performances des 2e et 3e cycles
Utilité pour: - le monitorage de l'éducation - le pilotage de l'éducation - la recherche en éducation	+ création d'une base de données exhaustive pour les 2e et 3e cycles de la scolarité obligatoire (compétences et contexte) • mesures d'entrée et de sortie complètes pour les études de tendances et les études par cohortes • livraison régulière de données • orientation vers la contextualisation de l'étude PISA • possibilités d'appariement avec d'autres ensembles de données des pouvoirs publics – en raison de la coordination à garantir avec PISA en 11e année, il n'est guère possible d'intégrer des items ou intérêts de recherche propres dans le test et/ou la contextualisation ± la vérification des objectifs de formation en langues étrangères s'effectue certes à une faible cadence, mais elles pourrait permettre d'intégrer des intérêts de recherche propres dans la contextualisation
Développement et maintien des compétences dans le domaine des évaluations à large échelle	+ des structures professionnelles permanentes sont indispensables pour une mise en œuvre réussie – exclusion de la didactique des sciences
Charge pour les écoles et les administrations occasionnée par les enquêtes	En 15 ans auraient lieu: • 10 enquêtes principales pour la vérification, 10 enquêtes pilotes, au max. 20 enquêtes pré-pilotes • 4 enquêtes principales pour PISA avec des échantillons qui sont plus vastes que ceux de la vérification en 11e année; plus 4 enquêtes pilotes (field trials)
Coûts moyens par an sur 15 ans à partir de la première enquête principale (pour la vérification, compter en plus 2 ans de préparation)	• pour la vérification: env. 2,3 millions de francs par an (coûts du projet) • PISA: coûts supplémentaires d'env. 0,5 million par an pour les échantillons représentatifs à l'échelle cantonale; (coûts pour l'échantillon national pris en charge à 50 % par la Confédération; part de la CDIP équivalant à env. 0,6 million par an)

- **Variante C**

Le *design* de la variante C correspond essentiellement à celui de la variante A, mais sans enquête en 4e année. Les enquêtes, qui se limitent à la 8e et à la 11e année, portent notamment sur la vérification des compétences fondamentales dans les domaines disciplinaires que sont la langue de scolarisation, les mathématiques et les langues étrangères. Les sciences n'entrent pas en ligne de compte.

Variante C: vérification en 8e et 11e années															
Année n°	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Cohorte 1	8			11											
Cohorte 2					8			11							
Cohorte 3									8			11			
Cohorte 4													8		

Les mesures effectuées en 8e et en 11e années pour une même cohorte permettent d'évaluer, de manière profitable pour le monitorage de l'éducation, uniquement les performances du troisième cycle (degré sec. I) de la scolarité obligatoire. Il manque les mesures de sortie pour le premier cycle, qui sont en même temps les mesures d'entrée pour le deuxième cycle.

Dans la variante C, les enquêtes commencent aussi avec une nouvelle cohorte tous les quatre ans, mais seulement depuis la 8e année. Il est alors possible de suivre les tendances qui se profilent au fil des ans pour la 8e et pour la 11e année.

Indépendamment de cela, la Suisse ne continue de participer à l'étude PISA qu'avec un échantillon représentatif à l'échelle nationale de jeunes de 15 ans, comme dans la variante A.

Évaluation globale: la variante C permet de mettre en œuvre les objectifs fixés dans le concordat HarmoS, avec toutefois de nettes restrictions par rapport à la variante A. Les résultats sont d'une grande pertinence pour la 8e et pour la 11e années, de même que pour le troisième cycle (sec. I) du fait que l'on teste trois domaines disciplinaires. Mais le fait de renoncer aux enquêtes en 4e année réduit considérablement la pertinence des résultats puisqu'il manque les mesures de sortie pour le premier cycle et donc également les mesures d'entrée pour le deuxième.

Tableau 4: Évaluation de la variante C sur la base de différents critères

Critère	**Évaluation**
Concordat HarmoS de 2007: vérification de l'atteinte des objectifs de formation qui servent à l'harmonisation	± vérification des objectifs de formation en langue de scolarisation en 8e et en 11e années ± vérification des objectifs de formation en mathématiques en 8e et en 11e années ± vérification des objectifs de formation en langues étrangères en 8e et en 11e années – l'atteinte des objectifs de formation n'est pas vérifiée en 4e année.
Concordat HarmoS de 2007: contribution à une évaluation systématique et continue des développements survenant dans le système et de ses performances	± il n'est possible de dégager qu'une partie des indications sur les tendances et des constatations sur les performances des cycles: • mesures de tendances sur les développements en 8e et en 11e année • évaluation des performances du 3e cycle

Critère	Évaluation
Utilité pour: - le monitorage de l'éducation - le pilotage de l'éducation - la recherche en éducation	± création d'une base de données exhaustive pour les 2e et 3e cycles de la scolarité obligatoire (compétences et contexte) • mesures d'entrée et de sortie pour les études de tendances et les études par cohortes seulement dans deux années de scolarité • livraison régulière de données • possibilité d'intégrer des intérêts de recherche propres dans la contextualisation • possibilités d'appariement avec d'autres ensembles de données des pouvoirs publics
Développement et maintien des compétences dans le domaine des évaluations à large échelle	+ des structures professionnelles permanentes sont indispensables pour une mise en œuvre réussie – exclusion de la didactique des sciences
Charge pour les écoles (et, au niveau de la mise en œuvre, pour les administrations cantonales) occasionnée par les enquêtes	En 15 ans auraient lieu: • 7 enquêtes principales pour la vérification, 7 enquêtes pilotes, max. 14 enquêtes pré-pilotes • 4 enquêtes principales pour PISA (avec échantillons nationaux), en plus 4 enquêtes pilotes (field trials)
Coûts moyens par an sur 15 ans à partir de la première enquête principale (pour la vérification, compter en plus 2 ans de préparation)	• pour la vérification: env. 1,95 million de francs par an • PISA: pas de coûts supplémentaires, car seulement échantillon national composé de jeunes de 15 ans (coûts pris en charge à 50 % par la Confédération; part de la CDIP équivalant à env. 0,6 million par an)

- **Variante D**

La variante D propose de renoncer complètement à la vérification de l'atteinte des compétences fondamentales. La Suisse ne participerait plus qu'à l'étude PISA, et ce, avec d'une part des échantillons représentatifs à l'échelle cantonale d'élèves de 11e année et, de l'autre, avec un échantillon représentatif à l'échelle nationale de jeunes de 15 ans, comme cela s'est fait jusqu'à présent.

Variante D: PISA en 11e année

Année n°	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Cohorte 1			PISA												
Cohorte 2							PISA								
Cohorte 3											PISA				
Cohorte 4															PISA

Jusqu'en 2012, la Suisse participait à PISA avec des échantillons représentatifs des régions linguistiques et, en partie, des cantons. Depuis 2015, avec l'introduction de la vérification, la Suisse ne participe plus à PISA qu'avec un échantillon représentatif à l'échelle nationale.

Renoncer complètement au programme équivaut à ne plus pouvoir vérifier les objectifs de formation, donc l'atteinte des compétences fondamentales. Des études de tendances peuvent uniquement être réalisées en lien avec la fin de la scolarité. En l'absence de mesures d'entrée, il est impossible d'évaluer les performances de la scolarité obligatoire au profit du monitorage de l'éducation.

Évaluation globale: la variante D ne permet pas de poursuivre la vérification ni de mettre en application l'art. 10 du concordat HarmoS relatif aux standards de formation (compétences fondamentales). Elle se concentre exclusivement sur les résultats de l'étude PISA et sur les comparaisons pouvant être obtenues en 11e année sur la base d'échantillons supplémentaires représentatifs au niveau cantonal pour les enquêtes PISA.

Tableau 5: Évaluation de la variante D sur la base de différents critères

Critère	**Évaluation**
Concordat HarmoS de 2007: vérification de l'atteinte des objectifs de formation qui servent à l'harmonisation	– pas de vérification des objectifs de formation (compétences fondamentales)
Concordat HarmoS de 2007: contribution à une évaluation systématique et continue des développements survenant dans le système et de ses performances	± mesures de tendances sur les développements en 11e année uniquement, sur la base d'enquêtes PISA – aucune évaluation des performances des cycles de la scolarité obligatoire
Décision de 2007 relative à la mise en œuvre: vérification pour la comparaison intercantonale et PISA avec échantillon national	– non, retour à la situation d'avant 2015-2016
Utilité pour: - le monitorage de l'éducation - le pilotage de l'éducation - la recherche en éducation	± création d'une base de données plus complète qu'à l'heure actuelle (compétences et contexte) concernant les élèves de 11e année scolaire, sur la base de l'étude PISA • livraison de données régulière, bien que relativement peu fréquente • pas d'étude par cohortes • possibilité seulement marginale d'intégrer des intérêts de recherche propres dans le test et/ou la contextualisation • possibilités d'appariement avec d'autres ensembles de données des pouvoirs publics
Développement et maintien des compétences dans le domaine des évaluations à large échelle	– Pas ou peu de possibilités, seulement réalisation et établissement de rapports ± Recherche locale possible sur la base des données PISA
Charge pour les écoles et les administrations occasionnée par les enquêtes	En 15 ans auraient lieu: 4 enquêtes principales pour PISA avec des échantillons qui sont chacun plus vastes que ceux de la vérification en 11e année puisque combinés avec un échantillon national de jeunes de 15 ans; plus 4 enquêtes pilotes (*field trials*)
Coûts moyens par an sur 15 ans à partir de la première enquête principale	PISA: coûts supplémentaires d'env. 0,5 million par an pour les échantillons représentatifs à l'échelle cantonale; (coûts pour l'échantillon national pris en charge à 50 % par la Confédération; part de la CDIP équivalant à env. 0,6 million de francs par an)

3.1.2 Élargissement à la totalité du spectre de performances des tests menés au titre de la vérification

Actuellement, la vérification permet, du fait qu'elle met l'accent sur les compétences fondamentales, de dégager des indications sur le degré d'atteinte des compétences fondamentales, mais ne livre que peu, voire pas de constatations sur les différences de performances entre les élèves qui se situent dans les parties moyenne et supérieure du spectre. De ce fait, les résultats ne se prêtent pas à certaines analyses importantes qui sont nécessaires pour permettre un pilotage qui soit fondé sur des données probantes (par ex. analyses sur l'évolution des performances de certains groupes d'élèves par rapport à l'évolution des performances d'autres groupes au sein d'un même cycle).

En tant qu'organe de pilotage du programme, le Bureau de coordination HarmoS est parvenu à la conclusion que les tests menés au titre de la vérification ne devraient à l'avenir pas seulement servir à vérifier l'atteinte des compétences fondamentales, mais aussi en même temps permettre de saisir les compétences sur l'ensemble du spectre de performances. Le rapport coût/bénéfice peut ainsi être nettement amélioré. C'est également le moyen de tenir la promesse qui a été faite de remplacer complètement, avec la vérification, les comparaisons entre régions et cantons qui se faisaient habituellement dans le cadre de l'étude PISA auparavant. PISA utilise des échelles de compétences à plusieurs niveaux. La distribution entre les niveaux joue un rôle important dans le travail de compte rendu des résultats. En ce qui concerne la vérification, ce dernier doit continuer de se concentrer en premier lieu sur la vérification de l'atteinte des compétences fondamentales. En même temps, il doit aussi tirer parti des connaissances sur la répartition des compétences pour formuler des conclusions.

Il est prévu d'étendre la vérification à l'ensemble du spectre de performances dans le cas des variantes A, B et C. Aucun nouvel objectif de formation n'est adopté pour ce faire. La mise en œuvre se fait dans le cadre du développement planifié des cadres de test (*test frameworks*) ainsi que des tâches et des tests. Elle n'entraîne pas de coûts de développement allant au-delà de ce qui était prévu.

3.2 Adaptations en matière de gouvernance

3.2.1 Rattachement de la Banque de données de tâches à une institution scientifique et abandon de l'idée de mettre en place une banque de données de tâches à d'autres fins

La Banque de données de tâches (BDT) fait partie du SG CDIP; son secrétariat est «domicilié» auprès du CSRE.

Aux termes de l'art. 3 de son règlement d'organisation du 23 janvier 2014, la BDT assume notamment les tâches suivantes:

a. mettre à disposition une structure technique multifonctions et plurilingue,
b. garantir une échelle édumétrique applicable dans toute la Suisse,
c. instaurer un processus pour le développement des tâches et des tests et pour l'étalonnage de tâches par rapport aux compétences fondamentales,
d. coordonner l'échange des tâches servant à la vérification de l'atteinte des objectifs nationaux de formation (compétences fondamentales) ainsi que l'échange des tâches servant aux projets cantonaux.

Selon l'art. 4 de son règlement d'organisation (*Services*), la BDT peut également accepter des mandats concernant le développement, la mise à disposition ou l'étalonnage de tâches (par rapport aux compétences fondamentales).

À ce jour, la BDT a uniquement assumé des tâches en lien avec les cycles d'enquête menés ou en préparation: développement des tâches et des tests ainsi que détermination des valeurs seuils (cf. domaines de compétence b et surtout c ci-dessus). La structure pour la gestion et la livraison des tests et des tâches a finalement été mise en place autrement. L'offre de services à l'intention des régions et des cantons (cf. art. 3 et 4) n'a pas trouvé d'écho.

L'expertise scientifique réalisée sur la vérification préconise en particulier de définir un rattachement institutionnel systématique de la BDT aux institutions scientifiques choisies pour la mise en œuvre de la vérification, parce qu'elle se charge de travaux de nature scientifique et qu'il s'agit de garantir sur la durée le maintien du savoir-faire acquis et parce qu'il y a de nombreux recoupements avec d'autres travaux de mise en œuvre, qu'il convient de gérer le plus efficacement possible.

Pour les raisons citées, le Bureau de coordination HarmoS se dit favorable à une intégration de la BDT dans les structures scientifiques qui sont coordonnées ou dirigées par l'*Interfaculty Centre for Educational Research* (ICER) de l'Université de Berne.

En même temps, il recommande de renoncer entièrement aux tâches qui, selon le règlement, ne sont pas directement associées à la vérification et qui n'ont jusqu'ici pas été accomplies par la BDT.

3.2.2 Transfert au Consortium scientifique de la compétence en matière de valeurs seuils et de questionnaires contextuels

Conformément au règlement d'organisation du 8 mai 2014, c'est à l'Assemblée plénière de la CDIP qu'il incombe d'approuver les valeurs seuils («compétences fondamentales atteintes») sur les échelles de test, que la BDT a au préalable déterminées par des méthodes scientifiques.

Selon les bases légales actuellement en vigueur, le questionnaire contextuel pour les élèves utilisé dans le cadre de la vérification doit être présenté au Comité de la CDIP pour approbation. Cette pratique s'est établie parce que les questions portant sur le contexte social de l'élève sont souvent jugées délicates. Mais le fait que cette approbation survienne juste avant les enquêtes n'est pas jugé satisfaisant par un certain nombre d'acteurs.

Dans l'expertise scientifique qui a été rendue, la détermination des valeurs seuils et le développement des questionnaires sont considérés comme constituant des travaux purement scientifiques; une crainte y est formulée, à savoir que la crédibilité des résultats ne soit amoindrie par la participation des instances politiques à ces travaux et la possibilité d'exercer une influence. L'expertise propose de ce fait de laisser les travaux scientifiques à la science et, pour cela, de définir en termes plus concrets que jusqu'à présent le mandat confié par la politique et l'administration de l'éducation aux institutions scientifiques chargées de la mise en œuvre, en vue d'une adoption conjointe. Pour ce faire, il s'agit d'adopter un cadre conceptuel (*framework*) pour la mise en œuvre de la vérification, et en particulier pour les domaines des tests et de la contextualisation (questionnaire). L'étude PISA travaille avec de tels cadres pour les tests et pour les questionnaires. Avant tout nouveau cycle, ces cadres sont mis à jour en collaboration avec des spécialistes externes, soumis aux pays mandataires pour évaluation et finalement adoptés. Ils forment alors la base sur laquelle se fondent l'attribution du mandat et la mise en œuvre de l'étude.

Le Bureau de coordination HarmoS est favorable à la séparation qui est proposée entre les tâches relevant de la politique et celles qui participent de la réalisation scientifique de la vérification; il recommande aussi que l'interface entre mandants et mandataires soit définie à l'aide de cadres conceptuels.

3.2.3 Consolidation de la position des didactiques des disciplines

Dans les années 2000, des spécialistes de la didactique des disciplines ont grandement contribué au projet HarmoS en participant au développement des modèles de compétences et à la description de standards de formation pour les domaines disciplinaires que sont la langue de scolarisation, les mathématiques, les langues étrangères et les sciences. Dans le cadre de la vérification, les didactiques disciplinaires jouent surtout un rôle important dans le développement des tâches, mais elles ne sont reliées à la structure de projet que par le biais de la BDT et elles n'ont, en tant que telles, pas de place à occuper dans les structures du projet. Selon l'expertise scientifique qui a été réalisée, la perspective des didactiques disciplinaires des domaines faisant l'objet des tests doit être intégrée de manière systématique dans la vérification, d'une part en raison de l'éclairage disciplinaire spécifique qu'elles peuvent apporter sur l'aménagement et l'utilité de la vérification et, de l'autre, du fait de leur proximité avec la formation initiale et continue des enseignantes et enseignants.

Le Bureau de coordination HarmoS adhère largement aux suggestions de l'expertise et propose deux mesures visant à mieux valoriser la perspective des didactiques disciplinaires dans la vérification, à savoir: 1) une implication directe des didactiques disciplinaires dans le groupe de travail Développement des tests de la direction du projet Vérification; 2) une représentation des didactiques disciplinaires dans le groupe

d'experts de la CDIP qu'il est proposé de mettre sur pied pour accompagner le projet Vérification (cf. point suivant). Il estime en outre qu'il vaut mieux intégrer au processus d'élaboration des tâches notamment des didacticiennes et didacticiens qui ont de l'expérience dans l'enseignement auprès d'élèves appartenant aux catégories testées, et ce, afin de garantir le lien avec la pratique.

3.2.4 Mise sur pied d'un groupe d'expertes et experts interrégional, international et interdisciplinaire

Un groupe chargé de constituer le futur consortium scientifique, composé de représentantes et représentants de l'Université de Berne, de l'Université de Genève et de la *Scuola universitaria professionale della Svizzera italiana* (SUPSI), a travaillé jusqu'à fin 2016 pour élaborer, sur mandat de la CDIP, un projet de *règlement d'organisation du Consortium scientifique pour la recherche et l'évaluation en éducation* (WIKO BFE). Ce projet de règlement est reproduit en annexe aux conventions de coopération conclues avec l'Université de Berne en vue de la réalisation des cycles actuels de vérification et de PISA. L'ancrage qui y est décrit pour les évaluations à large échelle (*large-scale assessments*) dans le paysage des hautes écoles suisses ainsi que l'accès à une expertise de haut niveau n'ont jamais été mis en œuvre tel que cela avait été anticipé. Voici ce qui était prévu: une chaire de professeur/e pour la recherche et l'évaluation en éducation à l'Université de Berne, un conseil scientifique assumant des fonctions de conseil, de direction et de décision (composé de représentantes et représentants des Universités de Berne et de Genève ainsi que de la SUPSI et év. d'autres membres), une Assemblée des délégués constituée de représentantes et représentants de hautes écoles suisses, intéressés par le domaine, de même qu'un conseil de spécialistes internationaux, constitué de trois expertes ou experts issus de régions de langues différentes.

En guise de remplacement, du moins partiel, le SG CDIP propose de constituer un groupe interrégional, international et interdisciplinaire composé d'expertes et d'experts chargés d'accompagner le projet Vérification et de représenter en particulier les intérêts de la CDIP pendant le développement, l'adoption et la mise en œuvre des cadres conceptuels. Pour sa constitution, il conviendra de veiller à garantir un large éventail de disciplines ainsi qu'une expertise disciplinaire étendue, mais aussi une représentation adéquate des différentes régions. Lors de l'élaboration des cadres conceptuels, il faudra également tenir compte des didactiques des disciplines, qui actuellement ne participent qu'au développement des tâches.

3.2.5 Organisation de la collaboration entre le Consortium scientifique, c'est-à-dire l'Université de Berne et les centres de coordination régionaux

Selon le *règlement d'organisation concernant la réalisation de la vérification de l'atteinte des compétences fondamentales*, l'Université de Berne est tenue de collaborer avec des institutions des différentes régions linguistiques pour la réalisation de la vérification. Le règlement précise aussi que les enquêtes doivent être menées de manière décentralisée. Comme le montre le point précédent, l'ancrage régional et décentralisé du projet Vérification a été mis en œuvre de façon moins complète que prévu notamment dans le projet de règlement d'organisation élaboré pour le WIKO BFE. Les centres de coordination régionaux rattachés à la PHSG, au SRED et à la SUPSI sont en revanche très impliqués dans la direction de projet et dans les groupes de travail interinstitutionnels créés par l'ICER pour la mise en œuvre de la vérification. Les trois centres de coordination entretiennent également des contacts avec les cantons et se chargent de diriger les collectes de données dans les écoles.

Si l'on renforçait la fonction de direction de l'ICER et que l'on centralise davantage le projet, il serait certainement possible de diminuer les coûts occasionnés par la vérification. Davantage de tâches liées au contenu pourraient être assumées, de manière centralisée, par l'ICER – plutôt que d'être effectuées par les groupes de travail interinstitutionnels (par ex. GT Questionnaire contextuel, GT Développement des tests, GT Livraison des tests, GT Gestion de l'enquête). Le réseau des référentes et référents cantonaux pourrait aussi être géré de manière centralisée. Si à l'avenir les tests pouvaient être livrés en ligne, les enquêtes pourraient éventuellement être réalisées dans les écoles par des organismes privés.

La probable diminution de coûts qu'apporterait le fait de centraliser s'accompagnerait néanmoins d'un affaiblissement de l'ancrage de la vérification dans les différentes régions linguistiques. De plus, le savoir-faire en matière de mesure de performances finirait avec le temps par disparaître en dehors du centre. Vu les expériences positives faites avec structures actuelles, l'ICER lui-même dit préférer renoncer à une centralisation systématique.

Afin ne pas affaiblir le soutien dont bénéficient actuellement les travaux au niveau des régions linguistiques, le Bureau de coordination HarmoS propose de maintenir pour l'instant la bonne collaboration qui est en place, puis d'examiner conjointement avec l'Université de Berne, au cours des prochaines années, les adaptations qui pourraient être apportées.

3.3 Conséquences financières

Comme présenté au point 1.4.3, le montant annuel (sans coûts supplémentaires occasionnés par le coronavirus) versé pour la vérification est de 1,13 million de francs et pour la BDT de 0,37 million de francs, sous forme de contributions cantonales. Les contributions correspondant à six ans, soit environ 9 millions de francs, permettent de financer entièrement deux enquêtes.

Il ressort des tableaux 1 à 5 du point 3.1 que les propositions de réaménagement et de revalorisation du projet Vérification entraîneraient une hausse des coûts annuels principalement liée à l'augmentation du nombre d'enquêtes. Cette hausse correspond aux coûts indiqués de 1,95 million de francs pour la variante C, de 2,25 millions de francs pour la variante A et de 2,8 millions de francs par année pour la variante B. Dans le cas de la variante D, le projet de vérification serait abandonné. La réalisation des enquêtes PISA auprès d'un échantillon d'élèves de 11e année, dont la taille serait à peu près équivalente à celle d'un échantillon utilisé pour la vérification de l'atteinte des compétences fondamentales pour ce degré de scolarité, engendrerait des coûts supplémentaires avoisinant 0,5 million de francs par année. La participation à l'étude PISA est considérablement moins chère que la réalisation des enquêtes de vérification, étant donné qu'elle ne nécessite pratiquement pas de travaux de développement pour la Suisse.

Bibliographie

Ambühl, H. (2011). L'accord intercantonal du 14.6.2007 sur l'harmonisation de la scolarité obligatoire (concordat HarmoS): Commentaire, genèse et perspectives, instruments. CDIP. https://edudoc.ch/record/96778/files/harmos-konkordat_f.pdf

Angelone, D., & Keller, F. (2019a). Überprüfung des Erreichens der Grundkompetenzen (ÜGK) im Fach Mathematik im 11. Schuljahr: Technische Dokumentation zur Testentwicklung und Skalierung. Aufgabendatenbank der EDK. http://uegk-schweiz.ch/wp-content/uploads/2019/05/%C3%9CGK2016_Technischer-Bericht_ADB.pdf

Angelone, D., & Keller, F. (2019b). Überprüfung des Erreichens der Grundkompetenzen (ÜGK) in den Fächern Schulsprache und erste Fremdsprache im 8. Schuljahr: Technische Dokumentation zur Testentwicklung und Skalierung. Aufgabendatenbank der EDK. http://www.vecof-svizzera.ch/wp-content/uploads/2019/06/%C3%9CGK2017_Technischer-Bericht_ADB.pdf

Angelone, D., Keller, F., Arnold, J., Jaun-Holderegger, B., Hauser, M., Eckstein, B., Erzinger, A., Leutwiler, L., Nell-Tuor, N., & Schönenberger, S. (2019). Machbarkeitsstudie zur Überprüfung der Grundkompetenzen im 4. Schuljahr [Unveröffentlichter Bericht im Auftrag der EDK]. Aufgabendatenbank der EDK.

Buschor, E., Gilomen, H., & McCluskey, H. (2003). PISA 2000: Synthèse et recommandations. OFS; CDIP. http://edudoc.ch/record/24196/files/8f.pdf

CDIP. (1970). Concordat sur la coordination scolaire du 29 octobre 1970. http://edudoc.ch/getfile.py?recid=1548

CDIP. (2001a). Lignes directrices de la Conférence suisse des directeurs cantonaux de l'instruction publique (CDIP) du 5 juillet 2001. CDIP. http://edudoc.ch/getfile.py?recid=24716

CDIP. (2001b). Programme de travail de la CDIP à partir de 2001 basé sur des points forts de ses activités du 7.6.2001. CDIP. http://edudoc.ch/record/112625/files/TP2001f.pdf

CDIP. (2004a). Enseignement des langues à l'école obligatoire: Stratégie de la CDIP et programme de travail pour la coordination à l'échelle nationale, décision du 25 mars 2004. http://edudoc.ch/record/30009/files/sprachen_f.pdf

CDIP. (2004b). Harmos: Finalités et conception du projet: Juin 2004. CDIP. http://edudoc.ch/record/33504/files/weissbuch_f.pdf

CDIP. (2007a). Accord intercantonal sur l'harmonisation de la scolarité obligatoire (concordat HarmoS) du 14 juin 2007. http://edudoc.ch/record/24710/files/HarmoS_f.pdf

CDIP. (2007b). Mise en oeuvre de l'accord intercantonal sur l'harmonisation de la scolarité obligatoire (concordat HarmoS) au niveau de la coordination intercantonale des 25/26 octobre 2007. http://edudoc.ch/getfile.py?recid=30024

CDIP. (2009). Participation à PISA 2012 (point fort mathématiques) décision de l'Assemblée plénière du 29/30.10.2009. http://edudoc.ch/record/102487/files/05_f.pdf

CDIP. (2011a). Compétences fondamentales pour la langue de scolarisation: Standards nationaux de formation adoptés par l'Assemblée plénière de la CDIP le 16.6.2011. CDIP. http://edudoc.ch/record/96790/files/grundkomp_schulsprache_f.pdf

CDIP. (2011b). Compétences fondamentales pour les langues étrangères: Standards nationaux de formation adoptés par l'Assemblée plénière de la CDIP le 16.6.2011. CDIP. http://edudoc.ch/record/96779/files/grundkomp_fremdsprachen_f.pdf

CDIP. (2011c). Compétences fondamentales pour les mathématiques: Standards nationaux de formation adoptés par l'Assemblée plénière de la CDIP le 16.6.2011. CDIP. http://edudoc.ch/record/96783/files/grundkomp_math_f.pdf

CDIP. (2011d). Compétences fondamentales pour les sciences naturelles: Standards nationaux de formation adoptés par l'Assemblée plénière de la CDIP le 16.6.2011. CDIP. http://edudoc.ch/record/96786/files/grundkomp_nawi_f.pdf

CDIP. (2012a). Participation de la Suisse à PISA 2015: Décision: Décision du 25.10.2012. http://edudoc.ch/record/105013/files/PB_pisa_f.pdf

CDIP. (2012b). Vérification de l'atteinte des compétences fondamentales; cadre conceptuel et financement de la base de données de tâches: Décision: Décision du 25.10.2012. http://edudoc.ch/record/105011/files/PB_aufgabendatenbank_f.pdf

CDIP. (2013a). Vérification de l'atteinte des compétences fondamentales; projet: Adoption: Décision du 20.6.2013. http://edudoc.ch/record/107771/files/PB_grundkompetenzen_f.pdf

CDIP. (2013b). Règlement du Bureau de coordination pour la mise en oeuvre de l'accord intercantonal sur l'harmonisation de la scolarité obligatoire (Bureau HarmoS) du 24 janvier 2008. http://edudoc.ch/record/38021/files/KostaHarmos_f.pdf

CDIP. (2014a). Règlement d'organisation de la Banque de données de tâches de la Conférence suisse des directeurs cantonaux de l'instruction publique du 23 janvier 2014. http://edudoc.ch/record/110606/files/fj_Organisationsreglement_f_140123.pdf

CDIP. (2014b). Règlement d'organisation concernant la réalisation de la vérification de l'atteinte des compétences fondamentales du 8 mai 2014. http://edudoc.ch/record/113129/files/Organisationsregl_UeGK_f.pdf

CDIP. (2017). Vérification de l'atteinte des compétences fondamentales: Planification des enquêtes à partir de 2020: Décision de l'Assemblée plénière du 22 juin 2017. http://edudoc.ch/record/127424/files/pb_grundkompetenzen_erhebungen_2020_f.pdf

CDIP. (2018). Vérification de l'atteinte des compétences fondamentales; domaines disciplinaires à tester dans le cadre de l'enquête 2020, élèves de 11e année: Décision de l'Assemblée plénière du 26 octobre 2018. http://edudoc.ch/record/133107/files/pb_fachbereiche_f.pdf

CDIP. (2019a). Vérification de l'atteinte des compétences fondamentales: Rapport sur les enquêtes 2016 et 2017. CDIP. http://edudoc.ch/record/204068/files/201905_bericht_Erhebungen_ugk_2016_2017_f.pdf

CDIP. (2019b). Brève information: Compétences fondamentales en langues et en mathématiques: Résultats des enquêtes 2016 et 2017. CDIP. https://edudoc.ch/record/216948/files/grundkomp_kurzinfo_f.pdf

CDIP. (2019c). Vérification de l'atteinte des compétences fondamentales; domaines disciplinaires à tester dans le cadre de l'enquête 2022: Décision: Décision de l'Assemblée plénière du 25 octobre 2019. CDIP. http://edudoc.ch/record/206950/files/PB_fachbereiche_erhebung_2022_f.pdf

CDIP. (2020a). Vérification de l'atteinte des compétences fondamentales; non-réalisation en 2020 de l'enquête prévue et élaboration d'un nouveau calendrier pour les enquêtes 2020 et 2022: Décision: Décision de l'Assemblée plénière par voie de correspondance du 1er avril 2020. CDIP. http://edudoc.ch/record/209007/files/PLE-2020-2_non-realisation-enquete-2020.pdf

CDIP. (2020b). Vérification de l'atteinte des compétences fondamentales; nouveau calendrier pour les enquêtes auprès d'élèves de 4e et de 11e année de scolarité: Décision. http://edudoc.ch/record/210794/files/AP_calendrier_uegk.pdf

CDIP. (2021a). Accord intercantonal du 14 juin 2007 sur l'harmonisation de la scolarité obligatoire (concordat HarmoS): État d'avancement des procédures cantonales d'adhésion. EDK. https://edudoc.ch/record/210894/files/liste_rat_df.pdf

CDIP. (2021b). Dispositif: Utilisation des données dans le cadre de la vérification de l'atteinte des compétences fondamentales. https://edudoc.ch/record/218326/files/VB070521-dispositif-cofo-f.pdf

CDIP/IDES. (2021). Enquête CDIP/IDES auprès des cantons; état: Année scolaire 2020-2021; Question No 121: Selon les réglementations cantonales, est-ce que des tests d'aptitudes standardisés obligatoires pour tous les élèves sont organisés (p. Ex.: Épreuves cantonales, épreuves de référence, épreuves communes romandes, examens, tests)? https://edudoc.ch/record/212326/files/121.pdf

Consortium HarmoS Langue de scolarisation. (2010). Langue de scolarisation: Rapport scientifique de synthèse et modèle de compétences: Version provisoire (avant adoption des standards de base): État: 2 février 2010. [s.n]. http://edudoc.ch/record/87021/files/L1_wissb_f_20100119.pdf

Consortium HarmoS Langues étrangères, Schneider, G., Lenz, P., & Studer, T. (2009). Langues étrangères: Rapport scientifique de synthèse et modèle de compétences: Version provisoire (avant adoption des standards de base): État: 18 juin 2009. [s.n]. http://edudoc.ch/record/87024/files/L2_wissB_25_1_10_f.pdf

Consortium HarmoS Mathématiques, Linneweber-Lammerskitten, H., Wälti, B., & Moser Opitz, E. (2009). Mathématiques: Rapport scientifique de synthèse et modèle de compétences: Version provisoire (avant adoption des standards de base): État: 13 décembre 2009. [s.n]. http://edudoc.ch/record/87029/files/math_kurzbericht_2009_f.pdf

Consortium Sciences naturelles+. (2009). Sciences naturelles: Rapport scientifique de synthèse et modèle de compétences: Version provisoire (avant adoption des standards de base): État juillet 2009. Consortium HarmoS sciences naturelles+. http://edudoc.ch/record/86400/files/sc_nat_rapp_synth_2009.pdf

Crotta, F., Salvisberg, M., & Zanolla, G. (2020). Note scolastiche e competenze fondamentali in matematica degli allievi dell'11° anno HarmoS in Ticino. Swiss Journal of Educational Research, 42(2), 323–342. https://doi.org/10.24452/sjer.42.2.3

DEFR, & CDIP. (2019). Valorisation optimale des chances: Déclaration 2019 sur les objectifs communs concernant l'espace suisse de formation. CDIP; DEFR. http://edudoc.ch/record/205823/files/erklaerung_20190627_f.pdf

Erzinger, A., Hauser, M., Dutrevis, M., Hascher, T., Keller, R., Lenz, P., & Soussis, A. (2019). Erläuterungen zu den Skalen des Kontextfragebogens der ÜGK Sprachen 2017: Theoretischer Hintergrund, Inhalte und Konstrukte. Universität Bern, Päd. Hochschule St. Gallen, Service de la recherche en éducation (SRED), Universität Fribourg. http://uegk-schweiz.ch/wp-content/uploads/2019/09/%C3%9CGK2017_Sprachen_KFB_def.pdf

Fischbach, A., & Ugen, S. (2018). ÜGK/COFO Mathematics 2016: Audit report: Commissioned by EDK/CDIP upon the request of KOSTA HarmoS. [s.n.]. http://edudoc.ch/record/204067/files/UeGK_Audit_Report_230218.pdf

IRDP. (2021). EPROCOM - banque d'items. https://www.irdp.ch/institut/eprocom-banque-items-3673/irdp-fromsearch-1.html

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.-E., & Vollmer, H. J. (2004). Le développement de standards nationaux de formation: Une expertise (Bd. 1). Bundesministerium für Bildung und Forschung (BMBF). http://edudoc.ch/record/33469/files/develop_standards_nat_form_f.pdf

Nidegger, C., Roos, E., Petrucci, F., Verner, M., Helbling, L. A., Arnold, E., Erzinger, A. B., Pham, G., Ambrosetti, A., Crotta, F., Salvisberg, M., Moretti, A., Angelone, D., Keller, F., Leutwiler, L., Lindauer,

T., Lenz, P., Aeppli, T., Karges, K., … Epars, F. (2019). Vérification de l'atteinte des compétences fondamentales: Rapport national COFO 2017: Langues 8e année scolaire. CDIP. http://edudoc.ch/record/204071/files/COFO_2017_Complet_FR.pdf

Nidegger, C., Roos, E., Petrucci, F., Verner, M., Helbling, L. A., Arnold, E., Erzinger, A. B., Pham, G., Schönenberger, S., Crotta, F., Ambrosetti, A., Salvisberg, M., Angelone, D., Keller, F., & Linneweber-Lammerskitten, H. (2019). Vérification de l'atteinte des compétences fondamentales: Rapport national COFO 2016: Mathématiques 11e année scolaire. CDIP. http://edudoc.ch/record/204074/files/COFO_2016_Complet_FR.pdf

Petrucci, F., & Roos, E. (2020). Atteinte des compétences fondamentales dans le canton de Genève: Que nous enseignent les enquêtes COFO 2016 et 2017 ? SRED. https://edudoc.ch/record/213034

Ramseier, E., Moser, U., Moreau, J., & Antonietti, J.-P. (2008). Schlussbericht der HarmoS-Methodologiegruppe. EDK. http://www.lehrplanforschung.ch/wp-content/uploads/2011/06/Methodologie-Schlussbericht-_HarmoS11.pdf

Roth, M., Ruf, I., Sánchez Abchi, V., Soussi, A., Weiss, L., & Marc, V. (2021). EpRoCom: Dispositif romand de mutualisation de tâches évaluatives (IRDP Focus). https://www.irdp.ch/data/secure/3586/document/2021_irdp_focus_eprocom.pdf

SKBF CSRE. (2018). Rapport sur l'éducation. http://www.skbf-csre.ch/fr/rapport-sur-leducation/rapport-education/

Annexe 1: Autres aspects liés à la mise en œuvre de la vérification

Développement des tests, détermination des valeurs seuils

Comme décrit au point 1.3.3, le rapport d'audit sur l'enquête 2016 (Fischbach & Ugen, 2018) contient déjà des propositions pour renforcer le développement des tâches et des tests. Ces propositions sont surtout mises en œuvre dans le cadre des travaux relatifs à l'enquête 2024. L'expertise voit d'un bon œil les mesures déjà prises, mais fait également état de problèmes et formule des recommandations à cet égard. Une partie des aspects concernés sont présentés et font l'objet d'une discussion dans ce qui suit. Le Bureau HarmoS attire en outre l'attention sur des changements qu'il faudra prévoir dans le cas où les tests de vérification ne viseraient plus à l'avenir à vérifier l'atteinte des compétences fondamentales uniquement mais également celle des compétences de l'ensemble du spectre.

Points faibles liés au cadre conceptuel de test (*test framework*) et aux bases sur lesquelles repose le développement des tests

Ancrage didactique: l'expertise relève qu'il n'existe pas de cadres conceptuels définis pour les tests (*test frameworks*) du projet de vérification, comme c'est le cas par exemple pour l'étude PISA. Selon l'expertise, de tels cadres décrivent notamment du point de vue de la didactique des disciplines ce qui est compris dans les compétences qui doivent être testées. Des tâches sont indiquées à titre d'exemple et pour faciliter la compréhension. Ces tâches et la description de leurs caractéristiques permettent de mieux comprendre et d'étudier les exigences associées aux compétences fondamentales. Le tout présente plusieurs avantages pour l'ancrage des objectifs de formation dans la didactique des disciplines, dans la formation du corps enseignant et dans la pratique de l'enseignement. Pour le groupe impliqué dans la réalisation de l'expertise, chaque test de vérification doit reposer sur un cadre conceptuel (*framework*) toujours actualisé. Il s'agit de développer et d'étoffer les cadres de test (*test frameworks*) existants, tels que ceux esquissés dans le rapport sur l'enquête 2017, et de décrire les domaines de compétences de sorte qu'on puisse en tirer un meilleur parti pour la formation disciplinaire et, finalement, pour la pratique dans l'enseignement.

Bases concrètes pour le développement des tests: les objectifs de formation validés pour les quatre domaines disciplinaires décrivent en substance ce que les élèves doivent savoir au minimum, et indiquent donc en quoi consistent les exigences associées. L'expérience montre toutefois que cela ne suffit pas pour garantir un développement systématique des tâches, étant donné qu'il manque dans les descriptions des points de repère importants pour la mise en œuvre dans les tâches des tests. Afin de disposer de bases fiables pour le développement des tâches et de pouvoir mieux contrôler dès le départ la difficulté des tâches, l'expertise propose d'analyser les tâches déjà utilisées et ayant des propriétés statistiques identifiées pour connaître l'incidence de leurs caractéristiques sur le degré de difficulté. Il s'agit également de déterminer les caractéristiques qui ont une incidence variable sur la difficulté des tâches en fonction du plan d'études et/ou de la région linguistique.

D'après le Bureau HarmoS, les propositions formulées dans l'expertise en vue de renforcer l'ancrage didactique de la vérification et de tirer parti d'analyses de tests et de tâches pour un développement ciblé de ces dernières sont pertinentes et de nature à produire l'effet recherché. Du fait que les prochains tests a) vérifieront l'atteinte des compétences fondamentales, b) devraient tester les compétences de l'ensemble du spectre (cf. notamment le point 2.1.4), et qu'on utilisera un *design* longitudinal (cf. point 2.2), le mandat confié aux didactiques des disciplines (ainsi qu'à la psychométrie) devra être élargi de la manière suivante:

- Les modèles de compétences fondamentales devront désormais comporter une dimension verticale grâce a) à une caractérisation du développement des compétences entre les différents degrés et b) à des échelles qui décrivent le spectre des performances pour chaque passation de test, y compris par des exemples de tâches.

– Des bases conceptuelles fiables doivent être créées pour le développement des tests[7], qui permettent d'évaluer avec fiabilité tant les compétences fondamentales que d'autres compétences du spectre. Pour le haut de l'échelle, il n'existe que peu de tâches déjà utilisées dans les cycles de vérification passés qui puissent être analysées pour fournir des caractéristiques concernant la difficulté. Les cadres (*frameworks*) prévus pour le développement des tests devront proposer une approche permettant d'évaluer efficacement et avec précision des élèves qui ont une maîtrise très variable des compétences à tester. Par exemple, il est possible d'avoir recours à une méthode de test adaptative en plusieurs étapes (*multi-stage adaptive testing*, MSAT) comme c'est le cas pour PISA. La difficulté des blocs de tâches est alors déterminée individuellement dès le deuxième bloc, sur la base des premiers résultats obtenus.

Adéquation de la difficulté des tâches

Comme on le sait, les tests de vérification avaient jusqu'ici principalement pour but de permettre de vérifier l'atteinte des compétences fondamentales. La décision quant à ce qu'on entend par «atteinte», en se référant aux résultats des tests, n'intervient de manière définitive qu'au moment où sont déterminées les valeurs seuils. Dans le cadre de la procédure standardisée reprenant la méthode *Bookmark* et utilisée pour la vérification, un organe de composition mixte regroupant plusieurs spécialistes reçoit un livret qui classe les différents items par ordre croissant de difficulté (sur la base des résultats des tests). Il incombe alors à cet organe de décider quelle tâche les élèves qui ont tout juste acquis les compétences fondamentales sont susceptibles de pouvoir résoudre avec une probabilité de 62 %. Sur la base de cette décision, il est ensuite possible de calculer la proportion des élèves testés qui ont atteint les compétences fondamentales.

Si l'une des principales fonctions des tests doit être de distinguer les élèves qui se situent au-dessus du seuil («compétences fondamentales atteintes») de ceux qui se situent en dessous, il faut alors disposer de suffisamment de tâches spécialement adaptées à cette fin. Dans la théorie des tests (*item response theory*), les tâches permettent particulièrement bien de distinguer entre eux les élèves testés qui ont des compétences similaires lorsque la probabilité qu'ils parviennent à résoudre les tâches en question se situe vers 50 % seulement. Cela vaut en principe aussi pour le groupe d'élèves dont les compétences se situent aux alentours des compétences minimales. À nouveau, il est essentiel pour la détermination des valeurs seuils que l'on dispose de suffisamment de tâches pouvant être résolues correctement, avec une probabilité d'environ 62 %, par les élèves se situant à proximité du seuil. Les tâches assez simples, pour lesquelles la probabilité de résolution va croissant, jusqu'à atteindre 80 % environ, permettent d'illustrer ce que les élèves qui disposent de compétences minimales sont à même de faire avec une confiance allant croissant également. Les tâches simples jouent également un rôle sur la motivation des élèves à passer le test: si trop de tâches sont perçues comme difficiles, les élèves risquent de perdre rapidement leur volonté de réussir. Les tâches dont la probabilité de résolution dépasse les 80 % servent surtout à distinguer entre eux les élèves qui ont beaucoup de difficultés à l'école. À l'inverse, les tâches d'un niveau assez exigeant, que les élèves disposant de compétences minimales ne sont que peu susceptibles de résoudre, aident à différencier entre eux les élèves plus performants. Pour les études qui visent à expliquer les différences de performances entre les élèves et autres études s'intéressant aux liens entre les résultats des tests et d'autres variables, il est important qu'il subsiste une certaine dispersion chez les élèves ayant des difficultés et ceux qui ont plutôt de la facilité, c'est-à-dire que le nombre d'élèves testés obtenant le minimum ou le maximum de points soit le moins élevé possible.

Vu que normalement seule une faible proportion des élèves testés devrait ne pas atteindre les compétences fondamentales, il ressort des considérations exposées ci-dessus que, dans un test ayant pour but la vérification de l'atteinte des compétences fondamentales, une majorité des tâches devraient être faciles, voire très faciles, pour la plupart des élèves.

[7] Dans la recherche menée en matière de tests, on parle parfois d'échelles orientées vers les tâches (en anglais: *constructor-oriented scales*), établies dans l'optique d'une construction ciblée des tâches. Ces échelles viennent préciser les caractéristiques à associer aux tâches en termes d'exigences, tandis que les échelles de compétences (*user-oriented scales*) décrivent les compétences constatées sur la base d'exigences données.

Le groupe chargé de réaliser l'expertise émet un avis critique sur les tests menés jusqu'ici pour plusieurs domaines de compétences, car leurs tâches ne sont pas assez en adéquation avec l'objectif visé. Trop de tâches se révèlent trop difficiles et trop peu suffisamment faciles pour être vraiment utiles à la vérification de l'atteinte des compétences fondamentales. Dans l'ensemble, le groupe estime que des mesures doivent impérativement être prises – et ce, d'autant plus qu'il a été établi via les analyses et les sondages menés aux fins de l'expertise qu'il n'existe pas de consensus parmi les acteurs impliqués sur ce en quoi doit consister la mesure de l'atteinte des compétences fondamentales. C'est pourquoi il est proposé dans l'expertise que la psychométrie et les didactiques des disciplines se mettent d'accord, pour les différents domaines disciplinaires, sur la façon dont il convient de mesurer l'atteinte des compétences fondamentales et sur la façon de parvenir, à partir de la description faite des objectifs de formation, à une dichotomie interprétable sous l'angle de la didactique des disciplines entre «compétences minimales atteintes» et «compétences minimales non atteintes».

Le Bureau HarmoS appuie la recommandation faite dans le cadre de l'expertise selon laquelle la psychométrie et les didactiques des disciplines doivent établir ensemble une vision générale de la façon dont il convient de vérifier l'atteinte des compétences fondamentales, précisant qu'il doit également être question, pour la suite du projet, de définir une compréhension commune afin de pouvoir tester l'ensemble du spectre des compétences.

Traduction des tâches

À la suite notamment des recommandations formulées dans le rapport d'audit sur l'enquête 2016, on accorde désormais davantage d'attention à la traduction des tâches. L'expertise recommande de poursuivre les efforts entrepris en ce sens et d'appliquer les standards internationaux reconnus en matière de traduction de tests.

Le Bureau HarmoS, quant à lui, estime que la priorité n'est pas de suivre les standards internationaux. L'expérience montre qu'il est possible d'atteindre un niveau de qualité élevé en favorisant la collaboration entre les régions linguistiques lors du développement des tests de même qu'en impliquant de manière ciblée les institutions appropriées du système éducatif au travers d'une collaboration minutieusement préparée et de vérifications régulières.

Études de validation

L'expertise fait état d'un urgent besoin de disposer d'études de validation (publiées par des tiers) à propos du projet de vérification. Elle cite en particulier les études qui vérifient dans quelle mesure les résultats obtenus aux tests de vérification concordent avec les résultats de tests reconnus ou avec ceux d'évaluations menées dans les écoles sur les mêmes domaines de compétences (*concurrent validity*). Elle propose en outre de réaliser des sondages représentatifs auprès du corps enseignant pour évaluer l'adéquation des tests avec l'enseignement dispensé (*curricular validity*). L'expertise considère que ce genre d'études doit avoir la priorité et suggère de ne pas se lancer dans une différenciation plus poussée des échelles de tests avant d'avoir réalisé de telles études de validation.

Le Bureau HarmoS est d'avis qu'il est tout à fait souhaitable de disposer d'études de validation (publiées) permettant d'examiner la validité des tests. Par exemple, l'étude de faisabilité évoquée plus haut concernant la vérification en 4^{e} année (Angelone et al., 2019) fait seulement partie de la littérature grise et les mesures prises pour la validation des tâches au cours du développement ne sont documentées que de manière synthétique dans la série de rapports établis sur les cycles de vérification. On dispose heureusement de premières études dans ce sens, lesquelles ont été réalisées en Suisse romande et au Tessin et qui portent sur les liens observés entre les résultats des tests de vérification et les notes que les élèves obtiennent à l'école (Crotta et al., 2020; Petrucci & Roos, 2020). Étant donné qu'il est question de réorienter en partie le projet de vérification, le Bureau de coordination HarmoS estime qu'il n'est pas pertinent d'effectuer d'abord des études de validation puis des travaux complémentaires pour la différenciation des échelles de test. Il sera d'ailleurs aussi nécessaire de mener des études de validation une fois seulement qu'on pourra disposer des données issues des tests conçus selon la nouvelle approche.

Calendrier appliqué au développement des tests

Comme exposé au point 1.3.3, le développement des tâches a été substantiellement renforcé en vue surtout de l'enquête devant être menée en 2024. Néanmoins, l'expertise rendue propose encore un certain nombre d'améliorations possibles. Il s'agit notamment d'ajouter, après la phase pré-pilote et l'enquête pilote, mais avant la détermination de la valeur seuil, une étude normative dans laquelle seront utilisées des tâches entièrement développées auprès d'échantillons représentatifs. La détermination de la valeur seuil n'interviendra alors qu'après obtention de l'ensemble des tâches terminées et des données statistiques relatives aux tâches et aux élèves. Il s'agit ainsi d'éviter de devoir attendre que l'enquête principale soit terminée pour déterminer la valeur seuil, dans l'éventualité où des adaptations seraient apportées aux tâches entre l'enquête pilote et l'enquête principale. En effet, le risque qu'une influence soit exercée, pour des motivations politiques, sur les valeurs seuils est, selon le groupe ayant réalisé l'expertise scientifique, plus élevé à ce moment-là.

Dans la planification faite pour le cycle 2024, il est prévu que l'enquête pilote fasse office d'étude normative. Cela s'est avéré possible après un développement important de la phase pré-pilote. Il a toutefois été nécessaire d'adapter encore, après la phase pré-pilote, un nombre relativement élevé de tâches, sans pouvoir ensuite les soumettre une nouvelle fois à un test pré-pilote. On risque alors de devoir apporter encore d'autres adaptations sur la base des résultats statistiques obtenus au terme de l'enquête pilote.

Le Bureau de coordination HarmoS est aussi d'avis que la détermination des valeurs seuils doit intervenir en amont de l'enquête principale. La nécessité de mener des études normatives doit être examinée. Si l'on applique un *design* impliquant une planification à long terme et comprenant des mesures de tendances, il s'avérera nécessaire de redéfinir la façon dont on détermine la valeur seuil. La comparabilité entre les valeurs seuils de mesures de tendances successives est principalement assurée par le fait que l'on utilise et réutilise des tâches issues du même ensemble (*pool*).

Ensemble de tâches (*pool*)

Jusqu'à présent, le secrétariat de la BDT s'est employé à développer les tests de manière ponctuelle pour les cycles de 2016, 2017, 2023 et 2024. Une partie des tâches utilisées dans les cycles passés ont été rendues publiques tandis que d'autres ont été gardées secrètes pour servir de tâches de référence pouvant être réutilisées dans des cycles ultérieurs.

L'expertise rendue précise que, si l'on opte pour un *design* visant à dégager des tendances, il faut veiller à disposer d'un nombre suffisant de tâches de référence pour pouvoir garantir des mesures de tendances sur le long terme.

Le Bureau HarmoS est conscient de la nécessité d'avoir à disposition et de tenir à jour suffisamment de tâches pouvant être réutilisées pour les mesures de tendances. Lors de la planification d'un *pool* de tâches, il s'agit notamment de vérifier que l'on puisse établir des liens, pour les *pools* destinés à la 8e et à la 11e année, au moyen de tâches communes, de manière à permettre de réelles mesures de progression dans les coupes longitudinales de cohortes prévues – ce qui ne serait guère réalisable si l'on mettait l'accent exclusivement sur les compétences fondamentales. Mais si l'on a recours à des tâches pour les élèves faibles et pour les élèves performants lors de ces deux prises de mesure, les chevauchements pourraient être suffisants pour justifier un ancrage vertical des échelles pour les deux degrés.

Détermination des valeurs seuils

Comme exposé dans le premier volet, les valeurs seuils entre «compétences fondamentales atteintes» et «compétences fondamentales non atteintes» ont été déterminées jusqu'à présent selon une méthode *Bookmark* adaptée.

Pour le groupe chargé de l'expertise, la procédure choisie correspond aux méthodes les plus récentes en la matière.

Si, à l'avenir, il n'y a plus besoin de séparer les résultats en deux catégories seulement (le nombre de catégories serait alors fonction des niveaux de compétences à distinguer), il conviendra de revoir en conséquence le processus de détermination des valeurs seuils. Il s'agira alors de déterminer avec un soin tout particulier les valeurs seuils à appliquer au *pool* de tâches duquel on continuera de prélever les tâches de référence.

Planification et mise en œuvre de la contextualisation

Ainsi que cela a été décrit dans le premier volet, on a eu recours, pour les précédentes enquêtes menées dans le cadre de la vérification, à des questionnaires destinés aux élèves et élaborés en vue essentiellement de montrer les liens entre les performances et certaines variables pouvant donner des éléments explicatifs, tels que le contexte socio-économique par exemple. S'y sont ajoutés certains intérêts propres à la didactique des disciplines, de même que des intérêts de recherche spécifiques. Il est primordial que la contextualisation soit de haute qualité et qu'elle permette de tirer des enquêtes des connaissances utiles et probantes pour le pilotage du système éducatif.

L'expertise rendue comporte à différents endroits des recommandations sur la manière de concevoir et de mettre en œuvre la contextualisation en tant que composante à part entière du projet Vérification:

- Les objectifs et les fonctions visés par le projet Vérification, et plus précisément par chaque cycle en particulier, constituent le point de départ du développement des questionnaires, car ils permettent de tirer les grandes questions à prendre en considération pour l'enquête. Ces questions peuvent être complétées par d'autres permettant une recherche plus approfondie.
- De ces questions découlent les exigences à remplir pour le *design* d'échantillonnage (par ex. participation de classes entières), le *design* d'enquête et les instruments à déployer (par ex. questionnaire destiné aux élèves, aux enseignantes et enseignants de certaines disciplines de même qu'aux directions d'école). On peut aussi envisager des appariements (via le NAVS13) avec des données collectées par d'autres biais.
- Une fois qu'un consensus a été trouvé quant aux objectifs, aux fonctions, aux questions, au *design* d'enquête, aux instruments et éventuellement aux appariements de données à viser dans le cadre du projet, les acteurs politiques approuvent ce qui a été convenu sous la forme d'une décision stratégique.
- Pour servir de base à la mise en œuvre de la contextualisation, on élabore un cadre de contextualisation détaillé (similaire au cadre conceptuel pour les tests, mentionné plus haut), dans une démarche consensuelle entre le niveau des mandants et celui de la mise en œuvre. Ce cadre décrit les concepts qui doivent être mis en œuvre dans le cadre de la contextualisation ainsi que les méthodes et les instruments à utiliser. En général, ces instruments sont des questionnaires conçus pour un ou plusieurs groupes cibles.
- Les étapes suivantes sont mises en œuvre par les acteurs scientifiques chargés du mandat, dans le respect des normes de qualité usuelles dans le monde scientifique. Les acteurs politiques sont les destinataires des résultats une fois ces derniers obtenus (pour plus de renseignements sur les rapports entre mandants et mandataires, cf. point 2.4.1).

Le Bureau HarmoS trouverait judicieux que les enquêtes reposent sur un cadre conceptuel global, à caractère contraignant pour toutes les parties, qui fixerait de manière générale les objectifs et les moyens associés à chaque enquête.

Utilisation des données pour des analyses secondaires

Les points 1.3.4 et 1.3.5 donnent un bref aperçu de la manière dont les données issues de la vérification sont mises à disposition et utilisées pour des analyses qui vont au-delà du cadre fixé pour la communication initiale des résultats. De manière générale, le nombre d'analyses de ce genre est assez limité.

Le groupe chargé de l'expertise scientifique s'est exprimé sur la manière dont les données issues de la vérification peuvent être demandées et utilisées. Partant du constat que l'accès à ces données est assez rarement demandé par des scientifiques qui ne sont pas impliqués dans le projet, il suggère de faire davantage la promotion de ces données auprès de la communauté scientifique, tout comme d'analyser et de simplifier les processus de demande et d'autorisation. Il recommande également de clarifier un certain nombre d'imprécisions constatées dans le dispositif relatif à l'utilisation de ces données et de créer une procédure d'élimination des divergences applicable aux cas de conflits entre les acteurs scientifiques intéressés et les autorités ayant compétence pour autoriser l'utilisation des données.

Dans les réponses données lors de la consultation écrite sur laquelle se fonde l'expertise, il est demandé que des améliorations et des simplifications soient apportées à la documentation relative aux données des enquêtes et aux procédés techniques, de façon à la rendre plus accessible, notamment pour les chercheurs et chercheuses en dehors du projet.

Comme exposé au point 2.2 traitant du *design* d'enquête, le Bureau HarmoS adhère à la suggestion du groupe qui a réalisé l'expertise, à savoir de mettre à disposition des données attrayantes pour la recherche. On a observé que les chercheurs et chercheuses souhaitant utiliser les données issues de la vérification mais qui n'ont pas participé au projet doivent en effet souvent se faire conseiller, et ne peuvent donc pas directement travailler à partir de la documentation établie sur les fichiers à usage scientifique. Si la documentation sur ces fichiers est certes assez complète, elle ne permet toutefois pas forcément aux personnes intéressées, mais qui n'ont pas de connaissances précises du programme, de s'y retrouver sans aide extérieure. Il s'agit de mieux guider les personnes qui souhaitent faire usage de ces données. Quant aux données des enquêtes qui ne sont pas contenues dans les fichiers à usage scientifique (par ex. les réponses fournies dans les items de test et les questionnaires ou encore les données liées aux processus enregistrées par les systèmes de test numériques), elles devraient être peu connues en dehors du projet. De manière générale, il faudrait que les données issues de la vérification soient mieux connues et plus facilement accessibles. Les expériences faites avec les dispositifs du 22 mars 2018 (CDIP, 2018a) et du 5 mai 2021 (CDIP, 2021b) relatifs à l'utilisation des données sont positives jusqu'à présent; on n'a enregistré aucun litige en rapport avec cette possibilité d'utiliser les données.

Documentation relative aux cycles de vérification

Pour ce qui est de la documentation établie en lien avec les cycles déjà menés (voir point 1.3.4), l'expertise rendue la juge trop hétérogène. Si les rapports techniques et la documentation rédigée étaient plus homogènes – une homogénéité qui se manifesterait dans toutes les régions linguistiques, tous les domaines testés et toutes les étapes de développement suivies – la manière dont ils peuvent être reçus et analysés s'en trouverait facilitée. L'expertise suggère de ne produire plus qu'un seul rapport technique par période de test, mais aussi d'utiliser un modèle uniforme pour la rédaction des futurs rapports. Et elle relève la nécessité de veiller alors à décrire de manière suffisamment précise les données et la manière de procéder, sur le plan technique, pour la réalisation de travaux de recherche de qualité.

Le Bureau HarmoS estime que ces recommandations sont judicieuses.

Compte rendu des résultats et communication sur les enquêtes

D'après les avis exprimés dans le cadre de l'expertise, la plupart des acteurs concernés trouvent que les rapports sur la vérification sont de bonne qualité. Mais il semble que des améliorations s'avèrent nécessaires dans la communication des résultats. L'expertise recommande d'inclure plus tôt et plus activement les cantons dans la communication, notamment dans le but de renforcer la visibilité du projet Vérification. Les cantons ne devraient pas être informés sur le tard en ce qui concerne la communication des résultats à l'échelle nationale.

Du côté des cantons, on souhaiterait également recevoir davantage de soutien spécifique de la part des acteurs du projet Vérification (ICER, centres de coordination), d'une part pour l'interprétation des résultats et, de l'autre, pour une compréhension approfondie du questionnaire contextuel. De bonnes explications peuvent aider à apporter une réponse satisfaisante au degré d'acceptation parfois plutôt faible que l'on observe, au sujet de certaines questions, dans les milieux politiques comme parmi la population.

Le SG CDIP s'est déjà activement rapproché des cantons en mai 2019, lors de la communication initiale faite sur les résultats des enquêtes 2016 et 2017, et leur a dans ce contexte proposé différentes offres d'information. Celles-ci ont toutefois reçu un accueil assez variable selon les cantons. Le Bureau HarmoS trouve très important que la communication initiale sur les résultats de la vérification soit préparée de manière à recevoir l'attention qu'elle mérite. À cet effet, les cantons doivent être systématiquement inclus en tant que partenaires privilégiés et il convient de leur fournir de manière proactive les renseignements et offres d'information en lien avec le projet (y compris les informations d'ordre technique et conceptuel de la part de la direction de projet). Les canaux de communication à privilégier sont en particulier les différents organes et groupes de la CDIP, en plus des contacts de l'unité Communication de la CDIP.

Les avis exprimés dans le cadre de l'expertise révèlent aussi un certain nombre de souhaits d'amélioration concernant la communication adressée au corps enseignant. Des supports d'information spécialement adaptés devraient en particulier être mis à la disposition du corps enseignant et des élèves, pour présenter les enquêtes menées dans le cadre de la vérification et permettre ainsi une bonne compréhension du projet. Ces supports devraient remplir les fonctions suivantes: renseigner sur les tests (notamment au travers d'exemples de tâches), préciser ce qui est attendu des enseignantes et enseignants de même que des élèves, et finalement expliquer pourquoi aucun feedback individuel n'est donné à la suite des enquêtes.

Le Bureau HarmoS estime qu'il est de plus en plus important, notamment en raison de l'augmentation du nombre d'enquêtes prévues pour la suite, de veiller à une bonne communication avec les directions d'établissement, le corps enseignant et les élèves. On peut s'attendre à ce qu'une bonne communication, suffisamment claire, permette finalement aussi de renforcer la motivation pour les tests. On sait en effet qu'un manque de motivation et de volonté de coopérer lors des enquêtes peut entraîner une distorsion des résultats.

Dans les avis exprimés, les référentes et référents cantonaux pour les évaluations à large échelle se disent dans l'ensemble bien renseignés sur le projet Vérification, en particulier grâce aux centres de coordination régionaux. Les séances qui leur sont destinées ne sont par contre pas toujours jugées utiles: les informations présentées sont parfois déjà connues; l'occasion de discuter n'est souvent pas saisie; les précisions données par la direction de projet sont parfois ressenties comme trop techniques.

Les besoins exprimés dans le cadre de l'expertise par les référentes et référents cantonaux pour les évaluations à large échelle sont connus du SG CDIP, qui cherche à adapter en conséquence la forme et le contenu des séances.

Le groupe chargé de l'expertise relève quant à lui que l'on ne devrait pas, au moment de préparer la communication initiale sur les résultats, travailler avec des valeurs (relatives à l'atteinte des compétences fondamentales) ajustées sur le plan statistique, mais bien avec les valeurs observées uniquement. La communication initiale a pour but de renseigner, pour chaque canton ou partie d'un canton (dans le cas des cantons bilingue ou plurilingues), sur la proportion d'élèves ayant acquis les compétences fondamentales. Elle présente également des comparaisons dites «ajustées», c'est-à-dire tenant compte des différences de composition sociale entre les populations scolaires des différents cantons. Or le groupe chargé de l'expertise craint que de telles adaptations entraînent une relativisation de la nécessité d'agir, en minimisant les besoins de changement, puisqu'il s'agit justement de mettre en place des mesures là où des facteurs sociaux ont des répercussions défavorables sur la réussite scolaire.

Pour le Bureau HarmoS, cet effet de minimisation n'est évidemment pas souhaitable. La façon de concevoir la communication initiale devra de toute manière être revue en fonction de la direction que prendra la suite du projet Vérification. Si le fait de proposer des approches explicatives en lien avec les différences

constatées dans les résultats devient, en plus de l'évaluation du système, un objectif à part entière de la vérification (cf. point 2.1.2), cet élément se répercutera aussi dans la communication initiale. Les facteurs sociodémographiques constitueront alors une composante parmi d'autres.

Recoupements avec le rapport sur l'éducation

Les résultats de la vérification peuvent être intégrés de trois manières différentes dans le rapport sur l'éducation: 1) de manière directe (en particulier par des comparaisons entre cantons, que l'on ne trouve nulle part ailleurs); 2) au travers d'études d'approfondissement réalisées, sur la base des données de la vérification, par le CSRE sur des thématiques en lien avec le rapport sur l'éducation; 3) via des travaux de recherche de tiers, qui se basent sur les résultats de la vérification.

Les résultats publiés lors de la communication initiale faite sur les précédents cycles de vérification n'ont pas pu et ne peuvent pas, en raison de problèmes d'incompatibilité entre les calendriers, être intégrés dans le rapport sur l'éducation. Cependant, des analyses approfondies menées à partir de données issues de la vérification sont prises en compte dans le rapport 2023 sur l'éducation.

Le directeur du CSRE considère que le *design* longitudinal par cohortes tel qu'il est proposé – avec une planification sur le long terme et des mesures de tendances régulières incluant l'ensemble du spectre de compétences – permet particulièrement bien de tenir compte des besoins constatés en matière de compte rendu dans le domaine de l'éducation et de la formation:

- Il devient possible de planifier l'utilisation des données issues de la vérification. Ces données sont précieuses parce qu'elles sont obtenues au travers d'échantillons représentatifs ou d'enquêtes exhaustives réalisées dans pratiquement tous les cantons.
- Les échelles de test livrant des mesures fiables sur l'ensemble du spectre de compétences permettent bien d'analyser les liens statistiques entre les données obtenues dans les tests et les données contextuelles.
- Les mesures de tendances sont conçues pour générer des connaissances utiles au pilotage.
- Les coupes longitudinales de cohortes comprennent aussi, dans les cantons de plus petite taille, bon nombre de coupes longitudinales individuelles. Celles-ci sont particulièrement utiles pour suivre des parcours de développement.

Pour que l'on puisse modéliser correctement les facteurs exerçant une influence dans le contexte scolaire, il serait de plus utile d'établir des échantillonnages de classes et de prélever des caractéristiques contextuelles au niveau de la classe et de l'enseignement.

Annexe 2: Expertise scientifique sur le projet Vérification (en allemand)

Universität Zürich[UZH]

Institut für Erziehungswissenschaft

Wissenschaftliches Gutachten zur „Governance der Überprüfung des Erreichens der Grundkompetenzen (ÜGK)“

Prof. Dr. Katharina Maag Merki (Universität Zürich), Leitung

Dr. Anja Giudici (Universität Oxford)

Prof. i.R. Dr. Eckhard Klieme (DIPF – Leibniz-Institut für Bildungsforschung und Bildungsinformation, Johann Wolfgang Goethe-Universität Frankfurt am Main.)

Prof. Dr. Hans-Anand Pant (Humboldt-Universität zu Berlin)

Prof. Dr. Claudia Schreiner (Universität Innsbruck)

Mitarbeit und Koordination: Kai Schudel (Universität Zürich)

31. Oktober 2021

Inhaltsverzeichnis

1 Einleitung **4**
1.1 Strukturierung der Fragen und Ergebnisdarstellung 4
1.2 Informationsgrundlage 5

2 Aufgaben und Ertrag der ÜGK **8**
2.1 Ziele und Funktionen der ÜGK (A1.2) 9
2.2 Nutzung der Ergebnisse für den eigenen Arbeits- und Zuständigkeitsbereich (A1.3) 11
2.3 Umfang der Evidenz im Gesamtsystem (A2.8; A2.9) 14
2.4 Kommunikation der Ergebnisse der ÜGK und Berichterstattung (A1.4; A1.5) 15
2.5 ÜGK und Bildungsbericht (A2.10) 16

3 Zielvorgaben der ÜGK **18**
3.1 Auswahl der Fächer und Jahrgangsstufen (B1.1; B1.2) 18
3.2 Mindeststandards oder differenzierte Leistungsstufen? (B1.4; B3.13; B3.15) 20

4 Design der ÜGK **29**
4.1 Erhebungsrhythmus und Stichprobenziehung (B1.3; B1.5; B1.6; B2.8) 29
4.2 Konzeption und Entwicklung der Kontextfragebögen (B3.16) 32
4.3 Prozesse der Aufgaben- und Testentwicklung (B3.11; B3.12) 34
4.4 Technische Fragen zur ÜGK (B3.14; B3.17; B3.18) 37

5 Institutionelle Gestaltung der ÜGK **39**
5.1 Organisation und Verantwortungsstruktur (B2.7; B.2.8; B2.9) 39
5.2 Verhältnis ÜGK und Wissenschaft (A1.6; B2.10; B3.16; B3.19) 43

6 Empfehlungen **47**
6.1 ÜGK im Gesamtsystem 49
6.1.1 Weiterführung der ÜGK bei substanziellen Anpassungen im Gesamtdesign, der institutionellen Verankerung und Gestaltung sowie in der Zuweisung von finanziellen und personellen Ressourcen 49
6.1.2 Langfristige Planung bezüglich Erhebungsdesign gewährleisten und Durchführung vertraglich verbindlich absichern 49
6.1.3 Verhältnis der ÜGK zu Monitoringinstrumenten in den Kantonen und Sprachregionen klären 50
6.1.4 Aufgaben und Rollen von Politik, Bildungsadministration und Forschung/Wissenschaft im Gesamtgefüge der ÜGK klären 50

6.2 Grundlagen der ÜGK 52
6.2.1 Funktionen der ÜGK klären 52
6.2.2 Frage nach den Grundkompetenzen klären 52
6.2.3 Frage nach der Harmonisierung der Grundkompetenzen klären 53
6.3 Zielvorgaben der ÜGK 54
6.3.1 An einer Orientierung an Mindeststandards und der Überprüfung des Erreichens der Grundkompetenzen festhalten sowie die 2011 formulierten Grundkompetenzen weiterhin als Basis der Aufgabenentwicklung akzeptieren 54
6.3.2 Frage klären, welche Bedeutung der soziale Vergleich zwischen den Kantonen hat und adäquate Lösungen für den Kantonsvergleich finden 55
6.4 Design der ÜGK 56
6.4.1 An der Wahl der Jahrgangsstufen und der Fächer / Fachbereiche sowie an einer standardisierten Erhebung über alle Kantone hinweg festhalten 56
6.4.2 Erhebungsdesign klären: Trend-, Kohorten- und Längsschnitterhebungen? 56
6.4.3 Inhalte der nationalen Berichte klären 57
6.4.4 Kontextinformationen erweitern, in Abhängigkeit der Funktionen der ÜGK 57
6.4.5 Stichprobenziehung entsprechend den Funktionen der ÜGK anpassen 58
6.4.6 Öffentlichen Kommunikation der Ergebnisse der ÜGK verbessern 58
6.4.7 Weitere Empfehlungen 58
6.5 Institutionelle Gestaltung der ÜGK 59
6.5.1 Klare sowie vereinfachte Zuständigkeits-, Entscheidungs- und Verantwortungsstrukturen implementieren 59
6.5.2 Testentwicklung als komplexes und anspruchsvolles wissenschaftliches Aufgabenfeld anerkennen, mit genügend personellen Ressourcen ausstatten und institutionell in die wissenschaftliche Projektstruktur der ÜGK einbinden 60
6.5.3 Fachdidaktik(en) stärker in den Gesamtprozess der ÜGK einbinden 61
6.5.4 Sprachregionale Unterschiede in Bezug auf fachliches Lernen und Aufgaben herausarbeiten sowie für die Testentwicklung nutzen 61

I. Anhang – Literaturverzeichnis 62

II. Anhang – Ergebnisse der standardisierten Befragung 64

III. Anhang – Fragekatalog der EDK 82

1 Einleitung

Im Januar 2021 hat die EDK beschlossen, wissenschaftliche Expert*innen mit einem Gutachten zu fachlichen Aspekten der „Governance der Überprüfung des Erreichens der Grundkompetenzen (ÜGK)“ zu beauftragen. Der Auftrag wurde einer internationalen Gutachtengruppe, bestehend aus Dr. Anja Giudici (Universität Oxford), Prof. i.R. Dr. Eckhard Klieme (DIPF), Prof. Dr. Hans Anand Pant (Humboldt-Universität zu Berlin) und Prof. Dr. Claudia Schreiner (Universität Innsbruck) unter der Leitung von Prof. Dr. Katharina Maag Merki (Universität Zürich) erteilt. Die Gutachtengruppe hat das vorliegende Gutachten im Zeitraum von April bis Oktober 2021 auf Basis eines Fragekatalogs der EDK erarbeitet.

1.1 Strukturierung der Fragen und Ergebnisdarstellung

Die Ausgangslage des wissenschaftlichen Gutachtens bildet der Fragekatalog der EDK (siehe Anhang III), der grob in einen Bereich A der „externen Governance“ und einen Bereich B der „internen Governance“ aufgeteilt ist. Der Bereich A wird in die zwei Teilbereiche „A1 Systemsteuerung, Zielkohärenz und Handlungskoordination“ und „A2 Die ÜGK und weitere Erhebungen im Bildungssystem“ unterteilt. Der Bereich B umfasst drei Teilbereiche: „B1 Erhebungsdesign“, „B2 Projektorganisation und institutionelle Verankerung“ sowie „B3 Fachliche und technische Aspekte“.
Eine Analyse der Fragestellungen hat gezeigt, dass einzelne Aspekte, so zum Beispiel der Kontextfragebogen oder das Verhältnis von Wissenschaft und Politik, an verschiedenen Stellen angeschnitten werden. Deshalb hat die Gutachtengruppe entschieden, die Fragen thematisch zu sortieren und in einem groben Modell darzustellen (Abbildung 1). In einem ersten Bereich sind Fragen zu den *Aufgaben der ÜGK im Bildungsmonitoring* angesiedelt. Die Definition der Aufgaben sowie die Positionierung der ÜGK im Gesamtsystem führt in der Folge zu *Zielvorgaben der ÜGK* bezüglich Fächerauswahl und Definition von Bildungsstandards. Diese Zielvorgaben wiederum münden in eine operative Umsetzungs-Ebene mit Implikationen für das *Design der ÜGK* und für *die institutionelle Gestaltung der ÜGK*. Der *Ertrag der ÜGK* hängt wiederum zusammen mit den allgemeinen Vorstellungen über die Aufgaben der ÜGK im Bildungsmonitoring und die damit verbundenen Erwartungen an die ÜGK. Darüber hinaus ist der wahrgenommene Ertrag abhängig von der konkreten Umsetzung der ÜGK.

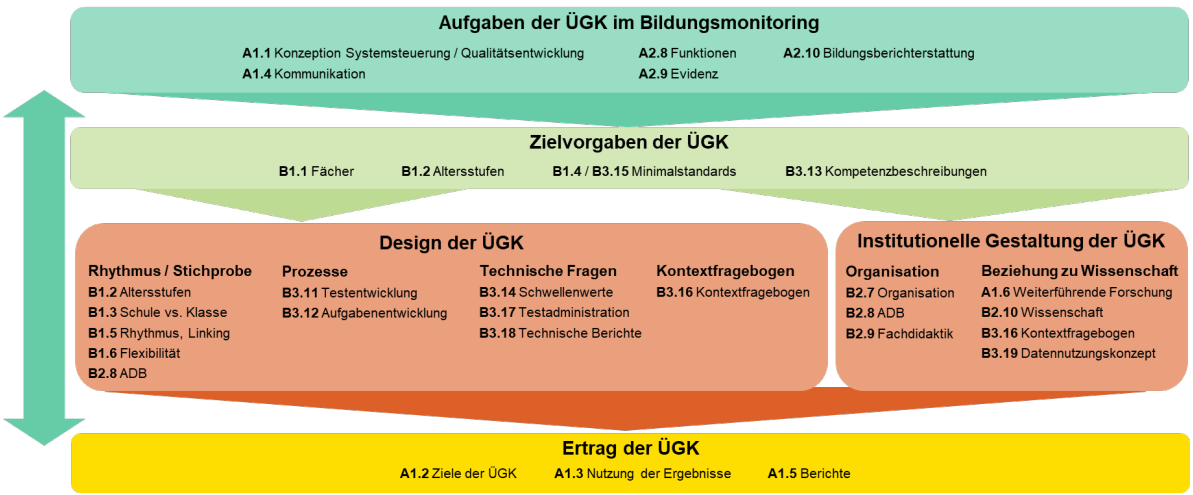


Abbildung 1: Modell zur Analyse der ÜGK im Gutachten

Die Darstellung der Befunde der Gutachtengruppe orientiert sich am obigen Modell (Abbildung 1). In Kapitel 2 werden zuerst Befunde zu *Aufgaben und Ertrag der ÜGK* zusammengefasst, Kapitel 3 präsentiert die Befunde zu den *Zielvorgaben der ÜGK*, Kapitel 4 berichtet die Befunde zum *Design der ÜGK* und in Kapitel 5 werden die Befunde zur *institutionellen Gestaltung der ÜGK* erläutert. Diese Kapitel werden

jeweils in Themen unterteilt, die in den Untertiteln beschrieben sind und gleichzeitig auf die Nummern/spezifischen Fragen des Fragekatalogs der EDK verweisen. Die Fragen A1.1 und A1.7 zielen auf Empfehlungen ab und werden entsprechend nicht explizit unter Befunde erörtert. [1] In Kapitel 6 schliesslich sind Empfehlungen der Gutachtengruppe zusammengefasst, die in die Bereiche *ÜGK im Gesamtsystem (Kapitel 6.1)*, *Grundlagen der ÜGK (Kapitel 6.2)*, *Zielvorgaben der ÜGK (Kapitel 6.3)*, *Design der ÜGK (Kapitel 6.4)* und *institutionelle Gestaltung der ÜGK (Kapitel 6.5)* aufgeteilt sind.

1.2 Informationsgrundlage

Die Ergebnisse und Empfehlungen des Gutachtens basieren auf Dokumenten der ÜGK, auf Einzel- und Gruppengesprächen mit Akteur*innen, einer standardisierten Befragung von Akteur*innen und auf der Fach-Expertise der Gutachter*innen. Das Generalsekretariat der Schweizerischen Konferenz der Erziehungsdirektoren (GS EDK) hat den Gutachter*innen für den Bearbeitungszeitraum Zugang zu allen relevanten Dokumenten zur ÜGK gewährt (z.B. Verträge, Berichte, Konzeptpapiere). Auf dieser Basis konnte eine *Dokumentenanalyse* vorgenommen werden.
Auf Grundlage der Dokumentenanalyse, sondierender Gespräche mit Akteur*innen der ÜGK und dem Fragekatalog der EDK wurden in einem nächsten Schritt für das Gutachten *relevante Akteur*innen definiert*. Orientiert an der im Fragekatalog angelegten Unterscheidung zwischen externer und interner Governance wurde zwischen externen und internen Akteur*innen unterschieden.
*Externe Akteur*innen* sind aktuelle oder mögliche Adressat*innen der ÜGK-Ergebnisse, die also Ansprüche an die Nutzung der ÜGK artikulieren können. Als wesentliche externe Akteur*innen wurden die Vertretungen der kantonalen Bildungsadministrationen und der EDK-Regionalkonferenzen sowie die Bildungsforschung und die Schulpraxis identifiziert. Die Bildungsadministrationen können als zentrale Stakeholder betrachtet werden, da das Bildungsmonitoring sowie die Qualitätsentwicklung im Bildungssystem in ihrem Verantwortungsbereich liegen. Hier wurden einerseits kantonale Vertreter*innen fokussiert, die Auskunft darüber geben können, ob und wie ÜGK-Ergebnisse für das kantonale Bildungsmonitoring und die Qualitätsentwicklung nutzbar sind, andererseits auch Personen, die mit der Umsetzung der ÜGK beauftragt sind (kantonale Referenzpersonen). Es hat sich gezeigt, dass diese beiden Funktionen oft personell dort zusammenfallen, wo kantonale Bildungsdirektionen (oder Volksschulämter) in Bezug auf Grösse und Ressourcen gering ausgestattet sind. Entsprechend wird die Perspektive dieser beiden Gruppen im Folgenden nicht mehr differenziert und als Perspektive der kantonalen Vertreter*innen (oder der „kantonalen Vertretungen“) zusammengefasst. Als *interne Akteur*innen* wurden Personen adressiert, die an der Gestaltung und Umsetzung der ÜGK beteiligt sind. Dies sind die Verantwortlichen der Projektleitung ÜGK, das Board und der Beirat des Interfaculty Centre for Educational Research (ICER) der Universität Bern, Mitglieder des Kosta HarmoS und externe Wissenschafter*innen, die im Rahmen des wissenschaftlichen Konsortiums an der ÜGK mitarbeiten. Teilweise sind diese Akteur*innen in mehreren Rollen in die ÜGK involviert.
Die Erfahrungen und Sichtweisen dieser Akteur*innen wurden für das Gutachten über zwei Wege eingeholt. Zum einen wurden Einzel- und Gruppengespräche geführt, dies im Zeitraum Juni bis Oktober 2021, zum anderen wurden dieselben Personen gebeten, an einer standardisierten Befragung teilzunehmen. Diese Befragung fand im Juli 2021 statt. Die *Einzel- und Gruppengespräche* wurden meistens von jeweils zwei Gutachter*innen geführt, protokolliert und aufgenommen, damit die gesamte Gutachtengruppe die Gesprächsinhalte nachträglich nachvollziehen konnte[2]. Dabei wurden die Gruppen so zusammengestellt, dass die Akteur*innen ähnliche Funktionen repräsentieren und sprachregional zusammenpassten. Es konnten 12 Einzel- oder Gruppengespräche mit externen Akteur*innen geführt werden: mit Vertreter*innen von Kantonen aus den vier Landessprachen sowie mit Vertreter*innen der verschie-

[1] Die Frage A2.11 nach dem optimalen Beitrag der ÜGK zur Dateninfrastruktur konnte im Gutachten nicht untersucht werden.

[2] Sämtliche Tonaufnahmen werden nach Ende des Auftrags an die Gutachtengruppe gelöscht.

denen Regionalkonferenzen beziehungsweise von Kosta HarmoS. Dabei ist es – mit Ausnahme der Kantone Zug, Schaffhausen und der französischsprachigen Vertretung des Kantons Bern – gelungen, die Perspektive sämtlicher Kantone und Sprachregionen mindestens über eine Vertretung einzuholen. Wichtig ist, dass auch kleinere Kantone und solche, die dem HarmoS-Konkordat nicht beigetreten sind, vertreten waren. Zusätzlich nahmen in sechs Einzel- oder Gruppengesprächen auch externe Akteur*innen der Bildungsforschung und der Schulpraxis unterschiedlicher Sprachregionen teil. Interne Akteur*innen, die Projektleitung ÜGK und das wissenschaftliche Konsortium, wurden in sechs Einzel- oder Gruppengesprächen befragt und mit Fachdidaktiker*innen konnten zusätzlich drei Gespräche geführt werden. Ebenfalls wurden zwei Gespräche mit Mitgliedern des Beirats und des Boards geführt. Die Zusammenstellung aller Akteur*innen, deren Perspektiven in die Gespräche einflossen (Tabelle 1), zeigt, dass die Bereitschaft der Akteur*innen, das Gutachten mit ihrer Erfahrung zu unterstützen, sehr gross war, und dass es den Gutachter*innen gelungen ist, eine grosse Breite und Vielfalt an wichtigen Akteur*innengruppen zu befragen.

Externe Akteur*innen	Anzahl		Institutionen
	fr/it	dt/ro	
Vertreter*innen der Kantone, Regionalkonferenzen oder Kosta HarmoS	8	27	Alle Kantone (ausser ZG, SH und französischsprachiges BE) und alle Regionalkonferenzen
Schulpraxis	2	2	VSLCH, LCH, SER, CLACESO
Bildungsforschung		1	SKBF
Interne Akteur*innen	**Anzahl**		**Organisationen**
Akteur*innen der Projektleitung ÜGK	9		ICER, ADB, SUPSI, SRED, PHSG, IBE
Externe Wissenschafter*innen des Wissenschaftlichen Konsortiums	1		Arbeitsgruppe Kontextfragebogen
Mitglied des Beirats	1		
Mitglieder des Boards	3		
Fachdidaktiker*innen	4		

Tabelle 1: Personenanzahl der Akteur*innen für Einzel- und Gruppengespräche; SKBF=Schweizerischen Koordinationsstelle für Bildungsforschung; VSLCH=Verband Schulleiterinnen und Schulleiter Schweiz; LCH=Dachverband Lehrerinnen und Lehrer Schweiz; SER=Syndicat des Enseignantes et Enseignants de Suisse romande; CLACESO=Conférence Latine des Chefs d'Etablissement de la Scolarité Obligatoire

Der *standardisierte Fragebogen* wurde per E-Mail an dieselben Akteur*innen geschickt, die schon für die Einzel- und Gruppengespräche angeschrieben worden waren. Zusätzlich diente die Befragung dazu, die Einschätzungen von Forschenden aus der Bildungsforschung (z.B. Erziehungswissenschaft, Fachdidaktik) und von Personen aus der Aus- und Weiterbildung von Lehrpersonen in standardisierter Form zu erhalten. Für den Zugang zu dieser Akteur*innengruppe konnte auf den E-Mail-Verteiler der Schweizerischen Gesellschaft für Bildungsforschung (SGBF) zurückgegriffen werden. I
nsgesamt haben 141 Personen den Online-Fragebogen ausgefüllt. Es haben 39 kantonale Vertreter*innen teilgenommen, wovon 21 aus Kantonen stammen, die im HarmoS-Konkordat sind, und 15 aus Kantonen, die dem HarmoS-Konkordat nicht beigetreten sind. Damit kann die Datenbasis für die Kantone als belastbar interpretiert werden. Weitere 77 Personen, die den Fragebogen ausgefüllt haben, gaben an, in der Forschung und/oder der Aus- und Weiterbildung von Lehrpersonen tätig zu sein[3]. Bezüglich

[3] Die restlichen 25 Personen haben andere Funktionen angegeben, darunter 3 Personen aus der Schulpraxis.

der sprachregionalen Verteilung konnten 103 deutsch- oder romanischsprachige, 25 französischsprachige und 13 italienischsprachige Akteur*innen gewonnen werden. Die Befunde der standardisierten Befragung werden teilweise im Bericht präsentiert und stehen im Anhang im ganzen Umfang zur Verfügung.

Dank

An dieser Stelle möchten wir unseren Dank allen Personen aussprechen, die die Erstellung dieses Gutachtens unterstützt haben.

Zunächst geht unser Dank an alle Personen, die ihre Erfahrungen und Einschätzungen mit uns geteilt und sich für ein Einzel- oder Gruppengespräch zur Verfügung gestellt haben. Ebenso gebührt unser Dank all jenen, die die standardisierte Befragung mitten in der Sommerpause bearbeitet haben. Sehr herzlich möchten wir zudem Peter Lenz vom Generalsekretariat der EDK danken, der uns mit allen notwendigen Unterlagen ausgestattet hat und uns in unserer Arbeit hervorragend unterstützt hat. Zu guter Letzt gebührt unser Dank Susanne Hardmeier, Generalsekretärin der EDK, sowie der EDK, die uns die Erstellung des Gutachtens anvertraut haben.

Wir würden uns sehr freuen, wenn das Gutachten einen Beitrag leistet, das Bildungsmonitoring in der Schweiz und die Realisierung der ÜGK qualitativ weiterzuentwickeln.

2 Aufgaben und Ertrag der ÜGK

Die ÜGK ist als Instrument der Systemevaluation im Rahmen des nationalen Bildungsmonitorings eingebettet in ein komplexes Set an weiteren Leistungsüberprüfungen und Standortbestimmungen, die mit unterschiedlicher Intensität an obligatorischen Schulen in der Schweiz durchgeführt werden.

- Die Schweiz nimmt seit 2000 an der internationalen PISA-Studie teil, seit 2015 – im Zusammenhang mit dem Entscheid zur ÜGK – ausschliesslich mit einer nationalen Stichprobe (mit Ausnahme des Kantons Tessin, der nach wie vor mit einer kantonal repräsentativen Stichprobe an den PISA-Erhebungen teilnimmt). Die Teilnahme an der PISA-Studie erlaubt die Positionierung der Schweiz im internationalen Vergleich, ermöglicht aber keine kantonsvergleichenden Analysen mehr.
- Mit der Einführung der ÜGK liegt nun erstmals ein nationales, auf schweizerische Bedürfnisse ausgerichtetes Instrument der Systemevaluation vor, welches diese kantonsvergleichenden Analysen mit explizitem Bezug zu den entwickelten Bildungsstandards möglich macht. Alle Kantone – mit Ausnahme des Kantons Zug ab 2023 – nehmen an diesen Erhebungen teil, auch jene, die nicht dem HarmoS-Konkordat beigetreten sind.
- Gemäss Umsetzungsstrategie des HarmoS-Konkordates (EDK, 2007, Art. 5.3.1 und 5.3.2) sollen, ebenfalls auf Basis der Bildungsstandards und der regionalen Lehrpläne, Instrumente zur individuellen Standortbestimmung der Schüler*innen auf sprachregionaler Ebene entwickelt werden. Dieses Ziel ist allerdings in der geplanten Form (noch) nicht realisiert worden.
- In einzelnen Fällen, beispielsweise in der Nordwestschweiz, kommen kantonsübergreifende Erhebungen und Leistungstests zum Einsatz, wobei der Bezug zu den Mindeststandards teils unklar ist.
- Zusätzlich haben rund die Hälfte der Kantone eigene standardisierte Leistungstests auf der Primarstufe und fast alle Kantone standardisierte Leistungstests auf der Sekundarstufe implementiert, wobei der Bezug zu den Mindeststandards teils unklar ist.[4]
- In verschiedenen Kantonen werden keine eigenen Leistungstests durchgeführt.

Damit variiert die Erhebungsdichte je nach Kanton stark. Zudem haben die in den Kantonen oder Regionen eingesetzten Instrumente unterschiedliche Ziele. Zwar dienen die meisten in den Kantonen und Regionen angewendeten Tests primär der Standortbestimmung der Schulen, Klassen oder Schüler*innen, wie beispielweise die in verschiedenen deutschsprachigen Kantonen angewendeten Stellwerk- oder Klassencockpit-Tests oder die Genfer EVACOM. Allerdings dienen verschiedene Leistungsevaluationen zusätzlich auch dem kantonalen oder regionalen Systemmonitoring, darunter beispielweise die Épreuves Cantonales de Référence (ECR) im Kanton Neuenburg oder die in der Nordwestschweiz zum Einsatz kommenden Checks.[5] Diese Situation trägt dazu bei, dass das Bedürfnis nach Daten und Analysen zur Systemevaluation auf nationaler Ebene je nach Kanton variiert. Die in den Gesprächen angesprochenen Problemlagen in Bezug auf Ziele und Funktionen der ÜGK in Abstimmung mit anderen Erhebungsinstrumenten müssen vor dem Hintergrund dieser heterogenen Erhebungs- und Datendichte gelesen werden.

[4] Siehe dazu die entsprechende EDK Kantonsumfrage von 2019/20: https://www.edk.ch/edk.ch/platform/de/de/bildungssystem/kantonale-schulorganisation/kantonsumfrage/b-35-standardisierte-leistungstests

[5] https://www.edk.ch/edk.ch/platform/de/de/bildungssystem/kantonale-schulorganisation/kantonsumfrage/b-35-standardisierte-leistungstests

2.1 Ziele und Funktionen der ÜGK (A1.2)

Breite Akzeptanz des Ziels „Überprüfung des Erreichens der Grundkompetenzen“ als Systemevaluation auf nationaler Ebene, unabhängig davon, ob der eigene Kanton dem HarmoS-Konkordat beigetreten ist oder nicht.

Die ÜGK ist heute das einzige Instrument, das eine interkantonal vergleichende Lernstandsanalyse erlaubt. Das entsprechende Ziel, die Überprüfung des Erreichens der Grundkompetenzen in der Schweiz im Rahmen der nationalen Systemevaluation, wird in den Gesprächen durchgehend in hohem Masse akzeptiert. Die standardisierte Befragung zeigt, dass seitens der kantonalen Vertreter*innen alle Personen, mit einer Ausnahme, es eher wichtig (44 %) oder sehr wichtig (51 %) finden, das Erreichen der Grundkompetenzen der Schüler*innen auf nationaler Ebene zu überprüfen, dies unabhängig davon, ob der eigene Kanton dem HarmoS-Konkordat beigetreten ist oder nicht. Bei Personen aus der Forschung und Aus- und Weiterbildung ist die Zustimmung mit 74 % („eher wichtig“ oder „sehr wichtig“) ebenfalls sehr hoch. Von den kantonalen Vertretungen wird die ÜGK vor allem als (eher) wichtig für das Fällen von bildungspolitischen Entscheiden auf nationaler Ebene (90 %) beurteilt; auch hier bestehen keine Unterschiede zwischen den HarmoS- und Nicht-HarmoS-Kantonen. Die ÜGK wird darüber hinaus als eher wichtig oder sehr wichtig für das Fällen von bildungspolitischen Entscheiden auf kantonaler (77 %) und sprachregionaler Ebene (64 %) angesehen. Von einzelnen Vertreter*innen der Schulpraxis wird allerdings darauf hingewiesen, dass die ÜGK tendenziell ein dysfunktionales Potenzial für Konkurrenzverhalten und Ranking habe und die Durchführung nationaler Leistungserhebungen somit grundsätzlich wenig sinnvoll sei.

Der ÜGK werden neben der Systemevaluation auf nationaler Ebene verschiedene weitere Funktionen zugeschrieben.

Die Umschreibung der Funktion der ÜGK im Rahmen der Systemevaluation auf nationaler Ebene in den Gesprächen kann drei Themenbereichen zugeordnet werden, wobei die ersten beiden typische Ziele des Bildungsmonitorings sind, die auch im Zusammenhang mit anderen Bildungssystemen genannt werden:

- Interkantonaler Vergleich, d.h. soziale Bezugsnormorientierung: „Wo steht der eigene Kanton im Vergleich zu den anderen?“
- Harmonisierung des Erreichens der Grundkompetenzen zwischen den Kantonen und Sprachregionen: „Inwiefern ist die Harmonisierung realisiert?“
- Kriteriale Bezugsnormorientierung: „Haben die Schüler*innen in der Schweiz und im eigenen Kanton die Grundkompetenzen erreicht?“

In einzelnen Fällen erhoffen sich Vertreter*innen der Kantone durch die ÜGK zusätzlich eine Überprüfung von Massnahmen, die im Kanton in den letzten Jahren eingeführt wurden. Dies auch im Zusammenhang mit den Ergebnissen der PISA-Studien. Zudem könne die ÜGK die Funktion haben, einen überkantonalen Meinungsaustausch zu Bildungsthemen zu ermöglichen. Die Publikation der ÜGK-Resultate hätte das Potential, als notwendig erachtete regionale oder gar schweizweite Diskussionen über die Ziele und Strukturen des Bildungswesens zu entfachen. Um diese zu erfüllen, fehle es der ÜGK allerdings an Visibilität.

Die Befunde der standardisierten Befragung zeigen darüber hinaus, dass neben der Systemevaluation auf nationaler und kantonaler Ebene auch erwartet wird, dass *mittels der ÜGK Ursachen für Leistungsunterschiede zwischen regionalen und kantonalen Bildungssystemen identifiziert werden können*. 80 % der befragten kantonalen Vertreter*innen sowie 68 % der Personen aus der Forschung und Aus- und Weiterbildung geben an, dass dies eine relevante Funktion der ÜGK sei.

Umstrittener ist, ob die ÜGK Handlungswissen bereitstellen kann und soll.

In der standardisierten Befragung ist die Mehrheit der kantonalen Vertreter*innen der Ansicht, dass die ÜGK auch Handlungswissen bereitstellen sollte, wie das Bildungssystem auf nationaler (85 %), sprachregionaler (62 %) oder kantonaler (74 %) Ebene verändert werden kann. Etwas weniger häufig, aber mit 54% immer noch mehr als die Hälfte der kantonalen Vertreter*innen, stimmen der Aussage zu, dass die ÜGK auch die Funktion haben sollte, Handlungswissen darüber zu generieren, wie die Schulpraxis verändert werden kann. In den Gesprächen wird allerdings deutlich, dass für die meisten Befragten die ÜGK kaum eine Funktion bezüglich der Generierung von Handlungswissen für die konkrete Schulpraxis, für die Weiterentwicklung des Unterrichts oder kantonaler Bildungssysteme hat beziehungsweise haben kann. Dies vor allem, da die Ergebnisse nicht einzelnen Lehrpersonen zurückgemeldet und für die konkrete Weiterentwicklung des Unterrichts genutzt werden könnten. Zudem können im Rahmen der ÜGK-Erhebungen keine Informationen gewonnen werden, wie allfällige Probleme am besten behoben werden könnten. Während ersteres nicht als Problem angesehen wird – das Geben einer individuellen Rückmeldung an die Lehrpersonen sei nicht die Funktion der ÜGK – wird das Fehlen konkreter bildungspolitisch nutzbarer Hinweise für die Veränderung der Bildungspraxis zum Teil bemängelt. Von den Vertreter*innen der Schulpraxis wird die hauptsächliche Ausrichtung der ÜGK auf die Funktion der Systemevaluation allerdings ausdrücklich begrüsst. Darüber hinaus argumentieren viele Kantone, dass sie eigene Instrumente hätten, die explizit die Funktion der Standortbestimmung und Unterrichtsentwicklung haben. Die konzeptionell angedachte Abgrenzung zwischen nationaler Systemevaluation und kantonalen Erhebungen zur individuellen Standortbestimmung gilt somit generell als sinnvoll. Fehlen hingegen diese Instrumente in einem Kanton, so wird oftmals formuliert, dass die ÜGK mehr Funktionen übernehmen sollte.

Die positive Einschätzung der ÜGK bezüglich der verfolgten Ziele hängt von der im eigenen Kanton vorhandenen Datenqualität und -dichte ab.

Insbesondere Kantone, die selbst oder im Verbund mit anderen Kantonen Instrumente zur Erfassung der Lernleistungen der Schüler*innen implementiert haben, sprechen der ÜGK eine *ergänzende Funktion* zu, ergänzend zu den eigenen innerkantonalen und sprachregionalen Analysen und Erhebungen. Kantonale Erhebungen werden von lokalen Fachexpert*innen erarbeitet und teilweise von den Lehrpersonen korrigiert, womit sie als weniger objektiv gelten können. Aus dieser Perspektive wird die ÜGK vor allem als externes, unabhängigeres und explizit auf das vergleichende Systemmonitoring ausgerichtetes Instrument interpretiert und geschätzt. Dabei ist die interkantonal vergleichende Perspektive zentral. Diese Gruppe ist mit der jetzigen Konzeption der ÜGK als Teil einer breiteren Erhebungslandschaft mehrheitlich zufrieden und warnt vor einem Ausbau dieses Instruments. Diese Kantone werten den Umstand, dass die ÜGK anders ausgerichtet ist als PISA (andere Stichproben, Messen von Grundkompetenzen anstatt von Niveaustufen), besonders positiv und als Hinweis darauf, dass die Instrumente gut aufeinander abgestimmt sind.

Demgegenüber sprechen Kantone, welche zwar an Vergleichen interessiert sind, jedoch nicht an regionalen Leistungserhebungen teilnehmen oder selbst über keine umfassende Datenbasis verfügen, der ÜGK eine *Ersatzfunktion* zu. Diese Kantone erwarten, dass die ÜGK als vollwertiger und massgeschneiderter Ersatz für die nicht mehr stattfindenden regional-vertiefenden PISA-Analysen auf Kantonsbasis dient. Zu dieser Gruppe gehören unter anderem auch die französischsprachigen Kantone, welche die Entwicklung einer sprachregionalen Leistungserhebung nach der Einführung der ÜGK auf Eis gelegt haben, um die Schulen nicht zu überfordern.[6] Aber auch Kantone, die selbst über keine Monitoringinstrumente verfügen, können dieser Gruppe zugeordnet werden. Diese Kantone würden sich von

[6] Anstatt sprachregionaler Tests bauen diese Kantone eine regionale Aufgabendatenbank auf, die Lehrpersonen verwenden können, um den Stand ihrer Schulklassen zu evaluieren. Der Kanton Genf hat anhand der sprachregionalen Lehrpläne eigene standardisierte Leistungstests entwickelt und eingeführt.

den ÜGK eine umfassendere Ergebnisdokumentation wünschen und sehen somit die Funktion der ÜGK in einem erweiterten Sinne. Allerdings besteht in dieser Gruppe momentan eine gewisse Skepsis darüber, ob die ÜGK diese Funktion erfüllen kann. Diese Skepsis liegt einerseits am (noch) fehlenden Vertrauen in das Instrument der ÜGK, da diesem im Vergleich zu PISA eine schwächere wissenschaftliche Glaubwürdigkeit attestiert wird. Andererseits sprechen gewisse Kantone grundsätzlichere Probleme an, so die Besorgnis, dass das eigentliche Design der ÜGK – das ausschliessliche Messen von Grundkompetenzen oder das Fehlen von Daten zu bestimmten Fachbereichen oder regionaler Vergleiche – das Gewinnen von vertieftem Steuerungswissen verhindere. Hier werden entsprechende Anpassungen in Design und Analyse gefordert (vgl. dazu Kapitel 4).

ÜGK als Selbstevaluation der Kantone oder als Fremdevaluation?

Problematisiert wird vereinzelt die Frage, ob die ÜGK eine Art Selbstevaluation der Kantone sei oder ob es sich um eine Fremdevaluation handeln sollte, die vom Bund verantwortet werden müsste. Die Konzipierung der ÜGK als Fremdevaluation durch den Bund sei zentral, um den Ertrag für die Analyse der Stärken und Schwächen des schweizerischen Bildungssystems zu steigern und gegenüber dem Bund (Bezug nehmend auf Art. 62 der Bundesverfassung) Rechenschaft abzulegen. Dies würde aber auch bedeuten, dass die EDK keinen direkten Zugriff auf die Aufgabendatenbank und das Setzen der Standards mehr haben dürfte.

2.2 Nutzung der Ergebnisse für den eigenen Arbeits- und Zuständigkeitsbereich (A1.3)

Nutzen für die Bildungspolitik und die Bildungsadministration, kaum aber für die Schulpraxis.

Die ÜGK wird von den Vertreter*innen der Kantone als relevant für die Bildungspolitik und Bildungsadministration eingestuft. Für die Schulpraxis wird der Nutzen als sehr gering beurteilt. Die Resultate dienten in erster Linie der Bildungspolitik und Bildungsadministration für die Einordnung des Leistungsstands der Schüler*innen im Kanton (und in der Schweiz). Dies zeigt sich beispielsweise darin, dass im Kanton die Diskussion hauptsächlich im Dreieck Bildungsdirektor*in, Amtsvorsteher*in und kantonale Referenzperson stattfindet, nicht aber unter Einbezug der Lehrer*innenverbände.
Der mögliche Nutzen wird von einzelnen Akteur*innen in mehrerer Hinsicht beschrieben: Neben dem konkreten Einleiten von Massnahmen (z.B. Weiterbildungen oder Entwicklung eines neuen Lehrmittels) oder der Zuweisung von mehr finanziellen Mitteln für bestimmte Volksschulbereiche, wird von Vertreter*innen der Kantone auch beschrieben, dass die ÜGK zu mehr internen Diskussionen geführt habe. Der Blick werde bezüglich spezifischer Problemzonen geschärft. Manche seien zwar allenfalls bereits seit langem bekannt, aber durch die Ergebnisse der ÜGK würde deutlich, dass sie nun endlich angegangen werden müssten. Massnahmen, die schon länger geplant seien, könnten einfacher ins Feld gebracht werden. Darüber hinaus würde der Fokus der Politik auf das Bildungssystem gestärkt.
Ein möglicher Nutzen für einzelne Lehrpersonen im Sinne von der Bereitstellung von Handlungsempfehlungen oder als Kompetenz-Diagnoseinstrument wird von den Befragten nicht beschrieben, aber auch nicht problematisiert. Hingegen wird in einzelnen Gesprächen auf den indirekten Einfluss der ÜGK verwiesen, der sich beispielsweise darin zeige, dass Lehrpersonen und Schulen durch allfällige positive Resultate eine Bestätigung ihrer Arbeit in der Praxis erhielten, oder aber, falls die kantonalen Ergebnisse als problematisch beurteilt würden, Massnahmen für die Schulpraxis einleiten würden.

Aufwand-Nutzen-Verhältnis wird als ungünstig beurteilt, Mehrwert der ÜGK gegenüber anderen Erhebungen und Informationen nur teilweise wahrnehmbar.

In verschiedenen Gesprächen wird betont, dass der Aufwand der ÜGK im Verhältnis zu ihrem Nutzen gross oder zu gross sei. Zudem wird der Mehrwert, den die ÜGK im Vergleich zu anderen Daten und

Informationen im Bildungssystem erbringt, nur von einer knappen Mehrheit der kantonalen Vertreter*innen als eher gross (44 %) oder sehr gross (13 %) eingeschätzt. Des Weiteren zeigt die standardisierte Befragung, dass die kantonalen Vertreter*innen für die Identifikation von Stärken und Schwächen des Bildungssystems mehrheitlich andere Quellen und Informationen nutzen als die Ergebnisse der ÜGK. Am häufigsten sind dies wissenschaftliche Publikationen im Allgemeinen, Informationen aus Vernehmlassungen und Beratungen im Kanton oder in der Sprachregion, die Bildungsberichte Schweiz sowie Ergebnisse aus Erhebungen/Tests im eigenen Kanton. Einzig 31 % respektive 38 % der kantonalen Vertreter*innen geben an, die nationalen Berichte oder die vertiefenden Analysen zur ÜGK eher häufig beziehungsweise sehr häufig zu nutzen.
Auch für Entscheide bezüglich der Qualitätsentwicklung des Bildungswesens nutzen die Vertreter*innen der Kantone gemäss der standardisierten Befragung vor allem Informationen aus Vernehmlassungen, Beratungen im eigenen Kanton oder in der Sprachregion, wissenschaftliche Publikationen im Allgemeinen und Ergebnisse aus Erhebungen/Tests im eigenen Kanton. Ergebnisse aus den nationalen Berichten der ÜGK oder aus den vertiefenden Analysen werden einzig von circa 35 % der kantonalen Vertreter*innen eher häufig oder sehr häufig genutzt.

Verschiedene Änderungen im Konzept und der Durchführung der ÜGK könnten den Nutzen steigern.

In den Gesprächen werden verschiedene Vorschläge gemacht, wie man den Nutzen der ÜGK für die Identifikation von Stärken und Schwächen sowie das Fällen bildungspolitischer Entscheidungen steigern könne. Einzelne Kantone sind der Meinung, dass das Missverhältnis zwischen Aufwand und Ertrag nur durch differenziertere Ergebnisse und Analysen korrigiert werden könne. Um aus dieser Sicht den Nutzen zu steigern, müsse die ÜGK nicht nur Grundkompetenzen, sondern das gesamte Leistungsspektrum erfassen und es müsse eine höhere Kadenz in den Erhebungen erreicht werden. Letzteres wird allerdings aufgrund der finanziellen Belastung, des teils fehlenden politischen Willens und aufgrund der hohen Belastung der Schulen, insbesondere in kleinen Kantonen, teilweise kritisch beurteilt – in kleinen Kantonen sind jeweils alle Schulen von den Erhebungen betroffen.
In den Gesprächen werden weitere Gründe genannt, die dazu geführt hätten, dass die Ergebnisse der ÜGK bislang nur begrenzt für die Bildungsplanung und Bildungspolitik nutzbar gewesen seien. So wurden die ersten Erhebungen eher als Piloterhebungen interpretiert, deren Validität noch nicht richtig abgeschätzt werden konnte. Die ÜGK hätte medial bisher keine Wellen ausgelöst, wie dies bei der Veröffentlichung der Ergebnisse der PISA-Studie der Fall war, und sei damit in der Politik und Öffentlichkeit weniger präsent. Während einige befragte Personen diesen Umstand kritisieren, merken andere an, dass dies teilweise durchaus gewollt sei, da man ein mögliches Ranking zwischen den Kantonen verhindern wollte.
Da im Ablauf der Erhebungen (Jahrgangsstufen, Fächerauswahl) bisher keine klare Logik und auch keine langfristige Perspektive sichtbar geworden seien, bestünde zudem keine Planungssicherheit. Die Interpretation der Befunde sei schwierig und eine langfristige Entwicklung des Bildungssystems vor Ort kaum möglich. Falls Veränderungen im Konzept der ÜGK in Zukunft vorgenommen werden könnten, so wird der ÜGK auf bildungspolitischer und bildungsadministrativer Ebene ein sehr viel grösserer Nutzen zugesprochen (dies unabhängig davon, ob die Schulen eine direkte Rückmeldung erhalten oder nicht). Drei Viertel der kantonalen Vertreter*innen, die die Frage in der standardisierten Befragung beantwortet haben, geben sodann auch an, dass Änderungen im Konzept oder in der Vorgehensweise der ÜGK vorstellbar seien, die dazu führen würden, dass sie die Ergebnisse in ihrem Aufgaben- und Zuständigkeitsbereich besser nutzen könnten.

Ob aufgrund der ÜGK-Ergebnisse Massnahmen und Konsequenzen für Bildungspolitik oder Bildungspraxis abgeleitet werden, hängt von verschiedenen Faktoren ab.

Das Fällen von Entscheiden aufgrund der Ergebnisse der ÜGK scheint *abhängig vom erreichten Resultat* zu sein. Falls das Ergebnis der ÜGK für den eigenen Kanton in Abhängigkeit von den eigenen Erwartungen positiv ausfalle, führe dies zu keinen spezifischen Massnahmen, da diese Ergebnisse als Indikator für den Erfolg des eigenen Bildungssystems interpretiert werden. Falls das Ergebnis allerdings negativ ausfalle, werde dies als „politisches Desaster“ erlebt, was zu einem hohen sozialen Druck sowie zu Massnahmen in der Praxis führe. In einem Kanton hat jedoch schon die (geplante) Durchführung der ÜGK zu Massnahmen geführt, bevor überhaupt die Ergebnisse vorlagen.

Zentral ist die von den Akeur*innen *eingeschätzte Validität der Befunde*. In mehreren Gesprächen wird den PISA-Erhebungen eine ausgeprägte Glaubwürdigkeit attestiert. Sie gelten als wissenschaftlich fundiert und nuanciert, man schätzt die Bereitstellung benutzerfreundlicher Daten, die tiefergehende Analysen erlauben. Die ÜGK geniesst nicht überall dasselbe Vertrauen. Falls deren Validität als gering eingeschätzt wird, sei es aufgrund der aus ihrer Sicht mangelnden Durchführungsqualität, der Testmethoden, der Schwellenwertsetzung oder weil die sozioökonomischen Unterschiede zwischen den Kantonen nicht genügend berücksichtigt worden seien, werden die Ergebnisse nicht weiter in die Planung von Massnahmen einbezogen. Ebenso wird vor allem in der französischsprachigen Schweiz moniert, dass die Leistungstests und die Kontextfragebögen hauptsächlich in der Deutschschweiz konzipiert würden, daher für die Bildungspraxis in der französischsprachigen Schweiz weniger curricular valide seien. Dies sei teilweise in den Aufgabenstellungen ersichtlich geworden. Zudem würden sich die Lernkulturen und die Lehrpläne zwischen den Sprachregionen unterscheiden, so dass die Aussagekraft der Erhebungsinstrumente teils angezweifelt wird.

Ebenfalls wird moniert, dass der Kontextfragebogen in der Ergebnisdarstellung der ÜGK in den Berichten kaum Eingang gefunden habe und die Daten nicht (genügend) ausgewertet worden seien. So kritisieren mehrere Kantone, dass *das Potenzial der via Kontextfragebogen erhobenen Daten ungenügend ausgeschöpft werde***.** Einige Kantone beschreiben sich gar als „frustriert“, weil in ihren Augen die aus dem Kontextfragebogen gewonnenen Daten nicht in die Analysen eingeflossen sind.

Zudem sei der *Zugang zu den Daten zu kompliziert***,** *um eigene Analysen durchführen zu können,* insbesondere für Kantone ohne eigene Forschungsinstitutionen. Ausserdem fehle in den meisten Kantonen das Fachwissen wie auch die Ressourcen, um die Daten im jetzigen Format verwenden, die Ergebnisse angemessen interpretieren und in Massnahmen transferieren zu können.

Verschiedene Kantone wünschen sich, dass *auf der Basis der ÜGK tiefergehende und spezifischere Analysen erarbeitet werden*. Dieses Bedürfnis ist besonders ausgeprägt in einigen kleineren Kantonen, die selbst nicht über spezialisierte Bildungsplanungs- und Bildungsforschungsstellen verfügen. Insbesondere hier wird die Kritik laut, es könne aus den von der ÜGK erhobenen Daten zu wenig Steuerungswissen gewonnen werden, vor allem angesichts des Aufwands und der Kosten. Hier wünscht man sich die Möglichkeit, anhand der erhobenen Daten spezifischere Fragestellungen oder gezieltere Vergleiche erarbeiten zu lassen (z.B. zwischen verschiedenen städtischen Gebieten oder ähnlichen Schülerpopulationen in verschiedenen Kantonen). Momentan brauche die Erarbeitung der Vertiefungsanalysen, die die Kantone in Auftrag geben können, zu viel Zeit. Eine Nutzung der Ergebnisse für Bildungspolitik und Bildungspraxis sei daher kaum zeitnah möglich. Gewünscht wird von verschiedenen Akteur*innen, dass die vorhandenen Daten vollständiger ausgewertet werden und dass diese Analysen zum Kern der ÜGK werden und nicht zusätzlich beantragt und finanziert werden müssen.

Schwierig bezüglich des Entscheids, ob Massnahmen ergriffen werden oder nicht, ist die Beantwortung der Frage**,** *wie viele der Schüler*innen die Grundkompetenzen erreicht haben müssen* beziehungsweise wie gross die Gruppe jener Schüler*innen sein dürfe, die die Grundkompetenzen nicht erreicht hat, ohne dass dies zu Druck in der Politik oder Praxis führe. Ebenfalls wird mehrfach darauf hingewiesen, dass, auch wenn die Ergebnisse der ÜGK nachvollziehbar seien oder andere bisherige Ergebnisse bestätigten, die *Frage offenbleibe, wie denn mögliche negative Befunde erklärt werden könnten und wie entsprechenden Herausforderungen begegnet werden könne*.

Da die Lehrpersonen und Schüler*innen keine spezifischen Rückmeldungen zu den Ergebnissen der eigenen Klasse oder zu den eigenen Leistungen erhalten, *seien die ÜGK zu weit weg von deren Alltag*. Allerdings wird dies nur dann kritisiert, wenn der ÜGK diese Funktion zugeschrieben wird.
Die *Nützlichkeit der Befunde für die Bildungspolitik und -administration wird dann besonders positiv eingeschätzt, wenn die Ergebnisse der ÜGK bisherige Befunde (z.B. PISA oder interne Tests/Erhebungen) bestätigen*. Im Falle des Vergleichs mit intern erhobenen Daten erhält die ÜGK eine Art *Legitimationsfunktion*, indem sie als externes Instrument, dem man besser trauen könne, die internen Ergebnisse bestätige und somit legitimiere. Falls allerdings die ÜGK-Befunde mit den Befunden aus anderen Erhebungen nicht oder nur in Teilen korrespondierten (weil die Schüler*innen je nach Erhebung besser oder schlechter abschneiden, oder die Ergebnisse fachspezifisch sehr stark variieren), sei die Interpretation und Kommunikation der Befunde für die Bildungspolitik und -administration anspruchsvoll und für manche politischen Akteur*innen schwierig zu akzeptieren.

2.3 Umfang der Evidenz im Gesamtsystem (A2.8; A2.9)

Unklarheit, ob ÜGK ausgebaut werden soll oder im aktuellen Umfang funktional ist.

Trotz der eben angesprochenen Probleme wird im Allgemeinen nicht gewünscht, neue Erhebungsinstrumente zu etablieren, vor allem nicht auf nationaler Ebene. So besteht generell die Meinung, dass die Anzahl standardisierter Leistungsmessungen so gering wie möglich gehalten werden müsse. Standardisierte Erhebungen erforderten (zu) viel Unterrichtszeit und sollten sowohl im Schulalltag als auch in der Bildungspolitik kein übermässiges Gewicht erhalten.
Allerdings gibt es auch Stimmen, sowohl bei den Kantonen als auch bei den Vertretungen der Lehrpersonen, die sich von einem gezielten Ausbau der ÜGK wichtiges Steuerungswissen versprechen. Konkrete Anliegen beziehen sich auf die getesteten Fachbereiche, die Reichweite der Kontextfragebögen (z.B. Einbezug der Lehrpersonen, Eltern oder Schulleitungen), die gemessenen Kompetenzstufen sowie die Kadenz.

Mehr Unterstützung bei der Interpretation der Ergebnisse und der weiterführenden Analysen gewünscht.

Es besteht ein Bedürfnis nach mehr Unterstützung bei der Interpretation vorhandener Daten sowie bei weiterführenden Analysen auf Basis der Daten, vor allem mit Bezug auf die durch den Kontextfragebogen erhobenen Kontextvariablen. Einige Kantone versprechen sich mehr Handlungswissen durch die bessere Verknüpfung der ÜGK mit existierenden kantonalen Instrumenten, entweder konzeptionell oder aber in der Durchführung. Die bessere Nutzung von Synergien zur gemeinsamen Arbeit an der Verbesserung von Tests, Fragebögen und Prozessen wird auch von den aktuell in der Gestaltung der ÜGK involvierten Expert*innen als potentiell wertvoll gesehen. So erhofft man sich, beispielsweise die Testdichte zu verringern und die Schulen zu entlasten sowie tiefergehende Analysen zu ermöglichen. Verschiedene Kantone spüren eine gewisse Überlastung von Seiten der Lehrpersonen und finden es teilweise schwierig, Schulen zum Mitmachen zu animieren. Allgemein lässt sich wiederum feststellen, dass Ansichten zu dieser Frage sehr stark mit der jeweiligen kantonalen und regionalen Test- und Dateninfrastruktur zusammenhängen.

2.4 Kommunikation der Ergebnisse der ÜGK und Berichterstattung (A1.4; A1.5)

Kantonale Vertretungen fühlen sich gut über die ÜGK informiert und setzen sich mit den Ergebnissen auseinander.

Über 87 % der kantonalen Vertreter*innen geben an, gut über die ÜGK und die Ergebnisse informiert zu sein. Etwas geringer, aber immer noch im hohen Ausmass, scheinen die kantonalen Vertreter*innen über die Art und Weise der Durchführung der Erhebungen (77 %) über die Verantwortlichkeiten zur strategischen Ausrichtung der ÜGK (74 %) und über die involvierten Institutionen, die den Auftrag haben, die ÜGK durchzuführen (69 %), informiert zu sein.

Fast alle kantonalen Vertreter*innen haben sich gemäss ihren Angaben mit den nationalen Berichten auseinandergesetzt, allerdings mit unterschiedlicher Intensität: Über 80 % geben an, dies (eher) vertieft gemacht zu haben. Die weiterführenden Berichte (z.B. zur Testentwicklung oder zum Kontextfragebogen), wurden von 41 % der kantonalen Vertreter*innen (eher) vertieft analysiert. Über 80 % der Befragten haben sich auch mit zusätzlichen wissenschaftlichen Publikationen, die mit Daten der ÜGK entstanden sind, auseinandergesetzt, circa ein Fünftel (eher) vertieft.

Sowohl von kantonalen Vertreter*innen wie auch von den Personen aus der Forschung und Aus- und Weiterbildung, die angegeben haben, die Berichte genutzt zu haben, sind die allermeisten der Ansicht, dass die Qualität der Berichte gut oder sehr gut sei.

Kommunikation zwischen Wissenschaft und Politik wird teilweise als schwierig beurteilt.

Über Kontakte mit Kosta HarmoS oder ICER wird in den Gesprächen kaum berichtet. Allerdings wird von den Vertreter*innen der Kantone thematisiert, dass die Kommunikation zwischen Wissenschaft und Politik vereinfacht werden müsse; die Wissenschaftler*innen hätten häufig ihre Hypothesen im Kopf, es brauche aber eine Sensibilität für die Praxis, die oftmals zu wenig sichtbar würde. Einzelne kantonale Referenzpersonen berichten in den Gesprächen, dass sie zweimal pro Jahr an einem Treffen in Bern teilnehmen, an dem sie über die wichtigsten Punkte informiert werden. Allerdings wird der Informationsgehalt nicht als sehr gross erlebt, da die Informationen oftmals bereits über die Bildungsdirektor*innen, die direkt über die EDK in die Diskussionen involviert sind, weitergeleitet werden. Zudem wird teilweise der als „zu wissenschaftlich“ wahrgenommene Präsentationsstil kritisiert und das Gefühl angesprochen, es werde an diesen Treffen kaum diskutiert – man werde eher vor vollendete Tatsachen gestellt. In der eigenen Region hingegen bestünden enge und gute Kontakte, so auch mit den Durchführungszentren.

Verstärkung der Kommunikation der Ergebnisse durch die EDK und bessere Abstimmung mit kantonalen Interessen gewünscht.

Knapp die Hälfte der kantonalen Vertreter*innen (44 %) ist gemäss der standardisierten Befragung der Meinung, dass die Kommunikation der Ergebnisse durch die EDK verstärkt werden müsste. 26 % der kantonalen Vertreter*innen wünschen sich zudem eine intensivere Kommunikation durch die kantonalen Behörden.

In den Gesprächen wird teilweise moniert, dass die Kantone der Publikation der Ergebnisse durch die EDK „hinterherhinken“, da sie erst aktiv werden könnten, wenn die Ergebnisse durch die EDK publiziert sind. Die Kantone könnten die Kommunikation der Ergebnisse so nicht gut vorbereiten und seien zwangsläufig in einem „Krisenkommunikationsmodus“.

Unterstützung in der Kommunikation der Ergebnisse gewünscht.

Sowohl verschiedene Kantone als auch Vertreter*innen der Lehrpersonen wünschen sich von den Verantwortlichen der ÜGK (vom ICER oder den regionalen Durchführungszentren) mehr Unterstützung in der Kommunikation der Erhebungsmethodik, der dahinterliegenden Konzepte sowie der Resultate.

Unterstützung brauche man zum Beispiel dabei, um allfällige Unterschiede zwischen der ÜGK und PISA sowie deren Implikationen zu erklären.
Aufgrund der teilweise fehlenden Akzeptanz des Kontextfragebogens wird konkret auch mehr Unterstützung gewünscht, um Politik, Bevölkerung und Lehrpersonen den Zweck dieses Instruments und der darin enthaltenen Items zu erklären. So sollte diesen Gruppen beispielsweise konkret aufgezeigt werden, warum spezifische Fragen gestellt werden oder weshalb die Resultate der ÜGK denjenigen von anderen Erhebungsinstrumenten zu widersprechen scheinen.
Zudem würde es sich aus Sicht einiger Teilnehmender an den Gesprächen lohnen, geeignete Kommunikationsmittel für die Lehrpersonen zu erstellen, damit sich diese besser eingebunden fühlten sowie besser verstehen könnten, was der Nutzen der Tests sei, was sie messen, was Mindesterwartungen an die Schüler*innen seien oder wie die Testaufgaben aussähen. Lehrpersonen hätten gerne individuelle Rückmeldungen. Es brauche eine klare Kommunikation, warum dies nicht möglich sei, damit keine falschen Erwartungen entstünden. Zudem wird angeregt, massgeschneiderte Kurzberichte zu erstellen, die auch von den Lehrpersonen gelesen werden könnten. Vor allem durch einen vertiefteren Einblick in die Testaufgaben könnten Lehrpersonen für die Kompetenzerfassung sensibilisiert werden.

2.5 ÜGK und Bildungsbericht (A2.10)

ÜGK-Daten und -Ergebnisse werden als relevant für den Bildungsbericht betrachtet, allerdings besteht ein zeitliches Passungsproblem, um diese in den aktuellen Bildungsberichten zu berücksichtigen.

Die durch die ÜGK generierten Daten werden als relevant für den Bildungsbericht beschrieben, auch weil die ÜGK die einzige Studie sei, die Kantonsvergleiche möglich macht.
Der Bildungsbericht habe zur Aufgabe, zu ausgewählten bildungspolitischen Fragestellungen eine Antwort zu finden, eine Beschreibung der Schweizer Bildungslandschaft anhand spezifischer Indikatoren zu ermöglichen sowie Erkenntnisse zu gewinnen, welches die Ursachen für bestimmte Phänomene seien oder wie etwas verändert werden könne. Diese Fragen werden anhand existierender Forschung sowie, an zweiter Stelle, anhand zusätzlicher Datenanalysen beantwortet. Die ÜGK könne also direkt und indirekt zum Bildungsbericht beitragen.
Die Daten der ÜGK 2016/2017 konnten bislang vor allem aus zeitlichen Passungsproblemen nicht im Bildungsbericht berücksichtigt werden. 2018 wurde die Publikation des Bildungsberichts zwar verschoben, um eine Integration der ÜGK-Ergebnisse doch noch zu ermöglichen, jedoch standen die ÜGK-Daten und entsprechende publizierte Ergebnisse vor dem Publikationszeitpunkt des Bildungsberichts nicht zur Verfügung. Ein ähnliches Szenario werde sich 2023 wiederholen, weil die ÜGK 2020 wegen der COVID-19-Pandemie verschoben worden sei und man nicht auf ältere ÜGK-Daten zurückgreifen möchte. Daher seien die Herausgeber*innen des Bildungsberichts 2023 auf Studien von Forscher*innen auf Basis der ÜGK-Daten angewiesen, die man dann allenfalls für den Bildungsbericht nutzen könnte.

Qualität der Daten und Analysen zentral, allerdings genüge das Erhebungsdesign nur teilweise den Anforderungen.

Ob die ÜGK-Erhebungen auch relevantes Erklärungswissen generieren können, hänge von der Qualität der Daten und Analysen ab. Die Verantwortlichen äussern in dieser Hinsicht eine gewisse Skepsis. Diese hänge vor allem mit dem Design der ÜGK-Erhebung zusammen. Konkret genannt werden drei Probleme: Erstens, dass die ÜGK Grundkompetenzen statt des gesamten Kompetenzspektrums erfasse. Dies mache statistische Analysen schwierig und verhindere, Erklärungswissen zu generieren. Zweitens erschwere die Anlage der ÜGK als Querschnittsuntersuchung und das Fehlen eines zuverlässigen Erhebungsrhythmus‘ die Identifikation des Effekts der Schule und des Bildungssystems auf die Entwick-

lung der Schüler*innen. Drittens verschleiere der starke Fokus auf Schüler*innen durch den Kontextfragebogen die Sicht auf andere, gemäss der Forschung relevante potentielle Erklärungsfaktoren wie zum Beispiel Lehrpersonen, Unterrichtsprozesse oder Interaktionen zwischen den Schüler*innen.

3 Zielvorgaben der ÜGK

Die nationalen Bildungsziele der EDK wurden vor zehn Jahren für die Schulsprache, die zweite Landessprache beziehungsweise Englisch als Fremdsprache sowie für Mathematik und Naturwissenschaften verabschiedet. Sie beziehen sich auf das Ende des 4., 8. und 11. Jahres der obligatorischen Schule. Genau hier müssen die ÜGK-Erhebungen ansetzen. Ebenfalls festgelegt wurde bereits im HarmoS-Projekt, dass die Bildungsziele als Grundkompetenzen (=Mindeststandards) benannt und operationalisiert werden sollen. Dementsprechend müssen alle bisherigen und in Entwicklung befindlichen ÜGK-Tests den jeweiligen Schwellenwert fokussieren, bei dem Grundkompetenzen als „vorhanden“ bezeichnet werden können.
Die Gutachtengruppe wurde jedoch von der EDK gebeten, zu prüfen, wie sinnvoll es sei, am Testen von Grundkompetenzen im Sinne der Überprüfung des Erreichens von Mindeststandards festzuhalten. Auch die Art der Beschreibung dieser Grundkompetenzen in den Bildungsstandards sowie die Festlegung auf bestimmte Fachbereiche und Jahrgänge stehen auf dem Prüfstand.
Die anstehende Gesamtschau zur ÜGK soll und muss also gewissermassen noch einmal die Prinzipien des HarmoS-Projekts, des daraus entstandenen Konkordats und die Grundlagen ihrer Umsetzung in der ÜGK prüfen. Das vorliegende Gutachten bereitet eine solche Evaluierung der Grundlagen und der Ausgestaltung der Erhebungen vor. Die Antworten und Empfehlungen stützen sich nicht zuletzt auf die Grundposition der Gutachtengruppe, dass die ÜGK eindeutig und in erster Linie als Instrument des Bildungsmonitorings und somit der bildungspolitischen Steuerung anzusehen ist, und erst nachrangig als Datenquelle für die Forschung.

3.1 Auswahl der Fächer und Jahrgangsstufen (B1.1; B1.2)

Auswahl der Fächer und Jahrgangsstufen findet Akzeptanz.

In der standardisierten Befragung wurde gefragt: „Welche der nachfolgenden Fachbereiche sollten aus Ihrer Sicht im Rahmen der ÜGK überprüft werden?“ Deutlich wurde, dass die Schulsprache und Mathematik grösste Akzeptanz finden (vgl. Tabelle 2): 90% der kantonalen Vertreter*innen sowie etwa 70% der Personen aus Forschung und Aus- und Weiterbildung hielten es für notwendig, diese beiden Fächer zu überprüfen. Etwas niedrigere Werte, aber immer noch mit einer Zustimmung von einer Mehrheit der Befragten, ergaben sich für die zwei weiteren in der ÜGK vorgesehenen Fächergruppen Fremdsprachen und Naturwissenschaften. Interessant ist, dass den Vertreter*innen der Kantone die Fremdsprachen etwas wichtiger sind, den Personen aus Forschung und Aus- und Weiterbildung hingegen die Naturwissenschaften. Nur knapp ein Drittel der Befragten votierte für die Ergänzung des ÜGK-Programms durch ein musisches Fach oder Sport; dies waren überwiegend Personen, die in der Aus- und Weiterbildung von Lehrpersonen tätig sind.

Dieses Bild wird durch die Gespräche gestützt. Zusätzliche Fächer spielten dort mit einer Ausnahme (eine Person aus der Schulpraxis wollte die ÜGK besser von PISA abgrenzen) keine Rolle, während die Naturwissenschaften mitunter kritisch gesehen wurden – weil es mit entsprechenden Tests in der Schweiz wenig Erfahrung gebe und weil es als schwierig angesehen wird, eine Konzeption zu entwickeln, die sowohl die deutschsprachige als auch die französisch- und italienischsprachige Tradition und die naturwissenschaftlichen Einzeldisziplinen angemessen berücksichtige[7]. Eine Person votierte expli-

[7] Die Sonderstellung der Naturwissenschaften ist schon in der Formulierung von Grundkompetenzen bei HarmoS angelegt: In der Schulsprache, Fremdsprachen und Mathematik sind die Grundkompetenz-Standards nach Kompetenzbereichen gegliedert, die in Mathematik „Themenbereiche“ heissen und tatsächlich auf einer Systematik von Fachinhalten beruhen, in den Sprachen auf der Systematik des Gemeinsamen Europäischen Referenzrahmens für Sprachen (GER). Diese Kompetenz- und Themenbereiche können prinzipiell gut in Aufgabenpakete und Tests übersetzt werden. Die Grundkompetenzen Naturwissenschaften hingegen sind nach relativ abstrakten „Handlungsaspekten“ gegliedert, weil, wie es einleitend heisst, die Festsetzung fachlicher Themen in den Kerncurricula der

zit für eine Beschränkung auf Mathematik und Schulsprache, alles andere sei „Verzettelung“. Viele Befragte betonten, dass es auf die Systematik der durchgängigen Fächerwahl über alle Kohorten hinweg im Rahmen eines langfristig geplanten Erhebungsprogramms mit angemessenem Rhythmus ankomme.

Fach / Fachgruppen	Kantonale Vertreter*innen (n = 39)	Personen aus Forschung und Aus- und Weiterbildung (n = 77)
Schulsprache	90%	73%
Mathematik	90%	71%
Naturwissenschaften	59%	64%
Fremdsprachen	74%	52%
Musik	21%	20%
Bewegung und Sport	23%	17%
Gestalten	13%	18%

Tabelle 2: Zustimmungsrate zur Überprüfung verschiedener Fächer im Rahmen der ÜGK; standardisierte Befragung;
* Keine Angaben: je 24%

Auch die von der EDK vorgenommene Auswahl der Jahrgangsstufen findet Zustimmung, allerdings mit einigen Abstrichen für die jüngeren Jahrgänge. Als „sehr sinnvoll“ bewerteten 24 % der Teilnehmenden an der standardisierten Befragung die Erhebung Ende des 4. Schuljahres nach HarmoS-Zählung (also bei circa 8-Jährigen), 41 % die Erhebung Ende des 8. Schuljahres (circa 12-Jährige) und 48 % die Erhebung Ende des 11. Schuljahres (circa 15-Jährige). Nimmt man diejenigen hinzu, welche die betreffende Erhebung als „eher sinnvoll“ akzeptieren, wächst die Zustimmung auf 50 % beziehungsweise 64 % und 67 %. Die Skeptiker*innen finden sich eher unter Personen aus Forschung und Aus- und Weiterbildung als unter kantonalen Vertreter*innen. Alternative Zeitpunkte (z.B. rund um den Schuleintritt, 6. statt 4. Schuljahr, 7. statt 8., 10. statt 11.) wurden sehr selten genannt; sie würden auch eher dann Sinn machen, wenn der Zweck der ÜGK grundlegend geändert würde, zum Beispiel in eine Diagnose individueller Lernrückstände rechtzeitig vor dem Abschluss einer Schulstufe, was aber nicht der aktuellen Zwecksetzung entspricht.
In den Gesprächen gab es ebenfalls Zweifel, ob eine Testerhebung bei 8-Jährigen realisierbar sei. Dazu wird die ÜGK 2024 mit dem Schwerpunkt 4. Schuljahr wertvolle Erfahrungen liefern; erste ermutigende Erkenntnisse erbrachte 2019 eine Machbarkeitsstudie (Angelone et al., 2019). In den beiden höheren Jahrgängen wurde vor allem aus der Schulpraxis auf das Problem hingewiesen, dass in denselben Schüler*innengruppen weitere Erhebungen durchgeführt werden, etwa in der 8. Jahrgangsstufe in den französischsprachigen Kantonen Orientierungstests und in der 11. Jahrgangsstufe in einzelnen deutschsprachigen Kantonen Stellwerk-Tests; hinzu kommt gegebenenfalls PISA. Über eine langfristige Planung und eine abgestimmte Stichprobenziehung, so die häufig formulierte Forderung, sollten Überlastungen für einzelne Gruppen vermieden werden (vgl. dazu Kapitel 4).

Sprachregionen erfolgt. Hinzu kommt, dass die sogenannte HarmoS-Validierungsstudie in den Naturwissenschaften u.a. mit aufwändigen Aufgabenformaten wie Schüler*innenexperimenten gearbeitet hat. Ein solches „Performance Assessment“ bietet Vorteile hinsichtlich Aussagekraft, Tiefe der Testanforderungen, Unterrichtsnähe und Testmotivierung, verlangt aber einen höheren Aufwand an Entwicklungsarbeit und mehr Zeit beim Testen und verkompliziert die Logistik.

3.2 Mindeststandards oder differenzierte Leistungsstufen? (B1.4; B3.13; B3.15)

Die Entwicklung eines validen Testsystems setzt im Allgemeinen voraus, dass das Konstrukt (z. B. Grundkompetenz in Mathematik oder Schulsprache auf einer bestimmten Schulstufe) in dreifacher Hinsicht spezifiziert ist:

a. Die inhaltlichen Standards sind klar definiert – wenn möglich durch die Benennung kognitiver Operationen/Leistungsmerkmale („Can-do“-Statements) – und beschrieben.
b. Die im Test verwendeten, von Expert*innen systematisch entwickelten Aufgaben lassen sich mit Anforderungsmerkmalen beschreiben, die hoch mit den Beschreibungen der inhaltlichen Standards korrespondieren.
c. Niveau-Deskriptoren („Proficiency Level Descriptors“) ermöglichen eine präzise Beschreibung von Kompetenzstufen („Proficiency Levels“), wobei die beiden zuvor genannten Aspekte (Inhaltsstandards und Testaufgabenmerkmale) kombiniert werden (vgl. Tiffin-Richards & Pant, 2017).

Im Fall der ÜGK liefern die 2011 verabschiedeten HarmoS-Dokumente Beschreibungen der inhaltlichen Standards (Komponente a), während Fachdidaktiker*innen in Zusammenarbeit mit der Aufgabendatenbank der EDK (ADB) Testaufgaben entwickelten und Beispielaufgaben vorstellten, deren Merkmale in den Berichten zur ÜGK 2016 und 2017 beschrieben wurden (Komponente b). Differenzierte Niveau-Deskriptoren (Komponente c) liegen jedoch für die Schweiz nicht vor.

Die Begrenzung der Bildungsstandards auf eine Niveaustufe (Grundkompetenz, Mindeststandard) ist weitgehend akzeptiert.

Alle schweizerischen Bildungsstandards beschränken sich auf die Beschreibung einer einzigen Niveaustufe, die „Grundkompetenz“ genannt wird und den „Mindeststandard“ darstellt, dessen Erreichen in der ÜGK erfasst werden soll. Die Grundkompetenzen nach HarmoS sind im Prinzip normativ festgelegte Niveaustufen innerhalb komplexerer Kompetenzmodelle (vgl. Ramseier, Moser, Moreau & Antonietti 2008, S. 14ff.). Die vollständigen Kompetenzmodelle werden jedoch nicht expliziert, und deshalb gibt es keine Deskriptoren, die verschiedenen Kompetenzstufen gegeneinander abgrenzen.
In der standardisierten Befragung bezeichnete nur eine sehr kleine Minderheit (6 von 141 Befragten) den Ansatz, ausschliesslich Mindeststandards zu beschreiben, als „überhaupt nicht sinnvoll“; zwei Drittel jener, die die Frage beantwortet haben (circa 30% haben keine Meinung geäussert), finden dies aber „eher sinnvoll” oder „sehr sinnvoll” (Abbildung 2). Mit etwa einem Drittel Antworten in der Kategorie „eher nicht sinnvoll“ gibt es allerdings ein gewisses Zweifel-Potenzial. Dennoch forderte in den Gesprächen niemand einen Neustart von HarmoS; die Bildungsstandards von 2011 werden weitgehend als Grundlage der ÜGK akzeptiert.

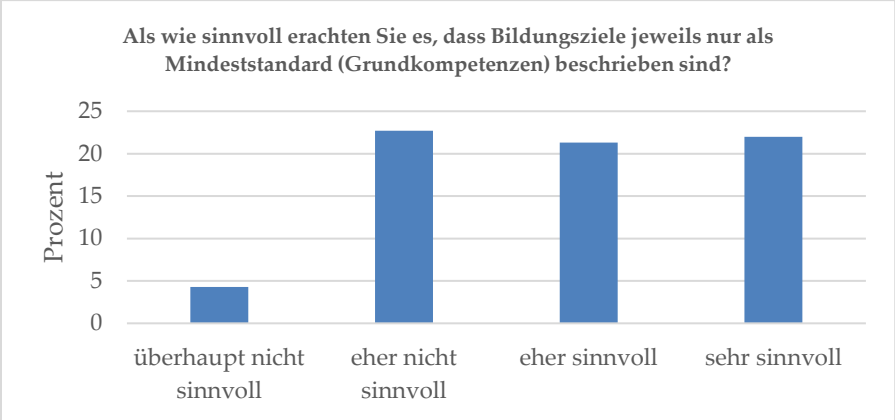


Abbildung 2: Ergebnisse standardisierte Befragung zur Fokussierung auf Mindeststandards; Prozentangaben; Keine Angaben = 29.8%

Differenzierte, das gesamte Leistungsspektrum umfassende Stufenmodelle sind umstritten.

Selbst wenn die Bildungsziele als Mindeststandards beschrieben sind, wäre es prinzipiell denkbar, bei der *Testung* im Rahmen der ÜGK ein breiteres Leistungsspektrum abzudecken und differenzierte Stufenmodelle zu formulieren, mit denen die Getesteten in hierarchische Kategorien eingeordnet werden können (z.B. „unter Mindeststandard“, „Mindeststandard“, „Regelstandard“, „über Regelstandard“ und „Optimalstandard“; vgl. Fuchs 2019, S. 7). Dies ist die Vorgehensweise in internationalen Vergleichsstudien sowie bei der Überprüfung von Bildungsstandards in Österreich und Deutschland. Dieser Ansatz wurde aber auch noch in der allerersten empirischen Studie zur Umsetzung der HarmoS-Kompetenzmodelle, der sogenannten Validierungsstudie, verwendet: Hier wurde in der Regel mit vier Stufen gearbeitet, die allerdings nicht mit einem formalen „Standard-Setting“-Verfahren abgegrenzt wurden, sodass „Methodik und Transparenz der Festlegungen insgesamt nicht befriedigend“ waren (Ramseier et al., 2008, S. 15). In der ÜGK hat man sich dann darauf beschränkt, „Grundkompetenz“ als eine einzige Stufe zu beschreiben.

Um zu klären, wie relevant die Erfassung des gesamten Leistungsspektrums ist, wurden in der standardisierten Befragung zwei Aussagen vorgelegt, welche die Teilnehmenden jeweils auf einer vierstufigen Antwortskala von „stimme überhaupt nicht zu“ bis „stimme voll und ganz zu“ bewerten sollten. Die Abbildungen 3 und 4 stellen die Ergebnisse dar, wobei circa ein Drittel der Befragten die Aussagen nicht bewertet hat.

- Bei Aussage (A) „Anhand der Tests soll einzig überprüft werden, ob die Grundkompetenzen erreicht worden sind“ teilen sich die Antworten in zwei praktisch gleich grosse Gruppen: 33 % antworten positiv, 35 % negativ.
- Der Aussage (B) „Die Tests sollen das gesamte Leistungsspektrum mit mehreren Kompetenzstufen abdecken (wie zum Beispiel bei PISA)“ stimmen 45 % der Befragten (eher) zu, insbesondere 51 % der kantonalen Vertreter*innen und 55 % der Personen aus Forschung und Aus- und Weiterbildung.

Da sich die beiden Aussagen widersprechen (jedenfalls, wenn man das HarmoS-Konzept von Grundkompetenzen voraussetzt), ist es nicht verwunderlich, dass die Antworten stark negativ korreliert sind (r=-.86). Wer Aussage A zustimmt, hat also in der Regel Aussage B abgelehnt, und umgekehrt. Die Befragten lassen sich daher zumeist als Befürwortende einer reinen Grundkompetenz-Messung oder als Befürwortende eines Stufenmodells einordnen. Das Stufenmodell hat mehr Befürwortende; die Präferenz ist aber unter Personen aus Forschung und Aus- und Weiterbildung (32:21) stärker ausgeprägt als unter kantonalen Vertreter*innen (17:13). Zugleich gibt es, vor allem unter den Personen aus Forschung und Aus- und Weiterbildung, viele mit fehlendem oder nicht eindeutigem Urteil. Befürwortet wird das Stufenmodell vor allem von jenen, die der ÜGK primär die Funktion „Systemevaluation“ zuordnen, und von jenen, die ÜGK-Berichte und nachfolgende Forschungsergebnisse selbst nutzen.

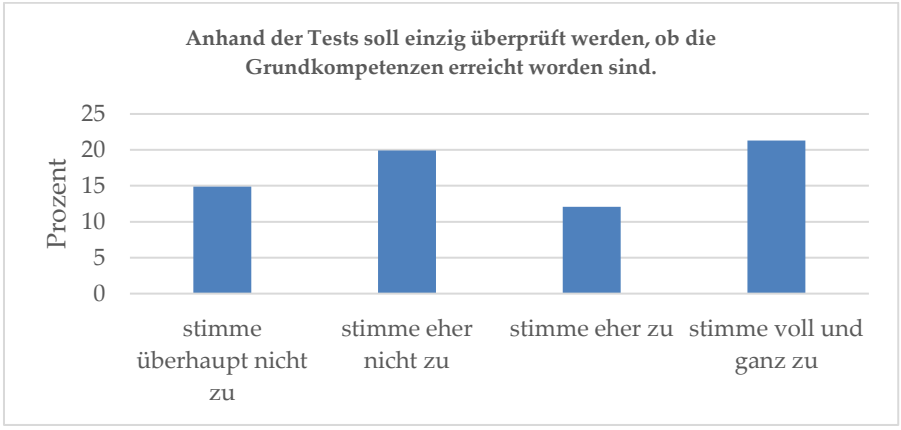


Abbildung 3: Ergebnisse standardisierte Befragung zur Frage A, ob einzig das Erreichen der Grundkompetenzen geprüft werden soll; Prozentangaben; Keine Angaben = 31.9%

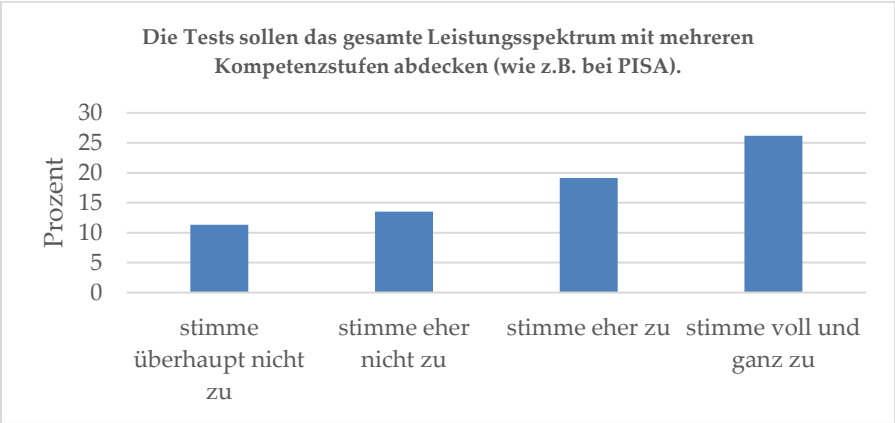


Abbildung 4: Ergebnisse standardisierte Befragung zur Frage B, Abdeckung des vollständigen Spektrums an Kompetenzstufen; Prozentangaben; Keine Angaben = 29.8%

Das Ergebnis der standardisierten Befragung ergibt somit tendenziell ein Votum für eine andersgeartete, das vollständige Spektrum umfassende, abgestufte Testskala. Auch mehrere Gesprächsteilnehmenden äusserten den Wunsch, mit der „Ressourcenverschwendung für einen 0/1-Indikator“ (Zitat aus einem Gespräch) aufzuhören und das ganze Leistungsspektrum zu testen. Diese Gesprächsteilnehmenden haben sich teils eher für eine Differenzierung unterhalb des Schwellenwerts interessiert (mit dem Anliegen, Förderanregungen zu erhalten), teils eher für Differenzierung oberhalb (mit dem Anliegen, auch etwas über leistungsstarke Schüler*innen zu erfahren). Sie erwarteten jedenfalls von differenzierten Stufenmodellen bessere Chancen, über kantonsbezogene Prozentangaben (Anteil derjenigen, die Grundkompetenzen erreicht haben) hinaus Analysen durchführen und somit Erklärungswissen aufbauen zu können. Dies führte auch zur vergleichsweise starken Befürwortung eines Stufenmodells unter Personen aus Forschung und Aus- und Weiterbildung. Einige von ihnen formulierten die Ansicht, die 0/1-Variable „Grundkompetenz nicht erreicht“/“erreicht“ sei für weiterführende Analysen und den Aufbau von Erklärungswissen ungeeignet.

In den Gesprächen wurden andererseits – vor allem von Vertreter*innen der Kantone und von Fachdidaktiker*innen – starke Argumente für die Ursprungsidee der ÜGK, also die Fokussierung auf Grundkompetenzen, vorgetragen. Eines der Argumente war hier, dass für differenzierte Messungen und Rückmeldungen in den Sprachregionen andere Instrumente verfügbar seien. Des Weiteren wurde argumentiert, dass durch den Fokus auf Mindeststandards vor allem die Leistungsschwächeren in den Blick kämen, um die sich die Schule beziehungsweise Bildungspolitik besonders kümmern müsse. Einige der Befragten waren zudem skeptisch, dass der Übergang zu einer vollständigen Skala über die Sprachgrenzen hinweg konsistent gelingen könne, weil es jenseits der Grundkompetenzen, auf die sich zu einigen schon schwer genug war, je nach Sprachregion und damit verbundener didaktischer Tradition unterschiedliche Kompetenzkonzepte und Erwartungshorizonte gebe. Schliesslich wurde darauf hingewiesen, ein Bruch mit der bisherigen Strategie impliziere einen vollständigen Neuanfang, der schon bei den 2011 veröffentlichten Beschreibungen der Grundkompetenzen ansetzen müsste. Jene Darstellungen zielten auf ein einziges Kompetenzniveau, das in den Bildungsstandards nach Themen- und Kompetenzbereichen sowie Handlungsarten aufgeschlüsselt und in den Aufgaben umgesetzt werde. Alle diese Schritte müssten erneut, ohne den Fokus auf Mindeststandards, durchlaufen werden. Ein solches Vorgehen würde grosse Kosten verursachen, sowohl die Messreihe der ÜGK als auch den politischen Prozess der Harmonisierung gleichsam auf „Start“ zurücksetzen.

Auch ohne Stufenmodelle sind Studien möglich, die über die erreicht/nicht erreicht-Klassifizierung hinausgehen.

Die nationalen (Erst-) Berichte zur ÜGK 2016 und 2017 behandeln konsequent nur das Erreichen oder Verfehlen des festgelegten Schwellenwerts, setzen dieses dichotome Ergebnis in Bezug zu anderen Variablen (z.B. Geschlecht und Migrationshintergrund) und vergleichen die Anteile der Schüler*innen, deren Testscores den Schwellenwert übersteigen, zwischen den Kantonen.
Das bedeutet aber nicht, dass Sekundäranalysen – beispielsweise für die Bildungsberichterstattung – sich ebenfalls auf eine dichotome Metrik beschränken müssen. Vielmehr liegen für alle Getesteten die Scores (= geschätzte Personen-Fähigkeiten) auf der kontinuierlichen Metrik des Rasch-Modells vor, in der Masseinheit „Logit“. In den Technischen Berichten werden die empirisch beobachteten Verteilungen als Balkendiagramme dargestellt (vgl. Abbildungen 5 und 6). Man sieht, dass insbesondere der Mathematik-Test der ÜGK 2016 und der Lesetest für Deutsch als Fremdsprache 2017 Testscore-Verteilungen ergaben, welche der Gauss’schen Normalverteilung sehr nahekommen, also für statistische Analysen bestens geeignet sind. Andere Tests aus der ÜGK 2017 haben eine leicht schiefe Verteilung, aber Deckeneffekte liegen auch hier nicht vor, sodass ausreichend Varianz abgedeckt wird. Zudem passt, wie in den Technischen Berichten dargestellt, das Rasch-Modell gut zu den beobachteten Daten. Es handelt sich also bei den ÜGK-Tests um Indikatoren für fachliche Lernergebnisse mit guten Messeigenschaften, die auch hohe Leistungsniveaus ausdifferenzieren und sehr gut in Sekundäranalysen ausgewertet werden können.
Vorbehalte gegen eine Verwendung der Testwerte in der Bildungsforschung können also höchstens mit Mängeln der Validität begründet werden. Tatsächlich präsentieren aber weder die HarmoS-„Validierungsstudie“ noch die Dokumente zu ÜGK 2016 und 2017 Daten zur Validität der Tests, abgesehen von der Modellanpassung und von Zusammenhangsmustern mit Schüler*innenmerkmalen. Es ist überfällig, empirische Daten zum Zusammenhang zwischen ÜGK-Scores einerseits, anderen Leistungsindikatoren (z.B. kantonale Tests, kommerzielle Leistungstests, PISA-Tests, Schulnoten) andererseits vorzulegen. Die EDK sollte entsprechende Untersuchungen – beispielsweise im Kontext von Pilotierungen – in Auftrag geben. Diese Untersuchungen würden dazu dienen, belastbare Hinweise auf die Validität der Testskalen zu erbringen. Zusätzliche Validitätsargumente könnten repräsentative (!) Befragungen von Lehrkräften über die Passung zu ihrem Unterricht erbringen sowie eine längsschnittliche Verknüpfung der ÜGK-Erhebungen (vgl. dazu Kapitel 4). Ein Wechsel zu neuen, breiter ausdifferenzierten Testskalen, ohne dass das bisherige Modell in solchen Studien genutzt und überprüft wurde, wäre aus Sicht der Gutachtengruppe vorschnell.
Somit ist das Argument, vertiefende Analysen (z.B. für den Bildungsbericht) verlangten eine andere Testskala, nicht stichhaltig. Zum einen stehen auch in der jetzigen Testlogik differenzierende, kontinuierliche Messwerte zur Verfügung. Zum anderen muss hervorgehoben werden, dass der Hauptzweck der ÜGK nicht solche erweiterten Analysen sind, sondern das Bildungsmonitoring bezüglich der Erreichung von fachbezogenen Grundkompetenzen. Anders als bei PISA handelt es sich bei der ÜKG nicht um allgemeine Fähigkeitsmessungen (ability, literacy), sondern um Messungen des in der Schule erworbenen Wissens und Könnens (achievement). Da sich die ÜGK unmittelbar auf nationale Bildungsstandards bezieht, kann von einer hohen curricularen Validität (Passung zu den vorgegebenen Lehrplänen und den im Unterricht behandelten Inhalten) ausgegangen werden, die für solche Tests von zentraler Bedeutung ist. Mehr noch: Gerade weil ÜGK-Tests verbindliche Grundkompetenzen erfassen, sollten sie über alle Regionen der Schweiz hinweg maximal curricular valide sein. Hätte man – wie es einige Kritiker*innen fordern – das gesamte Spektrum möglicher Kompetenzstufen aufgenommen, wären diese Aufgaben nur noch Teilen der Schüler*innenschaft zugänglich und die curriculare Validität ginge zurück.

Die Schwierigkeit der Aufgaben variiert schon jetzt breit, auch über das Niveau der Grundkompetenz hinaus.

Dass die Testscores eine ausreichende Varianz haben, liegt wiederum daran, dass die Schwierigkeiten der Aufgaben innerhalb der bislang eingesetzten ÜGK-Tests recht breit streuen. Dies ist aus den „Punktwolken“ der Abbildungen 5 und 6 ersichtlich, wo jeder Punkt eine der eingesetzten Testaufgaben darstellt, deren empirisch ermittelte Schwierigkeit rechts auf der sogenannten Logit-Skala ablesbar ist (Logit ist die Masseinheit der Schwierigkeit im verwendeten psychometrischen Ansatz, dem Rasch-Modell). Die Abbildungen – den Technischen Berichten zur ÜGK 2016 und 2017 entnommen – belegen, dass die Schwierigkeitsparameter in allen bisherigen Tests innerhalb eines beachtlichen Intervalls von circa vier Logits variieren.

Die Streubreite der tatsächlichen Aufgabenschwierigkeiten lässt sich nicht vermeiden, denn solche Schwierigkeiten „sind konstruktionsbedingt variabel. Man kann ‚denselben‘ Aufgabeninhalt so in Form einer Testaufgabe bringen, dass diese eher schwer oder eher leicht zu lösen ist“ (Rost, 2004, S. 663) [8]. Die HarmoS-Dokumente (Bildungsstandards) aus dem Jahr 2011 können nur die Art und den Inhalt der Anforderungen benennen, die zu den Grundkompetenzen zählen. Die Standards müssen in der Aufgabenentwicklung immer wieder re-interpretiert, ausdifferenziert und in konkretere Anforderungen übersetzt werden. Der exakte Schwierigkeitsgrad einer Aufgabe hängt von vielen Aufgabenmerkmalen und komplexen, nur teilweise verstandenen Antwortprozessen ab. Jene Befragten, die mit der Aufgabenentwicklung betraut waren, sind sich der Kluft zwischen Kompetenzformulierungen der Bildungsstandards einerseits und ÜGK-Aufgaben andererseits sehr bewusst. Insbesondere die Bildungsstandards in Mathematik wurden in mehreren Gesprächen als recht vage und abstrakt bezeichnet. Dies hat vermutlich zu den Problemen bei der ÜGK 2016 beigetragen, die in einem Audit von Fischbach und Ugen (2018) diskutiert wurden.

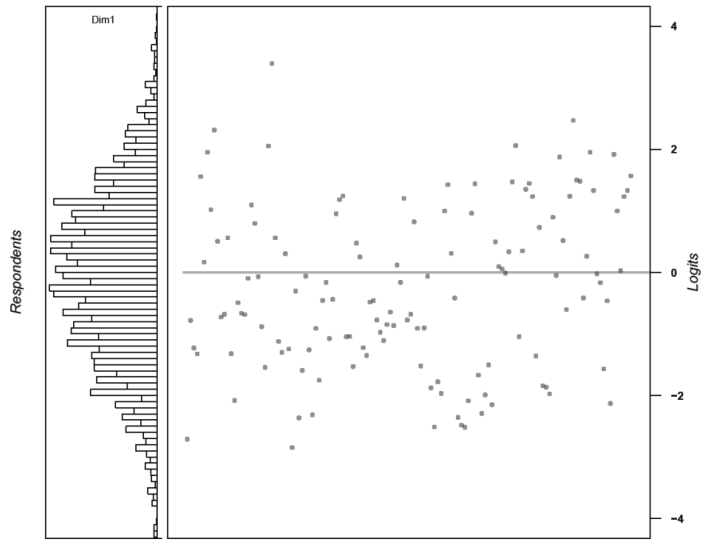
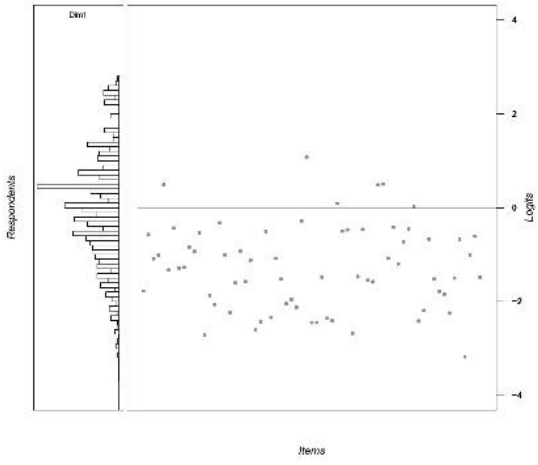


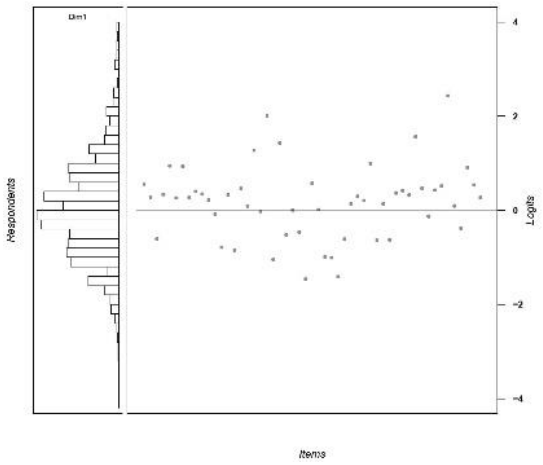
Abbildung 5: Verteilung der Personenparameter (Balkendiagramm, links) und Schwierigkeitsparameter aller Items (Punktwolke, rechts) aus der ÜGK 2016 zur Mathematik im 11. Schuljahr. Schwierigkeit eines Items wie in PISA berechnet als derjenige Score, bei welchem Probanden eine 62-prozentige Wahrscheinlichkeit haben, das entsprechende Item korrekt zu lösen (pr62). (Quelle: Angelone & Keller 2019a, S. 10)

[8] So auch das Audit von Fischbach und Ugen (2018, S. 21) zur ÜGK 2016: „ The auditors would like to emphasize that many theoretical descriptors allow for the construction of test items of very different—not to say diametrically opposed—mathematical demands and difficulties, as the empirical difficulties of the ÜGK/COFO items unmistakably testify. … The auditors' conclusion regarding the descriptors' room for interpretation is, again, not a Swiss specificity but a more general issue in the definition and evaluation of mathematics competence."

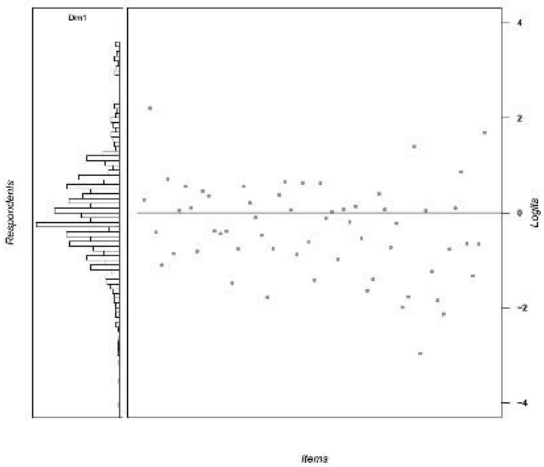
Schulsprache Lesen



Fremdsprache Deutsch Leseverstehen



Fremdsprache Deutsch Hörverstehen



Fremdsprache Englisch Hörverstehen

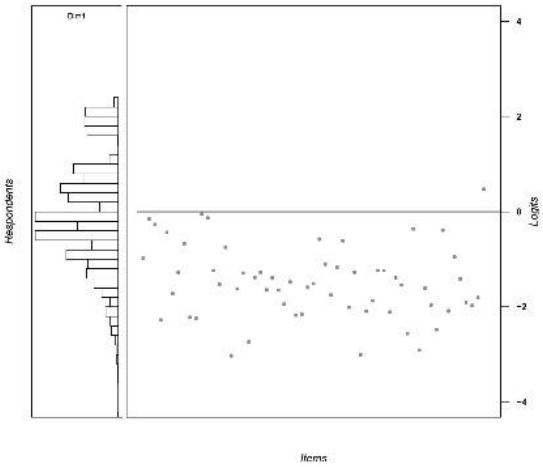


Abbildung 6: Verteilung der Personenparameter (Balkendiagramm, links) und Schwierigkeitsparameter aller Items (Punktwolke der pr62-Werte, rechts) für vier Tests aus der ÜGK 2017 zu Sprachen im 8. Schuljahr (Quelle: Angelone & Keller, 2019b, S. 17)

Wegen der unvermeidbar variierenden Schwierigkeiten muss man in der ÜGK (wie auch in allen anderen Testsystemen, die standardbasiert Bildungsmonitoring betreiben) zu dem Hilfsmittel greifen, gezielt eine gewisse Bandbreite von Aufgaben zu entwickeln, diese auf einer Testskala abzubilden und dann anhand der konkret vorliegenden Aufgaben über ein „Standard Setting“-Verfahren zu entscheiden, wo genau das angestrebte Niveau liegen soll. Dementsprechend geht der Messansatz, der der ÜGK zugrunde liegt, von einem *Kontinuum* aus, auf dem Aufgaben einerseits und Fähigkeiten der Getesteten andererseits verankert werden können. Das Rasch-Modell dient dazu, dieses Kontinuum darzustellen. Im Rahmen der Schwellenwert-Setzung („Bookmark-Verfahren“) werden die Aufgaben nach ihrer Position auf der kontinuierlichen Skala, d.h. nach ihrer Schwierigkeit, angeordnet, sodass Expert*innen eine Trennlinie ziehen können zwischen Aufgaben, die von „gerade eben grundkompetenten“ Schüler*innen noch sicher gelöst werden, und solchen, deren Anforderungen mehr oder weniger über Grundkompetenz hinausgehen[9]. Dieses Vorgehen *setzt also voraus*, dass die Tests etliche Aufgaben enthalten, die über das Anforderungsniveau der Grundkompetenz hinausgehen. Die Einbeziehung schwierigerer Aufgaben ist also keinesfalls als „Fehler“ in der Testentwicklung anzusehen!

[9]Im Technischen Bericht zur ÜGK Mathematik 11 beschreiben Angelone und Keller (2019a, S. 14) die Schwellenwertsetzung folgendermassen: „Anschliessend mussten die Experten in sprachregional zusammengesetzten Gruppen auf der Grundlage eines nach Itemschwierigkeiten gereihten Itembuchs an der Stelle ein Lesezeichen (Bookmark) setzen, an der sie die Schwelle zwischen mathema-

Allerdings liefern die Bildungsstandards selbst für solche anspruchsvolleren Aufgaben keine explizite Grundlage, denn sie beschreiben nur das Niveau der Grundkompetenz. Die fachdidaktischen Teams, die die Aufgaben entwickelt haben, sind mit dieser Herausforderung unterschiedlich umgegangen, was in den Gesprächen dargelegt wurde, aber in den offiziellen Berichten nur schwach angedeutet wird.

- Das Fremdsprachen-Team verweist darauf, dass dies nur möglich gewesen sei, indem man neben den Schweizer Bildungsstandards „die Systematik des Gemeinsamen Europäischen Referenzrahmens für das Sprachenlernen“ (GER) herangezogen habe. So habe man neben dem im EDK-Dokument beschriebenen GER-Niveau A1.2 auch Aufgaben einsetzen können, die A2 abbilden. „Der Einbezug eines Niveaus, das über den Grundkompetenzen liegt, war dadurch möglich, dass sich die Grundkompetenzen auf eine weiterführende Skala, die GER-Skala, beziehen. Der Test sollte insbesondere im kritischen Skalenbereich an der Nahtstelle zwischen den Niveaubändern A1.2 und A2.1 genügend Aufgaben aufweisen, um eine zuverlässige Grenzziehung zu ermöglichen“ (Lenz et al. 2019, S. 37f.).
- Das Mathematik-Team (Linneweber-Lammerskotten et al., 2019, S. 18) verweist darauf, dass den Bildungsstandards dieses Faches ein mehrdimensionales Kompetenzmodell mit mehreren Niveaustufen zugrunde läge, das zum Beispiel Erfahrungen aus PISA aufgreife, auch wenn dieses Modell weder in den Schweizer Bildungsstandards noch im Bericht zur ÜGK 2016 expliziert wird. Im Gespräch schlägt eine Vertretung dieser Fachgruppe vor, zukünftig für jeden ÜGK-Test ein „Framework“ zu veröffentlichen, welches die Bildungsstandards konkretisiert, neue fachdidaktische Entwicklungen aufgreift, und dann auch Beispielaufgaben sowie Aufgabenmerkmale benennt, die Anforderungen unterhalb und oberhalb der Schwelle unterscheiden.
- Das Team Schulsprache hat sich am ehesten darauf beschränkt, Aufgaben im unteren Schwierigkeitsbereich, um den möglichen Schwellenwert herum, zu entwickeln – durchaus erfolgreich, wie man in Abbildung 6 sieht. Dazu hätten, so eine interviewte Person, Hinweise von Psychometrikern beigetragen. In Vorbereitung der ÜGK 2024 sei diese Fokussierung noch besser möglich, weil man tatsächliche Schwierigkeiten und kognitive Anforderungen in den „Kognitive Labs“ prüfen könne.

Diese verschiedenen Strategien führten zu sehr unterschiedlichen Verteilungen der Aufgaben auf die Bereiche unterhalb und oberhalb der jeweils gesetzten Schwelle. Im Lesetest Schulsprache war das Verhältnis ausgewogen: 35 Aufgaben lagen darunter, 32 darüber[10]. In der ÜGK 2016 Mathematik aber gab es 45 Aufgaben unterhalb und 87 oberhalb der Schwelle, und bei den Fremdsprachen lagen beim Leseverstehen nur 11 bis 13 (je nach Sprache) von 53 Aufgaben unterhalb der Schwelle, im Hörverstehen 13 bis 21 von 59 Aufgaben[11]. Diese geringe Zahl ist tatsächlich problematisch, weil allein die Aufgaben, die unterhalb der Schwelle liegen und somit von „minimal grundkompetenten“ Schüler*innen mit mindestens 62 % Wahrscheinlichkeit gelöst werden können, das ausmachen, was empirisch mit „Grundkompetenz“ gemeint ist. Die Aufgaben oberhalb der Schwelle dienen hingegen zur Abgrenzung, denn sie stellen irgendwelche Anforderungen, die über das hinausgehen, was empirisch als „grundkompetent“ bezeichnet wird. Es geht bei der ÜGK eben *nicht* um ein „Mastery Learning“ wie in den 1960er und 1970er Jahren, wo man ein Aufgabenuniversum als Zielbereich festlegte und dafür die Erfolgswahrscheinlichkeit prüfte. Es geht vielmehr um die Abgrenzung zwischen zwei Stufen auf einem Kontinuum: „Grundkompetenz erreicht“ (Personen-Score liegt oberhalb des Schwellenwertes, sodass alle Aufgaben unterhalb der Schwelle beherrscht werden) versus „Grundkompetenz nicht erreicht“ (Personen-Score liegt unterhalb des Schwellenwertes).

tischen Grundkompetenzen und höheren Kompetenzen vermuteten.“ Diese Formulierung erscheint der Gutachtengruppe klärungsbedürftig, da nicht sicher ist, dass die Expert*innen des Standardsetting-Panels tatsächlich instruiert wurden, sich eine „*minimal* grundkompetente” Person vorzustellen.

[10] Dies lässt sich aus den im Anhang im Technischen Bericht (Angelone & Keller, 2019b) angegebenen Item-Kennwerten erschliessen, wenn man diese mit dem Schwellenwert von -1,260 Logits (ebenda, S. 20) abgleicht.

[11] Im Technischen Bericht zur ÜGK 2016 wurden diese Zahlen angegeben (Angelone & Keller 2019a, S. 14); für die Fremdsprachen lassen sie sich wiederum aus den Schwellenwerten und Item-Parametern rekonstruieren.

Aus Sicht der Gutachtengruppe ist also die Bandbreite der Tests an sich kein Problem – im Gegenteil: Um den Schwellenwert herum braucht es eine besondere Häufung von Aufgaben. Aber man braucht auch deutlich höhere und niedrigere Anforderungen, um ein Kontinuum aufzuspannen. Zum Problem wird die Breite erst, wenn zu wenige Aufgaben verbleiben, um zu definieren, was „grundkompetente“ Schüler*innen wissen und können. Probleme können auch entstehen, wenn die Logik der Vorgehensweise und die dabei entstehenden Differenzierungen intransparent bleiben. Das scheint bei den bisherigen ÜGK der Fall zu sein: Darüber, was die beiden Stufen „erreicht“ versus „nicht erreicht“ fachlich-inhaltlich unterscheidet und wie sich dies zur Definition der Grundkompetenzen in den Bildungsstandards verhält, liegen bislang weder aus den HarmoS-Dokumenten des Jahres 2011, noch aus den Pilotierungen, noch aus der psychometrischen Analyse der ÜGK-Daten oder aus der Schwellenwert-Setzung Spezifikationen vor[12]. Das erschwert nach Meinung der Gutachtengruppe die Kommunikation der Ergebnisse in die Praxis und Wissenschaft.

Die Gutachtengruppe sieht es als notwendig an, zu diesen Fragen eine transparente, fächerübergreifende Position zu entwickeln, indem Psychometriker*innen und Fachdidaktiker*innen darstellen, was im Rahmen der ÜGK unter „Messung von Grundkompetenzen“ verstanden wird, wie man also von der Beschreibung der Grundkompetenzen in den HarmoS-Dokumenten zu einer operational unterlegten und fachdidaktisch interpretierbaren Unterscheidung zwischen „erreicht“ und „nicht erreicht“ kommt. Möglicherweise würde sich dabei zeigen, dass der Begriff „Grundkompetenz“ in zwei unterschiedlichen Bedeutungen verwendet wird: Zum einen als das Insgesamt der Kompetenzen, die in den ÜGK-Tests erfasst werden, und zum anderen als die untere von zwei Stufen der Kompetenzskala (in Abgrenzung zu höheren Anforderungen); eine Klärung wäre dann von großer Bedeutung. Der Verweis auf das Verfahren der Schwellenwert-Bestimmung allein reicht nicht – es kommt darauf an, dessen Ergebnis inhaltlich verständlich zu machen, zum Beispiel exemplarisch durch Benennung von typischen Anforderungen der bisherigen Tests, die unterhalb und oberhalb der jeweiligen Schwelle lagen. Ein solches Positionspapier müsste deutlich machen, dass – trotz der Fokussierung auf Mindeststandards – Aufgaben mit einer gewissen Bandbreite benötigt werden, und dass man für eine valide Testkonstruktion Aufgabenmerkmale identifizieren muss, welche nachweislich die Schwierigkeit des Tests beeinflussen.

[12] Zu einigen Beispielaufgaben der ÜGK 2016 (Mathematik) werden im Bericht mögliche Varianten skizziert, die eine höhere Schwierigkeit hätten, aber es kommt zu keiner verallgemeinerten Differenzierung.

Es fehlt an systematischen Analysen zu Aufgabenmerkmalen.

Die an den Gesprächen beteiligten Fachdidaktiker*innen sind durchweg an solchen Aufgabenanalysen interessiert. Es hat bereits entsprechende Explorationen gegeben, diese wurden aber nicht systematisch ausgewertet und publiziert. Nach Ansicht der Gutachtengruppe sollten Aufgabenanalysen gezielt von der Politik beziehungsweise Bildungsadministration in Auftrag gegeben und budgetiert werden, denn es handelt sich hierbei nicht um Studien mit primär wissenschaftlichem Interesse, sondern um Massnahmen im Rahmen der Aufgaben- und Testentwicklung, die für eine solide Fundierung und eine transparente Kommunikation der Schwellenwerte unerlässlich sind. Die Neuerungen in der Aufgabenentwicklung, die als Antwort auf das Audit von Fischbach und Ugen (2018) zur ÜGK 2016 für 2024 umgesetzt werden, weisen den richtigen Weg: „Kognitive Labs“ mit Schüler*innen, Befragungen von Lehrpersonen und quantitative Pilotierungen können systematische Anforderungsanalysen auf den Weg bringen und somit genauere Beschreibungen dessen, was „Grundkompetenz“ ausmacht, ermöglichen.

Anhand der Messungen in Deutsch, Englisch und Französisch – jeweils als erster Fremdsprache – wird besonders deutlich, warum vertiefende Aufgaben- und Testanalysen wichtig sind. Hier wurden 2011 die Grundkompetenzen in den Bereichen Leseverstehen und Hörverstehen für alle drei Sprachen identisch formuliert und bei der ÜGK 2017 in einen einheitlichen Aufgabensatz übertragen. Empirisch verhielten sich die Aufgaben in den drei Sprachen, die jeweils in unterschiedlichen Kantonen als erste Fremdsprache geprüft wurden, jedoch durchaus verschieden, wie Abbildung 6 zeigt: Dieselben Testaufgaben zum Hörverstehen sind nämlich, gemessen an der jeweiligen Kompetenzverteilung der Schüler*innen (vgl. Häufigkeitsdiagramme im jeweils linken Teil der Grafiken), in der Fremdsprache Deutsch schwieriger als in der Fremdsprache Englisch. Darauf hat man bei der Schwellenwert-Setzung offenbar mit unterschiedlich anspruchsvollen Grenzwerten reagiert: Von den 59 identischen Aufgaben beherrscht ein „grundkompetenter“ Schüler oder eine „grundkompetente“ Schülerin im Englischen 21, im Deutschen aber nur 13 Aufgaben[13]. Die Problematik der etwas unterschiedlichen Testnormen trotz identisch formulierter Mindeststandards wurde im Bericht nicht erwähnt. Damit wurde eine Chance vertan, die Schwellensetzungen – auch mit der ihnen innewohnenden gewissen Willkür – transparent zu machen, sie fachdidaktisch zu interpretieren und wichtige Erkenntnisse über die Realität des Sprachraums Schweiz zu generieren. Die Daten der ÜGK 2017 könnten ferner über sogenannte DIF-Analysen (Differential Item Functioning) dazu genutzt werden, diese einmalige Realität besser zu verstehen, differenzierte Hinweise für den Sprachunterricht zu gewinnen und zu reflektieren, was dieser Sprachkontext für die Harmonisierung von Bildungszielen und -ergebnissen bedeutet[14]. Damit könnte das Anliegen von HarmoS konzeptuell und empirisch stark vorangebracht und mit dem fachdidaktischen Praxisdiskurs verknüpft werden.

[13] Dies ergibt sich wiederum, wenn man im Tabellenanhang zum Technischen Bericht die p62-Werte der Aufgaben mit dem jeweiligen Schwellenwert vergleicht. Die Schwelle liegt für Hörverstehen Englisch bei -1,767 Logits, für Hörverstehen Deutsch bei -0,877 Logits; der Nullpunkt der Logit-Skalen entspricht, wie in den Abbildungen 5 und 6 dargestellt, dem jeweiligen Mittelwert der getesteten Schülerpopulation.

[14] Für erste Überlegungen in dieser Richtung vgl. Schneider, Lenz, Studer (2009, S. 35)

4 Design der ÜGK

4.1 Erhebungsrhythmus und Stichprobenziehung (B1.3; B1.5; B1.6; B2.8)

Systematische und längerfristige Planung für den Erhebungsrhythmus wird dringend gewünscht.

Drei Zitate aus den offenen Schlussbemerkungen, um die in der standardisierten Befragung gebeten wurde, können die Position beschreiben, die mehrheitlich auch in den Gesprächen eingenommen wurde:

- „Die Frage nach den Testzeitpunkten erscheint mir zentral.“
- „Mir fehlt ein klare inhaltliche, terminlich langfristige (8-10 Jahre!) Konzeption, welche Fächer in welchen Schuljahren bezüglich Erreichung der Grundkompetenzen überprüft werden. Aktuell stelle ich beim Vorgehen seit Beginn eine gewisse Beliebigkeit fest.“
- „Es muss eine längerfristige, klarere Planung geben, welche Fächer für welche Schuljahre und wann. Es könnte sehr interessant sein, etwas im Längsschnitt zu planen (dieselben Schüler*innen werden in den Klassen 4, 8 und 11 in denselben Fächern getestet).“

Auch wenn sich die Vorstellungen darüber, was genau der optimale Erhebungsrhythmus sei, unterscheiden, stimmen die Befragten und Teilnehmenden an den Gesprächen darin überein, dass die EDK von einer ad-hoc-Ausschreibung einzelner Studien zu einer mittel- bis langfristigen Planung übergehen sollte. Dies hätte folgende Vorteile:

- Ein langfristiger Planungshorizont gibt allen Akteur*innen eine hohe Sicherheit: Den Kantonen und der EDK, die von immer wieder neuen Debatten über Einzelerhebungen verschont werden sowie auch den beauftragten Wissenschafter*innen. Dahingehende Wünsche wurden in den Gesprächen vor allem von der Projektleitung ÜGK geäussert.
- Regelmässige Erhebungen zum selben Fach in derselben Altersgruppe erlauben Trendaussagen, die eher geeignet sind, Bildungspolitik und -administration zu informieren[15].
- Allen Beteiligten, auch der Öffentlichkeit, würde klar, dass die ÜGK-Erhebungen keine einmaligen Schnappschüsse sind, sondern von der regelmässigen Wiederholung leben, die es erst ermöglicht, Aussagen über Trends zu machen. Das könnte die Aufregung um einzelne Erhebungen sowie die Fixierung der öffentlichen Debatte auf Rankings, die immer wieder kritisiert wurde, reduzieren.
- Man könnte klarer definieren, welches die Indikatoren für „Harmonisierung“ sind. Etwa folgendermassen: Harmonisierung ist kein Zustand, sondern ein *Prozess* der Angleichung, dessen Fortschritt daran ablesbar ist, ob die Kantone sich hinsichtlich (a) des Anteils der Schüler*innen, die mindestens Grundkompetenzen erreichen beziehungsweise (b) der entsprechenden Anteile in bestimmten Teilgruppen (nach Geschlecht, Zuwanderungshintergrund, sozialem Status und allenfalls Bildungsgang) *annähern*.

Die Vorteile von Trendmessungen sind unstrittig. Vielfach wird Trendmessung (d.h. die Wiederholung der Erhebung im selben Fach und in derselben Jahrgangsstufe nach einigen Jahren), um systemische Entwicklungen abzubilden, sogar als definitorischer Kern des Bildungsmonitorings verstanden. Umstrittener sind hingegen Kohortenlängsschnitte, also die wiederholte Testung eines Jahrgangs von Schüler*innen mit jeweils neuen Stichproben, oder gar individuelle Längsschnitte – die Erfassung der Bildungsverläufe einer Stichprobe etwa von Jahrgangsstufe 4 über 8 bis 11. Beide sind ungleich aufwändi-

[15] Aus internationalen Schulleistungsstudien und regionalen Vergleichen anderer Länder kann man lernen, dass es keine wirklich überzeugenden empirischen Erklärungen für das Ausmass der zu Beginn eines Erhebungsprogramms festgestellten Leistungsunterschiede zwischen den Bildungssystemen gibt. Dies liegt an der Vielzahl der Einflussfaktoren: zum Beispiel sozio-ökonomische und kulturelle Faktoren, Schulstrukturen und pädagogische Kulturen. Eher lassen sich unterschiedliche *Veränderungen* auf Systemebene erklären, insbesondere wenn Daten zu parallelen beziehungsweise vorgeschalteten Änderungen der Bedingungsfaktoren (z.B. Änderungen der sozialen Zusammensetzung der Schüler*innenschaft, Änderung der Schulstruktur oder der Stundentafel, grundlegende Änderungen der Lehrer*innenbildung) vorliegen. Daher sind Trendaussagen unerlässlich, wenn man Anregungen für die Bildungspolitik und für administrative Massnahmen erhalten möchte. Hinzu kommt, dass Trendaussagen – anders als Aussagen über den Anteil der Schüler*innen, welche eine bestimmte Schwelle überschreiten – nicht der Willkür unterliegen, wie hoch der Erwartungshorizont gesetzt wird.

ger und setzen eine entsprechende Koppelung der Erhebungen voraus, bei individuellen Längsschnitten auch eine Koppelung der Stichproben sowie das personenscharfe Zusammenspiel der Daten. Eine solche Erweiterung wurde fünfmal in den Gesprächen gefordert sowie sechsmal in den offenen Bemerkungen zur standardisierten Befragung. Erhofft werden erweiterte Analysen zu den individuellen, familialen und schulischen Bedingungen erfolgreicher und problematischer Bildungsverläufe – vor allem, wenn zusätzlich Klassenstichproben eingeführt werden. Die Gutachtengruppe sieht hier jedoch die Gefahr einer operativen Überforderung der ÜGK und einer Überlastung auf Seiten der betroffenen Schüler*innen und Schulen. Die Gutachtengruppe empfiehlt, das ÜGK-Erhebungsdesign im Kern als Trenddesign zu konzipieren. Fraglich bleibt, in welchen Abständen Erhebungen wiederholt werden sollen.

Engere Kadenzen zwischen den Erhebungen gewünscht, allerdings könnte dies zu Überforderungen im System führen.

Wiederholt wurden in den Gesprächen engere Kadenzen gefordert. Im Detail variierten die Vorschläge zwischen einem Rhythmus von drei Jahren, einem solchen von drei bis fünf Jahren oder sechs bis neun Jahren bis zur Aussage „alle neun Jahre reicht“. Spielt man die Vorschläge konkreter durch, werden schnell Grenzen deutlich: Wenn man alle Sprachen in einer Erhebung zusammenfasst, gibt es bei der ÜGK prinzipiell je eine Erhebung zur Mathematik und zu den Sprachen in den Jahrgangsstufen 4, 8, 11 sowie zwei Erhebungen zu Naturwissenschaften, also insgesamt acht Erhebungen. Bei einem Wiederholungsrhythmus von drei Jahren müsste man also meistens drei Erhebungen pro Jahr durchführen. Bei einem Wiederholungsrhythmus von neun Jahren käme man auf etwa eine Erhebung pro Jahr. Den Vorstellungen der Projektleitung ÜGK zufolge wäre letztgenanntes mit den in der Schweiz vorhandenen personellen Ressourcen machbar, die erstgenannte Variante aber nicht.
Aus Sicht der Schulforschung und der Governance-Forschung ist ein Zeitraum von neun Jahren allerdings zu lang, um systemische Veränderungen nachvollziehen und rechtzeitig politisch reagieren zu können. Besser geeignet wäre ein Abstand von sechs Jahren zwischen zwei Trend-Messungen. Deshalb werden hier zwei mögliche Planungsmodelle für einen Zyklus von sechs Jahren vorgelegt (vgl. Abbildung 7). Beide Modelle gehen davon aus, dass jährlich eine ÜGK-Erhebung durchgeführt werden kann. Entsprechend der Befunde zum Thema „Fächerwahl“ (vgl. Kapitel 3.1) stehen Schulsprache und Mathematik im Mittelpunkt, weil eine ebenbürtige Berücksichtigung von Fremdsprachen und Naturwissenschaften entweder zu einem unrealistischen Erhebungsaufwand geführt oder die Kadenz der Erhebungen sehr erhöht hätte. Denkbar wäre aber, dass Fremdsprachen gemeinsam mit der Schulsprache erhoben werden und Naturwissenschaften gemeinsam mit Mathematik (wie z.B. in Deutschland üblich).

- Modell 1 erlaubt eine Trendmessung jeweils im Abstand von sechs Jahren. Im Abstand von drei Jahren wird jeweils dieselbe Schulstufe untersucht, allerdings mit wechselnden Fachinhalten. Das Modell hat den Vorzug, dass die Belastungen vor Ort gering sind: Jede*r einzelne Schüler*in muss höchstens einmal in seiner/ihrer Schullaufbahn an einer ÜGK-Erhebung teilnehmen; Lehrpersonen sind frühestens nach drei Jahren erneut betroffen – es sei denn, sie unterrichten sowohl in Jahrgangsstufe 4 als auch in 8.
- Modell 2 erlaubt ebenfalls eine Trendmessung im Abstand von konstant sechs Jahren sowie zusätzlich Kohortenlängsschnitte zwischen Jahrgangsstufe 4 und 11 für Sprachen sowie 8 und 11 für Mathematik, falls solche gewünscht sind.

Beide Modelle können durch entsprechende Stichprobenziehung so gestaltet werden, dass zusätzlich Längsschnitt-Analysen auf Schulebene möglich werden. Hierzu müssten Trenderhebungen bei zwei der Untersuchungen (eine an Primarschulen, eine im Sekundarbereich – in Modell 2 beispielsweise M4 und S11) eine durchgehende Kernstichprobe von Schulen umfassen. Solche Veränderungsanalysen auf Schulebene (bei denen die einzelnen Schulen selbstverständlich anonym bleiben) könnten für die Schulentwicklung besonders wichtige Informationen liefern, wenn zusätzlich schulbezogene Daten von den Schulleitungen erfragt werden.

Diese Modelle sollen nur die prinzipiellen Möglichkeiten einer langfristigen Planung illustrieren. Eine konkrete Planung muss effektive Jahreszahlen einsetzen, dabei den Anschluss an bereits durchgeführte und geplante Erhebungen, mögliche Erweiterungen des Programms (vor allem durch Naturwissenschaften) sowie Verknüpfungen mit anderen Tests (z.B. PISA, regionale Erhebungen) definieren. Die Gutachtengruppe konnte sich bereits davon überzeugen, dass die Projektleitung ÜGK auf derartige Planungen vorbereitet ist.

	Jahr 1	Jahr 2	Jahr 3	Jahr 4	Jahr 5	Jahr 6	Zyklus beginnt neu wie bei Jahr 1	
Modell 1	S4			M4			S4	
		S11			M11			S11
			S8			M8		

	Jahr 1	Jahr 2	Jahr 3	Jahr 4	Jahr 5	Jahr 6	Zyklus beginnt neu wie bei Jahr 1, aber mit längsschnittlichen Links	
Modell 2	S4 →				S8 →			S11
		S11		M4				
			M8 →			M11		
							S4 →	

Abbildung 7: Planungsmodelle für ÜGK-Untersuchungen mit Trendmessungen sowie (in Modell 2) Kohorten-Längsschnitten; S = Sprachen (Schulsprache, gegebenenfalls mit Fremdsprache), M = Mathematik, eventuell mit Naturwissenschaften, 4 = Jahrgangsstufe 4, 8 = Jahrgangsstufe 8, 11 = Jahrgangsstufe 11, Pfeile = Längsschnitt einer Kohorte.

Eine Koppelung zwischen den ÜGK-Aufgaben und den Leistungstests in den Kantonen ist wissenschaftlich komplex und aktuell ohne langfristige Perspektive weder konzeptionell noch personell zu bewältigen.

Der ursprünglich zusätzlich vorgesehene Aufgabenbereich der ADB, eine Koppelung der Aufgaben aus der ÜGK mit kantonalen Erhebungen zu unterstützen, wird von den Vertretungen verschiedener Kantone unterschiedlich gesehen. Dies hängt insbesondere davon ab, ob die jeweiligen Kantone ähnliche Daten selbst generieren oder nicht.

Grundsätzlich ist festzuhalten, dass es sich bei diesem Aufgabenbereich um eine hochkomplexe Aufgabe handelt – technisch, methodisch und insbesondere unter den Rahmenbedingungen in der Schweiz auch organisatorisch. Um diesen Aufgabenbereich konzipieren und langfristig aufbauen zu können, bedarf es jedenfalls einer langfristigen Planung, die von einem verbindlichen längerfristigen Gesamtkonzept für die ÜGK und entsprechendem Commitment zur langfristigen Implementierung abhängig ist. Unter den aktuellen Rahmenbedingungen ist dieser Teil des ursprünglich vorgesehenen Funktionsumfangs nicht zu verwirklichen. Soll dies dennoch angegangen werden, würde es dafür neben einem langfristigen verbindlichen Plan für die ÜGK-Erhebungen eine entsprechend grössere personelle Ausstattung der ADB, eine bessere Einbindung der ADB in die Gesamtstruktur der ÜGK-Durchführung sowie eine institutionelle wissenschaftliche Anbindung der ADB benötigen.

Kein Änderungsbedarf bei Standardisierung zwischen den Kantonen

Zu den Fragen, die der Gutachtengruppe gestellt wurden, gehört die folgende: „Müssen die Erhebungen zur ÜGK in allen Kantonen zwingend in derselben Zeitperiode und jeweils in denselben Fächern durchgeführt werden?“ Die Antwort muss aus wissenschaftlicher Sicht ganz eindeutig „ja“ lauten. Flexiblere Modelle würden die Vergleichbarkeit zwischen den Kantonen gefährden, ohne die das Anliegen, den Grad der Harmonisierung zu erfassen, nicht eingelöst werden kann. Auch vertiefende Auswertungen würden an Aussagekraft verlieren. Es ist aus der internationalen Forschung über Schulleistungsmessungen hinlänglich bekannt, dass Ergebnisse mit dem Erhebungsfenster zusammenhängen (also der Position innerhalb des Schuljahres und dem genauen Alter der Proband*innen) und von Kohorte zu Kohorte schwanken können. Zudem wäre die Geheimhaltung der Testunterlagen gefährdet, wenn diese teils erst mit Verzögerung genutzt werden könnten. Dieser Position haben auch die in den Gesprächen befragten Akteur*innen einhellig zugestimmt.

Kein Änderungsbedarf bei Stichprobenziehung

Eine weitere, auf den ersten Blick „technische“ Frage betrifft die mögliche Erhebung in Klassenverbänden statt mit Schüler*innen-Stichproben, die in jenen Kantonen, in denen keine Vollerhebung stattfindet, quer zu den Klassen gezogen werden. Weder in der standardisierten Befragung noch in den Gesprächen wurde ein merklicher Bedarf nach solchen Veränderungen sichtbar. Allerdings haben drei an der standardisierten Befragung Teilnehmende sowie einige an den Gesprächen beteiligte Bildungsforscher*innen angemerkt, dass Klassenstichproben notwendig seien, wenn (a) Rückmeldungen an Lehrpersonen gegeben werden sollen als Einstieg in Unterrichtsentwicklung, oder/und (b) Analysen zum Zusammenhang zwischen Klassenkontext und Schüler*innenleistung erwünscht sind. Aus der Bildungsforschung ist bekannt, dass Befunde verzerrt werden können, wenn Analysen nicht auf der angemessenen Ebene stattfinden. Beispielsweise wird PISA oft kritisiert, weil das Design keine klassenbezogenen Analysen zulässt und somit keine zuverlässigen Aussagen über Korrelate und Effekte von Unterrichtsmerkmalen machen kann. Eine Entscheidung in dieser Sache wird also davon abhängen müssen, ob die Zielsetzung der ÜGK nach wie vor allein das Bildungsmonitoring auf kantonaler Ebene ist, oder ob Auswertungen beziehungsweise Analysen für die Unterrichtspraxis erwünscht sind. Zudem spielen logistische und Ressourcen-Argumente eine Rolle: Klassenstichproben würden mehr Proband*innen erfordern.

4.2 Konzeption und Entwicklung der Kontextfragebögen (B3.16)

Stellenwert des Kontextfragebogens ist über die Jahre gewachsen, allerdings keine einheitliche Meinung, ob diese Analysen Teil der ÜGK sein sollen.

Mehreren Gesprächen zufolge ist der Stellenwert des Kontextfragebogens gegenüber den ersten Planungen der ÜGK, die von einem reinen Leistungstest ausgingen, in den letzten Jahren gewachsen. So wird die Rolle des sozialen Hintergrunds, der in Schüler*innenfragebögen erfasst wird, für bildungspolitische Diskussionen, speziell zum Thema Bildungsgerechtigkeit, von Forschenden ebenso wie von Vertreter*innen der Kantone als relevanter Bestandteil der ÜGK betont, und einzelne Aufgabenentwickler*innen sind an einem besseren Abgleich zwischen Aufgabenentwicklung und Kontextfragebogen interessiert, um Lernbedingungen untersuchen zu können.
Wie bedeutsam der Kontextfragebogen ist, wird in den Gesprächen allerdings unterschiedlich beurteilt. Ein grosser Teil der Akteur*innen ist der Meinung, dass mögliche Bedingungen des Erreichens der Grundkompetenzen nicht ohne eine differenzierte Analyse der individuellen Voraussetzungen der Schüler*innen und somit der Zusammensetzung der Schüler*innen in einem Kanton beurteilt werden könnten. Die Informationen aus dem Kontextfragebogen seien fast wichtiger als die Leistungsdaten, da

sie es erst ermöglichten, die Leistungsergebnisse angemessen zu interpretieren. Andere wiederum argumentieren, dass dies nicht Aufgabe der ÜGK sei, sondern innerkantonal geleistet werden könne und die Hauptaufgabe der ÜGK die Überprüfung des Erreichens der Grundkompetenzen der Schüler*innen sei. Dritte regen an, dass der Kontextfragebogen auch dazu genutzt werden könnte, ähnliche Kantone bezüglich der Zusammensetzung der Schüler*innenschaft zu identifizieren, da Unterschiede zwischen den Kantonen bezüglich der Zusammensetzung der Schüler*innenschaft bestünden und es keinen Sinn mache, alle Kantone mit allen zu vergleichen, sondern nur jene, die relativ homogene Bedingungen bezüglich der Zusammensetzung der Schüler*innen aufwiesen.

Hinweise zur Relevanz der Kontextfragebögen geben die Ergebnisse der standardisierten Befragung. Die Befragten sollten angeben, wie wichtig die verschiedenen Informationen für ihren jeweiligen Aufgaben- und Zuständigkeitsbereich sind. Dabei wurden Leistungsinformationen, Informationen zum sozialen Hintergrund und Informationen zur Lernumgebung unterschieden. Alle drei Punkte sahen in beiden Gruppen mehr als 60 % der Befragten als wichtig bis sehr wichtig an. Dabei unterschieden sich jedoch Personen aus Forschung und Aus- und Weiterbildung von kantonalen Vertreter*innen. Dies wird besonders deutlich, wenn man auf die Kategorie „sehr wichtig“ schaut (vgl. Abbildung 8): Für Personen aus Forschung und Aus- und Weiterbildung stellen Merkmale der Lernumgebung (z.B. des Unterrichts) eine wichtige Information dar; bei den kantonalen Vertretungen sind sie weniger relevant. Im Gegensatz dazu sind Informationen zur Herkunft der Schüler*innen für mehr als die Hälfte der kantonalen Vertretungen sehr wichtig, während dies nur für etwa 40% der Personen aus Forschung und Aus- und Weiterbildung gilt.

In den Gesprächen gab es dementsprechend sehr unterschiedliche Positionierungen zu Umfang und Weiterentwicklung der Kontextfragebögen: Während einige Praktiker*innen und Vertreter*innen der Bildungsadministration für eine Kürzung votierten, verlangten Forschende oftmals mehr Systematik in der Fragebogenentwicklung, Chancen zur Einbettung analytisch wichtiger Fragen aus der Bildungsforschung sowie auch eine Ergänzung durch Befragungen von Schulleitungen, Lehrkräften oder/und Eltern.

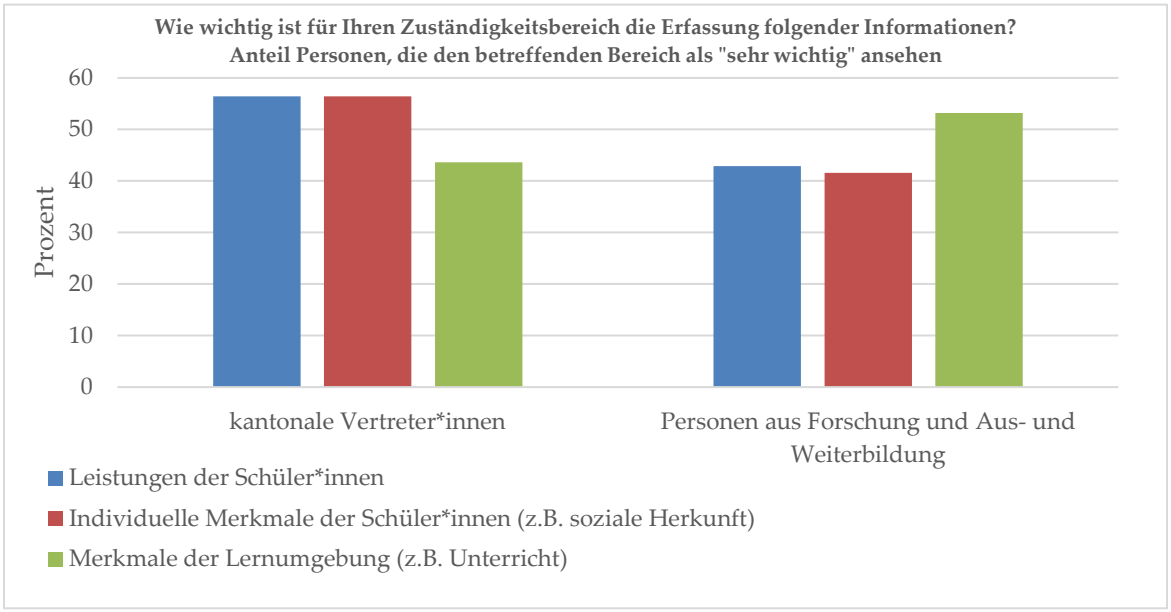


Abbildung 8: Ergebnisse standardisierte Befragung zur Bedeutung verschiedener Arten von Daten; prozentualer Anteil der jeweiligen Gruppe

Inhaltliche Rahmung des Kontextfragebogens wird kritisiert.

Eine professionelle Entwicklung des Kontextfragebogens sollte primär von vorgegebenen Fragestellungen für das Bildungsmonitoring und für vertiefende Forschung ausgehen. Prinzipiell wird der Kontextfragebogen im Large Scale Assessment dazu gebraucht, folgende Arten von Fragen zu beantworten (Kuger, Klieme, Jude & Kaplan, 2016):

1. Wie gross sind Leistungsunterschiede nach Geschlecht, sozialer Herkunft und Migrationshintergrund?
2. Welche „nichtkognitiven“ Lernergebnisse und schulbezogenen Dispositionen – etwa in Bezug auf fachliches Interesse, Lernmotivation, Selbstwirksamkeitserwartungen, Bildungsaspirationen – haben die Schüler*innen erreicht?
3. Wie hängen diese Dispositionen mit fachlichen Leistungen und dem sozialen Hintergrund zusammen?
4. Wie nehmen die Schüler*innen die Lernumgebungen und die erlebte Unterstützung in Elternhaus und Schule wahr?
5. Welchen Effekt haben die Lernumgebungen und die erlebte Unterstützung sowie Merkmale der Schule auf den Leistungszuwachs und die Entwicklung von Dispositionen?

Welche dieser Fragen in einer bestimmten Untersuchung beantwortbar sind und welche nicht, hängt wesentlich vom Stichprobendesign ab. Ein*e Gesprächsteilnehmer*in sagte dazu, Fragestellungen zur Rolle von Interaktionen zwischen den Schüler*innen könnten beispielsweise derzeit in der ÜGK nicht analysiert werden, weil keine Klassen beobachtet würden. Bevor man Themenfelder festlegt, muss daher klar sein, welches Design verfügbar ist. Ohne Klassenstichprobe können zum Beispiel für Merkmale des Unterrichts nur kantonsweite Verteilungen angegeben werden (Frage 4), aber keine sinnvollen Beziehungen zu Leistungen oder Dispositionen der Schüler*innen hergestellt werden (Frage 5). Mehrere Befragte monierten zudem, dass die ÜGK hinsichtlich der Konzeption von „Kontext“ sehr Schüler*innen-zentriert sei. Damit seien, fast zwangsläufig, am Schluss immer die Schüler*innen an den Resultaten „schuld“, weil man alle anderen potenziellen Erklärungsfaktoren ignoriere. Wie in anderen Ländern (z.B. Deutschland) könnten Eltern-, Lehrpersonen- und/oder Schulleitungsfragebögen zur Kontextdatenerhebung dazugehören, ggfs. als Option, die von Kantonen an- oder abgewählt werden kann.

4.3 Prozesse der Aufgaben- und Testentwicklung (B3.11; B3.12)

Empfehlungen des Audits von Fischbach und Ugen (2018) bei den Prozessen der Aufgaben- und Testentwicklung nur teilweise umgesetzt.

Der Audit-Report von Fischbach und Ugen (2018) kommt zusammenfassend zu Empfehlungen, die zur Verbesserung der im Audit detailliert festgestellten kritischen Aspekte der Testentwicklung im Rahmen der ÜGK vorgeschlagen wurden. Diese wurden im EDK-Bericht zu den Erhebungen 2016 und 2017 vom 21.05.2019 (EDK 2019) aufgegriffen und gliedern sich in folgende vier Hauptpunkte mit Unteraspekten[16]:

1. Die Aufgabenentwicklung aufzuwerten und zu überarbeiten, und hierbei
   - die Ressourcen für die Aufgabenentwicklung grundlegend zu überdenken
   - die Aufgabenentwicklung enger an die Testentwicklung und Datenanalyse anzubinden
   - das Koordinationsmandat der Aufgabendatenbank zu stärken
   - die Aufgabenentwicklungsgruppen um Lehrpersonen und anerkannte Assessment-Experten zu ergänzen
   - alle Aufgaben gründlich zu pilotieren
   - in rigorose Übersetzungsprozesse zu investieren

[16] Es wird hier die deutsche Übersetzung (EDK 2019) der Formulierungen von Fischbach und Ugen (2018) verwendet.

2. Die Grundkompetenzen in Mathematik zu überdenken
3. Die Leitung und die Koordination der ÜGK zu überdenken, und hierbei
   - eine eindeutige Trennung zwischen Auftraggeber *in und Projektleitung ÜGK zu schaffen
   - einen regelmässigen, formalisierten und transparenten Dialog zwischen den beiden Gremien zu implementieren
4. Die strategisch/politische ÜGK-Steuerung um eine wissenschaftliche Steuerung zu ergänzen, und hierbei
   - das geplante wissenschaftliche Konsortium (WiKo) zu implementieren
   - dieses Konsortium um einen wissenschaftlichen Beirat mit unabhängigen (inter)nationalen Expert*innen zu ergänzen.

Die Stellungnahme der EDK (vom 21.05.2019) zur Umsetzung der Empfehlungen in Massnahmen der Qualitätsentwicklung scheint die im Audit-Report benannten Empfehlungen nur in Teilen zu adressieren. So werden für die geplanten Aufgabenentwicklungen anscheinend keine zusätzlichen finanziellen Ressourcen zur Verfügung stehen. Stattdessen wird auf erhoffte oder vermutete Effizienzsteigerungen durch verbesserte Koordination, Erfahrungswerte und geringere Itemzielzahl verwiesen (EDK 2019, S. 23). Nach wie vor scheint nicht vorgesehen zu sein, Aufgabenentwicklung, Testzusammenstellung, Testdesign, Stichprobendesign, Standard-Setting und Datenanalyse *institutionell* stärker miteinander zu verzahnen, wie dies in Large-Scale-Assessment-Programmen anderer Länder (z.B. Luxemburg, Österreich oder Deutschland) der Fall ist. Hier liegen bisher eher unkonkrete Absichtsäusserungen seitens der EDK vor.

Ein Aufgreifen der Empfehlungen des Audits von Fischbach und Ugen (2018) im Bereich der Pilotierungen scheint hingegen besser realisiert worden zu sein. So sollen in künftigen ÜGK-Erhebungen grundsätzlich Pilotierungen, im Falle der erstmals administrierten ÜGK-4 auch Prä-Pilotierungen stattfinden (vgl. https://uegk-schweiz.ch/uegk-2024-harmos-4/). Hierbei wären aus Sicht des Gutachtenteams Budgetierungsfragen konkreter, frühzeitiger und verbindlicher zu klären, als dies derzeit von der EDK vorgeschlagen wird. Dies gilt ebenso für eine eventuelle Ausweitung auf naturwissenschaftliche Testdomänen oder die Testung produktiver Kompetenzen, die nach Aussage der Gesprächsteilnehmenden mit den bisherigen Mitteln nicht zu bewältigen seien. Grundsätzlich wird seitens der ADB der Wunsch geäussert, die technischen und operativen Aspekte der Skalierung extern zu vergeben, da andernfalls die Kapazitäten in der ADB nicht mehr ausreichten.

Die massiven sprachregionalen Unterschiede im Funktionieren einzelner Testaufgaben (Differential Item Functioning, „Sprachen-DIF“), die im Audit-Report von Fischbach und Ugen (2018) problematisiert werden und für die eine rigorose Qualitätssicherung des Übersetzungsprozesses empfohlen wird, werden seitens der EDK mit dem Verweis auf verbesserte iterative Checks innerhalb der Entwicklungsgruppe und stärkere Ko-Konstruktion der sprachregionalen Vertreter*innen der Entwicklungsgruppe adressiert. Unklar bleibt jedoch, ob formale Standards der Übersetzung von Testaufgaben berücksichtigt wurden, wie sie etwa in den Test Adaptation Guidelines (TAG) der International Test Commission (ITC 2005 und Hambleton & Zenisky 2010) oder im TRAPD-Ansatz (Translation, Review, Adjudication, Pretesting, and Documentation Approach) dargelegt werden. Darüber hinaus sollte bedacht werden, dass in einer mehrsprachigen Umgebung allfällige Unterschiede zwischen den Sprachen nicht nur als „Fehler“ interpretiert werden sollten, sondern wertvolle Hinweise auf unterschiedliche Kulturen geben können.

Es besteht ein Grundproblem in der Sequenz beziehungsweise im Timing des Testentwicklungsprozesses.

Insgesamt liegt trotz der genannten Anpassungen aus Sicht des Gutachtenteams ein Grundproblem in der Sequenz beziehungsweise im Timing des Testentwicklungsprozesses: Selbst wenn zukünftig vor jeder ÜGK-Hauptuntersuchung eine Pilotierung des Aufgabenpools an einer repräsentativen, grossen

Stichprobe stattfindet, können dennoch psychometrisch problematische Items nach eventuellen Korrekturen oder nachentwickelte Items nicht mehr vor dem Haupttest auf ihr „Funktionieren“ überprüft werden. In anderen Ländern (z.B. in Deutschland, siehe Pant & Stanat, 2015) ist hingegen zwischen Pilotierung (Fokus auf Qualität der Einzelaufgabe) und Haupterhebung eine Normierungsstudie (Fokus auf Eichung des gesamten pilotierten Aufgabenpools) geschaltet. Das Festlegen der Schwellenwerte für das Erreichen der Grundkompetenzen erfolgt so im Schweizer Konzept zwangsläufig anhand der Daten der Haupterhebung, während dies in Deutschland nach der Normierung passiert. Dies führt möglicherweise im Schweizer Design dazu, dass das Standard-Setting besonders anfällig dafür ist, „überraschende“ Verteilungen der Zielindikatoren (d.h. Anteil der Personen, die die Grundkompetenzen erreichen) quasi hinnehmen und erklären zu müssen oder politisch beeinflussen zu wollen.
Die alternativen Empfehlungen des Audits von Fischbach und Ugen (2018), entweder (a) die Standards selbst zu evaluieren, auf ihre (prognostische) Validität zu prüfen und allenfalls zu revidieren („Empirischer Ansatz“), (b) die Mindeststandards zu Regelstandards zu erklären („Terminologischer Ansatz“) oder (c) scheinbar zu „strenge“ Schwellenwerte stehen zu lassen und die resultierenden Verteilungen als Folge (noch nicht) umgesetzter Lehrplan-Implementation zu erklären („Rhetorischer Ansatz“) sind alle mit erheblichen Konsequenzen und Problemen behaftet.
Der – aus wissenschaftlicher Sicht zu favorisierende – „Empirische Ansatz“ setzt langwierige und teure Längsschnittstudien zur Beantwortung der Frage voraus, ob das (Nicht-)Erreichen der Grundkompetenzen tatsächlich zu differenziellen Bildungs- oder gar Erwerbsverläufen führt. Der „Terminologische Ansatz“ zieht die ebenfalls hochkomplexen Fragen der Entwicklung eines differenzierteren Standard- und Kompetenzstufenmodells nach sich und der „Rhetorische Ansatz“ ist ein möglicherweise ungedeckter Scheck auf die zukünftige Wirksamkeit gerade begonnener Curriculums-Umstellungen.
Gerade der „Rhetorische Ansatz” verweist auch auf eine Herausforderung, die weder im Audit von Fischbach und Ugen (2018) noch in den Gesprächen mit den Akteur*innen bisher ausreichend thematisiert erscheint: Wie soll für zukünftige Trendbetrachtungen (d.h. Vergleich des Anteils der grundkompetenten Schüler*innen im gleichen Fach und in der gleichen Jahrgangsstufe über die Zeit) im Bereich der Testentwicklung verfahren werden? Bisher wurden bis auf Anker-/Linkitems offenbar alle Testaufgaben veröffentlicht und sind somit für Folgeerhebungen „verbrannt“. Dies hätte für Trendbetrachtungen zur Folge, dass jeweils ein grösserer Teil des Aufgabenpools neu entwickelt werden muss. Bei den von Fischbach und Ugen (2018) aufgezeigten psychometrischen Problemen wäre eine inhaltliche Stabilität/Volatilität des Konstrukts und der Cut-Scores über die Erhebungswellen hinweg ein kritischer Validitätsaspekt, um Veränderungen über die Zeit sinnvoll interpretieren zu können.

Stärkung der Testentwicklung durch die Bildung multidisziplinärer Teams, durch eine stärkere fachdidaktische Fundierung oder Vermittlung von psychometrischen Grundkompetenzen gewünscht.

Moniert werden seitens der Gesprächsteilnehmer*innen aus der Testentwicklung und Umsetzung der ÜGK im Schulfeld, dass es nach wie vor keine nachhaltigen und personell ausreichend ausgestatteten sowie institutionell abgesicherten Strukturen der Item-/Testentwicklung gebe. Alle zeitkritischen Arbeiten seien damit in gewisser Weise weiterhin prekär angelegt und könnten sich im Falle einer Ausweitung des Domänenspektrums und/oder der Testfrequenz als dysfunktional erweisen. Die EDK (2019) hat als Reaktion auf die Empfehlungen des Audits von Fischbach und Ugen (2018) für die kommenden Erhebungen zwar den Anteil von Lehrpersonen in der Aufgabenentwicklung nach eigenen Angaben erhöht (vgl. EDK 2019, S. 23). Eine detaillierte Aufstellung dazu, wie viele Lehrpersonen in welchen Fächern mit welcher Expertise im Entwicklungsprozess beteiligt beziehungswiese eingeplant sind, lag dem Gutachtenteam nicht vor.
Zudem wird es in den durchgeführten Gesprächen seitens der ADB grundsätzlich als positiv bewertet, dass die relevanten Akteur*innengruppen aus verschiedenen Kantonen oder Sprachregionen an der

Aufgabenentwicklung beteiligt sind. Dennoch wird in den Gesprächen insbesondere die weitere Stärkung der Testentwicklung durch die Bildung echter multidisziplinärer Teams für die Testentwicklung (z.B. Pädagog*innen, Edumetriker*innen, Expert*innen für digitale Entwicklung) gewünscht. Andere Äusserungen gehen dahin, eine stärkere fachdidaktische Fundierung der Aufgabenformate vorzunehmen („Fachdidaktisch breit abgestützte Aufgabenentwicklung, die aktuellem fachdidaktischem Forschungsstand entspricht und nicht einfach der Erfassungsmethode geschuldet ist“).
Allerdings wird auch angeregt, dass den Beteiligten an der Aufgabenentwicklung mehr psychometrische Grundkenntnisse (IRT, Schwellenwerte) vermittelt werden. Ein universitärer Zertifikatslehrgang („CAS Kompetenzdiagnostik“) an den Universitäten Bern und St. Gallen kam bereits kurz nach dessen Einführung aufgrund mangelnden Interesses nicht mehr zustande, so dass eine Akquisition und systematische Grundausbildung neuer Aufgabenentwickler*innen auf diesem Wege bisher nicht erfolgreich war.

4.4 Technische Fragen zur ÜGK (B3.14; B3.17; B3.18)

Prozess des Standard-Setting ist State of the Art, teilweise bleiben aber Unklarheiten.

Das Audit von Fischbach und Ugen (2018) kommt explizit zum Ergebnis, dass der Prozess des Standard-Settings nach State of the Art erfolgt sei. Sowohl der gewählte Ansatz (modifizierte Bookmark-Methode) als auch die Panel-Zusammensetzung wie auch die Abläufe und die gewählten Response Probabilities (62.5 %) sind international üblich[17]. Mithin kann die teilweise kritische Rezeption der darauf basierenden Anteile von Getesteten, die die Grundkompetenzen (nicht) erreicht haben, nicht dem Verfahren oder seiner Implementation zugerechnet werden. In den Gesprächen wurden ebenfalls keine grundlegenden Zweifel an der Validität des Standard-Settings geäussert oder alternative Ansätze (z.B. Angoff-Verfahren) präferiert.
Es bliebe allerdings zu klären, ob die Panelteilnehmer*innen des Standard-Setting-Prozesses korrekt und über die verschiedenen Testdomänen gesehen einheitlich instruiert wurden (vgl. Fussnote 8 in Kapitel 3.2).

Ob auf ein digitales Assessment, durchgeführt von den Lehrpersonen, umgestellt werden soll, wird kontrovers gesehen.

Die Frage, ob die ÜGK mittelfristig auf ein digitales Assessment umgestellt werden sollte, *das von den Lehrpersonen selbst durchgeführt wird*, teilt sowohl Personen aus der Forschung und Aus- und Weiterbildung als auch kantonale Vertreter*innen in der standardisierten Befragung gleichermassen in zwei Gruppen: 46 % der kantonalen Vertreter*innen sind (eher) dafür, 36 % (eher) dagegen, bei Personen aus Forschung und Aus- und Weiterbildung fällt die Tendenz ähnlich, wenn auch in anderer Richtung aus, das heisst 36 % sind (eher) dafür beziehungsweise 40 % (eher) dagegen.
Zur Möglichkeit der kosteneffizienteren Testung mit schuleigenem Equipment äussern sich Vertreter*innen der ADB in den offenen Fragen grundsätzlich positiv, ohne dass dies jedoch konkreter ausgeführt würde („Weniger Geld ausgeben. Mit Testinfrastruktur der Schulen machbar.“).
Grundsätzlich ist zu bedenken, dass die Veränderung des Testmodus – von Testleitungs-administriert zu Lehrpersonen-administriert – zu erheblichen sogenannten „Mode-Effekten“ führen kann, die nur schwer zu kommunizieren sind. Die Erfahrungen aus dem Kontext der Vergleichsarbeiten (VERA)[18] in Deutschland zeigen, dass es für einzelne Aufgabentypen zu starken Schwierigkeitsschwankungen in

[17] Inhaltlich ist eine Response Probability von 62,5 % so zu interpretieren, dass die Teilnehmenden der Standard-Setting-Panels diejenige Testaufgabe in einem „Buch“ aller Testaufgaben, die nach aufsteigender empirischer Itemschwierigkeit geordnet wurden, identifizieren sollen, die ein/e „gerade so grundkompetente/r Schüler*in“ mit hinlänglicher Sicherheit (hier 62,5 %) richtig beantworten kann.

[18] https://www.kmk.org/themen/qualitaetssicherung-in-schulen/bildungsmonitoring/verfahren-zur-qualitaetssicherung-auf-schulebene/vera-faq.html [31. Oktober 2021]

Abhängigkeit davon kommen kann, ob eine Lehrperson oder eine externe Testleitung die Aufgaben administriert. Daher ist insbesondere darauf zu achten, dass in Pilotierung und Haupterhebung *der gleiche Modus* eingesetzt wird, um Inkonsistenzen zu vermeiden.

Verbesserte Dokumentation der Daten und der technischen Vorgehensweise gewünscht, sprachregionale Unterschiede in den Technischen Berichten und Dokumentationen sichtbar machen.

In den Antworten auf die (halb-)offenen Fragen der standardisierten Befragung finden sich mehrere Äusserungen von Forschenden, die sich eine verbesserte *Dokumentation der Daten* und der technischen Vorgehensweise wünschen, um Daten gegebenenfalls leichter für eigene Forschungsarbeiten nutzen zu können. Hervorgehoben wird explizit, dass zwecks besserer Verständlichkeit der Datenstrukturen die Dokumentation verbessert werden sollte, um zu einer externen Datennutzung zu motivieren („Datenformat, das nicht nur von Forscher*innen genutzt werden kann, die das Datenverarbeitungsmodell entwickelt haben [um zu vermeiden, dass die Daten nur von einigen wenigen Expert*innen genutzt werden]. Verschiedene Nutzungsebenen der Daten berücksichtigen“).
Insgesamt fällt auf, dass die Technischen Berichte und Dokumentationen[19] über die Sprachregionen, Testdomänen und Entwicklungsetappen hinweg teilweise nicht einheitlich gestaltet sind, was die vergleichende Rezeption und Beurteilung erschwert.

[19] Wir beziehen uns auf folgende Dokumente: Angelone & Keller (2019b), Erzinger et al. (2019), Pham et al. (2019a), Pham et al. (2019b), Salvisberg (2019), Verner & Helbling (2019).

5 Institutionelle Gestaltung der ÜGK

5.1 Organisation und Verantwortungsstruktur (B2.7; B.2.8; B2.9)

Trotz deutlicher Verbesserung nach wie vor verschiedene Probleme in der Projektorganisation sichtbar.

Aus den Gesprächen lassen sich als Ausgangspunkt in Bezug auf die Organisations- und Verantwortungsstruktur zwei Aussagen identifizieren, die mehrfach und aus unterschiedlichen Positionen eingebracht wurden: (1) Die Anfangsphase der ÜGK wurde als wenig koordiniert wahrgenommen. (2) Durch die Einrichtung des ICER, die übergreifende Koordination durch das ICER sowie die personelle Besetzung im Generalsekretariat der EDK hat sich die Situation deutlich gebessert. Diese Verbesserung wurde von mehreren Seiten wahrgenommen und in vielen Gesprächen explizit angesprochen. Nach wie vor bestehen allerdings Probleme hinsichtlich der Projektorganisation.

Es fehlt eine klare Abgrenzung der Kompetenzen zwischen dem Generalsekretariat der EDK und ICER.

Wie ein*e Gesprächsteilnehmer*in konkret formulierte, seien Abgrenzungen zwischen den Kompetenzen des GS EDK und ICER notwendig. Dieses Faktum wird auf unterschiedliche Weise immer wieder sichtbar, zum Beispiel in der Frage danach, wer denn eigentlich die ÜGK steuere. In anderer Form wurde dies in den Gesprächen evident, wenn das Fehlen einer „Identifikationsfigur“, die die ÜGK in der Öffentlichkeit repräsentiere, angesprochen wurde.
Es braucht klare Kompetenzdefinitionen für das GS EDK und das ICER innerhalb der ÜGK, so dass eine zentrale Steuerung des Gesamtvorhabens durch das GS EDK erhalten bleibt bei Gewährung der notwendigen Unabhängigkeit des ICER in der operativen und wissenschaftlichen Umsetzung. Dafür scheint es unumgänglich, die Ziele der ÜGK von Seiten der Politik klar zu definieren und unter Berücksichtigung des intendierten Wirkungsmodells einen langfristigen, verbindlichen Erhebungsplan auszuhandeln sowie mit wesentlichen Eckpunkten (etwa der Festlegung von Erhebungszeitpunkten, -zielgruppen und -inhalten) zu versehen. In diesem Rahmen benötigt das ICER eine entsprechend umfangreiche Unabhängigkeit für eine wissenschaftlich abgesicherte Implementation und Umsetzung.

Das ICER beschreibt seine Rolle als koordinierend. Die derzeitigen Strukturen führen zu einer Diffusion in der Leitung.

Das ICER beschreibt seine Rolle im Rahmen der ÜGK weder als Projektsteuerung noch als Projektleitung, sondern als koordinierend. Entscheidungen fallen gemäss den involvierten Akteur*innen entweder in der Projektleitung ÜGK als konsensuale Entscheide oder in der jeweils für einen Teilprozess zuständigen Arbeitsgruppe. Wer als Letztverantwortliche*r Entscheidungen trifft, wenn in der Projektleitung ÜGK kein Konsens erzielt werden kann, blieb in den Gesprächen offen. Damit ist die ÜGK durch eine sehr partizipative Entscheidungsstruktur geprägt. Dies mag den Vorteil von einem vermutlich teils intensiven entscheidungsvorbereitenden Diskurs haben, wobei anzumerken ist, dass ein entsprechend breiter Diskurs auch bei klaren Entscheidungsverantwortungen möglich ist. Jedenfalls erschwert diese Vorgehensweise die Koppelung von Entscheidungs- und Ergebnisverantwortung, den Umgang mit kurzfristig notwendigen Entscheidungen und einem eventuell erforderlichen Krisenmanagement. Ausserdem erschwert dies die Abstimmung von verschiedenen Teilprozessen und behindert effiziente, weitsichtige Entscheidungsprozesse. Koordination statt Projektsteuerung ist bei einem Unterfangen dieser Grösse und Komplexität als langfristig dysfunktional einzustufen. Durch diese Logik gibt es keine Möglichkeit, sichtbar zu werden, was wiederum auf das Problem der fehlenden Identifikationsfigur – die durchaus auch institutioneller Natur sein kann – verweist. Daher ist die Projektleitungsstruktur der ÜGK insbesondere unter dem Gesichtspunkt klarer Entscheidungs- und Ergebnisverantwortung zu diskutieren, zumal in der Kooperationsvereinbarung zwischen der EDK und der Universität

Bern (EDK, 2018) explizit steht, dass die Universität Bern beziehungsweise das ICER die *Gesamtleitung der Prozesse* innehabe (Absatz 3.2.2 in der Kooperationsvereinbarung). Im Gegensatz dazu stehen die Formulierungen zu den definierten Aufgaben des wissenschaftlichen Konsortiums (Art. 7, Abs. 2) im ÜGK-Reglement (EDK, 2014), auf welche im relevanten Abschnitt (Absatz 3.1 in der Kooperationsvereinbarung) einleitend verwiesen wird. Hier wird vorrangig von *Koordination* gesprochen. Diese Ambivalenz zwischen Leitung und Koordination ist demnach bereits in den Vertragsdokumenten grundgelegt.

Bestehende Reglemente sind mit den Organisationsstrukturen nicht immer im Einklang, insbesondere bezüglich des Wissenschaftlichen Konsortiums und der Entwicklung des Kontextfragebogens.

Besonders diffus erscheint die Position des wissenschaftlichen Konsortiums. Es existiert als zusätzliches Gremium und wird als loser Verbund beschrieben, der vor allem zur breiteren Information genutzt wird. Damit ist das wissenschaftliche Konsortium nicht das strategische, wissenschaftliche Gremium wie etwa im ÜGK-Reglement beschrieben und seine konkrete Rolle ist unklar. Die Aufgaben, die im ÜGK-Reglement einem wissenschaftlichen Konsortium zugeschrieben wurden, sind zwischenzeitlich im Rahmen der Kooperationsvereinbarung zwischen dem GS EDK und dem ICER Letzterem zugewiesen worden. Die Organisationsstrukturen für die ÜGK haben sich in den letzten Jahren weiterentwickelt, so dass sich der Bedarf ergibt, die (immer noch geltenden) Reglemente aus der Anfangszeit der ÜGK mit den zwischenzeitlich gewachsenen Strukturen in Einklang zu bringen. Dies dient der Klarheit in den Verantwortungsstrukturen und der Verwendung von Begriffen.
Die unklaren Leitungs- und Entscheidungsstrukturen sowie die unscharfe Abgrenzung zwischen Politik und Wissenschaft werden am Beispiel des Kontextfragebogens besonders deutlich. Der Ablauf zur Entwicklung des Kontextfragebogens – abgesehen von den für die Skalierung relevanten Basic-Variablen zum sozio-ökonomischen Hintergrund der Schüler*innen – wird in den Gesprächen als sehr offener Prozess beschrieben: Interessierte Institutionen könnten Wünsche und Ideen einbringen, einige funktionierende Skalen würden von vorausgegangenen ÜGK-Erhebungen übernommen. Dahinter stehe die Idee, dass interessierte Wissenschaftler*innen Skalen einbringen und diese dann auch selbst für Auswertungen nutzen können. Letzteres werde durch die Komplexität des resultierenden Datensatzes sowie die Einschränkungen im Umfang des Kontextfragebogens jedoch behindert. Insgesamt wirkt bei der Entwicklung des Kontextfragebogens die Organisations- und Entscheidungsstruktur besonders unklar.
Das ist zum einen im Zusammenhang mit den nicht genügend abgestimmten Vorgaben und Interpretationen zu sehen, wozu die Kontextfragebögen im Kontext der ÜGK dienen sollen. Den Ausgangspunkt zur Entwicklung des Kontextfragebogens bildet nicht ein gemeinsames Verständnis über den Zweck des Kontextfragebogens, sondern ein Konglomerat von Ansprüchen an den Kontextfragebogen, so beispielsweise von den Expert*innen zur Skalierung der Daten (wofür zentrale Variablen des Kontextfragebogens für das Hintergrundmodell benötigt werden) oder von Wissenschaftler*innen, die mittels des Kontextfragebogens eigene Forschungsinteressen realisieren möchten. Eine Perspektive, die in den Gesprächen nicht genannt wurde, aber eigentlich zentral erscheint, sind Bedürfnisse in Bezug auf die Ergebnisberichterstattung zur ÜGK.
Zum anderen werden in den Prozessen rund um den Kontextfragebogen jedoch auch die insgesamt diffusen Entscheidungsprozesse sichtbar. Wer trägt die Verantwortung dafür, dass die im Rahmen des Kontextfragebogens erhobenen Informationen die Bedürfnisse der ÜGK erfüllen? Die Tatsache, dass die Aufnahme einzelner Items in den Kontextfragebogen letztlich durch den EDK-Vorstand genehmigt wird, kann als ein Hinweis für die Vermengung von Verantwortung zwischen Politik und Wissenschaft gesehen werden, was dazu führt, dass die Politik – statt weitsichtiger strategischer Entscheidungen – operative Einzelentscheidungen trifft.

Die Trennung zwischen ICER und ADB und damit auch die Trennung der Verantwortung für Aufgabenentwicklung einerseits und Durchführung der ÜGK andererseits wird kritisiert.

Die Trennung zwischen den Verantwortlichkeiten für die Test- und Aufgabenentwicklung auf der einen Seite und für die Durchführung der Erhebungen auf der anderen Seite wird übereinstimmend von allen an den Gesprächen beteiligten Akteur*innen auf der operativen Ebene als problematisch gesehen. Dies ist insofern kritisch einzuschätzen, als die Aufgabenentwicklung das konzeptionelle Kernelement der ÜGK darstellt. Durch die institutionelle Verantwortungstrennung liegt dieser Kernprozess nicht in der Steuerungsverantwortung der mit der Durchführung beauftragen Institution. Entscheidungen in Bezug auf die Aufgaben- und Testentwicklung müssen sehr früh im Vorbereitungs- und Durchführungsprozess der ÜGK getroffen werden. Sie stehen mit verschiedenen (chronologisch später angesiedelten) Teilprozessen in Wechselwirkung. Das bedeutet, dass Entscheidungen im Kontext der Aufgabenentwicklung Auswirkungen auf die Prozesse der Erhebung, Verarbeitung und Analyse der Daten haben. Spezifische Anforderungen dieser später angesiedelten Prozesse müssen in Entscheidungen zur Test- und Aufgabenentwicklung Berücksichtigung finden. Die Verantwortungstrennung zwischen diesen und dem Aufgaben- und Testentwicklungsprozess beeinträchtigt daher sowohl Effektivität und Effizienz als auch die Qualität der Produkte.

Die ADB ist innerhalb der ÜGK isoliert, eine institutionelle Anbindung an eine wissenschaftliche Institution fehlt

Die ADB in der derzeitigen Verortung wird sowohl aus einer ÜGK-Innen- als auch aus einer -Aussenperspektive als isoliert beschrieben. Das betrifft sowohl eine strukturelle Isolation innerhalb der Projektorganisation der ÜGK als auch einen Mangel an institutioneller wissenschaftlicher Anbindung. Gleichzeitig ist die ADB mit der Aufgaben- und Testentwicklung für den Kernbereich der ÜGK zuständig, weil mit der Aufgabenentwicklung wesentliche inhaltliche Festlegungen erfolgen. Das verträgt sich weder mit ihrer isolierten Position noch mit der engen Anbindung an das GS EDK. Während das Festlegen von Anforderungen an das Bildungssystem (in Form von Bildungsstandards) in der Verantwortung der Politik liegt, betrifft das in der ADB angesiedelte Aufgabenportfolio die wissenschaftsbasierte Operationalisierung dieser Anforderungen. Sowohl die Entwicklung von Tests und Aufgaben als auch die Setzung von Schwellenwerten, die das Erreichen des inhaltlich festgelegten Anforderungsniveaus kennzeichnen, sind Aufgaben, die nach wissenschaftlichen Kriterien zu bearbeiten sind. Insbesondere vor dem Hintergrund von Glaubwürdigkeitsfragen ist eine formale Trennung zwischen den politischen Entscheiden (was sind die Grundkompetenzen, die Schüler*innen jedenfalls erreichen sollten?) und der wissenschaftlichen Operationalisierung dieser Entscheide (der operativen Umsetzung für die Erhebung des Erreichens der Grundkompetenzen) wichtig.
Darüber hinaus ist die ADB personell unterbesetzt und bietet qualifiziertem Personal wenig Perspektiven. Wesentliche Informationen zum Kernprozess der Test- und Aufgabenentwicklung liegen bei der ADB. Durch ihre vom Rest der ÜGK-Durchführung abgekoppelte Positionierung führt dies zu einer prekären Situation. Das Know-how liegt bei einigen wenigen Personen (derzeit bei einer Person!). Bei personellen Wechseln gibt es keine Strukturen, die den persönlichen Know-how-Verlust durch institutionelles Know-how abfedern können. Die isolierte Position der ADB (verstärkt durch deren geringe personelle Besetzung) birgt daher ein grosses Risiko in Bezug auf kritischen Know-how-Verlust.
Aus diesen Gründen sollte eine stärkere institutionelle Integration der ADB in die ÜGK überlegt werden. Relevant erscheinen insbesondere eine institutionelle Anbindung an eine wissenschaftliche Institution und eine Unabhängigkeit vom GS EDK. Letzteres auch, weil – wie oben angesprochen – die Test- und Aufgabenentwicklung die Operationalisierung der Festlegungen, die politisch zu treffen sind, darstellt. Ähnlich verhält es sich bei der Schwellenwertsetzung, die auf Basis einer inhaltlichen Definition der Grundkompetenzen (politische Entscheidung) ein operativer Prozess ist, der nicht politisch beeinflusst sein sollte. Jedenfalls bedarf es einer besseren institutionellen Anbindung an die wissenschaftliche Community, vorzugsweise verbunden mit einer besseren Integration in den Gesamtprozess der ÜGK.

Die Entscheidungen im Bereich der Aufgabenentwicklung sollten von Entscheidungen der Durchführung der ÜGK nicht entkoppelt sein. Es gilt umfangreiche Wechselwirkungen zwischen Konzeptions- und Durchführungsaspekten zu berücksichtigen, was durch die Trennung der Verantwortlichkeiten massiv erschwert wird. Dies könnte durch eine institutionelle Integration der ADB in das ICER (als für die Projektsteuerung verantwortliche Institution) unterstützt werden.

Die fehlende zeitliche Perspektive verhindert eine langfristige Planung und konzeptionelle Gestaltung der ÜGK.

Der Bedarf an einem langfristigen verbindlichen Plan für die ÜGK-Erhebungen wird in verschiedenen Kontexten angesprochen und ist wohl die in den Gesprächen am häufigsten genannte notwendige Veränderung für die Zukunft. Die fehlende langfristige Perspektive steht dem Aufbau einer effizienten Organisationsstruktur im Wege. Für eine koordinierte Abwicklung und effiziente Konzeption der Erhebungen ist es erforderlich, frühzeitig über verbindliche Informationen zu wesentlichen Eckpunkten (langfristige Festlegung des Erhebungsdesigns, insbesondere Zeitpunkte, Schulstufen, Fächer und deren Abfolge) zu verfügen. Sowohl der Effizienz als auch der Nachhaltigkeit sind durch die derzeitige Situation mit sehr kurzfristigen Entscheidungen, kurzen Vorlaufzeiten für einzelne Erhebungen sowie dem Fehlen eines längerfristigen Plans deutliche Grenzen gesetzt. Dies betrifft auch die vertragliche und damit finanzielle Situation. Eine längerfristige Planung sowie eine damit zusammenhängende vertragliche Beauftragung der Institutionen dienen der Planungssicherheit. Durch die kurzen Vertragslaufzeiten besteht in allen Bereichen die Gefahr, kompetente Personen zu verlieren. Dies führt – nicht nur in Bezug auf die ADB – zur permanenten Gefahr von Know-how-Verlust und behindert den Aufbau eines stabilen Kompetenzzentrums. Die Notwendigkeit eines langfristigen, verbindlichen Plans gilt nicht nur für die direkt in die Vorbereitung und Durchführung der ÜGK involvierten Institutionen, sondern auch für die Verantwortlichen der Bildungsplanung in den Kantonen oder für Wissenschaftler*innen, die Forschung im Kontext oder unter Einbeziehung der ÜGK planen.

Die Einbindung der Sprachregionen ist erst in Teilen realisiert.

Vor allem von französischsprachigen Personen wurde in den Gesprächen auf die Wahrnehmung mangelnder Einbindung aller Sprachregionen hingewiesen. Die ÜGK wird von manchen als „deutschsprachiges“ Projekt wahrgenommen, zu dem aus den französischsprachigen Regionen schwer Zugang gefunden werden könne, beziehungsweise das Personal aufgrund von Sprachkompetenzen anstatt Expertise ausgewählt werden müsse, um mitreden zu können.

Die systematische Einbindung der Fachdidaktiken ist auf die Aufgabenentwicklung beschränkt.

Vertretungen der Fachdidaktiken sind ausschliesslich in den Bereich der Aufgabenentwicklung systematisch eingebunden. In dieser Rolle sind sie über die ADB an die Projektleitung ÜGK angebunden. Die ADB ist eher technisch-psychometrisch ausgerichtet, dies sowohl operativ als auch konzeptionell in Bezug auf das Aufgabenportfolio. Dies ist wichtig für die Erfüllung ihrer Aufgaben im Bereich der Aufgaben- und Testentwicklung sowie Skalierung und Schwellenwertsetzung. Allerdings führt dies dazu, dass die Perspektive der Fachdidaktiken derzeit vorrangig in die Aufgabenentwicklung einfliesst, nicht aber in andere Phasen der ÜGK, etwa in gesamtkonzeptionelle Aspekte oder die Ergebnisanalyse und -interpretation, da die Fachdidaktiken nur über die ADB in der Projektleitung ÜGK vertreten sind und damit nur indirekte Verbindungen mit anderen Teilprozessen der ÜGK-Erhebungen bestehen. Zwar scheint es verschiedene persönliche Kanäle zum GS EDK, zu Kosta HarmoS oder zur Projektleitung ÜGK zu geben. Eine verlässliche, transparente, institutionalisierte Basis für die vollumfängliche Berücksichtigung der wichtigen fachdidaktischen Perspektive in der ÜGK besteht allerdings nicht.

Deshalb sollten Strukturen gefunden werden, die die Berücksichtigung der Perspektive der Fachdidaktiken in der Gesamtkoordination der ÜGK ermöglichen. Dies sollte systematisch erfolgen, so dass die Perspektive der Fachdidaktiken in der Gesamtkoordination gezielt einbezogen wird und dies nicht abhängig von der Mehrfachverantwortung oder persönlichen Kanälen einzelner Personen ist. Dies könnte auch dazu beitragen, die derzeit als erziehungswissenschaftlich-soziologisch fokussiert wahrgenommene Daten- und Ergebnisnutzung um eine fachdidaktische Perspektive anzureichern. Damit wird insbesondere eine Brücke zum Unterricht und zur Lehrer*innenaus- und -weiterbildung geschlagen. Die breitere Einbeziehung fachdidaktischer Positionen in der ÜGK ist darüber hinaus wichtig, um der Segmentierung von Entscheidungsprozessen entgegenzuwirken. In dieser Hinsicht ist mit positiven Effekten in beide Richtungen zu rechnen – für die ÜGK als Ganzes und als Rückwirkung auf die Arbeit der Fachdidaktiken im Bereich der Aufgabenentwicklung. Dabei ist relevant, dass es unterschiedliche Fachdidaktiken zu berücksichtigen gilt und nicht „die Fachdidaktik“. Strukturelle Anbindungsmöglichkeiten sollten also auf die Unterschiedlichkeit etwa zwischen Mathematik und den Sprachen (nicht nur, aber auch vor dem Hintergrund eines mehrsprachigen Landes) Rücksicht nehmen. Darüber hinaus ist auch die derzeit fehlende Planungssicherheit bezüglich der ÜGK-Erhebungen ein hinderlicher Faktor für den langfristigen Aufbau von Expertise in Bezug auf die Test- und Aufgabenentwicklung in den Fachdidaktiken und eine systematische Einbindung der Fachdidaktiken in den Gesamtprozess der ÜGK.

5.2 Verhältnis ÜGK und Wissenschaft (A1.6; B2.10; B3.16; B3.19)

Das Verhältnis zwischen ÜGK als politisches Instrument der Systemevaluation und der Wissenschaft ist an vielen Stellen unklar.

Im Kontext der Frage zu den Beziehungen zwischen Wissenschaft und ÜGK stellt sich die Grundsatzfrage, inwieweit eine Ausrichtung der ÜGK auf die Wissenschaft erstrebenswert ist. Aus der Sicht von vielen Personen, die in den Gesprächen Kantone oder Durchführungsinstitutionen repräsentiert haben, erscheint die vorrangige Erwartung an die ÜGK eine Orientierung an den Bedürfnissen der Kantone beziehungsweise am Zweck des Monitorings zu sein. Um diesem Wunsch zu entsprechen, bedarf es einer vorrangigen Ausrichtung der Konzeption und des Designs der Erhebungen an den Bedürfnissen des Monitorings. Das betrifft sowohl die Konzeption der Tests als auch der Kontextfragebögen. Grundsätzlich ist eine wissenschafts-basierte Konzeption und Durchführung für eine erfolgreiche Umsetzung der ÜGK unbedingt erforderlich. Ob und inwieweit die ÜGK allerdings auf die Wissenschaft und die Bedürfnisse von Forschenden allgemein zugeschnitten sein soll, ist zu diskutieren.
Auf Basis der standardisierten Befragung werden im Folgenden die Sichtbarkeit der ÜGK, die Vernetzung des ICER innerhalb der Wissenschaft in der Schweiz und die Nutzung von ÜGK-Berichten und Daten durch Wissenschaftler*innen beschrieben[20].
Die Sichtbarkeit des ICER als Koordinationszentrum der ÜGK im Kontext der Bildungsforschung wird von den befragten Wissenschaftler*innen in der Summe nicht besonders hoch eingeschätzt. Deutlich mehr als die Hälfte der Befragten halten das ICER im Kontext der Bildungsforschung für überhaupt nicht oder eher nicht sichtbar (19 % überhaupt nicht sichtbar, 41 % eher nicht sichtbar). 16 % nehmen das ICER als eher sichtbar wahr. Als sehr sichtbar wird das ICER nur von einer Person eingeschätzt. Diese Ergebnisse sind jedoch vor dem Hintergrund der noch jungen Geschichte des ICER zu sehen.
Für ihre eigene Forschung ist der Austausch mit den wissenschaftlichen Institutionen der ÜGK (ICER, ADB) für etwas weniger als die Hälfte der befragten Personen eher oder sehr wichtig (23 % beziehungsweise 20 %). 14% halten diesen Austausch für ihre eigene Forschung für überhaupt nicht wichtig, fast

[20] Es sei darauf verwiesen, dass hier nur die Ergebnisse der Gruppe der 64 Forschenden (ohne Personen von der Aus- und Weiterbildung) berichtet werden. Weiter hat ein Anteil von circa 23 % der Forschenden diese Fragen nicht beantwortet.

ein Fünftel für eher nicht wichtig. Dabei sind deutliche Unterschiede zwischen den Disziplinen zu beobachten. Während 24 % der Befragten, die sich (auch) dem Bereich Erziehungswissenschaft zuordnen, den Austausch mit den Institutionen der ÜGK als sehr wichtig bezeichnen, sind dies unter den Fachdidaktiker*innen nur 14 % und niemand von jenen Personen, die ausschliesslich andere Disziplinen angegeben haben.
Dieser Austausch soll aus Sicht der Befragten zu je etwa gleichen Teilen (eher) verstärkt werden oder wird als gerade richtig eingestuft. Nur eine Person möchte den Austausch (eher) reduzieren. Dabei zeigt sich erwartungskonform, dass jene Personen, die den Austausch bisher als wichtig angesehen haben, diesen auch eher verstärkt wissen möchten.
Etwa ein Drittel der Wissenschaftler*innen äussert im Rahmen der standardisierten Befragung Interesse an einer Mitarbeit. Etwas grösser ist das Interesse, sich an der Konzipierung des Kontextfragebogens zu beteiligen (39 % eher oder sehr interessiert) als an der Entwicklung der Leistungstests (30 % eher oder sehr interessiert). Interesse an einer Beteiligung an der Entwicklung von Leistungstests und/oder Kontextfragebögen wird verstärkt von Personen geäussert, die den Austausch mit den Institutionen der ÜGK bereits schätzen gelernt haben sowie Ergebnisse oder Daten der ÜGK bereits genutzt haben.

Der Zweck des Kontextfragebogens ist ungeklärt und die resultierenden Daten sind derzeit weder für Monitoring- noch Forschungszwecke gut nutzbar.

Wesentliche Designentscheidungen sollten vor dem Hintergrund des Hauptzwecks der ÜGK als Instrument des Bildungsmonitorings erfolgen. Konzeption und Durchführung der ÜGK-Erhebungen sollten auf diesen Zweck ausgerichtet werden. Dies hat vor allem Auswirkungen auf die Arbeiten in Bezug auf den Kontextfragebogen. Wenn dem Zweck des Monitorings entsprochen werden soll, sind die Inhalte des Kontextfragebogens auch vorrangig auf diese Bedürfnisse abzustimmen. Dann braucht es eine zentral verantwortliche (wissenschaftliche) Stelle, die den Kontextfragebogen unter diesen Gesichtspunkten konzipiert und optimiert. Das setzt jedoch voraus, die Leistungen der Wissenschaft als Dienstleistung zu organisieren und zu alimentieren und auch für diesen Bereich eine klare Verantwortungstrennung zwischen Politik und Wissenschaft vorzunehmen. Die Annahme, Wissenschafter*innen könnten Serviceleistungen aus wissenschaftlichem Interesse erbringen, ist nicht erfüllbar. Die derzeitige Vorgehensweise, Wissenschaftler*innen offen einzuladen und zu fragen, was sie interessiert, und damit den Inhalt des Kontextfragebogens in erster Linie unsystematisch vom Forschungsinteresse einzelner Personen abhängig zu machen, ist dem Zweck des Monitorings nicht dienlich. Diese Vorgehensweise scheint – so die Aussagen in den Gesprächen – dazu zu führen, dass die resultierenden Daten für niemanden wirklich gut nutzbar sind. Für Monitoringzwecke scheinen die Inhalte des Kontextfragebogens eher zufällig verwendbar; für wissenschaftliche Zwecke lassen sich die Ergebnisse nur begrenzt nutzen, weil – so eine an den Gesprächen beteiligte Person – durch den beschränkten Platz und ohne inhaltliche Fokussierung keine für hochrangige Publikationen brauchbaren Skalen resultierten.
Eine Fokussierung in Designfragen auf den Zweck des Monitorings schliesst jedoch nicht aus, dass die entstehende Datenbasis für eine breitere wissenschaftliche Nutzung von Interesse sein kann. Dies betrifft sowohl eine fachdidaktische Nutzung von Ergebnissen auf Itemebene, als auch breitere wissenschaftliche Fragestellungen, die von der soliden Messung von Schüler*innen-Leistungen profitieren können. Was die Attraktivität der resultierenden Daten für die Wissenschaft angeht, wird als Einschränkung auf die Messung des Erreichens des einen Schwellenwerts hingewiesen. Nachdem dieser 0/1-Indikator in Bezug auf das Erreichen der Grundkompetenzen mithilfe einer dahinterliegenden kontinuierlichen Leistungsskala erzeugt wird, stehen jedoch grundsätzlich sehr wohl Testscores zur Verfügung, die als Achievement-Daten nutzbar sind (wiewohl in den meisten Fächern nicht mit einer Abdekkung des gesamten Leistungsspektrums; vgl. Kapitel 3.2).

Eine Nutzung der ÜGK-Daten und Ergebnisse durch die Wissenschaft ist erst punktuell sichtbar.

Geht es um die Klärung der Frage nach der Nutzung von ÜGK-Berichten sowie Daten durch Wissenschaftler*innen, können Informationen aus zwei Quellen genutzt werden. Zum einen hat eine Nachfrage bei FORS ergeben, dass bis Juni 2021 für 56 Forschungsvorhaben die Nutzung der ÜGK-Daten beantragt wurde. Dabei sind 21 Anfragen aus dem ICER selbst oder aus Institutionen der Projektleitung ÜGK erfolgt. Zwei Anfragen erfolgten in Bezug auf die Erstellung des Bildungsberichts.
Zum anderen können wiederum die Antworten aus der standardisierten Befragung herangezogen werden. 17 % der befragten Wissenschaftler*innen haben die Berichterstattung zu den Ergebnissen der ÜGK bereits als Ausgangspunkt für eigene Forschungsvorhaben genutzt; weitere 5 % geben an, dies zu planen. Ähnlich klein ist die Gruppe an Personen mit 17 %, die angeben, FORS-Daten für eigene Analysen verwendet zu haben. Personen, die FORS-Daten bereits genutzt haben, ziehen das auch für die Zukunft wieder in Erwägung. Von jenen, die bisher von dieser Möglichkeit noch keinen Gebrauch gemacht haben, plant immerhin mehr als die Hälfte, das in Zukunft (eventuell) zu tun. Ähnlich verhält es sich auch bei Personen, die die Berichterstattung bereits als Ausgangspunkt für ihre Forschung verwendet haben. Auch sie schliessen die Nutzung von FORS-Daten in Zukunft nicht aus. Nur etwa ein Viertel der Befragten gibt an, eine Nutzung von FORS-Daten nicht zu planen. Das sind durchwegs Personen, die in der Vergangenheit weder die Berichterstattung zu den Ergebnissen noch die FORS-Daten im Rahmen ihrer Forschungstätigkeit genutzt haben.
Erwartungskonform gibt es deutliche Zusammenhänge zwischen den berichteten Merkmalen. So sind insbesondere Personen, die bereits mit FORS-Daten oder der Ergebnisberichterstattung gearbeitet haben, zur Mitarbeit in der Entwicklung der Leistungstests oder der Konzipierung der Fragebögen bereit. Sie beurteilen auch die Sichtbarkeit des ICER im Kontext der Bildungsforschung besser und schätzen den Austausch mit den Institutionen der ÜGK mehr.

Änderungen im Konzept oder in den Instrumenten der ÜGK könnten eine bessere Nutzung der ÜGK-Daten ermöglichen.

38 % der Forschenden können sich Änderungen im Konzept oder in den Instrumenten der ÜGK vorstellen, die dazu führen würden, dass sie die Daten der ÜGK häufiger für die eigene Forschung nutzen könnten und dass damit eine breitere Einbeziehung der Forschenden erreicht werden könnte. 39 % der Forschenden sind der Ansicht, dass es Änderungen geben müsste, damit mehr oder relevantere Fragen der Bildungsforschung bearbeitet werden könnten.
In den Antworten auf die (halb-)offenen Fragen des Fragebogens finden sich mehrere Äusserungen von Forschenden, die eine bessere Zugänglichkeit der Daten, eine Synopse, Gegenüberstellung beziehungsweise Verknüpfung der Daten und Ergebnisse der ÜGK mit weiteren Quellen (kantonalen Erhebungen, Schulstatistiken) sowie eine bessere Aufbereitung und Vergleichbarkeit der Berichte unter Transferaspekten wünschen, um die Datennutzung zu verbessern.
Die Gespräche haben verschiedene Punkte aufgezeigt, wo spezifische Herausforderungen bei der Datennutzung bestehen. So liegt erst seit Frühjahr 2018 ein Datennutzungskonzept vor. Zudem ist das Einholen von Bewilligungen schwerfällig und selbst das ICER musste für die Nutzung der Daten entsprechende Bewilligungen einholen. Darüber hinaus sei für sekundäranalytische Auswertungen nach Ansicht einiger Gesprächsteilnehmenden der Merkmalskranz, der im Kontextfragebogen erhoben wird, zu wenig ausdifferenziert (sowohl hinsichtlich der Zahl der Konstrukte als auch der Zahl der Items pro Konstrukt), um damit in wissenschaftlich angesehenen Journalen publizieren zu können. Dies betrifft auch die fachdidaktische Forschung.
Die Attraktivität der Datensätze für eine sekundäranalytische Nachnutzung, die nicht nur auf statistisch-methodische Fragestellungen beschränkt bleibt, steht und fällt mit dem Umfang, dem Design und der Qualität der erhobenen Kontextdaten. Die Beispiele aus anderen Ländern zeigen eindrücklich, dass die Erhebung über eine Lehrpersonen-Stichprobe, ein Klassenstichprobendesign zur Erfassung von

Lerngruppenmerkmalen und -effekten sowie ein Eltern- und Schulleitungsfragebogen nicht nur die Ergebnisse des Bildungsmonitorings in ihrer Aussagekraft deutlich ausweiten, sondern auch zu vielen erfolgreichen Forschungsanträgen und hochwertigen Publikationen führen.
Um theoriegeleitete Ergänzungsstudien an zukünftige ÜGK-Erhebungen anzudocken und Forschungsvorhaben damit ebenfalls für eine SNF-Förderung potenziell zu qualifizieren, bräuchte es nach Ansicht der Gesprächsteilnehmenden eine vertragliche Kooperation mit weiteren Hochschulen und deren Forschungsgruppen, um auch hochschulseitig ein Commitment zur besseren Sekundärnutzung der Daten zu erzeugen.

6 Empfehlungen

Die Frage nach der bestmöglichen Strategie für die zielorientierte Entwicklung eines Bildungssystems kann als eine der wesentlichen Fragen der letzten 50 Jahre bezeichnet werden. Dominierte in der Schweiz wie auch in anderen deutschsprachigen und europäischen Ländern lange die Strategie, den sogenannten „Input“, also die personellen oder finanziellen Ressourcen des Bildungswesens oder die Lehrpläne und Qualität der Ausbildung von Lehrpersonen zu fokussieren und gegebenenfalls anzupassen, so haben seit den 2000er-Jahren die sogenannten „Output“-Steuerungssysteme an Relevanz gewonnen (Klieme et al., 2003). Diese beinhalten die Festlegung von Bildungsstandards und die regelmässige Überprüfung der Kompetenzen der Schüler*innen und damit die Analyse der Frage, ob die Schüler*innen zu einem bestimmten Zeitpunkt auch tatsächlich die in den Bildungsstandards festgelegten Ziele erreicht haben.

Auch wenn in der Schweiz (oder in den anderen deutschsprachigen Ländern) kein radikaler Paradigmen-Wechsel von einer Input- zu einer Output-Steuerung stattgefunden hat (Altrichter & Maag Merki, 2016), sondern vielmehr die bereits früher existierenden Input-Steuerungsinstrumente durch Elemente der Leistungserreichungsüberprüfung (Output) erweitert wurden, so zeigen dennoch viele Studien, dass die neuen respektive erweiterten Monitoringsysteme einen – teils fundamentalen – Einfluss auf die Bildungspraxis, die Bildungsadministration und -politik haben können (Maag Merki, 2016). Sie haben das Potenzial, die Interaktionen zwischen den Akteur*innen sowie teilweise auch die Prozesse im Unterricht und in der Schule zu verändern. Darüber hinaus kann in vielen Ländern beobachtet werden, dass neue Institutionen entstehen oder Fachstellen implementiert werden, um ein stärker auf die Lernergebnisse ausgerichtetes Monitoringsystem überhaupt umsetzen zu können. Für die Schweiz sind dies auf nationaler Ebene am offensichtlichsten das ICER und die ADB, aber auch die Verantwortlichen für das Bildungsmonitoring in den Kantonen oder „kantonale Referenzpersonen“, die die Umsetzung der ÜGK in der Praxis unterstützen sollen. Diese müssen ihren eigenen Aufgaben- und Zuständigkeitsbereich im Verhältnis zu den neuen Aufgaben sowie zu bereits existierenden Personen und Institutionen entwickeln. Entscheidungszuständigkeiten und Aufgabenbereiche müssen aufeinander abgestimmt und neu ausgehandelt werden.

Ein Blick in die umfangreiche Forschungsliteratur im Zusammenhang mit der Effektivität der neuen oder angepassten Monitoringstrategie macht deutlich, dass die neue Form des Bildungsmonitorings sehr anspruchsvoll und voraussetzungsreich ist. Keinesfalls bestehen ‚einfache Wege‘, die die Entwicklung eines guten Bildungssystems, in dem (möglichst) alle Schüler*innen die gesetzten Bildungsstandards und Bildungsziele erreichen können, garantieren (Maag Merki, 2016). Diese Komplexität wird zudem dadurch gesteigert – und das kann im internationalen Vergleich, mit wenigen Ausnahmen, durchaus als ein Alleinstellungsmerkmal der Schweiz bezeichnet werden – dass das Erreichen der Grundkompetenzen durch die Schüler*innen nicht nur zu verschiedenen Jahrgangsstufen und in verschiedenen Fächern/Fachbereichen zu überprüfen ist, sondern dass dies auch sprachübergreifend zu geschehen hat.

Noch ein weiterer Aspekt kann als spezifisch schweizerisch betrachtet werden: Im Zuge der Umsetzung des Art. 62 der Bundesverfassung und dem Auftrag zur Harmonisierung von zentralen Eckwerten der obligatorischen Schule steht die Überprüfung des Erreichens der Grundkompetenzen im Kontext der Harmonisierung, auch wenn bislang wenig klar ist, inwiefern diese Harmonisierung über das Festlegen gemeinsamer Bildungsziele hinausgehen soll.

Für die Schweiz bedeuten die Einführung von Bildungsstandards und die regelmässige Überprüfung des Erreichens der Grundkompetenzen der Schüler*innen zudem eine zweifache Neuerung. Das erste Mal werden auf nationaler Ebene Bildungsstandards für unterschiedliche Jahrgangsstufen in unterschiedlichen Fächern/Fachgruppen festgelegt und empirisch überprüft. Zugleich wird mit der ÜGK das erste Mal überhaupt eine nur für die Schweiz entwickelte Bildungsmonitoring-Strategie auf nationaler Ebene implementiert. Dies in einem Land, in dem das Bildungswesen in der obligatorischen Schule in hohem Masse föderalistisch geprägt ist und bisherige Angleichungen – wenn überhaupt – in den Sprachregionen oder zwischen einzelnen Kantonen realisiert worden sind.

Die Analysen der durchgeführten Gespräche, die Auswertung der Rückmeldungen aus der standardisierten Befragung sowie die Analyse bestehender Dokumente und Reglemente ermöglichen es, einen differenzierten Blick in die Strukturen, Prozesse und Potenziale der aktuellen Umsetzung der ÜGK zu erhalten. Aus dem Zusammenführen dieser Analyse werden in der Folge Empfehlungen für die Weiterentwicklung der ÜGK formuliert.

Die folgenden Themenbereiche werden von der Gutachtengruppe als essenziell für eine zukunftsträchtige Umsetzung der ÜGK erachtet. Wir haben uns zum einen eng an den formulierten Fragen der EDK orientiert. Zum anderen haben wir, aufgrund von Überschneidungen zwischen den Blöcken A1 und A2 beziehungsweise B1 bis B3 entschieden, die Empfehlungen entlang von thematischen Bereichen zu formulieren. Darüber hinaus ist zu erwähnen, dass die einzelnen Empfehlungen eng miteinander verwoben sind. Dies bedeutet, dass die Weiterverfolgung einer Empfehlung (z.B. in Bezug auf die Festlegung der Funktionen der ÜGK) auch Konsequenzen für die Weiterentwicklung der ÜGK in einem anderen Themenbereich hat (z.B. in Bezug auf die Festlegung der Stichprobenstrategie). Insbesondere sind jene Empfehlungen, die in 6.1 „ÜGK im Gesamtsystem“ formuliert sind, als zentrale Rahmung für die nachfolgenden Empfehlungen zu verstehen. Die Empfehlungen orientieren sich an der folgenden thematischen Struktur:

- Kapitel 6.1: ÜGK im Gesamtsystem
- Kapitel 6.2: Grundlagen der ÜGK
- Kapitel 6.3: Zielvorgaben der ÜGK
- Kapitel 6.4: Design der ÜGK
- Kapitel 6.5: Institutionelle Gestaltung der ÜGK

6.1 ÜGK im Gesamtsystem

6.1.1 Weiterführung der ÜGK bei substanziellen Anpassungen im Gesamtdesign, der institutionellen Verankerung und Gestaltung sowie in der Zuweisung von finanziellen und personellen Ressourcen

Die Gesamtsicht der vorhandenen Daten zeigt deutlich, dass das Ziel der „Überprüfung des Erreichens der Grundkompetenzen“ im Sinne einer Systemevaluation auf nationaler Ebene breit akzeptiert ist. Ebenso wird die aktuelle Auswahl der getesteten Fachbereiche sowie die Auswahl der Jahrgangsstufen mit einer sehr hohen Zustimmung als sinnvoll erachtet, mit einem noch deutlicheren positiven Votum für Schulsprache und Mathematik (im Vergleich zu Fremdsprachen und Naturwissenschaften) und für die Überprüfung des Erreichens der Grundkompetenzen am Ende des 8. und 11. Schuljahres HarmoS (im Vergleich zum 4. Schuljahr).
Die Analysen zeigen aber ebenso deutlich, dass es grosse Unterschiede in den Erwartungen und Bedürfnissen der Kantone in Abhängigkeit der lokal und regional vorhandenen Qualitätsentwicklungsstrategien und Datengrundlagen gibt. Darüber hinaus kann zwar in Bezug auf die institutionelle Verankerung und Umsetzung der ÜGK eine deutliche Verbesserung im Vergleich zur Startphase beobachtet werden, allerdings sind noch verschiedene Herausforderungen zu bearbeiten. Ebenso verdeutlichen die Ergebnisse klar, dass der Nutzen der ÜGK (auch im Vergleich zum Aufwand) für die Identifikation von Stärken und Schwächen des Bildungssystems in Kantonen oder Bildungsregionen sowie für die Qualitätsentwicklung noch eingeschränkt ist, so dass das Potenzial der ÜGK für die Systemevaluation und die Qualitätsentwicklung des Bildungssystems noch nicht optimal genutzt wird. Besonders zentral ist dabei, dass bislang nicht deutlich geworden ist, was denn unter „Harmonisierung“ im Zusammenhang mit der Überprüfung des Erreichens der Grundkompetenzen verstanden beziehungsweise woran diese gemessen wird.
Die Gutachtengruppe empfiehlt, die ÜGK mit einer langfristigen Perspektive weiterzuverfolgen, jedoch mit substanziellen Anpassungen im Gesamtdesign, der institutionellen Verankerung und Gestaltung sowie in der Zuweisung von finanziellen und personellen Ressourcen.

6.1.2 Langfristige Planung bezüglich Erhebungsdesign gewährleisten und Durchführung vertraglich verbindlich absichern

Ein zentraler Kritikpunkt, der durchgängig formuliert wird, und den die Gutachtengruppe teilt, ist die bislang ungenügende Planungssicherheit. Dies verhindert in kritischem Masse, dass die ÜGK als Instrument für die nationale Systemevaluation effektiv werden kann. Zwar können zum Erreichen der Grundkompetenzen auf nationaler und kantonaler Ebene punktuell gewisse Aussagen gemacht werden, aber die Qualität und Aussagefähigkeit der Erhebungen sind durch die Kurzfristigkeit der Planung gefährdet. Zum einen würde mehr Kontinuität in der Entwicklung von Aufgaben, Tests und Kontextfragebogen vermutlich bessere Instrumente ermöglichen (z.B. ein Fragebogen, der über die Jahrgangsstufen und die Fachbereiche hinaus konzeptionell belastbar und umsetzbar ist); zum anderen würden Daten entstehen, die auch etwas zur Qualitäts*entwicklung* aussagen und damit der Bildungspolitik mehr Hinweise für Stärken und Veränderungsbedarf geben können.
Von besonderer Bedeutung ist die Planungssicherheit zudem für die bildungspolitischen Diskussionen und Strategien in den Kantonen und Sprachregionen, die ihr eigenes Monitoringsystem in Relation zu den ÜGK entwickeln beziehungsweise umsetzen können müssen. Ein längerer Planungshorizont stärkt die Relevanz der ÜGK im öffentlichen Diskurs. Darüber hinaus ist Planungssicherheit absolut zentral in Bezug auf das Gewinnen und Halten von Fachexpert*innen, ohne die die ÜGK nicht unter Berücksichtigung der wissenschaftlichen Standards realisiert werden kann. Für eine langfristige, nachhaltige Implementation der ÜGK braucht es somit einen klaren Auftrag, verbunden mit angemessenen Ressourcen und vertraglicher Sicherheit für jene Institutionen und Personen, die die ÜGK umsetzen.

Die Gutachtengruppe empfiehlt, eine langfristige Planung bezüglich Erhebungsdesign zu gewährleisten und die Durchführung vertraglich verbindlich abzusichern.

6.1.3 Verhältnis der ÜGK zu Monitoringinstrumenten in den Kantonen und Sprachregionen klären

In den Kantonen und Sprachregionen sind in unterschiedlicher Dichte eigene Monitoring- und Qualitätsentwicklungsinstrumente implementiert. Der wahrgenommene Nutzen der ÜGK in den Kantonen hängt von dieser lokalen Datenqualität und -infrastruktur sowie von der fachlichen Expertise in den Kantonen und Sprachregionen ab. Je nachdem erhält die ÜGK eine *ergänzende Funktion* (wenn im eigenen Kanton eine umfangreiche Datendichte und -qualität vorhanden ist) oder eine *Ersatzfunktion*, wenn das kantonale Bildungsmonitoring zentral von der Durchführung der ÜGK abhängt.
Um den unterschiedlichen Bedürfnissen der Kantone gerecht zu werden, **empfiehlt die Gutachtengruppe, eine modulare Struktur der ÜGK mit einem Kernmodul und einem Zusatzmodul festzulegen, wobei die Ergebnisse beider Module im nationalen Bericht dokumentiert werden sollten.**
Die genaue Ausgestaltung des Zusatzmoduls, welches aus Kapazitätsgründen für alle interessierten Kantone gleich sein müsste, könnte am besten im Gespräch zwischen dem ICER und den interessierten Kantonen entwickelt werden. Keinesfalls wird empfohlen, die ÜGK dergestalt zu modularisieren, dass einzelne Lehrpersonen und Schüler*innen eine eigene fachliche Rückmeldung zu ihren Leistungen erhalten, so dass sie darauf aufbauend im Sinne eines direkten Feedbacks ihren Unterricht oder das Lernen reflektieren können. Ein Systemevaluationsinstrument muss strikt von Instrumenten der individuellen Kompetenzdiagnostik getrennt werden, da für die wissenschaftlichen Voraussetzungen fehlen, diese reliabel und valide realisieren zu können.
Die Gutachtengruppe empfiehlt, die Frage zu klären, mit welchen kantonalen Instrumenten und auf welcher Ebene (z.B. auf Ebene Schule) eine Verknüpfung mit den Daten der ÜGK sinnvoll und gewünscht ist.
Wichtig dabei ist, dass die Klärung dieser Fragen nicht nur technisch (psychometrisch), sondern auch aus fachlich-inhaltlicher Sicht sehr anspruchsvoll ist. So beruhen beispielsweise die kantonalen Leistungstests und die Instrumente der ÜGK kaum auf gleichen Konzepten. Dadurch ist die Gefahr der Überlappung zwar sehr gering, aber ebensowenig können Synergien zwischen den verschiedenen Instrumenten genutzt werden. Beispielsweise würde der Nachweis positiver Zusammenhänge zwischen unterschiedlichen Instrumenten beide ein Stück weit validieren. Dieses grosse Feld zu bearbeiten, ist somit nur mit grossem zusätzlichem Aufwand leistbar – falls überhaupt.

6.1.4 Aufgaben und Rollen von Politik, Bildungsadministration und Forschung/Wissenschaft im Gesamtgefüge der ÜGK klären

Entsprechend den Analysen zu den implementierten Prozessen und Strukturen der ÜGK gibt es verschiedene Hinweise, dass die Aufgaben und Rollen von Politik, Bildungsadministration und Forschung/Wissenschaft noch nicht optimal festgelegt sind. Dies zeigt sich beispielsweise in der formalen Verortung der ADB, bei der Festlegung der Schwellenwerte, bei der Entwicklung des Kontextfragebogens, beim Datennutzungskonzept oder bei der Vorstellung, dass die Forschenden ein intrinsisches Interesse haben sollten, ihr Fachwissen für die Realisierung der ÜGK einzubringen, was zur Folge hat, dass (in Teilen) Forschende weniger als Dienstleister*innen, sondern als Kooperationspartner*innen betrachtet werden, die somit in gewissem Ausmass Eigenleistungen erbringen sollten. Diese Beispiele verweisen auf ein Spannungsfeld zwischen der ÜGK als politisches Systemevaluationsinstrument und der ÜGK als (mögliches) Forschungsinstrument.
Die Gutachtengruppe empfiehlt, dieses Spannungsfeld zu klären und bei der Gesamtstruktur, aber auch bei den einzelnen Teilprozessen – zum Beispiel bei der Entwicklung des Kontextfragebogens und der formalen Verortung der ADB – klare Aufgabenteilungen einzuführen und diese rechtlich zu verankern.

Dabei sollte berücksichtigt werden, dass die ÜGK ein bildungspolitisches Systemevaluationsinstrument ist. Die bildungspolitischen Akteur*innen legen die Ziele, das Erhebungsdesign und die Erhebungsinstrumente im Sinne eines strategischen Entscheids fest (z.B. welche Themen im Kontextfragebogen integriert werden sollen). Die Eigeninteressen der Forschenden sind dem politischen Interesse nachgelagert. Im Gegenzug muss die Umsetzung der politischen Entscheide nach wissenschaftlichen Standards und ohne politischen Einfluss erfolgen (z.B. anhand welcher Skalen und Items diese Themen operationalisiert werden können). Dabei sind die Forschenden Dienstleister*innen, die mittels entsprechender Aufträge finanziert werden. In Bezug auf die Entwicklung des Kontextfragebogens bedeutet dies, dass – analog zur Beschreibung der Grundkompetenzen und deren Ausformulierung für die jeweilige Testentwicklung – ein Rahmenkonzept der Befragung der Schüler*innen entwickelt werden sollte, ähnlich dem Questionnaire Framework bei PISA. Dieses Rahmenkonzept könnte in einem mehrstufigen Dialog zwischen Wissenschaft (wissenschaftliches Konsortium, vertreten durch ICER) und Politik (Kosta HarmoS und EDK) entstehen. Die Wissenschaft kann mögliche Fragestellungen und Themenfelder vorschlagen, die von der Politik priorisiert werden; die operative Ausgestaltung nach dem State of the Art der Sozialforschung wäre dann eindeutig Aufgabe der Wissenschaft.
Wenn anhand des politisch festgelegten Erhebungsdesigns auch wissenschaftlich interessante Fragestellungen bearbeitet werden können, wäre dies als Zusatznutzen zu begrüssen und dort, wo bildungspolitische mit wissenschaftlichen Interessen kompatibel sind, idealerweise umzusetzen. Durch die Ausarbeitung eines Datennutzungskonzeptes ist eine vielversprechende Grundlage geschaffen, dass die ÜGK auch als Forschungsinstrument genutzt werden kann. So sollen – etwa im Unterschied zu Deutschland – Längsschnittstudien und Verknüpfungen mit anderen Datensätzen auf Individualebene ausdrücklich ermöglicht werden. Dies stellt aus Sicht der Gutachtengruppe einen potenziell starken Anreiz für relevante Forschung dar. Allerdings bleiben im Beschlussdokument die genauen Kriterien und Verfahren für eine Genehmigung des Datenzugangs etwas im Dunkeln.
Die Gutachtengruppe empfiehlt, die wissenschaftliche Nachnutzung der aufbereiteten Datensätze innerhalb der Scientific Community noch stärker zu bewerben und das Antrags- und Genehmigungsgeschehen systematisch zu evaluieren und zu vereinfachen. Es ist zu überlegen, ob für Konfliktfälle zwischen den Interessen der Politik und den Interessen der Forschung ein geregeltes Clearing-Verfahren konzeptioniert werden sollte.

6.2 Grundlagen der ÜGK

6.2.1 Funktionen der ÜGK klären

Gemäss aktuellen Vorgaben hat die ÜGK die Funktion der Systemevaluation auf nationaler Ebene. In den Gesprächen und in der standardisierten Befragung wird aber mehrfach darauf hingewiesen, dass die ÜGK noch weitere Funktionen zu übernehmen habe, insbesondere auch das Bereitstellen von Erklärungswissen („Warum gibt es die identifizierten Unterschiede?“) oder Handlungswissen für Politik und Schulpraxis („Was können wir tun?“).
Ein gut belegter Befund ist, dass ein Systemevaluationsinstrument nicht gleichzeitig ein Instrument für die Schulpraxis sein kann (z.B. als Diagnoseinstrument in der Hand einzelner Lehrpersonen), weil dies zu Widersprüchen in der Handhabung des Instruments führt und keinen der Zwecke gut erfüllen lässt. Auch das Potenzial, unmittelbare Hinweise für die Veränderung der Schulpraxis zu generieren, ist in der Regel begrenzt. Allerdings ist es möglich, durch eine kluge Kombination von Stichprobendesign und Erhebungsdesign, auch Erkenntnisse für die Erklärung von Leistungsunterschieden zwischen Kantonen und Sprachregionen zu erhalten. Dazu ist es notwendig, weitergehende „Kontext“informationen zu erfassen, so auch zur Qualität der Lernumgebungen der Schüler*innen, und Einschätzungen der Lehrpersonen und Schulleitungen oder Informationen durch die Eltern zu berücksichtigen.
Die Gutachtengruppe empfiehlt, die Funktionen der ÜGK zu klären und allenfalls das Erhebungsdesign so zu erweitern, dass mittels der Daten nicht nur Unterschiede zwischen den Kantonen und Sprachregionen festgestellt werden können, sondern dass auch die Möglichkeit besteht, diese Unterschiede empirisch detaillierter zu untersuchen und Erklärungsansätze zu entwickeln.

6.2.2 Frage nach den Grundkompetenzen klären

Während bei den Befragten ein deutlicher Konsens besteht, die ÜGK auf Grundkompetenzen auszurichten, ist sichtbar geworden, dass meist unklar ist, wie viele der Schüler*innen denn „grundkompetent“ sein müssen, damit die Kantone mit dem Ergebnis zufrieden sind oder nicht. Grundkompetenzen, die auf Minimalstandards beruhen, wie sie im Rahmen der ÜGK festgelegt sind, bedeuten im Prinzip, dass alle Schüler*innen diese Grundkompetenzen erreichen müssen. Für das Etablieren eines nachhaltigen Bildungsmonitorings, das tatsächlich auch dazu dient, Schwächen im Bildungssystem zu identifizieren, ist es somit zentral festzulegen, ab wann Handlungsbedarf (auf nationaler Ebene oder in den einzelnen Kantonen) besteht, zumal es – vermutlich – kaum möglich sein wird, das Ziel des Erreichens der Grundkompetenzen für alle Schüler*innen zu realisieren.
Bislang besteht der Eindruck, dass in der Diskussion der Ergebnisse das (Nicht-)Erreichen des Mindeststandards nicht im Zentrum war, sondern stattdessen eher durch den Vergleich mit anderen Kantonen relativiert wurde.
So scheinen vor allem Kantone, die unter dem kantonalen Durchschnitt lagen, aufgrund der Ergebnisse Massnahmen für das Bildungssystem entwickelt zu haben; lagen hingegen die Kantone im schweizerischen Durchschnitt oder darüber, wurden häufig die Resultate zur Kenntnis genommen und dann ad acta gelegt. Eine Orientierung an einem Mindeststandard würde aber bedeuten, dass das Ausmass des Erreichens der Grundkompetenzen nach spezifischen Kriterien festgelegt wird, und dies zunächst unabhängig vom sozialen Vergleich mit anderen Kantonen (z.B. mindestens x % der Schüler*innen müssen die Grundkompetenzen erreichen). Falls dieses Kriterium nicht erreicht wird, wären entsprechende Massnahmen notwendig, auch wenn der Kanton im Durchschnitt oder sogar darüber liegen würde.
Angesichts der herausragenden Relevanz der Frage nach dem Erreichen der Grundkompetenzen und ihrer Klärung für die Weiterentwicklung der ÜGK sowie angesichts des – sogar reglementarisch – festgelegten Wunsches, nicht den sozialen Vergleich und ein mögliches Ranking zwischen den Kantonen ins Zentrum zu rücken (EDK, 2014), empfiehlt die Gutachtengruppe somit, diese Frage nach den Grundkompetenzen prioritär zu klären.

6.2.3 Frage nach der Harmonisierung der Grundkompetenzen klären

In Bezug auf das Thema der Harmonisierung verweisen die Ergebnisse der Analysen darauf, dass unklar ist, worauf sich der Anspruch auf Harmonisierung, wie er im Bildungsartikel formuliert ist, bezieht und welche weitergehenden Harmonisierungsziele die EDK erreichen möchte. Idealtypisch sind deren drei denkbar: Harmonisierung des Anspruchs, Harmonisierung des Monitorings und Harmonisierung der Ergebnisse.

Gemäss Bildungsartikel ist die *Harmonisierung des Anspruchs* zentral. Dieses Harmonisierungsziel kann mit der Formulierung gemeinsamer Bildungsstandards und Grundkompetenzen als erfüllt betrachtet werden. Eine über den Bildungsartikel hinausgehende *Harmonisierung des Monitorings* ist einerseits durch den Entscheid der EDK, das Erreichen der Grundkompetenzen zu überprüfen und andererseits mit der Durchführung der ÜGK-Erhebungen ebenfalls gegeben.

Unklar ist allerdings, inwiefern eine *Harmonisierung der Ergebnisse* angestrebt wird, anhand welcher Kriterien die Ergebnisse im Sinne der Zielerreichung positiv interpretiert werden können oder anhand welcher Indikatoren sich der Weg zu einer grösseren Harmonisierung abbilden lassen. So ist beispielsweise nicht nur zu klären, wie viele Schüler*innen denn grundkompetent sein müssen, sondern auch, wie gross die Unterschiede zwischen den Kantonen sein dürfen, damit eine Harmonisierung der Ergebnisse auf nationaler oder sprachregionaler Ebene vorhanden ist, und wie genau dieser Unterschied beziffert wird. Prinzipiell lassen sich diese Fragen auch innerhalb eines Kantons stellen, beim Vergleich etwa zwischen Gemeinden oder zwischen Stadt und Land.

Die Gutachtengruppe empfiehlt zu klären, anhand welcher Kriterien die Ergebnisse der ÜGK positiv zu interpretieren sind oder anhand welcher Indikatoren sich der Weg zu einer grösseren Harmonisierung abbilden lässt.

6.3 Zielvorgaben der ÜGK

6.3.1 An einer Orientierung an Mindeststandards und der Überprüfung des Erreichens der Grundkompetenzen festhalten sowie die 2011 formulierten Grundkompetenzen weiterhin als Basis der Aufgabenentwicklung akzeptieren

Die Gutachtengruppe empfiehlt, weiterhin an der Bestimmung des Anteils des Erreichens/Nicht-Erreichens der Grundkompetenzen festzuhalten, die Berichte zur ÜGK (wie schon 2016 und 2017) ausschliesslich darauf auszurichten und auf weiter ausdifferenzierte Kompetenzstufenmodelle zu verzichten. Vier Gründe sprechen dafür:

1. Die Grundkompetenzen sowie der Fokus auf Mindeststandards werden allgemein akzeptiert.
2. Auch der jetzige Modus der Festlegung auf Grundkompetenzen und die empirische Überprüfung des Erreichens ermöglicht es anhand der breit variierenden Testscores, das Spektrum der gemessenen Leistungen differenzierter als ausschliesslich mit dem Indikator „Grundkompetenzen erreicht/nicht erreicht“ zu analysieren.
3. Ein Wechsel auf weiter ausdifferenzierte Kompetenzstufen über den gesamten Leistungsbereich hätte zur Folge, dass der gesamte ÜGK-Prozess (Definierung von Kompetenzen auf weiteren Stufen, Definierung von Schwellenwerten für die weiteren Stufen sowie Entwicklung entsprechender Testaufgaben) neu konzipiert werden müsste, was zu erheblichen Verzögerungen führen würde.
4. Akteur*innen aus der ADB haben Zweifel geäussert, ob eine Angleichung aller Erwartungsniveaus über die Sprachregionen hinweg, mit ihren zum Teil grossen fachdidaktischen Traditionsunterschieden, überhaupt möglich wäre; der Fokus auf Mindeststandards erleichtert daher die Harmonisierung sehr.

Dessen ungeachtet wird empfohlen, die Grundkompetenzen Mathematik zu überprüfen und allenfalls den Erwartungshorizont zu adjustieren. So wurden diese möglicherweise zu anspruchsvoll formuliert. Das Audit von Fischbach und Ugen (2018) schlug daher eine Revision vor, welche die EDK an Kosta HarmoS delegiert hat. Sie wurde nach unserer Kenntnis noch nicht in Angriff genommen.

Weitere Empfehlungen

- Die Aufgabenentwicklung soll ein durchaus breit gefächertes Spektrum von Aufgabenschwierigkeiten anpeilen, auch deutlich unterhalb und oberhalb der Grenze zur Grundkompetenz, wobei um den Schwellenwert herum mehr Aufgaben zu platzieren sind als in anderen Abschnitten des Spektrums. Zudem sollte bei der Aufgabenentwicklung und -pilotierung explizit eingeplant werden, dass Aufgaben sowohl unterhalb als auch oberhalb des Schwellenwertes zur Veröffentlichung in der Praxis benötigt werden, um Grundkompetenzen zu illustrieren und von anderen Leistungsniveaus abzugrenzen.
- Jeder ÜGK-Test sollte mit einem Framework begründet werden, welches das Kompetenzkonstrukt auf fachdidaktischer Basis aktualisiert, Beispielaufgaben benennt und so den fachdidaktischen Diskurs – auch mit Lehrpersonen in der Praxis – vorantreibt. Dieses Framework sollte Ergebnisse der umfangreichen Pilotierung der Aufgaben nutzen und parallel zum jeweiligen Ergebnisbericht publiziert werden. Dies macht insbesondere den gesetzten Schwellenwert transparent: Anhand von Beispielaufgaben und Aufgabenmerkmalen soll dargestellt werden, was „grundkompetente“ Schüler*innen können, und welche Anforderungen über die Grundkompetenzen hinausgehen.
- Die EDK sollte zudem Validierungsstudien in Auftrag geben, gegebenenfalls in Zusammenhang mit Pilotierungen. Diese Validierungsstudien sollten sowohl die Scores auf der kontinuierlichen Testskala verwenden als auch die dichotomisierten Ergebnisse (0 = erreicht/1 = nicht erreicht), sodass Forschende und politische Entscheidungsträger*innen mehr Sicherheit bekommen in der Interpretation der ÜGK-Befunde.
- Psychometriker*innen, Fachdidaktiker*innen und Bildungsforscher*innen aus ÜGK-Projekten sollten in einem Grundlagenpapier fachübergreifend darstellen, was „Messung von Grundkompetenzen“ genau bedeutet.

- Die Aufgabenentwickler*innen sollten den zusätzlichen Auftrag erhalten, gemeinsam mit den Psychometriker*innen der ADB systematische Aufgabenanalysen vorzunehmen, das heisst Anforderungsmerkmale zu spezifizieren und deren Auswirkung auf das Schwierigkeitsniveau und auf differenzielle Schwierigkeitswerte in Sprachregionen oder Kantonen zu ermitteln. Dazu sollen primär die Daten der Pilotierungen verwendet werden. Ergebnisse sollen bei der Schwellenwertsetzung genutzt, in die Ausbildung von Lehrpersonen und in den fachdidaktischen Diskurs hineingetragen werden. Diese Arbeiten stellen eine Dienstleistung im Kontext der Testentwicklung dar, sind damit Teil von Serviceverträgen und erfordern entsprechende Ressourcen.
- Die vollständigen Testdaten (Antworten auf Item-Niveau und Testscores als „Plausible Values“) sollten für Sekundäranalysen zur Verfügung gestellt werden. Die Testscores in der gesamten Varianzbreite können allenfalls im schweizerischen Bildungsbericht genutzt werden, um Zusammenhänge mit nicht-kognitiven Merkmalen sowie familialen und schulischen Faktoren zu untersuchen. Der Erstbericht zur ÜGK sollte sich jedoch – sofern die politische Zielsetzung nach HarmoS unverändert bleibt – weiterhin darauf beschränken, Anteile derjenigen zu berichten, welche den Schwellenwert überschreiten, und diese Anteile mit anderen Variablen zu verknüpfen.

6.3.2 Frage klären, welche Bedeutung der soziale Vergleich zwischen den Kantonen hat und adäquate Lösungen für den Kantonsvergleich finden

Sowohl in den bisherigen nationalen Berichten wie auch in den Gesprächen wird deutlich, dass ein „sozialer“ Vergleich der Kantone in der Erreichung der Grundkompetenzen für die Befragten eine zentrale Bedeutung hat. In den nationalen Berichten wird der Vergleich mit unterschiedlichen Auswertungen realisiert.

Die Gutachtengruppe empfiehlt, die weitgehend deskriptive Ausrichtung der Berichte auf Anteile „grundkompetenter“ Schüler*innen nach Kantonen und Teilgruppen fortzusetzen. Sogenannte adjustierte Vergleiche sollten hingegen vertiefenden Publikationen vorbehalten bleiben.

So kann in Bezug auf das Erreichen der Grundkompetenzen in den Kantonen diese Adjustierung auch zu Problemen führen, da sie dazu verleitet, ein mangelndes Erreichen des gewünschten Ergebnisses im eigenen Kanton beispielsweise mit den geringeren familiären Bildungsressourcen oder der Häufung von Schüler*innen mit Migrationshintergrund zu erklären und Differenzen zwischen den Schüler*innen als „naturgegeben“ anzuerkennen. Gerade das Setzen von Mindeststandards beinhaltet aber die Idee, dass (fast) allen Schüler*innen durch das Bildungssystem ermöglicht werden soll, diese Ziele zu erreichen, unabhängig von familiären Merkmalen.

6.4 Design der ÜGK

6.4.1 An der Wahl der Jahrgangsstufen und der Fächer / Fachbereiche sowie an einer standardisierten Erhebung über alle Kantone hinweg festhalten

Wie bereits erwähnt, besteht ein breiter Konsens bezüglich der Auswahl der Fächer/Fachgruppen sowie der Jahrgangsstufen, am Ende derer die Grundkompetenzen überprüft werden sollen.
Die Gutachtengruppe empfiehlt, die Erhebungen auf jeden Fall für Mathematik und Sprachen (Schulsprache, in den höheren Jahrgangsstufen gekoppelt mit der ersten Fremdsprache) fortzuführen.
Ob Naturwissenschaften einbezogen werden können, gegebenenfalls in einer Erhebung gemeinsam mit der Mathematik, sollte nach sorgfältiger Prüfung der verfügbaren Ressourcen (von der Expertise in der Aufgabenentwicklung bis zu Ausstattung und Personal für Experimentiertests, wie sie in der HarmoS-Validierungsstudie verwendet wurden) im Rahmen einer systematischen Gesamtplanung entschieden werden. Im Anschluss an die ÜGK 2024, die Mathematik und Schulsprache in der 4. Klasse testet, sollte auch noch einmal geprüft werden, wie belastbar Untersuchungen in diesem Alter sind. Gerade das Wegfallen der 4. Jahrgangsstufe hätte aber zur Folge, dass Kohortenlängsschnittdesigns nicht mehr gut realisiert werden könnten. Für ein nationales Systemevaluationsinstrument zwingend notwendig hingegen ist eine zeitliche und inhaltliche Standardisierung der Erhebungen über alle Kantone hinweg.
Die Gutachtengruppe empfiehlt, die Erhebungen in allen Kantonen zur gleichen Zeit und mit der gleichen inhaltlichen Ausrichtung durchzuführen.

6.4.2 Erhebungsdesign klären: Trend-, Kohorten- und Längsschnitterhebungen?

Bislang wurden zwei ÜGK-Erhebungen durchgeführt, eine weitere Erhebung musste aufgrund der COVID-19-Pandemie verschoben werden. In Bezug auf die weitere Sequenzierung und inhaltliche Ausrichtung der einzelnen Befragungen stellt sich somit die Frage, in welchem zeitlichen Abstand, in welcher Jahrgangsstufe und in welchem/welchen Fachbereich(en) die weiteren Erhebungen durchgeführt werden sollen. Hierzu sind folgende grundsätzlich möglichen Erhebungsdesigns zu klären, da sie die Beantwortung von je unterschiedlichen Fragestellungen ermöglichen: Trenddesign, Kohortenlängsschnittdesign und Individual-Längsschnittdesign.[21]
Trenddesign: Trenderhebungen sind unter Bildungsmonitoringperspektive sehr wichtig. Dabei werden im regelmässigen Abstand (alle 3 bis 6 Jahre) die Leistungen der Schüler*innen einer bestimmten Jahrgangsstufe erfasst und mit den Leistungen der Schüler*innen gleicher Jahrgangsstufe in den nachfolgenden Erhebungen über die Jahre verglichen. Dies entspricht beispielsweise dem Design der PISA-Studien.
Kohortenlängsschnittdesign: Dieses Erhebungsdesign wird bei TIMSS (Trends in International Mathematics and Science Study) oder beim Bildungspanel (NEPS) verwendet. Dabei wird eine Kohorte, beispielsweise die Kohorte der Schüler*innen in der vierten Klasse zum Zeitpunkt X, erneut im achten Schuljahr zum Zeitpunkt X+4 und gegebenenfalls im elften Schuljahr zum Zeitpunkt X+7 befragt. Im Gegensatz zum „Individual-Längsschnittdesign“ (siehe nächster Punkt) wird dabei jedesmal eine neue Stichprobe gezogen. Denkbar sind ebenfalls Mehr-Kohorten-Designs, indem beispielsweise zusätzlich zur Kohorte mit Start im Jahr X eine weitere mit Start im Jahr X+6 untersucht wird. Um Erkenntnisse für das Bildungsmonitoring zu gewinnen, ist ein langfristiger Zeitplan unerlässlich.
Individual-Längsschnittdesign: Individuelle Längsschnitterhebungen sind sehr komplex, ermöglichen es aber am besten, Erklärungen für Bildungsverläufe und mögliche Differenzen in der Entwicklung der Kompetenzen der Schüler*innen im kantonalen Vergleich zu gewinnen. Sie beinhalten zusätzlich zum

[21] Je nach Erhebungsdesign muss der Umgang mit den bestehenden Testaufgaben geklärt werden. In einem Trenddesign dürfen beispielsweise die zentralen Testaufgaben nicht veröffentlicht werden, damit sie in den nachfolgenden Trenderhebungen weiterverwendet werden können.

Kohortenlängsschnittdesign die Verfolgung der einzelnen Schüler*innen im Bildungsverlauf. Individuelle Längsschnitterhebungen sind darüber hinaus für Fragen der Bildungsforschung sehr zentral, bedingen aber eine Beobachtung der einzelnen Schüler*innen anhand eines Identifikationscodes über die Bildungslaufbahn hinweg. Angesichts der verschiedenen Klassen-, Schul- und Kantonswechsel ist dies sehr voraussetzungsreich.

Die Gutachtengruppe empfiehlt, das ÜGK-Erhebungsdesign im Kern als Trenddesign zu konzipieren, da dies die Grundvoraussetzung für ein systematisches langfristiges Bildungsmonitoring ist. Inwiefern weitergehende Designs gewünscht und unter welchen Voraussetzungen sie möglich sind, müsste im Austausch aller relevanten Akteur*innengruppen geklärt werden.

6.4.3 Inhalte der nationalen Berichte klären

In der Berichterstattung der Ergebnisse erfolgte bislang eine Trennung zwischen den zentralen Ergebnissen zum Erreichen der Grundkompetenzen, die im nationalen Bericht publiziert werden, und den Vertiefungsstudien, die die Kantone in Auftrag geben können, um kantonal interessante Fragestellungen zu untersuchen. In den Äusserungen der von uns Befragten wird allerdings moniert, dass die Auswertungen, die im nationalen Bericht publiziert sind, die Informationen aus dem Kontextfragebogen zu wenig berücksichtigten, der grosse Aufwand für die Schulen und kantonale Systeme somit nur zu umfangmässig geringen Erkenntnissen führe. Zum anderen haben einzelne Kantone Mühe, Vertiefungsstudien zu finanzieren und die Dauer, bis diese Analysen vorliegen, wird als zu lang kritisiert.

Die Gutachtengruppe empfiehlt, das Konzept der Kernauswertungen und der Vertiefungsstudien zu überdenken sowie die Kontextinformationen, die bedeutsam sind für die Interpretation der Befunde, systematisch in die Kernanalysen einzubeziehen und damit den nationalen Bericht auszubauen.

6.4.4 Kontextinformationen erweitern, in Abhängigkeit der Funktionen der ÜGK

Im Zusammenhang mit der Frage nach der Identifikation möglicher Einflussfaktoren auf die Leistungen der Schüler*innen können im jetzigen System einzig individuelle Faktoren der Schüler*innen (z.B. Geschlecht, familiärer Hintergrund) berücksichtigt werden, da parallel zu den Leistungstests lediglich ein Fragebogen für die Schüler*innen eingesetzt wird. Bisherige nationale und internationale Forschungsergebnisse zeigen aber deutlich, dass Unterschiede in den Leistungen der Schüler*innen nicht nur durch diese Merkmale der Schüler*innen erklärt werden können, sondern dass weitere Faktoren, insbesondere der Unterricht und die Kompetenzen der Lehrpersonen, aber auch schulische Variablen (z.B. Kooperation im Kollegium, Schulleitungshandeln) bedeutsam sind. Diese sind zudem die wichtigsten Anknüpfungspunkte für die Qualitätsentwicklung. Darüber hinaus tragen insbesondere bei den jüngeren Schüler*innen Informationen durch die Eltern wesentlich zum besseren Verständnis der Unterschiede zwischen den Leistungen der Schüler*innen bei.

Die Gutachtengruppe empfiehlt, den konzeptuellen Rahmen des Kontextfragebogens in den HarmoS-Gremien grundsätzlich zu diskutieren, wenn Einigkeit über die Zielsetzung und die Grundlagen des Stichprobendesigns besteht.

- Bleibt es bei einer blossen Deskription des Anteils von Schüler*innen pro Kanton, die die Grundkompetenzen erreichen, werden sinnvollerweise nur diejenigen Informationen erfragt, die für das sogenannte „Hintergrundmodell“ der Schätzung von Testscores (Plausible Values) benötigt werden.
- Zielt die ÜGK jedoch auf weitergehende Fragen, empfiehlt die Gutachtengruppe, den Einsatz zusätzlicher Erhebungsinstrumente für Lehrpersonen, Schulleitungen und Eltern zu prüfen, damit ein besseres Verständnis für mögliche Unterschiede zwischen den Kantonen gewonnen werden kann. Dies wäre auch im Einklang mit dem State of the Art von international vergleichenden Studien.

6.4.5 Stichprobenziehung entsprechend den Funktionen der ÜGK anpassen

Entsprechend der aktuellen Zielsetzung der ÜGK, die Grundkompetenzen der Schüler*innen im Sinne eines Systemmonitorings auf nationaler Ebene zu überprüfen, ist die Stichprobenziehung in Bezug auf eine repräsentative Auswahl der Schüler*innen pro Kanton funktional. Sobald allerdings Erklärungswissen generiert werden soll, welches mehr als den Einfluss individueller Merkmale wie Geschlecht oder den familiären Hintergrund der Schüler*innen einzubeziehen hat, ist eine Anpassung des Stichprobendesigns notwendig. Denkbar sind hierzu zwei Ansätze: Stichprobenziehung auf Klassenebene oder Stichprobenziehung auf Klassen- und Schulebene, falls nicht nur Klassenvariablen, sondern auch Schulvariablen als mögliche Erklärungsindikatoren untersucht werden sollen.
Die Gutachtengruppe empfiehlt, die Stichprobenziehung entsprechend den Funktionen der ÜGK anzupassen.

6.4.6 Öffentlichen Kommunikation der Ergebnisse der ÜGK verbessern

Die Gutachtengruppe empfiehlt, die (politische) Kommunikation der Ergebnisse auf nationaler und kantonaler Ebene zu verbessern und dabei insbesondere auch den Bedürfnissen der Kantone für die innerkantonale Kommunikation angemessen Rechnung zu tragen. Damit könnten die Ergebnisse der ÜGK auch in der Öffentlichkeit noch sichtbarer werden. Aus fachlicher Perspektive könnten die wissenschaftlichen Akteur*innen bei der Interpretation der Ergebnisse stärker Unterstützung anbieten. Dies wäre allerdings als verbindlicher Auftrag zu formulieren und entsprechend zu budgetieren.

6.4.7 Weitere Empfehlungen

Es wird davon abgeraten, die Piloterhebungen in die Hand der Lehrpersonen zu geben, dies insbesondere deshalb, weil Pilotierungs- und Haupttesterhebungen unter den gleichen Bedingungen durchgeführt werden müssen (z.B. beide durch externe Testleitungsadministration oder beide durch schulseitige Testadministration).
Die Technischen Berichte sollten einheitlich ausfallen, wofür ein Template sinnvoll erscheint, damit diese Berichte und Dokumentationen über alle Testdomänen und Jahrgangsstufen hinweg so weit wie möglich harmonisiert sind. Pro Erhebungszeitpunkt sollte zudem ein einziger umfassender Technischer Bericht publiziert werden.

6.5 Institutionelle Gestaltung der ÜGK

6.5.1 Klare sowie vereinfachte Zuständigkeits-, Entscheidungs- und Verantwortungsstrukturen implementieren

Aufgrund der Komplexität der ÜGK und der Notwendigkeit, Teilprozesse aufeinander abzustimmen, braucht es klare Strukturen für Entscheidungen. Derzeit ist in der konkreten Praxis eine Diffusion in Zuständigkeiten und Verantwortlichkeiten zu beobachten, die dadurch verstärkt wird, dass auch vertragliche Unklarheiten bestehen.

Dies zeigt sich zum einen in der Unklarheit, ob die Universität Bern tatsächlich „Kooperationspartnerin“ der EDK ist oder eine Institution, die einen bestimmten Auftrag der EDK als dienstleistende Institution bearbeitet. Zum andern, ob die Universität Bern beziehungsweise das ICER nur eine Koordinations- oder auch eine Leitungsaufgabe zu übernehmen hat. So soll zwar die Universität Bern gemäss der Kooperationsvereinbarung zwischen der EDK und der Universität Bern die „Gesamtleitung der Prozesse“ übernehmen und mit verschiedenen Institutionen aus allen Sprachregionen „kooperieren“ (Absatz 3.2.2)[22]. Gemäss Absatz 3.2.3 wird dann allerdings einzig der Auftrag zur Koordination von verschiedenen Bereichen festgelegt. Die Mischung zwischen Leitungs- und Koordinationsfunktionen akzentuiert sich des Weiteren dadurch, dass die Universität Bern gemäss Vereinbarung mit den verschiedenen Institutionen „Kooperationsvereinbarungen“ abzuschliessen und darin die zu übernehmenden Rollen zu definieren hat (Absatz 3.2.2). In diesem Vorgehen zeigt sich ebenfalls eine Vermengung zwischen Leitungs- und Koordinationsfunktionen.

Unklarheiten manifestieren sich auch an anderen Stellen. So stellt sich beispielsweise die Frage nach den Funktionen des „wissenschaftlichen Konsortiums“, das in verschiedenen Reglementen beschrieben wird. Zum einen scheint, dass mit dem ICER das wissenschaftliche Konsortium umgesetzt wurde, zum anderen ist gemäss Unterlagen noch ein „wissenschaftliches Konsortium Bildungsforschung und -evaluation“ in Diskussion. Momentan scheint es sich eher um ein Verbund von interessierten Institutionen und Personen zu handeln, die sich vor allem sporadisch zum Zweck des Informationsaustauschs treffen, so dass es kaum eine tragende und strategische Rolle in langfristiger Perspektive ausüben kann. Genau dies wäre aber bedeutsam, um der wissenschaftlichen Komplexität der ÜGK gerecht werden zu können.

Darüber hinaus bestehen nicht nur vertragliche Vereinbarungen zwischen der EDK und der Universität Bern, sondern auch mit der Schweizerischen Koordinationsstelle für Bildungsforschung (SKBF) zur „Domizilierung der ADB“ und der FH Graubünden zur Realisierung von für die ÜGK zentralen technischen Aufgaben. In diese Aufgabengebiete ist die Universität Bern beziehungsweise das ICER nicht involviert. Sie kann – respektive darf – dafür aktuell keine Leitungsfunktionen übernehmen, was zu einem erhöhten Koordinationsaufwand und bedeutsamen Schnittstellen-Risiken führt. Gerade die Software, die von der FH Graubünden bereitgestellt wird, hat massive Folgen für den gesamten ÜGK-Prozess, wobei die Befragten auf verschiedene Probleme in der bisherigen Praxis hingewiesen haben. Darüber hinaus bleibt unklar, warum mit der SKBF und der FH Graubünden separate Verträge bestehen und die entsprechenden Aufgaben nicht auch mittels Kooperationsvereinbarungen über die Universität Bern geregelt werden.

Die Gutachtengruppe empfiehlt, die komplexen vertraglichen Strukturen mit Parallelverträgen der EDK zu überdenken und zusammenzuführen, da ansonsten die „Einheit der Materie“ verletzt wird und wichtige Teilaufgaben und -prozesse in unterschiedlichen Verantwortungsbereichen liegen.

Zudem wird eine Koordinationsfunktion (ohne die explizite Übernahme einer Leitungsfunktion) durch die Universität Bern beziehungsweise das ICER als nicht ausreichend und langfristig dysfunktional angesehen. Es braucht Klarheit bezüglich der Entscheidungs- und Prozessstrukturen und ein eindeutiges, vertraglich festgelegtes Mandat für die Universität Bern beziehungsweise das ICER zur Gesamtleitung

[22] Kooperationsvereinbarung betreffend die Zusammenarbeit zur Durchführung der ÜGK-Erhebungen in 2020 und 2022 und zur kontinuierlichen wissenschaftlichen Koordination der Überprüfung des Erreichens der Grundkompetenzen (ÜGK)

der Prozesse in den zentralen Bereichen. Am augenfälligsten ist dies bei der Aufgabenentwicklung und der Entwicklung des Kontextfragebogens. Um ein Gesamtkonzept für einen tragfähigen Kontextfragebogen zu entwickeln, der über die Jahrgangsstufen hinweg wesentliche Kontextinformationen zum Lernen der Schüler*innen erfasst, reicht ein rein koordinierendes Vorgehen nicht. Es braucht eine klare inhaltliche Gesamtstrategie und demzufolge strategische Entscheide. Eine Gesamtleitung ist sodann bezüglich der einzelnen Teilbereiche in der Praxis so zu konkretisieren, dass die verschiedenen involvierten Institutionen und Sprachregionen ihre Aufträge bestmöglich und im Sinne der Gesamtzielerreichung realisieren können.

Die Gutachtengruppe empfiehlt, die Leitungsfunktion der Universität Bern beziehungsweise des ICER zu stärken und die Entscheidungs- und Prozessstrukturen klarer aufzubauen. Darüber hinaus sollte der Punkt „Kooperation“ zwischen der Universität Bern und der EDK geklärt und ausdifferenziert werden: Welche Aufgaben liegen in der Kompetenz der EDK und welche im Sinne eines Auftrags in der Kompetenz der Universität Bern beziehungsweise des ICER?

6.5.2 Testentwicklung als komplexes und anspruchsvolles wissenschaftliches Aufgabenfeld anerkennen, mit genügend personellen Ressourcen ausstatten und institutionell in die wissenschaftliche Projektstruktur der ÜGK einbinden

Die Aufgabendatenbank der EDK (ADB) ist eine der Kernakteur*innen im Prozess der Überprüfung des Erreichens der Grundkompetenzen. Bei den entsprechenden Testentwicklungen und psychometrischen Analysen handelt es sich um komplexe und anspruchsvolle wissenschaftliche Aufgaben, die einzig von wissenschaftlichen Fachexpert*innen unter Berücksichtigung des State of the Art in Forschungsmethodik und Fachdidaktik realisiert werden können. Ohne einen hohen Standard in der Aufgaben- und Testentwicklung können die ÜGK ihre Funktion bezüglich des Systemmonitorings nicht erfüllen und die Ziele eines systematischen Monitorings auf nationaler Ebene nicht erreichen.

Bislang ist die ADB allerdings nicht als wissenschaftliche, sondern als politische Akteurin interpretiert und formal-rechtlich positioniert worden. Damit zusammenhängend erfolgte eine formal-rechtliche Trennung der Verantwortungsträger*innen der Test- und Aufgabenentwicklung einerseits und Durchführung der ÜGK andererseits. Auch die fehlende institutionelle Anbindung der ADB an eine wissenschaftliche Institution sowie ihre isolierte Position innerhalb der ÜGK stellen ein substanzielles Risiko dar. Zudem ist die ADB personell massiv unterbesetzt und bietet qualifiziertem Personal – auch aufgrund der Kurzfristigkeit der Verträge – wenig Perspektiven. Damit besteht nicht nur ein erhebliches Risiko in Bezug auf kritischen Know-how-Verlust, sondern auch die Gefahr, dass die komplexen Anforderungen an die Testentwicklung – mehrere Jahrgangsstufen, Fächer, Sprachen – mittel- und langfristig nicht adäquat bearbeitet werden können. Dies trotz des hohen Engagements der aktuellen Fachperson(en) in der ADB.

Die EDK argumentierte in Antwort auf die im Audit-Report von Fischbach und Ugen (2018) formulierten Kritikpunkte, dass durch eine Professionalisierung der Prozesse und durch den Erfahrungsgewinn real mehr Zeit für die Entwicklung der einzelnen Aufgaben zur Verfügung stünde. Diese Entscheide können aufgrund der Komplexität der Aufgaben- und Testentwicklung in drei Sprachen nicht überzeugen.

Die Gutachtengruppe empfiehlt,

- **eine ausreichende personelle Ausstattung mit längerfristiger Perspektive, eine systematische und institutionelle Anbindung an wissenschaftliche Institutionen sowie eine Loslösung vom GS EDK und eine Einbindung in die Projektstruktur der ÜGK zu ermöglichen.**
- **die Vorschläge des Audits von Fischbach und Ugen (2018) konsequent aufzugreifen, indem die Teilprozesse von Aufgabenentwicklung, Testzusammenstellung, Testdesign, Stichprobendesign, Standard-Setting sowie Datenanalyse technisch und institutionell stärker verzahnt werden, wie dies in Large-Scale-Assessment-Programmen anderer Länder (Luxemburg, Österreich, Deutschland) der Fall ist.**

6.5.3 Fachdidaktik(en) stärker in den Gesamtprozess der ÜGK einbinden

Vertretungen der Fachdidaktiken sind systematisch ausschliesslich in der Aufgabenentwicklung und des Standard-Settings eingebunden. In dieser Rolle sind sie über die ADB an die Projektleitung ÜGK angebunden. Damit fliesst die Perspektive der Fachdidaktiken derzeit vorrangig in die Aufgabenentwicklung ein, nicht aber in andere Phasen der ÜGK, etwa in gesamtkonzeptionelle Aspekte oder die Ergebnisanalyse und -interpretation. Eine verlässliche, transparente, institutionalisierte Basis für die umfängliche Berücksichtigung der wichtigen fachdidaktischen Perspektive in der ÜGK besteht nicht.

Die Gutachtengruppe empfiehlt,

- **den begonnenen Weg einer multidisziplinären Aufgaben- und Testentwicklung zu forcieren und in allen Phasen der Entwicklung Vertreter*innen aus Fachdidaktik, Unterrichtspraxis und Edumetrie einzubeziehen. Zu diesem Zweck sollte ein integriertes und langfristiges Konzept der Aufgaben- und Testentwicklung erstellt werden, aus dem der Professionalisierungsanspruch der Aufgabenentwicklung erkennbar wird.**
- **Strukturen zu finden, die die Berücksichtigung der Perspektive der Fachdidaktiken in der Gesamtkoordination der ÜGK ermöglichen. Dies sollte systematisch erfolgen, so dass die Perspektive der Fachdidaktiken in der Gesamtkoordination gezielt einbezogen wird und dadurch eine Brücke zum Unterricht wie auch zur Aus- und Weiterbildung von Lehrpersonen geschlagen werden kann. Zudem ermöglicht dies, die derzeit als erziehungswissenschaftlich-, psychologisch- sowie soziologisch-fokussiert wahrgenommene Daten- und Ergebnisnutzung um eine fachdidaktische, unterrichts- und schulnähere Perspektive anzureichern.**

6.5.4 Sprachregionale Unterschiede in Bezug auf fachliches Lernen und Aufgaben herausarbeiten sowie für die Testentwicklung nutzen

In den Gesprächen zeigen sich verschiedene Hinweise, dass in der Entwicklung des Kontextfragebogens oder der Testaufgaben die sprachregional verschiedenen fachdidaktischen Kulturen zu wenig berücksichtigt und sichtbar geworden sind.

Die Gutachtengruppe empfiehlt, die Daten der ÜGK zu den Testaufgaben in Bezug auf allfällige Differenzen zwischen den Sprachregionen zu analysieren (bzw. soweit solche Analysen stattfinden, sichtbar zu publizieren). Dies kann zum einen als Validitätsprüfung, zum anderen aber auch als Anregung für die fachdidaktische Weiterentwicklung der (eventuell vorhandenen) unterschiedlichen Lern- und Unterrichtskulturen genutzt werden kann.

Empfohlen wird sodann, die (allfällig vorhandenen) unterschiedlichen Lern- und Unterrichtskulturen sorgfältig herauszuarbeiten, Differenzen und Überschneidungen zu identifizieren und in einem fachlichen Diskurs für die Entwicklung der Erhebungsinstrumente zu nutzen.

I. Anhang – Literaturverzeichnis

Altrichter, H., & Maag Merki, K. (2016). Steuerung der Entwicklung des Schulwesens. In H. Altrichter & K. Maag Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem*, 2., erweiterte Ausgabe (S. 1-28). Springer VS.

Angelone, D., & Keller, F. (2019a). *Überprüfung des Erreichens der Grundkompetenzen (ÜGK) im Fach Mathematik im 11. Schuljahr: Technische Dokumentation zur Testentwicklung und Skalierung.* Aufgabendatenbank der EDK.

Angelone, D., & Keller, F. (2019b). *Überprüfung des Erreichens der Grundkompetenzen (ÜGK) in den Fächern Schulsprache und erste Fremdsprache im 8. Schuljahr: Technische Dokumentation zur Testentwicklung und Skalierung.* Aufgabendatenbank der EDK.

Angelone, D., Keller, F., Arnold, J., Jaun-Holderegger, B., Hauser, M., Eckstein, B., Erzinger, A., Leutwiler, L., Nell-Tuor, N., & Schönenberger, S. (2019). *Machbarkeitsstudie zur Überprüfung der Grundkompetenzen im 4. Schuljahr.* Aufgabendatenbank der EDK.

EDK (2007). *Interkantonale Vereinbarung über die Harmonisierung der obligatorischen Schule (HarmoS-Konkordat); Umsetzung auf der Ebene der interkantonalen Koordination vom 25./26. Oktober 2007.*

EDK (2014). *Organisationsreglement über die Durchführung der Überprüfung des Erreichens der Grundkompetenzen*. Bern: EDK.

EDK (2018). Kooperationsvereinbarung zwischen der EDK und der Universität Bern betreffend die Zusammenar*beit zur Durchführung der ÜGK-Erhebungen in den Jahren 2020 und 2022 und zur kontinuierlichen wissenschaftlichen Koordination der Überprüfung des Erreichens der Grundkompetenzen (ÜGK) gemäss Artikel 4 Absatz 2 des ÜGK-Reglements*. Bern: EDK.

EDK (2019). *Überprüfung des Erreichens der Grundkompetenzen: Bericht zu den Erhebungen 2016 und 2017*. Bern: Generalsekretariat der EDK.

Erzinger, A., Hauser, M., Dutrevis, M., Hascher, T., Keller, R., Lenz, P., & Soucis, A. (2019). *Erläuterungen zu den Skalen des Kontextfragebogens der ÜGK Sprachen 2017: Theoretischer Hintergrund, Inhalte und Konstrukte*. Bern & St. Gallen: Universität Bern, Pädagogische Hochschule St. Gallen, Service de la recherche en éducation (SRED), Universität Fribourg.

Fischbach, A., & Ugen, S. (2018). *ÜGK/COFO Mathematics 2016: Audit report: commissioned by EDK/CDIP upon the request of KOSTA HarmoS.* http://edudoc.ch/record/204067/files/UeGK_Audit_Report_230218.pdf

Fuchs, G. (2019). Prognosegüte bildungsstandardisierter Tests. Dissertation, FU Berlin. http://dx.doi.org/10.17169/refubium-1687

Hambleton, R.K. & Zenisky, L. (2010) Translating and adapting tests for cross-cultural assessments, in D. Matsumoto and F. Van de Vijer (eds.) Cross-cultural research methods in psychology, (pp. 46–70). Cambridge: Cambridge University Press.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.-E., & Vollmer, H. J. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise.* Bundesministerium für Bildung und Forschung.

Kuger, S., Klieme, E., Jude, N. & Kaplan, D. (Hrsg.), *Assessing contexts of learning: An international perspective*. Cham: Springer. doi:10.1007/978-3-319-45357-6_1.

ITC = International Test Commission (2005) International Test Commission guidelines for translating and adapting tests. http://www.intestcom.org/files/guideline_test_adaptation.pdf. [Accessed 01 September 2021].

Lenz, P., Aeppli, T., Karges, K., Wiedenkeller, E., Angelone, D. & Keller, F. (2019): Nationale Bildungsstandards in der ersten Fremdsprache. In Konsortium ÜGK (Hrsg.) (2019). *Überprüfung der Grundkompetenzen. Nationaler Bericht der ÜGK 2017: Sprachen 8. Schuljahr* (S. 35-49). Bern und Genf: EDK und SRED. https://doi.org/10.18747/PHSG-coll3/id/385

Linneweber-Lammerskitten, H., Angelone, D. und Keller, F. (2019). Überprüfte Kompetenzen in Mathematik. In Konsortium ÜGK (Hrsg.) (2019). *Überprüfung der Grundkompetenzen. Nationaler Bericht der ÜGK 2016: Mathematik 11. Schuljahr* (S. 17-27). Bern und Genf: EDK und SRED, https://doi.org/10.18747/PHSG

Maag Merki, K. (2016). Theoretische und empirische Analysen der Effektivität von Bildungsstandards, standardbezogenen Lernstandserhebungen und zentralen Abschlussprüfungen. In H. Altrichter & K. Maag Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem*, 2., erweiterte Ausgabe (S. 151-182). VS Verlag für Sozialwissenschaften.

Pant, H. A. & Stanat, P. (2015). Wie kommt man zu guten Testaufgaben? Die Konstruktion von Aufgaben und Tests für die Ländervergleiche des IQB. *SEMINAR: Aufgaben. Im Leistungskontext. In Schule und Lehrerausbildung. 48. Seminartag Berlin (II), 21*(1), 5–18.

Pham, G., Helbling, L., Verner, M., Petrucci, F., Angelone, D. & Ambrosetti, A. (2019a). *ÜGK– COFO – VeCoF 2016 results: Technical appendices.* St.Gallen & Genf: Pädagogische Hochschule St.Gallen (PHSG) & Service de la recherche en éducation (SRED).

Pham, G., Helbling, L., Verner, M., & Ambrosetti, A. (2019b). *ÜGK – COFO – VeCoF 2017 results: Technical appendices.* St.Gallen & Genève: Pädagogische Hochschule St.Gallen (PHSG) & Service de la recherche en éducation (SRED).

Ramseier, E., Moser, U., Moreau, J. & Antonietti, J.-P. (2008). *Schlussbericht der HarmoS-Methodologiegruppe. Bern*. http://www.lehrplanforschung.ch/wp-content/uploads/2011/06/Methodologie-Schlussbericht-_HarmoS11.pdf

Rost, J. (2004): Psychometrische Modelle zur Überprüfung von Bildungsstandards anhand von Kompetenzmodellen. *Zeitschrift für Pädagogik 50(5)*, S. 662-678.

Salvisberg, M. (2019). *Verifica delle competenze fondamentali nazionali 2017. Sviluppo degli esercizi nella lingua di scolarizzazione nell'8° anno scolastico.* Locarno: Scuola universitaria professionale delle Svizzera italiana (SUPSI).

Schneider, G., Lenz, P. & Studer, T. [Konsortium HarmoS Fremdsprachen] (2009). *Fremdsprachen: wissenschaftlicher Kurzbericht und Kompetenzmodell.* Bern: EDK. http://edudoc.ch/record/87025/files/L2_wissB_25_1_10_d.pdf

Tiffin-Richards, S.P. & Pant, H.A. (2017) Arguing Validity in Educational Assessment. In: D. Leutner, J. Fleischer, J. Grünkorn & E. Klieme (eds). *Competence Assessment in Education. Methodology of Educational Measurement and Assessment*. Springer, Cham. https://doi.org/10.1007/978-3-319-50030-0_27.

Verner, M., & Helbling, L. (2019). *Sampling ÜGK 2017. Technischer Bericht zu Stichprobendesign, Gewichtung und Varianzschätzung bei der Überprüfung des Erreichens der Grundkompetenzen 2017*. Zürich: Institut für Bildungsevaluation, assoziiertes Institut der Universität Zürich.

II. Anhang – Ergebnisse der standardisierten Befragung

Angaben zu Ihrem Arbeitsbereich

1. Was ist Ihr aktueller Arbeitsbereich? (mehrere Antworten möglich)

	Alle Befragten	Kantonale Vertreter*innen	Personen aus Forschung und Aus- und Weiterbildung
	Häufigkeit	Häufigkeit	Häufigkeit
a) Forschung	64 (45.4%)	1 (2.6%)	64 (83.1%)
b) Aus- und Weiterbildung von Lehrpersonen	44 (31.2%)	1 (2.6%)	44 (57.1%)
c) Schulpraxis	5 (3.5%)	0 (0%)	2 (2.6%)
d) Bildungsadministration: Leitende Position	24 (17%)	24 (61.5%)	2 (2.6%)
e) Bildungsadministration: Umsetzung der ÜGK	20 (14.2%)	20 (51.3%)	1 (1.3%)
f) Andere	10 (7.1%)	0 (0%)	2 (2.6%)
Subtotal	125	39	77
Fehlend	16 (11.3%)	0 (0%)	0 (0%)
Total	141	39	77

1a. Ist Ihr Kanton dem HarmoS-Konkordat beigetreten? (nur für Personen, die bei Frage 1 angegeben haben: d oder e

	Kantonale Vertreter*innen	
	Häufigkeit	Prozent
Ja	21	53.8%
Nein	15	38.5%
Noch offen	2	5.1%
Subtotal	38	97.4%
Fehlend	1	2.6%
Total	39	100.0%

1b. Welcher Disziplin bzw. welchen Disziplinen ordnen Sie Ihr Forschungsgebiet zu? (mehrere Antworten möglich)

	Personen aus Forschung und Aus- und Weiterbildung
	Häufigkeit
a) Erziehungswissenschaft	42 (54.5%)
b) Fachdidaktik(en)	21 (27.3%)
c) Psychologie	7 (9.1%)
d) Soziologie	15 (19.5%)
e) Politikwissenschaft	4 (5.2%)
f) (Bildungs-)Ökonomie	3 (3.9%)
g) Andere	4 (5.2%)
Subtotal	64
Fehlend	13 (16.9%)
Total	77

2. In welcher Sprachregion arbeiten Sie? (mehrere Antworten möglich)

	Alle Befragten	Kantonale Vertreter*innen	Personen aus Forschung und Aus- und Weiterbildung
	Häufigkeit	Häufigkeit	Häufigkeit
a) Romandie	24 (17%)	9 (23.1%)	14 (18.2%)
b) Italienische Schweiz	11 (7.8%)	2 (5.1%)	9 (11.7%)
c) Deutschschweiz	92 (65.2%)	28 (71.8%)	57 (74%)
Subtotal	123	39	76
Fehlend	18 (12.8%)	0 (0%)	1 (1.3%)
Total Antwortende	141	39	77

A. Grundsätzliche Bedeutung der ÜGK für die Qualitätsentwicklung des Bildungswesens

*3. Als wie wichtig erachten Sie es, dass das Erreichen der Grundkompetenzen der Schüler*innen auf nationaler Ebene überprüft wird?*

	Alle Befragten		Kantonale Vertreter*innen		Personen aus Forschung und Aus- und Weiterbildung	
	Häufigkeit	Prozent	Häufigkeit	Prozent	Häufigkeit	Prozent
Überhaupt nicht wichtig	2	1.4%	1	2.6%	2	2.6%
Eher nicht wichtig	10	7.1%	0	0.0%	10	13.0%
Eher wichtig	46	32.6%	17	43.6%	26	33.8%
Sehr wichtig	54	38.3%	20	51.3%	31	40.3%
Subtotal	112	79.4%	38	97.4%	69	89.6%
Fehlend	29	20.6%	1	2.6%	8	10.4%
Total	141	100.0%	39	100.0%	77	100.0%

4. Unabhängig von der aktuellen Durchführung: Wie wichtig ist die ÜGK aus Ihrer Sicht für das Fällen von bildungspolitischen Entscheiden:
a) auf nationaler Ebene?
b) auf sprachregionaler Ebene?
c) auf kantonaler Ebene?

		Alle Befragten		Kantonale Vertreter*innen		Personen aus Forschung und Aus- und Weiterbildung	
		Häufigkeit	Prozent	Häufigkeit	Prozent	Häufigkeit	Prozent
a)	Überhaupt nicht wichtig	4	2.8%	1	2.6%	4	5.2%
	Eher nicht wichtig	20	14.2%	2	5.1%	17	22.1%
	Eher wichtig	38	27.0%	20	51.3%	16	20.8%
	Sehr wichtig	47	33.3%	15	38.5%	29	37.7%
	Subtotal	109	77.3%	38	97.4%	66	85.7%
	Fehlend	32	22.7%	1	2.6%	11	14.3%
	Total	141	100.0%	39	100.0%	77	100.0%
b)	Überhaupt nicht wichtig	6	4.3%	1	2.6%	6	7.8%
	Eher nicht wichtig	26	18.4%	12	30.8%	14	18.2%
	Eher wichtig	45	31.9%	17	43.6%	24	31.2%
	Sehr wichtig	31	22.0%	8	20.5%	21	27.3%
	Subtotal	108	76.6%	38	97.4%	65	84.4%
	Fehlend	33	23.4%	1	2.6%	12	15.6%
	Total	141	100.0%	39	100.0%	77	100.0%
c)	Überhaupt nicht wichtig	6	4.3%	2	5.1%	5	6.5%
	Eher nicht wichtig	20	14.2%	6	15.4%	13	16.9%
	Eher wichtig	41	29.1%	14	35.9%	26	33.8%
	Sehr wichtig	42	29.8%	16	41.0%	22	28.6%
	Subtotal	109	77.3%	38	97.4%	66	85.7%
	Fehlend	32	22.7%	1	2.6%	11	14.3%
	Total	141	100.0%	39	100.0%	77	100.0%

5. Unabhängig von der aktuellen Durchführung: Wie wichtig ist die ÜGK aus Ihrer Sicht für das Fällen von Entscheiden für die Schulpraxis?

	Alle Befragten		Kantonale Vertreter*innen		Personen aus Forschung und Aus- und Weiterbildung	
	Häufigkeit	Prozent	Häufigkeit	Prozent	Häufigkeit	Prozent
Überhaupt nicht wichtig	11	7.8%	3	7.7%	8	10.4%
Eher nicht wichtig	33	23.4%	11	28.2%	20	26.0%
Eher wichtig	42	29.8%	19	48.7%	20	26.0%
Sehr wichtig	24	17.0%	5	12.8%	19	24.7%
Subtotal	110	78.0%	38	97.4%	67	87.0%
Fehlend	31	22.0%	1	2.6%	10	13.0%
Total	141	100.0%	39	100.0%	77	100.0%

*6. Die ÜGK umfasst zwei Elemente: a) Leistungsmessungen in einzelnen Fächern und b) Kontextfragebögen für die Erfassung von Merkmalen der Schüler*innen und der Lernumgebung (z.B. soziale Herkunft, Motivation, Unterricht). Wie wichtig ist für Ihren Aufgaben- und Zuständigkeitsbereich...*
*a) die Erfassung der Leistungen der Schüler*innen?*
*b) die Erfassung von individuellen Merkmalen der Schüler*innen (z.B. soziale Herkunft)?*
c) die Erfassung von Merkmalen der Lernumgebung (z.B. Unterricht)?

		Alle Befragten		Kantonale Vertreter*innen		Personen aus Forschung und Aus- und Weiterbildung	
		Häufigkeit	Prozent	Häufigkeit	Prozent	Häufigkeit	Prozent
a)	Überhaupt nicht wichtig	7	5.0%	0	0.0%	7	9.1%
	Eher nicht wichtig	11	7.8%	0	0.0%	11	14.3%
	Eher wichtig	35	24.8%	15	38.5%	16	20.8%
	Sehr wichtig	56	39.7%	22	56.4%	33	42.9%
	Subtotal	109	77.3%	37	94.9%	67	87.0%
	Fehlend	32	22.7%	2	5.1%	10	13.0%
	Total	141	100.0%	39	100.0%	77	100.0%
b)	Überhaupt nicht wichtig	3	2.1%	0	0.0%	3	3.9%
	Eher nicht wichtig	9	6.4%	1	2.6%	7	9.1%
	Eher wichtig	41	29.1%	14	35.9%	25	32.5%
	Sehr wichtig	56	39.7%	22	56.4%	32	41.6%
	Subtotal	109	77.3%	37	94.9%	67	87.0%
	Fehlend	32	22.7%	2	5.1%	10	13.0%
	Total	141	100.0%	39	100.0%	77	100.0%
c)	Überhaupt nicht wichtig	1	0.7%	0	0.0%	1	1.3%
	Eher nicht wichtig	6	4.3%	3	7.7%	3	3.9%
	Eher wichtig	42	29.8%	17	43.6%	22	28.6%
	Sehr wichtig	60	42.6%	17	43.6%	41	53.2%
	Subtotal	109	77.3%	37	94.9%	67	87.0%
	Fehlend	32	22.7%	2	5.1%	10	13.0%
	Total	141	100.0%	39	100.0%	77	100.0%

7. Aktuell hat die ÜGK die Funktion der Systemevaluation auf nationaler Ebene. Welche Funktion sollte aus Ihrer Sicht die ÜGK haben?
a) Systemevaluation auf nationaler Ebene
b) Systemevaluation auf sprachregionaler Ebene
c) Systemevaluation auf kantonaler Ebene
d) Identifikation von Ursachen allfälliger Unterschiede zwischen regionalen oder kantonalen Bildungssystemen
e) Handlungswissen, wie das Bildungssystem auf nationaler Ebene verändert werden kann
f) Handlungswissen, wie das Bildungssystem auf sprachregionaler Ebene verändert werden kann
g) Handlungswissen, wie das Bildungssystem auf kantonaler Ebene verändert werden kann
h) Handlungswissen, wie die Schulpraxis verändert werden kann
i) Andere

		Alle Befragten		Kantonale Vertreter*innen		Personen aus Forschung und Aus- und Weiterbildung	
		Häufigkeit	Prozent	Häufigkeit	Prozent	Häufigkeit	Prozent
a)	Nein	0	0.0%	0	0.0%	0	0.0%
	Eher nein	7	5.0%	0	0.0%	7	9.1%
	Eher ja	25	17.7%	7	17.9%	17	22.1%
	Ja	77	54.6%	30	76.9%	43	55.8%
	Subtotal	109	77.3%	37	94.9%	67	87.0%
	Fehlend	32	22.7%	2	5.1%	10	13.0%
	Total	141	100.0%	39	100.0%	77	100.0%
b)	Nein	4	2.8%	1	2.6%	3	3.9%
	Eher nein	19	13.5%	6	15.4%	13	16.9%
	Eher ja	33	23.4%	14	35.9%	16	20.8%
	Ja	54	38.3%	16	41.0%	36	46.8%
	Subtotal	110	78.0%	37	94.9%	68	88.3%
	Fehlend	31	22.0%	2	5.1%	9	11.7%
	Total	141	100.0%	39	100.0%	77	100.0%

c)	Nein	7	5.0%	1	2.6%	6	7.8%
	Eher nein	19	13.5%	7	17.9%	10	13.0%
	Eher ja	34	24.1%	11	28.2%	22	28.6%
	Ja	50	35.5%	18	46.2%	30	39.0%
	Subtotal	110	78.0%	37	94.9%	68	88.3%
	Fehlend	31	22.0%	2	5.1%	9	11.7%
	Total	141	100.0%	39	100.0%	77	100.0%
d)	Nein	4	2.8%	1	2.6%	3	3.9%
	Eher nein	21	14.9%	5	12.8%	13	16.9%
	Eher ja	38	27.0%	20	51.3%	18	23.4%
	Ja	47	33.3%	11	28.2%	34	44.2%
	Subtotal	110	78.0%	37	94.9%	68	88.3%
	Fehlend	31	22.0%	2	5.1%	9	11.7%
	Total	141	100.0%	39	100.0%	77	100.0%
e)	Nein	8	5.7%	0	0.0%	8	10.4%
	Eher nein	17	12.1%	4	10.3%	12	15.6%
	Eher ja	36	25.5%	18	46.2%	15	19.5%
	Ja	49	34.8%	15	38.5%	33	42.9%
	Subtotal	110	78.0%	37	94.9%	68	88.3%
	Fehlend	31	22.0%	2	5.1%	9	11.7%
	Total	141	100.0%	39	100.0%	77	100.0%
f)	Nein	11	7.8%	1	2.6%	10	13.0%
	Eher nein	27	19.1%	12	30.8%	14	18.2%
	Eher ja	40	28.4%	18	46.2%	17	22.1%
	Ja	31	22.0%	6	15.4%	26	33.8%
	Subtotal	109	77.3%	37	94.9%	67	87.0%
	Fehlend	32	22.7%	2	5.1%	10	13.0%
	Total	141	100.0%	39	100.0%	77	100.0%
g)	Nein	15	10.6%	2	5.1%	13	16.9%
	Eher nein	16	11.3%	6	15.4%	8	10.4%
	Eher ja	32	22.7%	15	38.5%	16	20.8%
	Ja	47	33.3%	14	35.9%	31	40.3%
	Subtotal	110	78.0%	37	94.9%	68	88.3%
	Fehlend	31	22.0%	2	5.1%	9	11.7%
	Total	141	100.0%	39	100.0%	77	100.0%
h)	Nein	20	14.2%	3	7.7%	16	20.8%
	Eher nein	28	19.9%	13	33.3%	12	15.6%
	Eher ja	27	19.1%	15	38.5%	11	14.3%
	Ja	35	24.8%	6	15.4%	29	37.7%
	Subtotal	110	78.0%	37	94.9%	68	88.3%
	Fehlend	31	22.0%	2	5.1%	9	11.7%
	Total	141	100.0%	39	100.0%	77	100.0%
i)	Nein	16	11.3%	6	15.4%	7	9.1%
	Eher nein	4	2.8%	1	2.6%	3	3.9%
	Eher ja	3	2.1%	2	5.1%	1	1.3%
	Ja	4	2.8%	1	2.6%	3	3.9%
	Subtotal	27	19.1%	10	25.6%	14	18.2%
	Fehlend	114	80.9%	29	74.4%	63	81.8%
	Total	141	100.0%	39	100.0%	77	100.0%

8. Welchen Mehrwert, alles in allem, erbringt die ÜGK für Ihren Aufgaben- und Zuständigkeitsbereich im Vergleich zu anderen Daten und Informationen im Bildungssystem?

	Alle Befragten		Kantonale Vertreter*innen		Personen aus Forschung und Aus- und Weiterbildung	
	Häufigkeit	Prozent	Häufigkeit	Prozent	Häufigkeit	Prozent
Überhaupt keinen Mehrwert	8	5.7%	1	2.6%	7	9.1%
Einen eher kleinen Mehrwert	47	33.3%	15	38.5%	29	37.7%
Einen eher grossen Mehrwert	39	27.7%	17	43.6%	21	27.3%
Einen sehr grossen Mehrwert	16	11.3%	5	12.8%	11	14.3%
Subtotal	110	78.0%	38	97.4%	68	88.3%
Fehlend	31	22.0%	1	2.6%	9	11.7%
Total	141	100.0%	39	100.0%	77	100.0%

B. Nutzung der ÜGK für den eigenen Arbeits- und Zuständigkeitsbereich

9. Wie häufig nutzen Sie in Ihrem Arbeits- und Zuständigkeitsbereich die nachfolgenden Informationen für die Identifikation von Stärken und Schwächen des Bildungssystems?
a) Ergebnisse aus den nationalen Berichten zur ÜGK
b) Vertiefende Analysen zur ÜGK
c) Ergebnisse aus Erhebungen / Tests im eigenen Kanton
d) Ergebnisse aus Erhebungen / Tests in der eigenen Sprachregion
e) Wissenschaftliche Publikationen im Allgemeinen
*f) Informationen aus Vernehmlassungen oder Beratungen im Kanton oder in der Sprachregion (z.B. 'Runder Tisch', regelmässige Treffen mit Akteur*innen aus der Schulpraxis)*
g) Bildungsberichte Schweiz
h) Andere

		Alle Befragten		Kantonale Vertreter*innen		Personen aus Forschung und Aus- und Weiterbildung	
		Häufigkeit	Prozent	Häufigkeit	Prozent	Häufigkeit	Prozent
a)	Nie oder fast nie	25	17.7%	4	10.3%	20	26.0%
	Selten	48	34.0%	20	51.3%	25	32.5%
	Eher häufig	24	17.0%	11	28.2%	15	19.5%
	Sehr häufig	10	7.1%	1	2.6%	7	9.1%
	Subtotal	107	75.9%	36	92.3%	67	87.0%
	Fehlend	34	24.1%	3	7.7%	10	13.0%
	Total	141	100.0%	39	100.0%	77	100.0%
b)	Nie oder fast nie	38	27.0%	7	17.9%	28	36.4%
	Selten	33	23.4%	14	35.9%	18	23.4%
	Eher häufig	29	20.6%	13	33.3%	17	22.1%
	Sehr häufig	8	5.7%	2	5.1%	5	6.5%
	Subtotal	108	76.6%	36	92.3%	68	88.3%
	Fehlend	33	23.4%	3	7.7%	9	11.7%
	Total	141	100.0%	39	100.0%	77	100.0%
c)	Nie oder fast nie	31	22.0%	3	7.7%	28	36.4%
	Selten	34	24.1%	9	23.1%	22	28.6%
	Eher häufig	27	19.1%	13	33.3%	12	15.6%
	Sehr häufig	14	9.9%	10	25.6%	5	6.5%
	Subtotal	106	75.2%	35	89.7%	67	87.0%
	Fehlend	35	24.8%	4	10.3%	10	13.0%
	Total	141	100.0%	39	100.0%	77	100.0%
d)	Nie oder fast nie	32	22.7%	6	15.4%	25	32.5%
	Selten	48	34.0%	16	41.0%	29	37.7%
	Eher häufig	21	14.9%	11	28.2%	10	13.0%
	Sehr häufig	7	5.0%	3	7.7%	4	5.2%
	Subtotal	108	76.6%	36	92.3%	68	88.3%
	Fehlend	33	23.4%	3	7.7%	9	11.7%
	Total	141	100.0%	39	100.0%	77	100.0%
e)	Nie oder fast nie	8	5.7%	1	2.6%	7	9.1%
	Selten	11	7.8%	5	12.8%	4	5.2%
	Eher häufig	35	24.8%	17	43.6%	15	19.5%
	Sehr häufig	54	38.3%	13	33.3%	42	54.5%
	Subtotal	108	76.6%	36	92.3%	68	88.3%
	Fehlend	33	23.4%	3	7.7%	9	11.7%
	Total	141	100.0%	39	100.0%	77	100.0%
f)	Nie oder fast nie	25	17.7%	1	2.6%	22	28.6%
	Selten	31	22.0%	7	17.9%	24	31.2%
	Eher häufig	33	23.4%	15	38.5%	16	20.8%
	Sehr häufig	17	12.1%	12	30.8%	5	6.5%
	Subtotal	106	75.2%	35	89.7%	67	87.0%
	Fehlend	35	24.8%	4	10.3%	10	13.0%
	Total	141	100.0%	39	100.0%	77	100.0%

g)	Nie oder fast nie	5	3.5%	0	0.0%	5	6.5%
	Selten	19	13.5%	7	17.9%	10	13.0%
	Eher häufig	58	41.1%	21	53.8%	36	46.8%
	Sehr häufig	26	18.4%	8	20.5%	17	22.1%
	Subtotal	108	76.6%	36	92.3%	68	88.3%
	Fehlend	33	23.4%	3	7.7%	9	11.7%
	Total	141	100.0%	39	100.0%	77	100.0%
h)	Nie oder fast nie	7	5.0%	0	0.0%	6	7.8%
	Selten	13	9.2%	6	15.4%	6	7.8%
	Eher häufig	6	4.3%	2	5.1%	4	5.2%
	Sehr häufig	7	5.0%	0	0.0%	5	6.5%
	Subtotal	33	23.4%	8	20.5%	21	27.3%
	Fehlend	108	76.6%	31	79.5%	56	72.7%
	Total	141	100.0%	39	100.0%	77	100.0%

10. Wie häufig nutzen Sie in Ihrem Arbeits- und Zuständigkeitsbereich die nachfolgenden Informationen für Entscheide bezüglich der Qualitätsentwicklung des Bildungswesens?
a) Ergebnisse aus den nationalen Berichten zur ÜGK
b) Vertiefende Analysen zur ÜGK
c) Ergebnisse aus Erhebungen / Tests im eigenen Kanton
d) Ergebnisse aus Erhebungen / Tests in der eigenen Sprachregion
e) Wissenschaftliche Publikationen im Allgemeinen
*f) Informationen aus Vernehmlassungen oder Beratungen im Kanton oder in der Sprachregion (z.B. 'Runder Tisch', regelmässige Treffen mit Akteur*innen aus der Schulpraxis)*
g) Bildungsberichte Schweiz
h) Andere

		Alle Befragten		Kantonale Vertreter*innen		Personen aus Forschung und Aus- und Weiterbildung	
		Häufigkeit	Prozent	Häufigkeit	Prozent	Häufigkeit	Prozent
a)	Nie oder fast nie	31	22.0%	4	10.3%	25	32.5%
	Selten	45	31.9%	19	48.7%	24	31.2%
	Eher häufig	20	14.2%	9	23.1%	12	15.6%
	Sehr häufig	8	5.7%	4	10.3%	4	5.2%
	Subtotal	104	73.8%	36	92.3%	65	84.4%
	Fehlend	37	26.2%	3	7.7%	12	15.6%
	Total	141	100.0%	39	100.0%	77	100.0%
b)	Nie oder fast nie	37	26.2%	8	20.5%	27	35.1%
	Selten	39	27.7%	14	35.9%	24	31.2%
	Eher häufig	20	14.2%	10	25.6%	10	13.0%
	Sehr häufig	7	5.0%	4	10.3%	4	5.2%
	Subtotal	103	73.0%	36	92.3%	65	84.4%
	Fehlend	38	27.0%	3	7.7%	12	15.6%
	Total	141	100.0%	39	100.0%	77	100.0%
c)	Nie oder fast nie	24	17.0%	3	7.7%	21	27.3%
	Selten	33	23.4%	6	15.4%	25	32.5%
	Eher häufig	32	22.7%	15	38.5%	16	20.8%
	Sehr häufig	14	9.9%	12	30.8%	3	3.9%
	Subtotal	103	73.0%	36	92.3%	65	84.4%
	Fehlend	38	27.0%	3	7.7%	12	15.6%
	Total	141	100.0%	39	100.0%	77	100.0%
d)	Nie oder fast nie	26	18.4%	5	12.8%	21	27.3%
	Selten	47	33.3%	16	41.0%	28	36.4%
	Eher häufig	25	17.7%	11	28.2%	15	19.5%
	Sehr häufig	5	3.5%	4	10.3%	1	1.3%
	Subtotal	103	73.0%	36	92.3%	65	84.4%
	Fehlend	38	27.0%	3	7.7%	12	15.6%
	Total	141	100.0%	39	100.0%	77	100.0%

e)	Nie oder fast nie	13	9.2%	0	0.0%	13	16.9%
	Selten	14	9.9%	6	15.4%	6	7.8%
	Eher häufig	34	24.1%	18	46.2%	16	20.8%
	Sehr häufig	41	29.1%	11	28.2%	30	39.0%
	Subtotal	102	72.3%	35	89.7%	65	84.4%
	Fehlend	39	27.7%	4	10.3%	12	15.6%
	Total	141	100.0%	39	100.0%	77	100.0%
f)	Nie oder fast nie	23	16.3%	1	2.6%	21	27.3%
	Selten	27	19.1%	5	12.8%	23	29.9%
	Eher häufig	30	21.3%	16	41.0%	12	15.6%
	Sehr häufig	22	15.6%	14	35.9%	8	10.4%
	Subtotal	102	72.3%	36	92.3%	64	83.1%
	Fehlend	39	27.7%	3	7.7%	13	16.9%
	Total	141	100.0%	39	100.0%	77	100.0%
g)	Nie oder fast nie	14	9.9%	0	0.0%	13	16.9%
	Selten	25	17.7%	10	25.6%	14	18.2%
	Eher häufig	48	34.0%	18	46.2%	31	40.3%
	Sehr häufig	15	10.6%	7	17.9%	7	9.1%
	Subtotal	102	72.3%	35	89.7%	65	84.4%
	Fehlend	39	27.7%	4	10.3%	12	15.6%
	Total	141	100.0%	39	100.0%	77	100.0%
h)	Nie oder fast nie	11	7.8%	0	0.0%	10	13.0%
	Selten	14	9.9%	6	15.4%	7	9.1%
	Eher häufig	4	2.8%	2	5.1%	2	2.6%
	Sehr häufig	3	2.1%	0	0.0%	3	3.9%
	Subtotal	32	22.7%	8	20.5%	22	28.6%
	Fehlend	109	77.3%	31	79.5%	55	71.4%
	Total	141	100.0%	39	100.0%	77	100.0%

11. Können Sie sich Änderungen im Konzept oder in der Vorgehensweise der ÜGK vorstellen, die dazu führen würden, dass Sie die Ergebnisse in Ihrem Aufgaben- und Zuständigkeitsbereich besser nutzen könnten?

	Alle Befragten		Kantonale Vertreter*innen		Personen aus Forschung und Aus- und Weiterbildung	
	Häufigkeit	Prozent	Häufigkeit	Prozent	Häufigkeit	Prozent
Nein	42	29.8%	9	23.1%	31	40.3%
Ja	61	43.3%	26	66.7%	34	44.2%
Subtotal	103	73.0%	35	89.7%	65	84.4%
Fehlend	38	27.0%	4	10.3%	12	15.6%
Total	141	100.0%	39	100.0%	77	100.0%

C. Kommunikation zwischen verschiedenen Akteur*innen

12. Wie gut sind Sie über die ÜGK informiert?
a) über die Ziele der ÜGK
b) darüber, wer insgesamt strategisch für die ÜGK verantwortlich ist
c) über die involvierten Institutionen, die den Auftrag haben, die ÜGK durchzuführen
d) über die Art und Weise der Durchführung der Erhebungen (z.B. Rhythmus, Auswahl Fächer)
e) über die Ergebnisse der ÜGK

		Alle Befragten		Kantonale Vertreter*innen		Personen aus Forschung und Aus- und Weiterbildung	
		Häufigkeit	Prozent	Häufigkeit	Prozent	Häufigkeit	Prozent
a)	Überhaupt nicht gut	6	4.3%	0	0.0%	6	7.8%
	Eher nicht gut	17	12.1%	3	7.7%	14	18.2%
	Eher gut	37	26.2%	11	28.2%	25	32.5%
	Sehr gut	45	31.9%	21	53.8%	20	26.0%
	Subtotal	105	74.5%	35	89.7%	65	84.4%
	Fehlend	36	25.5%	4	10.3%	12	15.6%
	Total	141	100.0%	39	100.0%	77	100.0%
b)	Überhaupt nicht gut	11	7.8%	0	0.0%	11	14.3%
	Eher nicht gut	21	14.9%	6	15.4%	16	20.8%
	Eher gut	34	24.1%	13	33.3%	19	24.7%
	Sehr gut	39	27.7%	16	41.0%	19	24.7%
	Subtotal	105	74.5%	35	89.7%	65	84.4%
	Fehlend	36	25.5%	4	10.3%	12	15.6%
	Total	141	100.0%	39	100.0%	77	100.0%
c)	Überhaupt nicht gut	12	8.5%	2	5.1%	10	13.0%
	Eher nicht gut	26	18.4%	6	15.4%	18	23.4%
	Eher gut	28	19.9%	12	30.8%	16	20.8%
	Sehr gut	39	27.7%	15	38.5%	21	27.3%
	Subtotal	105	74.5%	35	89.7%	65	84.4%
	Fehlend	36	25.5%	4	10.3%	12	15.6%
	Total	141	100.0%	39	100.0%	77	100.0%
d)	Überhaupt nicht gut	12	8.5%	0	0.0%	12	15.6%
	Eher nicht gut	24	17.0%	5	12.8%	18	23.4%
	Eher gut	38	27.0%	17	43.6%	20	26.0%
	Sehr gut	31	22.0%	13	33.3%	15	19.5%
	Subtotal	105	74.5%	35	89.7%	65	84.4%
	Fehlend	36	25.5%	4	10.3%	12	15.6%
	Total	141	100.0%	39	100.0%	77	100.0%
e)	Überhaupt nicht gut	9	6.4%	0	0.0%	9	11.7%
	Eher nicht gut	23	16.3%	1	2.6%	20	26.0%
	Eher gut	35	24.8%	16	41.0%	20	26.0%
	Sehr gut	38	27.0%	18	46.2%	16	20.8%
	Subtotal	105	74.5%	35	89.7%	65	84.4%
	Fehlend	36	25.5%	4	10.3%	12	15.6%
	Total	141	100.0%	39	100.0%	77	100.0%

13. Mit welchen der nachfolgenden Berichte zur ÜGK haben Sie sich in welcher Tiefe auseinandergesetzt?
a) Mit dem nationalen Bericht zur ÜGK 2016, Mathematik 11. Schuljahr
b) Mit einzelnen oder allen weiterführenden Berichten zur ÜGK 2016 (z.B. zur Testentwicklung oder zum Kontextfragebogen)
c) Mit dem nationalen Bericht zur ÜGK 2017, Sprachen 8. Schuljahr
d) Mit einzelnen oder allen weiterführenden Berichten zur ÜGK 2017 (z.B. zur Testentwicklung oder zum Kontextfragebogen)
e) Mit zusätzlichen wissenschaftlichen Publikationen, die mit Daten der ÜGK entstanden sind

		Alle Befragten		Kantonale Vertreter*innen		Personen aus Forschung und Aus- und Weiterbildung	
		Häufigkeit	Prozent	Häufigkeit	Prozent	Häufigkeit	Prozent
a)	Nicht	29	20.6%	2	5.1%	25	32.5%
	Oberflächlich	12	8.5%	1	2.6%	11	14.3%
	Eher oberflächlich	8	5.7%	3	7.7%	3	3.9%
	Eher vertieft	25	17.7%	15	38.5%	10	13.0%
	Vertieft	30	21.3%	14	35.9%	15	19.5%
	Subtotal	104	73.8%	35	89.7%	64	83.1%
	Fehlend	37	26.2%	4	10.3%	13	16.9%
	Total	141	100.0%	39	100.0%	77	100.0%
b)	Nicht	28	19.9%	3	7.7%	23	29.9%
	Oberflächlich	14	9.9%	3	7.7%	10	13.0%
	Eher oberflächlich	22	15.6%	13	33.3%	10	13.0%
	Eher vertieft	23	16.3%	11	28.2%	10	13.0%
	Vertieft	17	12.1%	5	12.8%	11	14.3%
	Subtotal	104	73.8%	35	89.7%	64	83.1%
	Fehlend	37	26.2%	4	10.3%	13	16.9%
	Total	141	100.0%	39	100.0%	77	100.0%
c)	Nicht	23	16.3%	1	2.6%	20	26.0%
	Oberflächlich	10	7.1%	1	2.6%	9	11.7%
	Eher oberflächlich	16	11.3%	3	7.7%	11	14.3%
	Eher vertieft	24	17.0%	14	35.9%	10	13.0%
	Vertieft	30	21.3%	16	41.0%	13	16.9%
	Subtotal	103	73.0%	35	89.7%	63	81.8%
	Fehlend	38	27.0%	4	10.3%	14	18.2%
	Total	141	100.0%	39	100.0%	77	100.0%
d)	Nicht	25	17.7%	2	5.1%	22	28.6%
	Oberflächlich	20	14.2%	4	10.3%	13	16.9%
	Eher oberflächlich	21	14.9%	13	33.3%	9	11.7%
	Eher vertieft	22	15.6%	10	25.6%	11	14.3%
	Vertieft	16	11.3%	6	15.4%	9	11.7%
	Subtotal	104	73.8%	35	89.7%	64	83.1%
	Fehlend	37	26.2%	4	10.3%	13	16.9%
	Total	141	100.0%	39	100.0%	77	100.0%
e)	Nicht	28	19.9%	6	15.4%	19	24.7%
	Oberflächlich	17	12.1%	9	23.1%	8	10.4%
	Eher oberflächlich	24	17.0%	12	30.8%	13	16.9%
	Eher vertieft	20	14.2%	5	12.8%	13	16.9%
	Vertieft	13	9.2%	3	7.7%	10	13.0%
	Subtotal	102	72.3%	35	89.7%	63	81.8%
	Fehlend	39	27.7%	4	10.3%	14	18.2%
	Total	141	100.0%	39	100.0%	77	100.0%

14. Wie beurteilen Sie die Qualität der folgenden Informationen zu den Ergebnissen der ÜGK?
a) Nationaler Bericht Mathematik 11. Schuljahr
b) Nationaler Bericht Sprachen 8. Schuljahr
c) Medienkonferenz
d) Presse
e) soziale Medien
f) Website der EDK
g) Website des Kantons
h) Website der ÜGK Schweiz http://uegk-schweiz.ch

		Alle Befragten		Kantonale Vertreter*innen		Personen aus Forschung und Aus- und Weiterbildung	
		Häufigkeit	Prozent	Häufigkeit	Prozent	Häufigkeit	Prozent
a)	Nicht genutzt / nicht vorhanden	32	22.7%	2	5.1%	28	36.4%
	Ungenügende Qualität	0	0.0%	0	0.0%	0	0.0%
	Genügende Qualität	16	11.3%	8	20.5%	8	10.4%
	Gute Qualität	41	29.1%	21	53.8%	19	24.7%
	Sehr gute Qualität	10	7.1%	4	10.3%	4	5.2%
	Subtotal	99	70.2%	35	89.7%	59	76.6%
	Fehlend	42	29.8%	4	10.3%	18	23.4%
	Total	141	100.0%	39	100.0%	77	100.0%
b)	Nicht genutzt / nicht vorhanden	28	19.9%	1	2.6%	25	32.5%
	Ungenügende Qualität	0	0.0%	0	0.0%	0	0.0%
	Genügende Qualität	14	9.9%	8	20.5%	6	7.8%
	Gute Qualität	44	31.2%	21	53.8%	22	28.6%
	Sehr gute Qualität	11	7.8%	4	10.3%	5	6.5%
	Subtotal	97	68.8%	34	87.2%	58	75.3%
	Fehlend	44	31.2%	5	12.8%	19	24.7%
	Total	141	100.0%	39	100.0%	77	100.0%
c)	Nicht genutzt / nicht vorhanden	51	36.2%	7	17.9%	42	54.5%
	Ungenügende Qualität	5	3.5%	1	2.6%	4	5.2%
	Genügende Qualität	18	12.8%	13	33.3%	6	7.8%
	Gute Qualität	19	13.5%	13	33.3%	5	6.5%
	Sehr gute Qualität	3	2.1%	0	0.0%	0	0.0%
	Subtotal	96	68.1%	34	87.2%	57	74.0%
	Fehlend	45	31.9%	5	12.8%	20	26.0%
	Total	141	100.0%	39	100.0%	77	100.0%
d)	Nicht genutzt / nicht vorhanden	38	27.0%	5	12.8%	32	41.6%
	Ungenügende Qualität	16	11.3%	4	10.3%	12	15.6%
	Genügende Qualität	27	19.1%	16	41.0%	11	14.3%
	Gute Qualität	12	8.5%	9	23.1%	2	2.6%
	Sehr gute Qualität	3	2.1%	0	0.0%	0	0.0%
	Subtotal	96	68.1%	34	87.2%	57	74.0%
	Fehlend	45	31.9%	5	12.8%	20	26.0%
	Total	141	100.0%	39	100.0%	77	100.0%
e)	Nicht genutzt / nicht vorhanden	68	48.2%	18	46.2%	45	58.4%
	Ungenügende Qualität	12	8.5%	7	17.9%	5	6.5%
	Genügende Qualität	10	7.1%	5	12.8%	6	7.8%
	Gute Qualität	3	2.1%	3	7.7%	0	0.0%
	Sehr gute Qualität	1	0.7%	0	0.0%	1	1.3%
	Subtotal	94	66.7%	33	84.6%	57	74.0%
	Fehlend	47	33.3%	6	15.4%	20	26.0%
	Total	141	100.0%	39	100.0%	77	100.0%
f)	Nicht genutzt / nicht vorhanden	34	24.1%	4	10.3%	27	35.1%
	Ungenügende Qualität	7	5.0%	3	7.7%	4	5.2%
	Genügende Qualität	19	13.5%	6	15.4%	11	14.3%
	Gute Qualität	34	24.1%	20	51.3%	15	19.5%
	Sehr gute Qualität	3	2.1%	0	0.0%	1	1.3%
	Subtotal	97	68.8%	33	84.6%	58	75.3%
	Fehlend	44	31.2%	6	15.4%	19	24.7%
	Total	141	100.0%	39	100.0%	77	100.0%

g)	Nicht genutzt / nicht vorhanden	57	40.4%	14	35.9%	39	50.6%
	Ungenügende Qualität	8	5.7%	3	7.7%	4	5.2%
	Genügende Qualität	15	10.6%	7	17.9%	8	10.4%
	Gute Qualität	12	8.5%	8	20.5%	5	6.5%
	Sehr gute Qualität	1	0.7%	0	0.0%	1	1.3%
	Subtotal	93	66.0%	32	82.1%	57	74.0%
	Fehlend	48	34.0%	7	17.9%	20	26.0%
	Total	141	100.0%	39	100.0%	77	100.0%
h)	Nicht genutzt / nicht vorhanden	34	24.1%	7	17.9%	23	29.9%
	Ungenügende Qualität	3	2.1%	2	5.1%	1	1.3%
	Genügende Qualität	23	16.3%	8	20.5%	15	19.5%
	Gute Qualität	33	23.4%	17	43.6%	16	20.8%
	Sehr gute Qualität	4	2.8%	0	0.0%	3	3.9%
	Subtotal	97	68.8%	34	87.2%	58	75.3%
	Fehlend	44	31.2%	5	12.8%	19	24.7%
	Total	141	100.0%	39	100.0%	77	100.0%

15. Müsste die Kommunikation der Ergebnisse durch die Behörden zukünftig verstärkt oder eher reduziert werden?
a) Kommunikation durch die EDK
b) Kommunikation durch die kantonalen Behörden

		Alle Befragten		Kantonale Vertreter*innen		Personen aus Forschung und Aus- und Weiterbildung	
		Häufigkeit	Prozent	Häufigkeit	Prozent	Häufigkeit	Prozent
a)	(Eher) verstärkt	63	44.7%	17	43.6%	44	57.1%
	Gerade richtig	33	23.4%	17	43.6%	14	18.2%
	(Eher) reduziert	3	2.1%	1	2.6%	2	2.6%
	Subtotal	99	70.2%	35	89.7%	60	77.9%
	Fehlend	42	29.8%	4	10.3%	17	22.1%
	Total	141	100.0%	39	100.0%	77	100.0%
b)	(Eher) verstärkt	53	37.6%	10	25.6%	39	50.6%
	Gerade richtig	41	29.1%	23	59.0%	18	23.4%
	(Eher) reduziert	4	2.8%	2	5.1%	2	2.6%
	Subtotal	98	69.5%	35	89.7%	59	76.6%
	Fehlend	43	30.5%	4	10.3%	18	23.4%
	Total	141	100.0%	39	100.0%	77	100.0%

16. Wie stark sind Sie in die Diskussion der Ergebnisse der ÜGK in Bezug auf das Ableiten von bildungspolitischen Entscheiden involviert?

	Alle Befragten		Kantonale Vertreter*innen		Personen aus Forschung und Aus- und Weiterbildung	
	Häufigkeit	Prozent	Häufigkeit	Prozent	Häufigkeit	Prozent
Ich bin nicht involviert	48	34.0%	1	2.6%	43	55.8%
Ich bin involviert, aber nur wenig	25	17.7%	10	25.6%	13	16.9%
Ich bin relativ stark involviert	18	12.8%	13	33.3%	6	7.8%
Ich bin sehr stark involviert	13	9.2%	11	28.2%	2	2.6%
Subtotal	104	73.8%	35	89.7%	64	83.1%
Fehlend	37	26.2%	4	10.3%	13	16.9%
Total	141	100.0%	39	100.0%	77	100.0%

17. Wie stark sind Sie in die Diskussion der Ergebnisse der ÜGK in Bezug auf das Ableiten von Entscheiden für die Schulpraxis involviert?

	Alle Befragten		Kantonale Vertreter*innen		Personen aus Forschung und Aus- und Weiterbildung	
	Häufigkeit	Prozent	Häufigkeit	Prozent	Häufigkeit	Prozent
Ich bin nicht involviert	50	35.5%	3	7.7%	42	54.5%
Ich bin involviert, aber nur wenig	32	22.7%	16	41.0%	15	19.5%
Ich bin relativ stark involviert	16	11.3%	10	25.6%	6	7.8%
Ich bin sehr stark involviert	6	4.3%	6	15.4%	1	1.3%
Subtotal	104	73.8%	35	89.7%	64	83.1%
Fehlend	37	26.2%	4	10.3%	13	16.9%
Total	141	100.0%	39	100.0%	77	100.0%

18. Möchten Sie stärker in die Diskussion der Ergebnisse in Bezug auf das Ableiten von bildungspolitischen Entscheiden involviert sein?

	Alle Befragten		Kantonale Vertreter*innen		Personen aus Forschung und Aus- und Weiterbildung	
	Häufigkeit	Prozent	Häufigkeit	Prozent	Häufigkeit	Prozent
Nein	54	38.3%	20	51.3%	30	39.0%
Ja	49	34.8%	14	35.9%	34	44.2%
Subtotal	103	73.0%	34	87.2%	64	83.1%
Fehlend	38	27.0%	5	12.8%	13	16.9%
Total	141	100.0%	39	100.0%	77	100.0%

19. Möchten Sie stärker in die Diskussion der Ergebnisse in Bezug auf das Ableiten von Entscheiden für die Schulpraxis involviert sein?

	Alle Befragten		Kantonale Vertreter*innen		Personen aus Forschung und Aus- und Weiterbildung	
	Häufigkeit	Prozent	Häufigkeit	Prozent	Häufigkeit	Prozent
Nein	53	37.6%	18	46.2%	30	39.0%
Ja	50	35.5%	16	41.0%	34	44.2%
Subtotal	103	73.0%	34	87.2%	64	83.1%
Fehlend	38	27.0%	5	12.8%	13	16.9%
Total	141	100.0%	39	100.0%	77	100.0%

D. Konzept der ÜGK

20. Welche der nachfolgenden Fachbereiche sollten aus Ihrer Sicht im Rahmen der ÜGK überprüft werden? (mehrere Antworten möglich)

	Alle Befragten	Kantonale Vertreter*innen	Personen aus Forschung und Aus- und Weiterbildung
	Häufigkeit	Häufigkeit	Häufigkeit
Schulsprache	96 (68.1%)	35 (89.7%)	56 (72.7%)
Fremdsprachen	74 (52.5%)	29 (74.4%)	40 (51.9%)
Mathematik	95 (67.4%)	35 (89.7%)	55 (71.4%)
Natur, Mensch, Gesellschaft bzw. Naturwissenschaften	77 (54.6%)	23 (59%)	49 (63.6%)
Gestalten	20 (14.2%)	5 (12.8%)	14 (18.2%)
Musik	24 (17%)	8 (20.5%)	15 (19.5%)
Bewegung und Sport	23 (16.3%)	9 (23.1%)	13 (16.9%)
Subtotal	98	35	58
Fehlend	43 (30.5%)	4 (10.3%)	19 (24.7%)
Total	141	39	77

21. Als wie sinnvoll erachten Sie es, dass die Überprüfung des Erreichens der Grundkompetenzen auf das Ende des 4., 8. und 11. Schuljahres festgelegt ist?
a) Ende des 4. Schuljahres (ca. 8-Jährige)
b) Ende des 8. Schuljahres (ca. 12-Jährige)
c) Ende des 11. Schuljahres (ca. 15-Jährige)

		Alle Befragten		Kantonale Vertreter*innen		Personen aus Forschung und Aus- und Weiterbildung	
		Häufigkeit	Prozent	Häufigkeit	Prozent	Häufigkeit	Prozent
a)	Überhaupt nicht sinnvoll	10	7.1%	2	5.1%	7	9.1%
	Eher nicht sinnvoll	18	12.8%	7	17.9%	12	15.6%
	Eher nicht sinnvoll	37	26.2%	13	33.3%	23	29.9%
	Sehr sinnvoll	34	24.1%	12	30.8%	18	23.4%
	Subtotal	99	70.2%	34	87.2%	60	77.9%
	Fehlend	42	29.8%	5	12.8%	17	22.1%
	Total	141	100.0%	39	100.0%	77	100.0%
b)	Überhaupt nicht sinnvoll	2	1.4%	0	0.0%	2	2.6%
	Eher nicht sinnvoll	7	5.0%	0	0.0%	5	6.5%
	Eher nicht sinnvoll	32	22.7%	10	25.6%	22	28.6%
	Sehr sinnvoll	58	41.1%	24	61.5%	31	40.3%
	Subtotal	99	70.2%	34	87.2%	60	77.9%
	Fehlend	42	29.8%	5	12.8%	17	22.1%
	Total	141	100.0%	39	100.0%	77	100.0%
c)	Überhaupt nicht sinnvoll	1	0.7%	0	0.0%	1	1.3%
	Eher nicht sinnvoll	3	2.1%	0	0.0%	3	3.9%
	Eher nicht sinnvoll	27	19.1%	9	23.1%	17	22.1%
	Sehr sinnvoll	68	48.2%	25	64.1%	39	50.6%
	Subtotal	99	70.2%	34	87.2%	60	77.9%
	Fehlend	42	29.8%	5	12.8%	17	22.1%
	Total	141	100.0%	39	100.0%	77	100.0%

22. Als wie sinnvoll erachten Sie es, dass Bildungsziele jeweils nur als Mindeststandard (Grundkompetenzen) beschrieben sind?

	Alle Befragten		Kantonale Vertreter*innen		Personen aus Forschung und Aus- und Weiterbildung	
	Häufigkeit	Prozent	Häufigkeit	Prozent	Häufigkeit	Prozent
Überhaupt nicht sinnvoll	6	4.3%	2	5.1%	4	5.2%
Eher nicht sinnvoll	32	22.7%	12	30.8%	18	23.4%
Eher nicht sinnvoll	30	21.3%	10	25.6%	16	20.8%
Sehr sinnvoll	31	22.0%	10	25.6%	22	28.6%
Subtotal	99	70.2%	34	87.2%	60	77.9%
Fehlend	42	29.8%	5	12.8%	17	22.1%
Total	141	100.0%	39	100.0%	77	100.0%

23. Reicht es aus Ihrer Sicht, wenn die Tests einzig feststellen, ob die Grundkompetenzen erreicht werden, oder sollten die Tests das gesamte Leistungsspektrum mit mehreren Kompetenzstufen abdecken?
a) Anhand der Tests soll einzig überprüft werden, ob die Grundkompetenzen erreicht worden sind.
b) Die Tests sollen das gesamte Leistungsspektrum mit mehreren Kompetenzstufen abdecken (wie z.B. bei PISA).

		Alle Befragten		Kantonale Vertreter*innen		Personen aus Forschung und Aus- und Weiterbildung	
		Häufigkeit	Prozent	Häufigkeit	Prozent	Häufigkeit	Prozent
a)	Stimme überhaupt nicht zu	21	14.9%	8	20.5%	13	16.9%
	Stimme eher nicht zu	28	19.9%	9	23.1%	19	24.7%
	Stimme eher zu	17	12.1%	5	12.8%	10	13.0%
	Stimme voll und ganz zu	30	21.3%	10	25.6%	17	22.1%
	Subtotal	96	68.1%	32	82.1%	59	76.6%
	Fehlend	45	31.9%	7	17.9%	18	23.4%
	Total	141	100.0%	39	100.0%	77	100.0%
b)	Stimme überhaupt nicht zu	16	11.3%	5	12.8%	10	13.0%
	Stimme eher nicht zu	19	13.5%	8	20.5%	9	11.7%
	Stimme eher zu	27	19.1%	6	15.4%	19	24.7%
	Stimme voll und ganz zu	37	26.2%	14	35.9%	23	29.9%
	Subtotal	99	70.2%	33	84.6%	61	79.2%
	Fehlend	42	29.8%	6	15.4%	16	20.8%
	Total	141	100.0%	39	100.0%	77	100.0%

24. Sollte mittelfristig an einer weitgehenden Umstellung der ÜGK auf ein digitales Assessment, das von den Lehrkräften selbst durchgeführt wird, gearbeitet werden?

	Alle Befragten		Kantonale Vertreter*innen		Personen aus Forschung und Aus- und Weiterbildung	
	Häufigkeit	Prozent	Häufigkeit	Prozent	Häufigkeit	Prozent
Nein	24	17.0%	6	15.4%	17	22.1%
Eher nein	25	17.7%	8	20.5%	14	18.2%
Eher ja	32	22.7%	11	28.2%	20	26.0%
Ja	14	9.9%	7	17.9%	8	10.4%
Subtotal	95	67.4%	32	82.1%	59	76.6%
Fehlend	46	32.6%	7	17.9%	18	23.4%
Total	141	100.0%	39	100.0%	77	100.0%

E. Einbindung der ÜGK in Wissenschaft und Forschung[23]

25. In der Schweiz wurde das Interfaculty Centre for Educational Research (ICER) an der Universität Bern mit der Koordination der ÜGK betraut. Wie sichtbar ist aus Ihrer Sicht das ICER im Kontext der Bildungsforschung in der Schweiz?

	Personen aus Forschung und Aus- und Weiterbildung	
	Häufigkeit	Prozent
Überhaupt nicht sichtbar	12	18.8%
Eher nicht sichtbar	26	40.6%
Eher sichtbar	10	15.6%
Sehr sichtbar	1	1.6%
Subtotal	49	76.6%
Fehlend	15	23.4%
Total	64	100.0%

26. Wie wichtig ist es Ihnen, im Zusammenhang mit Ihrer eigenen Forschung mit den wissenschaftlichen Institutionen der ÜGK (ICER, Aufgabendatenbank) in den Austausch zu treten?

	Personen aus Forschung und Aus- und Weiterbildung	
	Häufigkeit	Prozent
Überhaupt nicht wichtig	9	14.1%
Eher nicht wichtig	13	20.3%
Eher wichtig	15	23.4%
Sehr wichtig	13	20.3%
Subtotal	50	78.1%
Fehlend	14	21.9%
Total	64	100.0%

27. Soll dieser Austausch in Zukunft eher verstärkt oder reduziert werden, oder ist er gerade richtig?

	Personen aus Forschung und Aus- und Weiterbildung	
	Häufigkeit	Prozent
(Eher) verstärkt	24	37.5%
Gerade richtig	24	37.5%
(Eher) reduziert	1	1.6%
Subtotal	49	76.6%
Fehlend	15	23.4%
Total	64	100.0%

[23] Diese Fragen wurden nur den Personen gestellt, die als ihren aktuellen Arbeitsbereich ‘Forschung an Hochschulen oder Forschungsstellen’ angegeben haben: N = 64.

28. Sind Sie interessiert, an der Entwicklung der Leistungstests oder der Konzipierung des Kontextfragebogens mitzuarbeiten?
a) Leistungstests
b) Kontextfragebogen

		Personen aus Forschung und Aus- und Weiterbildung	
		Häufigkeit	Prozent
a)	Überhaupt nicht interessiert	16	25.0%
	Eher nicht interessiert	15	23.4%
	Eher interessiert	8	12.5%
	Sehr interessiert	11	17.2%
	Subtotal	50	78.1%
	Fehlend	14	35.1%
	Total	64	100.0%
b)	Überhaupt nicht interessiert	12	18.8%
	Eher nicht interessiert	12	18.8%
	Eher interessiert	12	18.8%
	Sehr interessiert	13	20.3%
	Subtotal	49	76.6%
	Fehlend	15	23.4%
	Total	64	100.0%

29. Haben Sie die Berichterstattung zu den Ergebnissen der ÜGK als Ausgangspunkt für eigene Forschungsvorhaben genutzt?

	Personen aus Forschung und Aus- und Weiterbildung	
	Häufigkeit	Prozent
Nein	36	56.3%
Ja	11	17.2%
Ist geplant	3	4.7%
Subtotal	50	78.1%
Fehlend	14	21.9%
Total	64	100.0%

30 Haben Sie über FORS Daten der ÜGK für eigene Auswertungen bezogen?

	Personen aus Forschung und Aus- und Weiterbildung	
	Häufigkeit	Prozent
Nein	39	60.9%
Ja	11	17.2%
Subtotal	50	78.1%
Fehlend	14	21.9%
Total	64	100.0%

31. Planen Sie in Zukunft, über FORS Daten der ÜGK für eigene Auswertungen zu beziehen?

	Personen aus Forschung und Aus- und Weiterbildung	
	Häufigkeit	Prozent
Nein	16	25.0%
Eventuell	22	34.4%
Ja	11	17.2%
Subtotal	49	76.6%
Fehlend	15	23.4%
Total	64	100.0%

32. Können Sie sich Änderungen im Konzept oder in den Instrumenten (Tests, Kontextfragebogen) der ÜKG vorstellen, die dazu führen würden...

		Personen aus Forschung und Aus- und Weiterbildung	
		Häufigkeit	Prozent
a) dass mehr oder relevantere Fragen der Bildungsforschung bearbeitet werden können?	Nein	23	35.9%
	Ja	25	39.1%
	Subtotal	48	75.0%
	Fehlend	16	25.0%
	Total	64	100.0%
b) dass eine breitere Einbeziehung der Forscher*innen erreicht werden kann?	Nein	24	37.5%
	Ja	24	37.5%
	Subtotal	48	75.0%
	Fehlend	16	25.0%
	Total	64	100.0%
c) dass Sie die Daten der ÜGK für Ihre eigene Forschung häufiger nutzen können?	Nein	24	37.5%
	Ja	24	37.5%
	Subtotal	48	75.0%
	Fehlend	16	25.0%
	Total	64	100.0%

III. Anhang – Fragekatalog der EDK

Auszug aus:
Gesamtschau zur Überprüfung des Erreichens der Grundkompetenzen (ÜGK). Konzept. Fassung vom 12. Mai 2021

2. Hauptteil: Analyse

Die Grundlage für den zweiten Hauptteil der Gesamtschau bilden

- die Fragen an die Gutachter/-innen,
- die darauf basierende Analyse und die Antworten der Gutachter/-innen,
- Fachliche Stellungnahmen von weiteren externen Expert/-innen, insbesondere von Stefan Wolter, Direktor der Schweiz. Koordinationsstelle für Bildungsforschung SKBF,
- die Diskussionen des Gutachtens, der Stellungnahmen und allgemein von Fragen zur ÜGK im Kosta HarmoS.

Die Auftragsskizze an die Gutachter/-innen (Stand 12.05.2021) umfasst die im Folgenden abgedruckten Fragen, unterteilt nach den Bereichen externe und interne Governance. Sie decken im Wesentlichen die Fragen ab, die bei der Vorbereitung der Gesamtschau aufgeworfen wurden. Im Folgenden sind sie für die Gutachter/-innen kurz kontextualisiert.

2.1 Bereich externe Governance (Bereich A)

A1 Systemsteuerung, Zielkohärenz und Handlungskoordination

Hintergrund: Die bildungspolitischen Diskussionen der Nullerjahre führten zur Festlegung von Bildungszielen («Grundkompetenzen») und zur Konzipierung der ÜGK. Die ÜGK soll die Steuerung bestimmter Aspekte des Bildungssystems ermöglichen sowie insbesondere die gesamtschweizerische Harmonisierung der Bildungsziele über Kantone und Regionen hinweg überprüfen helfen.
Das Projekt ÜGK betrifft zahlreiche Akteur/-innen unterschiedlicher Ebenen: die EDK, die EDK-Regionalkonferenzen, die Kantone, die Schulen; die Geschäftsstelle Aufgabendatenbank (GS ADB) der EDK, das wissenschaftliche Konsortium Bildungsforschung und -evaluation; das Interfaculty Centre for Educational Research (ICER; Uni Bern) und dessen Projektpartner; die Bildungsforschung, die Berufsbildung und weitere. Im Zusammenhang mit dem Ansatz der ÜGK und deren Akteur/-innen stellen sich v.a. folgende Fragen:

Übergeordnete Frage:

A1.1 Ist der gewählte *Systemsteuerungsansatz* aufgrund des Kenntnisstandes des Jahres 2021 *grundsätzlich* noch aktuell? Wenn ja: Was kann dieser Steuerungsansatz konkret für die Qualitätsentwicklung des Bildungssystems leisten, was nicht? Wie kann seine Wirksamkeit gesteigert werden? Welche Mittel zur Systemsteuerung müssten ggf. ergänzt werden? Wenn nein: Was müsste an seine Stelle treten?

Weitere Fragen:

A1.2 Erkennen die Akteur/-innen in den *Zielen der ÜGK*[24] eine Relevanz – insgesamt und für den eigenen Arbeits- oder Zuständigkeitsbereich, die eigenen Aktivitäten? Welchen anderen/alternativen Zielen von Studien im schulischen Bereich räumen die Akteur/-innen ggf. Priorität ein?

A1.3 Inwiefern und aus welchen Gründen werden die Ergebnisse der ÜGK genutzt bzw. nicht genutzt?

A1.4 Wie gut gelingt die *Kommunikation* zwischen dem Projekt ÜGK und verschiedenen anderen Akteur/-innen, namentlich der Bildungspolitik und -administration auf unterschiedlichen Ebenen des

[24] aus dem Organisationsreglement ÜGK, Art. 2 Ziel der ÜGK:

1 Die ÜGK liefert den Kantonen auf der Ebene des Bildungssystems Informationen darüber, inwieweit die Grundkompetenzen auf bestimmten Schulstufen erreicht worden sind. Ziel der ÜGK ist die Evaluation von Leistungen des Bildungssystems der obligatorischen Schule und nicht die Beurteilung einzelner Schulen oder Lehrpersonen.

2 Die Ergebnisse der ÜGK fliessen in die gemeinsame Berichterstattung von Bund und Kantonen im Rahmen des Bildungsmonitorings ein (Artikel 61a BV).

föderalistischen Systems; der Bildungsforschung; den Schulen; der Öffentlichkeit? Wird das Wissen über die ÜGK, deren Funktion und Potential nachhaltig vermittelt? Gibt es genügend Austausch zwischen den Akteur/innen, damit die jeweiligen Anliegen gegenseitig wahrgenommen werden?

A1.5 Als wie gut und nützlich wird die Berichterstattung zu den Ergebnissen der ÜGK (ÜGKs 2016 und 2017) wahrgenommen?

A1.6 Werden Berichterstattung und Daten genutzt (auch für eigene weiterführende, z.B. lokale, Forschungsvorhaben)?

A1.7 Wie können ggf. *Zielkohärenz und Kommunikation* verbessert werden, um den Nutzen der ÜGK als Instrument zur Steuerung des Systems obligatorische Schule zu optimieren?

A2 Die ÜGK und weitere Erhebungen im Bildungssystem

Hintergrund: Die Schweiz beteiligt sich seit dem Jahr 2000 an den PISA-Studien (Kontextfragebögen für Schüler/-innen und Schulleitungen). Eine Beteiligung an PISA-Optionen wie *Financial Literacy* oder *Foreign Language Assessment* oder der *Study of Social and Emotional Skills* (SSES) sowie weiteren internationalen Studien im Bereich obligatorische Schule ist nicht grundsätzlich ausgeschlossen. Die ÜGK umfasst jeweils neben den kognitiven Tests auch einen Schülerfragebogen, der allerdings wegen der Belastung, v.a. der jüngeren Schüler/-innen, im Umfang recht beschränkt ist. Die Variablen und Konstrukte des Fragebogens dienen einerseits der Erklärung der Testergebnisse (Bsp. SES, Geschlecht) und sind andererseits (auch) für sich selbst von Interesse (Bsp. Einstellung zur Schule, IT-Literacy).

Als die ÜGK konzipiert wurde (vgl. Umsetzungsbeschluss der EDK zum HarmoS-Konkordat von 2007), sah man vor, in gesamtschweizersicher Absprache einerseits Referenztests zur Systemevaluation auf nationaler, regionaler und kantonaler Ebene einzuführen (Ziel: Bildungsmonitoring) und andererseits Tests zur individuellen Standortbestimmung (Ziel: individuelle Förderung). Die Instrumente für die Systemevaluation sollten auf nationaler Ebene erarbeitet werden, die Instrumente für die individuelle Standortbestimmung auf sprachregionaler Ebene. Zurzeit ist in den Regionen und Kantonen der Schweiz eine Reihe von externen Schulleistungstests (z.B. *Stellwerk, Checks*, kantonale Prüfungen, Orientierungsarbeiten, Abschlussprüfungen) im Einsatz. Die Kantonsumfrage von IDES (EDK) ermittelt jeweils, welche dieser Tests in den einzelnen Kantonen im Einsatz sind. Die ursprünglich vorgesehene Art von individuellen Standortbestimmungen hat sich nicht umfassend und gleichmässig durchgesetzt.

Im Hinblick darauf stellen sich insbesondere die folgenden Fragen:

A2.8 Welche Leistungen erbringen die Erhebungen und Erhebungsteile auf verschiedenen Ebenen des Bildungswesens (Individuum, Einzelschule, Kanton, Region, Schweiz) für die Systemsteuerung? Die bisherigen Konzeptionen des Zusammenwirkens der Ebenen stammen aus der Zeit vor 2007. Was sind neue Erkenntnisse dazu aufgrund des Forschungsstands von 2021? Welche Folgerungen ergeben sich für den künftigen Einsatz der verschiedenen Erhebungen?

A2.9 Wo bestehen Lücken bei der Beschaffung von Evidenz für die Steuerung des Bildungssystems, und mit welchen Massnahmen bzw. Erhebungen können diese Lücken gefüllt werden?

A2.10 Wie kann der Beitrag der verschiedenen Erhebungen zum schweizerischen Bildungsbericht optimiert werden? Dieser beleuchtet periodisch pro Bildungsstufe die Aspekte *Effectiveness, Efficiency* und *Equity* des Systems.

A2.11 Wie kann die ÜGK optimal zur Dateninfrastruktur beitragen, wie kann sie von der Dateninfrastruktur (auch von den Registerdaten des Bundesamtes für Statistik BFS) profitieren?

2.2 Bereich interne Governance (Bereich B)

Im Bereich der internen Governance sollen strukturell-organisatorische, fachliche und technische Aspekte der Umsetzung des ÜGK-Projekts fokussiert werden.

Der Auditbericht von Fischbach und Ugen (2018) zur ÜGK 2016 im Fachbereich Mathematik konzentrierte sich auf einen Teilbereich, nämlich auf fachliche Aspekte des Testens: Aufgaben- und Testentwicklung; Bezug zwischen der Beschreibung der Grundkompetenzen und dem Test; Prozess der Schwellenwertsetzung. Die Empfehlungen aufgrund der Befunde werden bei der Entwicklung der Aufgaben und Tests für das 4. Schuljahr heute bereits umgesetzt. Die Aufgabenentwicklung wird durch eine Ausweitung des Feedbackprozesses und eine Systematisierung, Standardisierung und ebenfalls eine quantitative

Vergrösserung der Prä-Pilotierung und Pilotierung deutlich verstärkt. Dazu kommen Arbeiten zur Konkretisierung der Beschreibung der Grundkompetenzen im Fachbereich Mathematik. Eine erneute Analyse der Aspekte, die im Auditbericht behandelt wurden, anhand der ÜGKs 2016 und 2017 erscheint deshalb nicht sinnvoll. Vielmehr sollte das Augenmerk auf das Vorgehen bei der Vorbereitung der ÜGK 2024 im 4. Schuljahr gerichtet werden.

B1 Erhebungsdesign

Hintergrund: Bisher hat die EDK für vier Fachbereiche (Schulsprache, Mathematik, Fremdsprachen, Naturwissenschaften) und jeweils das Ende der drei Schulstufen, welche die obligatorische Schule kennt, Bildungsziele in Form von Grundkompetenzen festgelegt (bei den Fremdsprachen nur für die zweite und dritte). Bei den Grundkompetenzen handelt es sich um Leistungsstandards in Form von Mindeststandards. Das Erreichen der Grundkompetenzen wird in grösseren Abständen stichprobenbasiert im Sinn eines schweizweiten Monitorings überprüft. Dabei sollen die Leistungen nicht auf einer Kompetenzskala verortet werden, sondern es wird überprüft, welcher Anteil der Schülerinnen und Schüler die Grundkompetenzen erreicht.

B1.1 Ist es sinnvoll, *vier Fachbereiche* (Schulsprache, Mathematik, Naturwissenschaften und Fremdsprachen), und gerade diese vier Fachbereiche für die Definition von gemeinsamen Bildungszielen und die Überprüfung des Erreichens der Bildungsziele auszuwählen?

B1.2 Ist es sinnvoll, für das Ende des *4., 8. und 11. Schuljahres* (nach HarmoS-Zählung, also für ca. 8-, 12- und 15-Jährige) Ziele zu beschreiben und deren Erreichung auf jeder dieser Stufen zu überprüfen?

B1.3 Ist es sinnvoll, in Kantonen, in denen keine Vollerhebung stattfindet, mit Schulstichproben und darin Schülerstichproben zu arbeiten anstatt mit Klassenstichproben?

B1.4 Ist es sinnvoll, als Bildungsziele jeweils *Mindeststandards* zu beschreiben und nur die Erreichung der Mindeststandards (ja/nein) zu überprüfen?

B1.5 Was ist ein optimaler *Erhebungsrhythmus* für jeden der Fachbereiche? Soll mit Längsschnitten gearbeitet werden (für die ÜGK 2020 im 11. Schuljahr war eine Stichprobe aus derselben Population vorgesehen, die bereits 2017 im 8. Schuljahr getestet wurde).

B1.6 Müssen die Erhebungen zur ÜGK in allen Kantonen zwingend in derselben Zeitperiode und jeweils in denselben Fächern durchgeführt werden? Sind flexiblere Modelle denkbar?

B2 Projektorganisation und institutionelle Verankerung

Hintergrund: Im Jahr 2018 wurde die Governance der ÜGK grundlegend verändert, sodass sich eine Begutachtung der früheren Strukturen (vgl. dazu v.a. das Konzept zur ÜGK von 2013) für sich allein genommen erübrigt. Das GS EDK zog sich aus der Umsetzungsebene zurück, indem es mit der Universität Bern eine umfassende, projektbezogene Kooperationsvereinbarung «betreffend die Zusammenarbeit zur Durchführung der ÜGK-Erhebungen in den Jahren 2020 und 2022 und zur kontinuierlichen wissenschaftlichen Koordination der Überprüfung des Erreichens der Grundkompetenzen (ÜGK)» abschloss. Das von der Universität Bern für die Umsetzung der Kooperation eingerichtete Interfaculty Centre for Educational Research (ICER) kooperiert seinerseits im Rahmen des Organisationsreglements für die Durchführung der ÜGK aus dem Jahr 2014 mit verschiedenen Institutionen eines wissenschaftlichen Konsortiums, um die ÜGK durchzuführen und wissenschaftlich einzubetten. Um Synergien nutzen zu können, wurde die Umsetzung und Koordination des PISA-Zyklus 2021 bzw. 2022 ebenfalls dem ICER übertragen. Dieses kooperiert auch für PISA mit dem wissenschaftlichen Konsortium und mit seinen Projektpartnern (Durchführungsebene). Die Koordination eines wissenschaftlichen Konsortiums zur Einbindung der Large-Scale-Assessments in die Wissenschaft sowie für die wissenschaftliche Vernetzung gehört zu den Aufgaben des ICER. Für die Umsetzung der Studien koordiniert das ICER die Projektleitung ÜGK bzw. PISA, die sich aus Vertreter/-innen der regionalen Durchführungszentren, des IBE (zuständig für die Stichprobenziehung), der PH GR (Datenmanagement) und der GS ADB (nur ÜGK) zusammensetzt.

Die GS ADB hat bei der Umsetzung der ÜGK eine Sonderstellung inne, indem sie strategisch und fachlich direkt dem GS EDK unterstellt ist. Zu ihren Aufgaben gehören insbesondere die Aufgaben- und Testentwicklung in Zusammenarbeit mit Fachdidaktiker/-innen und weiteren Expert/-innen, die Skalierung der Testergebnisse und die Schwellenwertsetzung. Die fachliche Anbindung an das GS EDK hat insbesondere damit zu tun, dass die ADB «Hüterin der Grundkompetenzen» ist. Das Organisationsreglement

für die ADB aus dem Jahr 2014 sieht im Zusammenhang mit Testaufgaben, Testskalen und Schwellenwerten auch weitergehende Dienstleistungen im Bildungssystem Schweiz vor (vgl. die Bezeichnung Aufgabendatenbank). Diese Funktionen sind allerdings gegenüber dem Beitrag zur ÜGK in den Hintergrund getreten und werden nicht ausgeübt. Organisatorisch und räumlich ist die GS ADB der Schweizerischen Koordinationsstelle für Bildungsforschung (SKBF) zugeordnet.

In Bezug auf die organisatorischen Aspekte stehen die folgenden Fragen im Vordergrund:

B2.7 Wie sind die Effektivität, Effizienz und Nachhaltigkeit der gewählten Organisationsstruktur im Hinblick auf die Durchführung der ÜGK einzuschätzen?

B2.8 Wie sind speziell die Positionierung und Anbindung der GS ADB einzuschätzen? Sollen Alternativen zur aktuellen Struktur gesucht werden? Welche wird ggf. empfohlen? Soll der ursprünglich (im Organisationsreglement) vorgesehene Funktionsumfang beibehalten und praktisch umgesetzt werden?

B2.9 Wie ist die Einbindung der Fachdidaktiken der getesteten Fachbereiche einzuschätzen, die nicht direkt Teil des wiss. Konsortiums und der Projektleitung ÜGK sind, sondern (bloss) bei der Aufgaben- und Testentwicklung mit der GS ADB zusammenarbeiten?

B2.10 Wie gut gelingen die Einbindung der ÜGK in Wissenschaft und Forschung sowie die Zusammenarbeit mit wissenschaftlichen Institutionen (vgl. Koordinationsauftrag des ICER)? Wie könnte die ÜGK von wissenschaftlichen Netzwerken und wissenschaftlicher Zusammenarbeit weiter profitieren?

B3 Fachliche und technische Aspekte

In Bezug auf die Projektumsetzung stehen verschiedene fachliche und technische Fragen im Raum.

B3.11 Bei der Testentwicklung für die ÜGK 2024 im 4.Schuljahr wurden aufgrund des Gutachtens von Fischbach & Uger (2018) Verbesserungen vorgenommen. Kann dieses Vorgehen als Muster für zukünftige ÜGK-Tests gelten?

B3.12 Die ADB leistet die Aufgaben- und Testentwicklung in Zusammenarbeit mit Fachdidaktiker/-innen aus den verschiedenen Sprachregionen (oft PH-Dozent/-innen). Anderswo leisten Personen mit anderem Hintergrund die Testentwicklung (z.B. Lehrpersonen). Gibt es für die ÜGK empfehlenswerte, glaubwürdige Alternativen zur aktuellen Praxis?

B3.13 Die Grundkompetenzen für die verschiedenen Fachbereiche sind auf unterschiedliche Art und Weise beschrieben, mehr oder weniger abstrakt, mit mehr oder weniger Bezug zu Aufgaben- und Leistungsmerkmalen. Welche Art der Beschreibung eignet sich besonders, um die erwarteten Grundkompetenzen allgemein verständlich und zugleich auf die Testskala beziehbar zu machen?

B3.14 Bei der Schwellenwertsetzung wird das Bookmarkverfahren eingesetzt. Verschiedentlich wurden die gesetzten Schwellen in Zweifel gezogen. Was wird empfohlen, um die Glaubwürdigkeit der Schwellenwertsetzungen zu erhalten bzw. zu erhöhen?

B3.15 Die ÜGK-Tests fokussieren auf den Schwellenwert, d.h. die Items zielen idealerweise auf Anforderungen, die noch im Bereich der Grundkompetenzen und leicht darüber liegen. Da in der ÜGK das Erreichen der Grundkompetenzen überprüft wird, werden die Kompetenzen von viel stärkeren Schüler/-innen nicht zuverlässig getestet (*ceiling effect*), was es beispielsweise erschwert, Aussagen zu Bezügen zwischen Testergebnissen und erklärenden Variablen zu machen. Inwiefern ist ein Festhalten am Testen von Grundkompetenzen angezeigt? Was sind ggf. sinnvolle Alternativen (vgl. Frage B 1.4)?

B3.16 Schülerfragebögen müssen oft in kurzer Erhebungszeit bei Schüler/-innen im Alter von ca. 8 bis 15 Jahren alle Informationen erfassen, die für die Kontextualisierung der Testergebnisse relevant sind und nicht anders besorgt werden können. Im Hinblick auf die Systemsteuerung und aufgrund von legitimen Interessen von Forschenden besteht ein Interesse an weiteren Variablen und Konstrukten. Zudem möchte man gleiche bzw. analoge Informationen über die ganze Altersspanne hinweg erheben können. Wie ist das Vorgehen einzuschätzen, das im ÜGK-Konsortium zurzeit verfolgt wird? Welche bewährten Alternativen gibt es?

B3.17 Bei den ÜGK-Erhebungen gehen Testadministrator/-innen mit Tablets in die Schulen und leiten die Testsitzungen. Das ist relativ teuer und von den Gruppendynamiken her nicht immer günstig. Wie sind bei Schüler/-innen unterschiedlichen Alters die Alternativen dazu einzuschätzen (besonders

die Durchführung durch Lehrpersonen an schuleigenen Computern)? Was sind die Gelingensbedingungen für Alternativen? Was für ein Testsetting führt zu hoher Testmotivation und Leistungsbereitschaft der Schüler/-innen, obschon keine individuelle Rückmeldung vorgesehen ist?

B3.18 Zu den ÜGK-Erhebungen wurden technische Berichte und Dokumentationen erstellt. Sind diese Berichte geeignet, die wissenschaftliche Güte der eingesetzten Instrumente aufzuzeigen?

B3.19 Für die ÜGK-Daten gibt es ein Datennutzungskonzept, das letztmals im Mai 2021 erneuert wird. Die in der ÜGK erhobenen Daten und die freigegebenen Instrumente werden (über FORSbase) der wissenschaftlichen Community zur Verfügung gestellt. Befriedigen das Datennutzungskonzept und die zur Verfügung gestellten Daten und Instrumente die Bedürfnisse der Bildungsforschung und der Testforschung?

Anpassung je nach Gebiet sinnvoll. Sie erinnert zudem daran, dass die Empfehlungen Nr. 1 bis Nr. 4 den Kindergarten / vorschulischen Bereich betreffen.

Zu Empfehlung Nr. 5

Die *Präsidentin* weist darauf hin, dass auch hier eine Unterscheidung nach Gebiet erforderlich ist.

Die *Generalsekretärin* entschuldigt sich für die kurze Konsultationsfrist, das Generalsekretariat ist fälschlicherweise davon ausgegangen, dass – wie bei der SODK – keine Konsultation bei den Mitgliedern notwendig sei. Die zeitliche Dringlichkeit ist auf die parallele Behandlung der parlamentarischen Initiative zurückzuführen. Sie ist dankbar, wenn die Frist eingehalten werden könne. Der politische Druck sei nämlich gross.

C. Schäli ergänzt, dass vor diesem Hintergrund eine kurze Frist sinnvoll ist.

Beschluss | *Décision*

Der Beschlussentwurf wird einstimmig gutgeheissen. | *Le projet de décision est accepté à l'unanimité.*

4. Überprüfung des Erreichens der Grundkompetenzen; Gesamtschau: Information und Aussprache | *Vérification de l'atteinte des compétences fondamentales; vision d'ensemble: information et discussion*

Die *Präsidentin* erinnert daran, dass die Plenarversammlung den Koordinationsstab HarmoS beauftragt hat, eine Gesamtschau zur Überprüfung der Grundkompetenzen zu erstellen. Aufgrund dieser Auslegeordnung sollen dann die weiteren Entscheide getroffen werden. Offenbar ist im Kosta HarmoS nun eine grundlegende Debatte entbrannt, weshalb die Generalsekretärin um eine Aussprache bittet. Die ÜGK ist seit Beginn ein umstrittenes Instrument. Einerseits soll damit die Entwicklung der Harmonisierungsbestrebungen der Kantone dokumentiert werden. Andererseits sollen die Erhebungen als Teil des Bildungsmonitorings die Qualitätsentwicklung vorantreiben. Nach den zwei ersten Durchgängen haben wir gesehen, dass mit dem heutigen Instrument eigentlich nur Kantonsvergleiche möglich sind. Das ist nicht befriedigend. Nun wird es darum gehen, aufgrund von Varianten zu entscheiden, wie es mit ÜGK weitergehen soll. Die Verwaltung muss wissen, in welche Richtung sie arbeiten soll. Das Ziel der heutigen Aussprache ist, dass wir der Generalsekretärin, die auch Präsidentin des Kosta HarmoS ist, eine Stossrichtung mitgeben können.

Die *Generalsekretärin* ruft mit Verweis auf Art. 62 Abs. 4 der BV sowie Art. 7 und 10 (Standards und Grundkompetenzen) des Konkordats die Grundlagen und Eckwerte der ÜGK in Erinnerung. Anlässlich der letzten Kosta HarmoS Sitzung hätte die Gesamtschau zuhanden des Vorstands verabschiedet werden sollen. Aufgrund der Diskussion im Kosta HarmoS stellte sich jedoch grundsätzlich die Frage, in welche Richtung die EDK gehen wolle.

Die *Präsidentin* fasst die zu klärenden Fragen zusammen: Wie ist die politische Haltung? Wie kommen wir auf vernünftige Daten, ohne das Schulsystem zu überhitzen? In welchem Umfang sind solche Erhebungen erforderlich?

C. Schäli betont, dass die Schweiz eines der teuersten Bildungssystemen weltweit hat. Gleichzeitig sind wir nicht im Stande die Frage zu beantworten, ob wird die Bildungsziele erfüllen. Es braucht daher eine solche nationale Bildungsmessung. Wenn wir «nein» zur ÜGK sagen, muss man sich auch bewusst sein, dass 20 Jahre Arbeit in den Sand gesetzt werden.

S. Kölliker betont, rückblickend müsse man nicht allzu viel ausbreiten. Er hat das Gefühl, dass zu Beginn Fehler gemacht wurden, insbesondere bei den Grundlagenarbeiten. Es wurde bisher mit viel Einsatz und Geld viel geleistet. Die Grundfrage der Leistungserhebungen stellt sich sowohl auf internationaler als auch auf nationaler Ebene. Im internationalen Vergleich empfindet er diese als nicht so wichtig. National haben sich Monitoringberichte etabliert. Diese werden sowohl von der Politik als auch von der Gesellschaft gefordert. Es braucht daher einen Vergleich. Die Auswahl der Varianten ist einzugrenzen. Ein massiver Ausbau kommt nicht in Frage. Dies könne man gegenüber der Bevölkerung nicht verantworten. Bisher war

die Aussagekraft klein. Mit Bezug auf ein vom Kanton Zürich zur Verfügung gestelltes Dokument kommt er zum Schluss, dass nur zwei Varianten denkbar sind: Eine minimale Anpassung der ÜGK oder den Vorschlag von *S. Wolter*, welcher eine Variante Pisa Plus vorschlägt.

C. Häsler unterstützt die vorgehenden Voten und dankt für die Gesamtschau, welche sehr hilfreich sei. Auch der Kanton Bern unterstützt den Vorschlag des Kantons Zürich. Die Erhebung ist für die Qualitätssicherung erforderlich. Es stellt sich jedoch die Frage nach dem Nutzen. Varianten mit massiven Kostensteigerungen kommen nicht in Frage. Aus Sicht des Kantons Bern kommen daher auch Variante 2 und 3 des Zürcher Papiers nicht in Frage. Sie betont, dass gute Anpassungen wichtig sind, das nächste Mal soll dies aber nicht im letzten Moment erfolgen. Den Kontextfragebogen hatte der Vorstand diskutiert, konnte dann aber keine Anpassungen beschliessen. Ob die Pisa-Variante eine Möglichkeit darstellt, muss mit Blick auf die Kosten geprüft werden. Die politische Diskussion über die Vermessung der Bildung entsteht immer wieder. Dem muss man Rechnung tragen. Daher müssen Alternativen sorgfältig abgewogen und geprüft werden. Der Kanton Bern kann sich mit Bezug auf das Zürcher Papier eine Variante 1 oder eine Variante Pisa Plus mit Stichproben, wie von *S. Wolter* vorgeschlagen, vorstellen.

C. Amarelle remercie pour le travail, l'audit et les recommandations faites. Elle rappelle que la vérification des compétences fondamentales (cofo) permet d'affiner le dispositif propre au canton. Ce monitorage fonctionne comme aiguillon d'amélioration de l'école au niveau romand et cantonal. Elle indique que grâce à la vérification des cofo, son canton a pu perfectionner son système d'enseignement en mathématiques. Elle se prononce en faveur de la variante 2 ou 3 avec les tests PISA qui resteraient complémentaires aux tests cofo. Elle tient à souligner que les cantons romands font figure de bons élèves en ce qui concerne les tests cofo et qu'en cas d'abandon ils se verraient dans l'obligation de refaire des épreuves communes romandes, ce qui seraient un véritable retour en arrière. Concernant les coûts, elle remarque que la Suisse peut se permettre une telle dépense et indique que ce n'est pas le système le plus cher si l'on compare à l'international.

M. Bertoli se dit toujours en faveur d'avoir des données comparables pour permettre au système de s'améliorer. Concernant les variantes, il souhaiterait des propositions plus détaillées. Il remarque que son canton a maintenu le système PISA afin d'avoir des comparaisons avec les pays proches. Que ce soit des comparaisons internes, ou avec l'utilisation de PISA ou PISA Plus, les deux propositions sont à son avis très bien.

M. Knill stellt fest, dass ihr das Dokument aus dem Kanton Zürich nicht vorliegt. Sie könne sich somit auf keine Varianten beziehen. Sie teile die Meinung, dass man auf die ÜGK nicht verzichten kann. In einer sich wandelnden Welt muss man sich überlegen, welchen Nutzen wir daraus ziehen. Der Kontext der Schule und der Schüler und Schülerinnen verändert sich auch aufgrund der Digitalisierung. Müsste man die Grundkompetenzen dynamischer gestalten? Anderenfalls müssten Unterschiede ständig erklärt werden. Es sei auch eine Überlegung wert, ob man nicht auf eine Onlineversion einschwenken könnte, analog zu Pisa, auch um Kosten zu sparen. Sie begrüsst daher irgendeine Form der Überprüfung, aber sieht noch nicht, was das Endprodukt sein sollte.

M. Courtet indique qu'il soutient les propos de ses collègues de la CIIP. Il se prononce en faveur de la variante 2 qui est ambitieuse, ou alors éventuellement la variante 3 mais dans tous les cas en maintenant les tests PISA.

Die *Präsidentin* stellt fest, dass alle der Meinung sind, es brauche ein Instrument zur Leistungsmessung. Sie habe kein Votum gehört, wonach die ÜGK beerdigt werden soll. Es seien einige kritische Stimme im Zusammenhang mit dem Kontextfragebogen entstanden und es besteht breiter Konsens darüber, dass die ÜGK optimiert werden muss. Es muss geklärt werden, ob beim Verfassen der Kontextfragebögen die politische Sicht gefragt ist. Sie ist der Meinung es handelt sich dabei um wissenschaftliche Fragen. Auch der Begriff der «Grundkompetenzen» muss neu definiert werden. Dies würde gemäss dem Zürcher Papier der Variante 1 entsprechen.

S. Kölliker fügt mit Verweis auf Variante 4 des Zürcher Papiers hinzu, dass diese Auslegung nicht dem HarmoS Konkordat entspricht. Er bitte darum, dass die Überlegungen von *S. Wolter* vertieft überprüft werden. Erst dann sei ein Beschluss möglich.

Die *Präsidentin* ergänzt, dass es sich bei Variante 1 des Zürcher Papiers um minimale Anpassungen der ÜGK handelt, Variante 2 sei ein Vorschlag der Universität Zürich und Variante 4 entspricht dem Vorschlag von *S. Wolter*, wobei hier noch das Problem der übrigen Stufen und der Fremdsprachen zu klären wäre.

F. Modoux betont als Mitglied der Kosta HarmoS, dass *S. Wolter* nicht die Variante 4 befürworte. Er habe vielmehr gesagt, dass Variante 2 die beste sei, aber auch die kostspieligste. Das Generalsekretariat hat aber noch keine genauen Berechnungen vorgenommen. Weiter habe er die Haltung vertreten, dass wenn man nicht konsequent in ÜGK investieren will, eine Erweiterung von Pisa in Frage käme. *S. Wolter* habe bedauert, dass die Schweiz so wenig investiert hat.

Die *Präsidentin* bedankt sich für diese Ergänzungen und bestätigt, dass sie *S. Wolter* auch so verstanden habe.

Die *Generalsekretärin* erinnert daran, dass es darum gehe, das weitere Vorgehen zu diskutieren. Die Varianten, die nun diskutiert werden, liegen ihr nicht vor und sie könne dazu nicht Stellung nehmen. Alle werden jedoch Kostenfolgen haben. Sie entnehme der Diskussion, dass diese Varianten substantiiert und den Mitgliedern vorgelegt werden sollen. Im Zusammenhang mit dem Kontextfragebogen werde sie einen Vorschlag unterbreiten.

Die *Präsidentin* betont, dass eine Mitgliedschaft im Kosta HarmoS die Verwaltung nicht von der Mitsprache entbinde. Eine solche Vernetzung ist erwünscht.

A. Gwerder weist darauf hin, dass heute eine Orientierung gelungen ist. Aus Sicht eines Kantons, der dem HarmoS Konkordat nicht beigetreten ist, steht er hinter der ÜGK. Der Kanton NW befürwortet zwei Reduktionen (Fächer und Stufen) und eine Erweiterung bei der Taktung (Testung).

Die *Präsidentin* weist darauf hin, dass im Generalsekretariat nur eine Person mit der Aufgabe betraut ist. Wenn detaillierte Analysen erforderlich sind, ist zu bedenken, dass das System überhitzt wird.

G. Küpfer bittet darum, bei der Überprüfung der Varianten den Nutzen in den Vordergrund zu stellen.

Die *Präsidentin* stimmt dem Votum zu. Die Steuerung sei ein wichtiger Punkt. Das Generalsekretariat muss sich von den Varianten lösen und andere erarbeiten können.

S. Kölliker fragt nach, wie nun der Auftrag für die weitere Zusammenarbeit laute. Es gab eine klare Mehrheit und nun ist der Varianten-Fächer wieder offen. Variante 2 ist offensichtlich keine Option.

Die *Generalsekretärin* betont, sie habe verstanden, dass eine Variante mit grossen Kostenfolgen keine Option sei. Die übrigen Varianten würden nun geprüft. Sie befürchtet, dass ÜGK ohne Mehrkosten gleichbedeutend ist mit dem Status quo. Sie weist auf die unterschiedlichen sprachregionalen Haltungen hin, die bei der Lösungsfindung berücksichtig werden müssen.

Auf die Rückmeldung von *S. Kölliker*, dass die Kantone minimale Anpassungen wollen, fügt die *Präsidentin* hinzu, dass die Anpassungen von *S. Wolter* nicht so minimal sind. Zum jetzigen Zeitpunkt ist man nicht entscheidungsfähig, das war auch nie das Ziel dieser Diskussion.

5. Teilnahme der Schweiz an PISA 2025: Beschlussfassung zuhanden der Plenarversammlung | *Participation de la Suisse à PISA 2025: décision à l'attention de l'Assemblée plénière*

Die *Generalsekretärin* führt in das Geschäft ein.

M. Knill unterstützt fast alle Anträge. Zur Option Englisch muss man sich jetzt im Klaren sein, ob man diese will oder nicht und was man daraus machen kann. Eine interessante Frage ist, ob unsere Kinder mit anderen Ländern mithalten können. Wenn wir unterdurchschnittlich sind, dann wird man Englisch ausbauen wollen zum Nachteil von Französisch. Das muss uns bewusst sein. Unsere Schulen sind durch Testungen schon genug belastet. Diese Option ist nicht hochprioritär, man kann darauf verzichten.

Die *Präsidentin* teilt diese Einschätzung. Nicht zu vergessen ist, dass es ein Schülersegment gibt für das Englischkenntnisse wichtig sind. Es ist auch die Sprache der Wissenschaft. Allerdings geht das zu Lasten unserer Mehrsprachigkeit.