J'en suis aussi sûr que vous, mais pas avec le même pourcentage de chances, que ce soit hors contexte ou en contexte

Deux études sur la variabilité inter-individus des significations métriques données aux degrés de certitude verbaux

Dieudonné Leclercq
Université de Liège, Département Education et Formation

Résumé
La façon la plus répandue d'utiliser les degrés de certitude dans les évaluations scolaires consiste à demander aux étudiants de le faire avec des degrés définis par des mots tels que « pas sûr du tout », « très peu sûr », « peu sûr », « sûr », « très sûr » et « extrêmement sûr ». Cette recherche a étudié les significations en pourcentages de chances (en %) d'être correct que diverses personnes adultes donnent à de telles expressions verbales. Pour ce faire, deux expériences très faciles à répliquer ont été menées, l'une hors contexte et l'autre en contexte. Les résultats permettent de conforter 5 hypothèses : (1) une préférence pour les multiples de 10%, (2) une répétabilité intra-individuelle très élevée, (3) une sensibilité à la difficulté des questions, (4) de fortes différences interindividuelles et (5) une similitude des résultats hors contexte et en contexte. Les sources théoriques et expérimentales de chacune de ces hypothèses sont exposées avant la présentation des résultats.

Mots-clés
Degrés de certitude en pourcentages, sûr, expressions verbales du doute, contexte, répétabilité

Keywords
Confidence degrees in percentages, sure, verbal expressions of doubt, context, repeatability

Pour citer cet article : Leclercq, D. (2016). J'en suis aussi sûr que vous, mais pas avec le même pourcentage de chances, que ce soit hors contexte ou en contexte. *Evaluer. Journal international de Recherche en Education et Formation*, *2*(1), pp. 89-125.

1. Les contenus, les buts et la structure de cet article

1.1 Les degrés de certitude pour exprimer le doute ou l'assurance

Dans le présent article, j'appelle « recourir aux degrés de certitude en évaluation scolaire » le fait, pour un enseignant, d'inviter chaque étudiant à accompagner chacune de ses réponses à un test par une estimation (subjective) de la probabilité que cette réponse sera jugée correcte par le correcteur (souvent l'enseignant lui-même). C'est cette situation qui a été appliquée dans l'expérience en contexte.

On peut s'étonner de ce que le recours aux degrés de certitude soit une pratique assez rare dans l'enseignement. Et ce, en dépit des bénéfices potentiels que pourraient en retirer les apprenants et les enseignants, que ce soit dans des évaluations à visée formative ou à visée sanctionnante (certificative ou sélective). Bon nombre d'enseignants n'ont jamais pensé les utiliser, notamment parce qu'ils ne les ont jamais vécus eux-mêmes. D'autres s'en méfient, et avec de bonnes raisons. D'autres enfin – les plus rares- y recourent, mais bien souvent avec des méthodes que j'ai pratiquées moi-même pendant de nombreuses années et que je trouve aujourd'hui inadéquates, mais heureusement, corrigeables.

Ma thèse est que les situations de méfiance et de mésusage sont causées par une série d'erreurs conceptuelles et méthodologiques. Parmi elles, l'absence de préoccupation pour le réalisme (ou non) des degrés de certitude avancés (par comparaison avec la réalité) ou encore les barèmes de tarifs, c'est-à-dire les règles d'attribution de points qui peuvent amener les étudiants à fournir des degrés de certitude qui ne correspondent pas à leur intime conviction. Je ne m'attaquerai ici qu'à une autre erreur encore, que je considère comme la toute première, à savoir recueillir les degrés de certitude via des échelles de mots au lieu d'échelles métriques comme les probabilités (subjectives évidemment) ou les pourcentages de chance (tout aussi subjectifs) de fournir la réponse correcte.

1.2 Le but de cet article

Je fais cinq hypothèses à propos de ces traductions de mots vers des pourcentages :

1. les répondants ont une préférence pour les multiples de 10% ;
2. la répétabilité intra-individuelle des traductions est très élevée ;
3. en contexte, la distribution des degrés de certitude (verbaux et en pourcentages) est sensible à la difficulté des questions ;
4. pour chacun des 6 degrés verbaux, l'ampleur des différences interindividuelles est importante, introduisant ainsi dans les données une erreur aléatoire que l'on pourrait appeler un coefficient de brouillard ou de brouillage ;
5. il existe une forte similitude des résultats hors contexte et en contexte.

La justification de chacune de ces hypothèses sera fournie dans les sections qui se rapportent à chacune d'entre elles.

L'enjeu majeur, pour le recours aux degrés de certitude en contexte scolaire, est que, même si se confirmait une grande stabilité intra-individus, la grande variabilité inter-individus remet en cause la formulation des degrés de certitude en mots. Autrement dit, quand une personne (un étudiant par exemple) dit à une autre (un enseignant par exemple) « j'en suis sûr », le récepteur ne sait pas comment interpréter ce message de l'émetteur.

On trouve une abondance de recherches expérimentales sur les différences inter-individus de correspondance entre les mots et les pourcentages. A la revue de cette littérature, qui occupera un autre article en préparation, les publications existantes me sont apparues lacunaires sous trois aspects.

1.3 Lacune 1 : concernant le mot « sûr » et ses variations

Je n'ai trouvé aucune de ces recherches (sur la traduction de mots vers des pourcentages) qui porte sur l'usage, dans l'échelle verbale, du seul adjectif « sûr », accompagné d'adverbes modérateurs (« pas sûr du tout », « peu sûr », « très peu sûr ») et amplificateurs (très sûr », extrêmement sûr »). Or c'est probablement la modalité verbale la plus pratiquée en contexte scolaire par les enseignants qui découvrent (ou réinventent) les degrés de certitude, mais ne publient pas leurs travaux.

1.4 Lacune 2 : Le manque d'expériences en contexte « scolaire »

La majorité de ces recherches sur les correspondances mots-pourcentages ont été menées soit hors contexte, soit dans des contextes assez différents de la situation scolaire où un étudiant accompagne une réponse à une question par un degré de certitude, et où cette question a une solution correcte.

Ainsi, plusieurs auteurs ont étudié l'impact du contexte sur les traductions de degrés de certitude verbaux en pourcentage. Par « contexte », dans les deux expériences ci-après, il faut entendre le contenu de l'événement dont la probabilité est à estimer.

Par exemple, Johnson (1973) a étudié la traduction mots-pourcentages dans 3 contextes (3 contenus) différents : une prévision météo, une prédiction de succès personnel et une note émanant d'un service (secret) de renseignement. Il n'a pas observé de différence dans les traductions selon ces trois contextes. Par contre, Fabre (1991, 1993a et 1993b) et Fares (2006), eux, ont observé des différences. Fares (2006, p. 103) a démontré, dans une étude menée auprès de 59 étudiants, qu'« une expression comme '*il est sûr que*…' est traduite en nombres (allant de 0 à 10) différemment selon les trois contextes suivants : (1) une invasion imminente d'extra-terrestres, (2) la direction (à gauche ou à droite) que va prendre une voiture ou (3) le destin d'un arbre (il mourra un jour). »

A la différence de ces deux études, dans la deuxième expérience (du jeu-test Compar-Aires) que je décris ci-après, il existe une réponse correcte à la question posée. Ce qui rend ce jeu-test comparable avec bien des situations scolaires.

1.5 Lacune 3 : La répétabilité des traductions

La plupart des recherches sur la fiabilité des degrés de certitude prennent le terme « fiabilité » dans le sens de «calibration » (conformité à la réalité). Or, une autre forme de fiabilité mérite d'être investiguée d'abord, à savoir la « répétabilité **»** (ou stabilité dans le temps). Il ne s'agit pas ici du genre de stabilité qu'a étudiée Simpson (1963), à savoir celle de traductions, par des personnes différentes (on parle alors de « reproductibilité »), à 20 ans d'intervalle (en 1942 et en 1962), de 20 expressions verbales (il obtint des résultats très similaires). Il s'agit ici de savoir si, quelques heures, quelques jours ou quelques semaines après avoir donné un degré de certitude (ici en pourcentage), la même personne redonne le même degré de certitude pour la même question quand on lui rend la réponse (et, dans Compar-Aires, la certitude verbale) qu'elle avait donnée auparavant. Autrement dit, quand une personne déclare que, pour elle, « sûr » équivaut à « 60% », cela signifie-t-il aussi bien « 40% » que « 80% » ? Le terme (et le concept) de répétabilité est opposé à **«** reproductibilité **»** (obtenir la

même valeur par des mesures opérées par des personnes différentes). Dans la présente recherche, les deux ont été étudiés. Dans la littérature, je n'ai rencontré que peu d'expériences de répétabilité de jugements en contexte. Elles sont décrites en section E1.

C'est pour pallier ces trois lacunes qu'ont été conçues les deux recherches qui vont être présentées ci-après, et qui sont faciles à reproduire.

1.6 La structure du présent article

Cet article présente deux expériences avec l'échelle verbale de certitudes construite autour de l'adjectif « sûr ».

1. Dans l'expérience hors contexte (section B) c'est-à-dire sans contenu, les participants sont invités à traduire chacun des 6 échelons verbaux en un pourcentage de chances (qu'une réponse soit correcte).
2. Dans l'expérience en contexte (section C), les participants doivent répondre à 15 questions d'un jeu (Compar-Aires), en ajoutant pour chaque réponse un degré de certitude de deux façons différentes : d'une part en mots (une des 6 expressions verbales de l'échelle) et d'autre part en pourcentage (entre 0 et 100%).

2. Le dispositif de l'expérience hors contexte

2.1 La consigne de l'expérience « hors contexte »

Dans l'expérience rapportée ci-après, c'est volontairement (comme l'avait fait Hillson, 2005) que les expressions verbales n'ont pas été présentées en ordre (croissant ou décroissant) de certitude. Et ce, pour éviter que les participants « divisent » simplement le continuum allant de 0% à 100% en 6 parties égales puisqu'on leur présente 6 degrés verbaux de certitude. Cependant, les tableaux et graphiques de résultats présenteront, eux, les degrés dans l'ordre croissant en certitude. Ma consigne était :

Tableau 1 : La consigne de l'expérience « hors contexte » (2013)

« Traduisez chacun de ces six termes **en pourcentage de chance** (de 0% à 100%) que vous ayez raison quand vous l'employez pour parler de votre degré de certitude dans l'exactitude de votre réponse. »	Nom :		
	Sûr		
	Pas sûr du tout		
	Moyennement sûr		
	Extrêmement sûr		
	Peu sûr		
	Très sûr		

2.1 Pourquoi 6 degrés ?

L'expérience qui va être décrite aurait pu être menée à propos de 2 expressions verbales seulement (par exemple « pas sûr » et « sûr ») ou de 3 expressions (par exemple « peu sûr », « sûr » et « très sûr ») ou d'un autre nombre de degrés exprimés en variations de « sûr ». Le nombre de degrés, 6, est inspiré de mes recherches expérimentales (Leclercq, 1982, p. 241-256) et de celles de Bockhlisch, Bocklisch, Baumann, Scholz & Krems. (2010). Ces deux recherches font penser que 6 est la précision ou granulosité maximale que peut gérer fiablement (en termes de répétabilité et de calibration) un adulte bien scolarisé (ayant réussi au moins des études secondaires) quand il exprime ses degrés de certitude en pourcentage.

2.2 Pourquoi « sûr » et pas « certain » ?

Fares (2006, p. 98-101) a observé (avec 59 participants), que les expressions « *sûr* » et « *certain* » reçoivent les mêmes significations sur une échelle métrique allant de 0 à 10. Je n'ai donc pas répliqué avec l'expression « certain » l'expérience ci-après qui utilise l'expression « sûr ».

2.3 Pourquoi « je suis sûr que » et non « il est sûr que… » ?

On remarque que je n'ai pas écrit « je suis sûr (que) » (formulation interne) ou « il est sûr (que) » (formulation externe). En effet, Fares (2006, pp. 79-83) a demandé à 81 étudiants d'exprimer leur sentiment de « convenance » (sur une échelle ordinale allant de « A = pas du tout » à « F = parfaitement ») de chaque degré d'une échelle allant de 0 (incertitude totale) à 10 (certitude totale) pour des expressions probabilistes linguistiques (notamment l'expression « sûr »). Et ce, « pour mieux cerner et décrire numériquement le sens [qu'avaient, pour des sujets] les expressions proposées » (p. 81). La consigne de cette expérience est rapportée en Annexe 1. Sur base de ses résultats, Fares conclut (p. 83) que ses « analyses ne montrent aucune différence entre les expressions '*je suis sûr que*' et '*il est sûr que*' ». C'est « je suis sûr que… » que j'ai utilisé oralement, mais « je suis » n'a pas été repris à l'écrit sur les formulaires de réponses ni dans les résultats.

2.4 Pourquoi des pourcentages ?

Dans leurs expériences, Fabre (1993a et 1993b) et Fares (2006) demandent aux participants de traduire des expressions verbales dans un des degrés d'une échelle numérique allant de 1 à 10 ou de 0 à 10, sans faire référence à des pourcentages. Par contre, dans mes deux expériences, les valeurs numériques demandées font explicitement référence à une échelle de probabilités sous la forme de pourcentages de chance. Demander aux participants de répondre non pas par des probabilités (allant de 0 à 1), mais par des pourcentages de chance (allant de 0% à 100%), c'est adopter une consigne basée sur les fréquences : un choix de consigne qui est conforté par les travaux de Gigerenzer et Hoffrage (1998). Ceux-ci posent à des médecins la question sous la forme « Dans combien de cas sur 100 un diagnostic est-il correct s'il est déclaré… » (suit alors un terme comme « improbable », ou « assuré » ou «plausible », etc.). La consigne « fréquentiste » que j'ai adoptée (les pourcentages de chance signifiant « sur toutes les fois que je serai dans cet état de doute, mon taux de réponses correctes sera ;.. » me paraît approprié à la situation scolaire. En effet, durant les dix mois d'une année scolaire, l'étudiant a l'occasion de fournir de nombreuses réponses à des questions et de vérifier les taux d'exactitude, degré par degré, souvent même à l'intérieur d'un même test. A condition qu'un même degré de certitude ait été utilisé un nombre suffisant de fois (au moins 5 fois par exemple).

2.5 Les participants et la situation de la recherche « hors contexte »

Cette expérience a été réalisée à Bobigny en janvier 2013 lors d'un cours de trois jours consécutifs, intitulé « Psychologie des Apprentissages appliquée aux professions de la santé et à l'Education Thérapeutique du Patient » (DPSS – UFR Médecine – Université de Paris 13 Sorbonne Paris Cité) dans le cadre de deux Masters (en Ingénierie de la formation et en Pédagogie de la Santé).

A la fin de la première journée, les 33 participants ont été invités à répondre à un test portant sur la matière qui venait d'être enseignée. Ce test était constitué de 10 QCM, et les réponses devaient être accompagnées chacune d'un des 6 degrés de certitude suivants :5%, 20%, 40%, 60%, 80% et 95%. Le but était d'initier ces étudiants aux raisons et méthodes décrites dans

mon article « La connaissance partielle : Pourquoi et comment la mesurer chez le patient » (Leclercq, 2009). Après le test, une participante a demandé, à haute voix, « Pourquoi insistez-vous tant pour que l'on demande aux étudiants (et aux patients) d'exprimer leur degré de certitude en pourcentage de chance et non en mots (tels que « peu sûr », « moyennement sûr », « très sûr », etc.) ? ». J'ai répondu que j'expliquerais cela le lendemain. J'ai donc préparé, pour le lendemain matin, la petite grille ci-dessus (section B1) en 66 exemplaires : 33 pour le test (ou prétest), 33 pour le re-test (ou posttest) qui aurait lieu quelques heures plus tard.

Le lendemain, à 9h, j'ai rappelé la question qui m'avait été posée la veille et déclaré que (1) mon meilleur argument serait constitué par des résultats expérimentaux récoltés auprès d'eux-mêmes et (2) que cette stratégie démontrerait que l'on peut procéder à de telles micro-expériences avec ses propres étudiants (la majorité des participants étant des adultes formateurs). J'ai demandé si l'un(e) d'entre eux voyait une objection à cette procédure et ils ont, au contraire, manifesté leur vif intérêt. J'ai invité chacun des 33 participants à écrire, dans chacune des 6 cellules (voir section B1), le % de chance (entre 0% et 100%) correspondant, pour lui, à chaque expression, puis j'ai repris ces grilles remplies (de 6 nombres), en leur annonçant que je fournirais les statistiques des réponses le lendemain. A 14h, sans les en avoir avertis, j'ai rendu les mêmes grilles (vides) aux mêmes personnes en leur demandant de reproduire la même tâche que 5 heures plus tôt, sans leur rappeler les réponses qu'ils avaient données à 9h.

3. Le dispositif de l'expérience en contexte

3.1 Les raisons de développer un jeu ad hoc

Le jeu-test que j'ai créé pour l'occasion s'appelle Compar-Aires parce qu'il porte sur la comparaison visuelle des superficies (aires) de pays du Monde (ou d'une partie du Monde) présentés sur une carte. L'expérience ci-après a porté sur une carte de l'Europe, mais on peut imaginer des variantes du jeu (portant sur d'autres continents ou sur des régions d'un pays, etc.). Il ne s'agit que d'un des contextes possibles, et les résultats ci-après ne sont donc qu'indicatifs. Les contextes, en effet, peuvent varier à l'infini : par les contenus, par l'âge des participants, par leurs connaissances sur le sujet, par leur familiarisation avec l'auto-évaluation sous forme de degrés de certitude (et la façon dont ceux-ci ont été exploités), par les enjeux (gain ou retrait de points par exemple), etc. Ce jeu Compar-Aires a été conçu pour présenter plusieurs avantages : (1) son côté perceptif et ludique qui réduit l'impression, chez le participant, d'être jugé en tant que personne sur base de ses connaissances ou de sa logique par exemple ; (2) il est très facile de donner, ensuite, les réponses correctes aux participants, en projetant la même carte, mais avec les numéros d'ordre des tailles décroissantes des pays (le plus vaste portant le numéro 1).

3.2 Les consignes du jeu Compar-Aires

La consigne était la suivante : « Dans cette carte de l'Europe projetée à l'écran, les pays sont délimités (par leurs frontières) et colorés. Par analyse visuelle, répondez aux 15 questions. Elles comptent chacune 6 de ces pays à comparer entre eux. ».

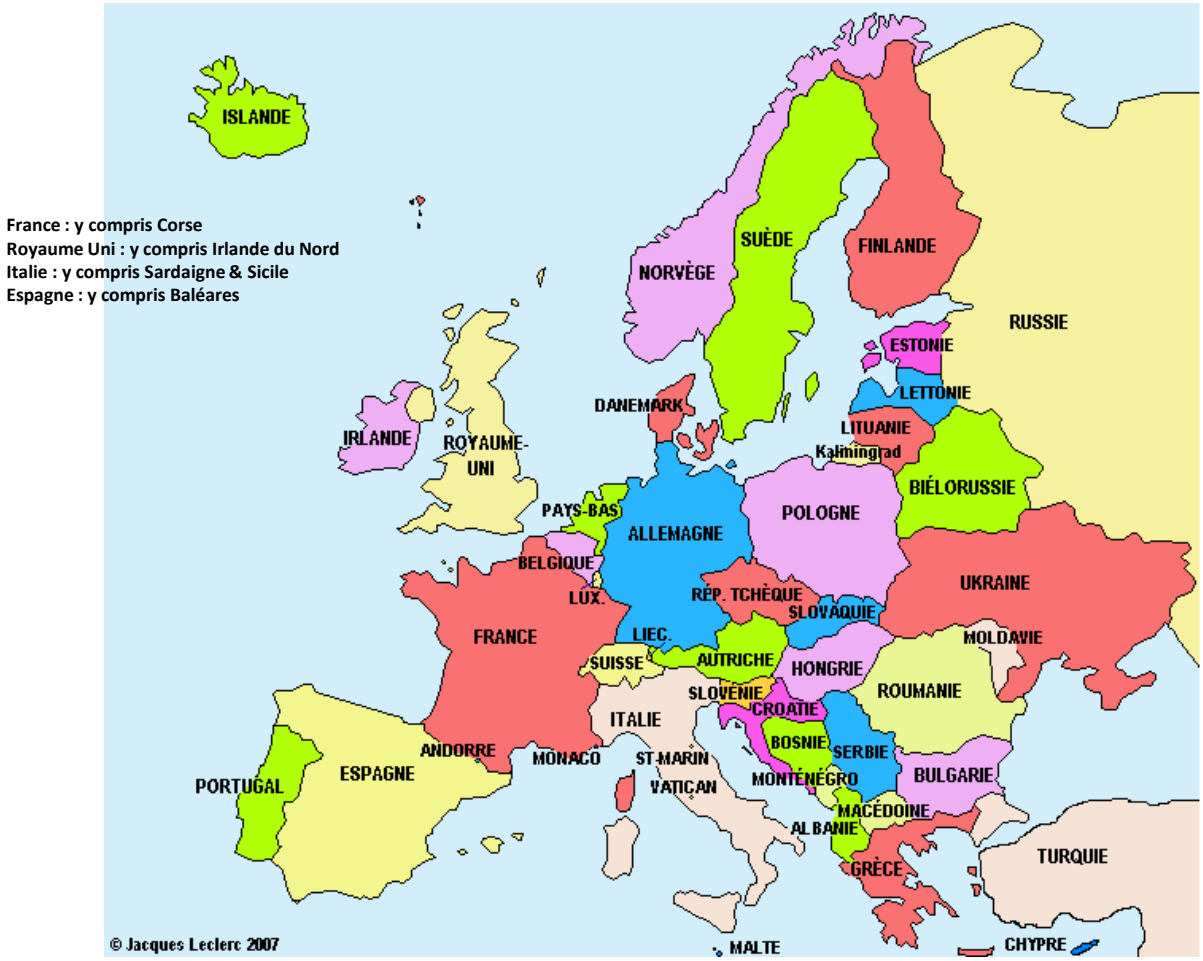


Figure 1 : Carte de l'Europe projetée pendant toute la durée du jeu (15 minutes)

« Répondez sur la feuille suivante en

- *entourant l'un des 6 pays ;*
- *entourant l'une des 6 expressions en mots exprimant le degré de certitude que votre réponse soit correcte ;*
- *fournissant (dans la colonne de droite) votre degré de certitude en % de chance que votre réponse soit correcte.*

Vous êtes invité(e) à laisser sur la feuille les traces de vos raisonnements (barrer les solutions rejetées, entourer celles entre lesquelles vous hésitez, entourer plusieurs fois la solution finalement choisie) ».

Voici les quatre premières questions :

	Lequel de ces pays a la plus **grande** surface en km² ?												
1	Slovaquie	Rép. Tchèque	Pays-Bas	Danemark	Royaume Uni	Suisse	Pas sûr du tout	très peu sûr	peu sûr	sûr	très sûr	Extrêmt sûr	%
2	Belgique	France	Allemagne	Norvège	Autriche	Suède	Pas sûr du tout	très peu sûr	peu sûr	sûr	très sûr	Extrêmt sûr	%
3	Suède	Allemagne	Espagne	Norvège	Finlande	Royaume Uni	Pas sûr du tout	très peu sûr	peu sûr	sûr	très sûr	Extrêmt sûr	%
4	Allemagne	France	Espagne	Suède	Finlande	Grèce	Pas sûr du tout	très peu sûr	peu sûr	sûr	très sûr	Extrêmt sûr	%

Figure 2 : Les 4 premières questions de *Compar-Aires*

Pour les 10 premières questions, il s'agit de déterminer quel est le plus grand (en superficie) des six pays et pour les 5 dernières (de11 à 15) quel est le plus petit (en superficie) des six. On remarquera que les degrés 2 et 3 (« très peu sûr » et « peu sûr ») ne sont pas les mêmes que dans l'expérience hors contexte (qui étaient « peu sûr » et « moyennement sûr »). Dans l'expérience hors contexte de Bobigny (section B), les 6 expressions verbales avaient été présentées « dans le désordre » et des participants (éliminés des données) n'avaient pas classé « moyennement sûr » en-dessous de « sûr » mais au-dessus. Je me suis rendu compte que cette expression « moyennement sûr » était ambiguë et ai donc changé la consigne. Par ailleurs, Hamm (1991), qui lui aussi a dû éliminer des sujets, a observé des variations inter-individus amplifiées quand les expressions verbales sont présentées dans le désordre plutôt que dans l'ordre de probabilité.

3.3 La construction des questions

Comme dans les expériences de psychophysique en laboratoire sur les seuils différentiels de perception, les questions ont été construites de manière à ce qu'elles soient de différents niveaux de difficulté. Une question est d'autant plus difficile que ses solutions ont des numéros d'ordre proches les uns des autres (les pays ont des superficies proches). La figure 3 présente les numéros d'ordre au-dessus de chacune des solutions des 4 premières questions.

	Lequel de ces pays a la plus **grande** surface en km^2 ?					
	26	20	29	28	10	30
1	Slovaquie	Rép. Tchèque	Pays-Bas	Danemark	Royaume Uni	Suisse
	32	2	5	7	19	4
2	Belgique	France	Allemagne	Norvège	Autriche	Suède
	4	5	3	7	6	10
3	Suède	Allemagne	Espagne	Norvège	Finlande	Royaume Uni
	5	2	3	4	6	13
4	Allemagne	France	Espagne	Suède	Finlande	Grèce

Figure 3 : Numéro d'ordre de chacune des solutions des 4 premières questions du jeu

On voit que la question 1 est facile car le numéro d'ordre (10) de la solution correcte (Royaume Uni, en 10^e position) est très éloigné des autres solutions. Par contre, la question 4 est difficile car le numéro d'ordre de la solution correcte (France, en 2e position) est proche de celui de 4 des 5 autres solutions. En réalité, pour les participants, les questions apparaissent (comme dans la figure 2) uniquement en encre noire et sans aucun surlignage jaune. C'est pour indiquer au lecteur où sont les réponses correctes qu'elles ont été mises en encre rouge et pour lui signaler les distracteurs les plus plausibles que ceux-ci ont été surlignés en jaune

3.4 La population expérimentale

En 2016, j'ai présenté ce jeu (de 15 questions) à 20 étudiants (âge modal 24 ans) en Master à l'Université de Liège (ULg). Ces étudiants suivaient un cours intitulé « Conception et Analyse de Messages Multi-Médias – CAMMM ». Je m'étais fixé comme critère d'inclusion (rétrospectif) le fait d'avoir utilisé au moins 4 des 6 degrés de certitude sur les 15 réponses. Les données d'un des participants ont été éliminées parce qu'il n'avait utilisé que 2 degrés de certitude (très peu sûr et peu sûr), avec un pourcentage de 100% pour « très peu sûr » et de 40% pour « peu sûr ». Restent donc les données de 19 participants, soit un nombre total de 285 réponses (19 x 15) et donc 285 certitudes verbales et 2 fois 285 certitudes en pourcentage (285 pour le test et 285 pour le retest).

3.5 Données descriptives de base

3.5.1 Le sens de la traduction

Dans le jeu, chaque participant pouvait choisir d'abord un degré de certitude verbal puis ensuite, le traduire en %, mais il pouvait tout aussi bien faire l'inverse, en fixant d'abord un pourcentage et en le traduisant ensuite vers l'échelle verbale. C'est pourquoi il serait plus neutre de parler de « Correspondance Mots-Pourcentages » et non de traduction, quand on veut ne pas préciser dans quel sens la (les) traduction(s) a (ont) été faite(s).

A la fin du jeu, juste avant de reprendre les feuilles, j'ai demandé aux participants d'indiquer par une flèche dans quel sens ils avaient fait la correspondance. Sur les 19 participants, un seul a déclaré être parti des %. Un autre a déclaré avoir fait tantôt l'un tantôt l'autre. Les 17 autres ont tous déclaré être partis des mots. La présentation du questionnaire a probablement

eu une influence déterminante sur cette démarche. En effet, l'échelle des mots est lue avant la case « % » car elle est imprimée immédiatement après la question, alors que la case devant recevoir le pourcentage est à l'extrême droite, à la fin de la ligne, donc lue en dernier lieu.

3.5.2 Les indices de difficulté de chacune des 15 questions

Les taux de réussite des questions ont pour moyenne 78,9% et un écart-type de 17,5% (voir Annexe 2 - Indices de Difficulté - taux de réponses correctes - des 15 questions).

La certitude moyenne a été exactement la même la première et la deuxième fois : 65,3%, les Ecarts-Types des certitudes étant 26,5% au Pré et 27,2% au Post (voir Annexe 3 - Certitudes moyennes chez les 19 participants en contexte au test ou PRE dans l'expérience en contexte de 2016).

3.5.3 Les taux de réussite de chacun des 19 étudiants

Le moins bon résultat individuel a été, pour un participant, de 5 réussites sur 15 questions. Un seul participant a fourni les 15 réponses correctes. Le mode (7 étudiants) est de 13 réussites (voir détails en Annexe 4 - Taux de réussite pour les 19 participants de l'expérience en contexte de 2016).

4. L'hypothèse 1 : Préférences pour les multiples de 10

4.1 L'origine de cette hypothèse

Cette hypothèse est basée sur l'observation que j'avais faite (Leclercq, 1975, 1982) avec un jeu-test créé pour l'occasion et que j'ai appelé *Confidence Guessing Game* (CGC) car il est inspiré du *Shannon Guessing Game* (Attneave, 1959). Il s'agit de deviner la lettre suivante de mots dans un texte tronqué au hasard. Dans cette situation, les participants avaient donc 27 possibilités (les 26 lettres de l'alphabet de la langue française + le « joker », c'est-à-dire n'importe quel caractère non lettre, y compris l'espace inter-mots). Le CGC ne porte pas sur les variations autour du mot « sûr », mais sur les pourcentages de chance d'avoir fourni la réponse correcte (deviner la lettre suivante dans des textes tronqués au hasard), et ce, pour trois échelles de granularités différentes : 4, 10 et 40 degrés. En 1975, 37 personnes avaient été invitées à répondre à ce jeu-test comportant 100 questions. La figure 4 montre la distribution des 7400 degrés de certitude, donc les deux séries de 100 certitudes données à un mois d'intervalle pour les mêmes questions et réponses (rappelées lors du retest).

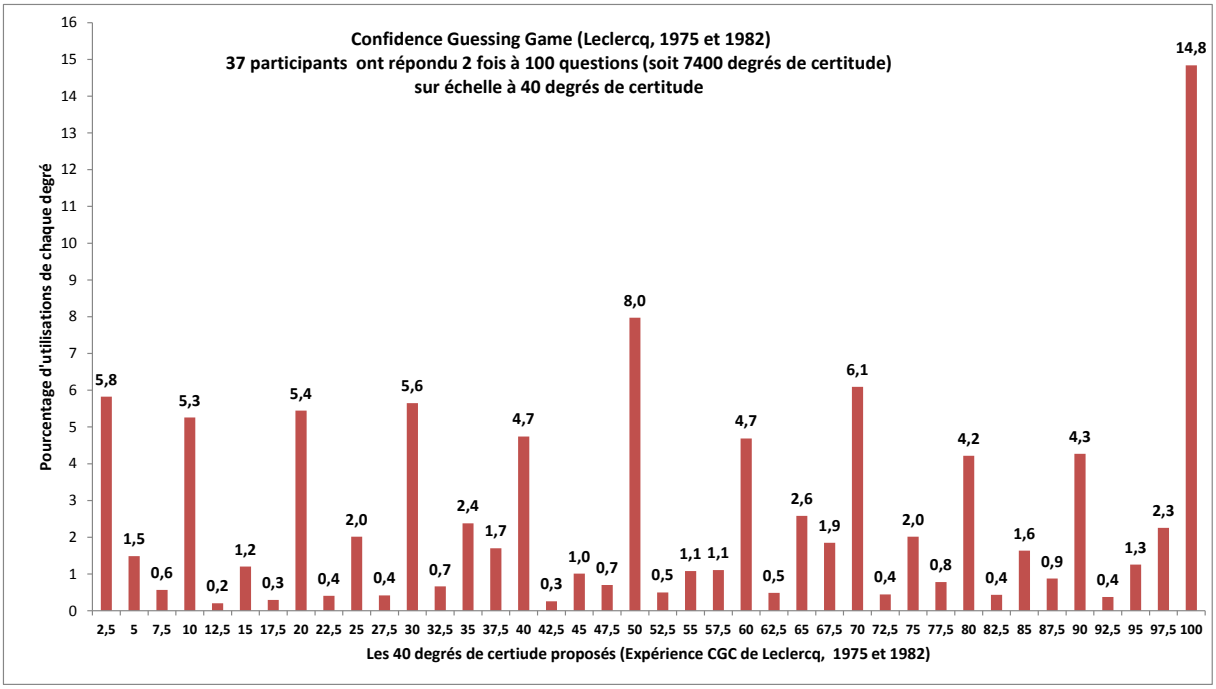


Figure 4 : Distribution des 7400 degrés de certitude dans l'expérience CGC 1975

On constate que les certitudes sont concentrées sur les multiples de 10. Mon explication de ce phénomène tient dans une hypothèse qui a fait l'objet d'une vérification expérimentale détaillée (Leclercq, 1975, 1982). Cette hypothèse est que la sensibilité (ou granularité) de répondants adultes, c'est-à-dire leur capacité à distinguer de façon fiable (en répétabilité et en calibration c'est-à-dire en correspondance avec la réalité) est largement inférieure à 40 degrés, et même à 10 degrés. Alors à quoi bon être plus précis ? Je fais donc l'hypothèse que, même inconsciemment, les répondants feraient leur une devise qui pourrait être :

Nul ne peut prétendre à une précision supérieure
à l'erreur de mesure de son instrument.

4.2 La répartition des pourcentages dans l'expérience hors contexte

La figure 5 présente la répartition des 198 pourcentages (toutes catégories verbales confondues) données par les 33 participants aux 6 degrés de certitude verbaux.

Des regroupements ont été opérés pour deux catégories :

- la colonne 0% ne contient que des réponses 0%, mais la colonne à sa droite (5%) regroupe 1 certitude 2% et 4 certitudes 5%.
- les 21,7% de certitudes (en réalité 43 réponses) regroupées dans la colonne « 100% » se décomposent en réalité en 18% de 100% (35 réponses), 2,5% de 99% (5 réponses) et 1,5% de 98% (3 réponses).

Pour tous les autres pourcentages, ce sont les valeurs telles qu'elles ont été fournies par les participants qui sont présentées dans la figure 5.

On y voit que les 33 participants ont utilisé des pourcentages répartis sur toute la gamme allant de 0 à 100%.

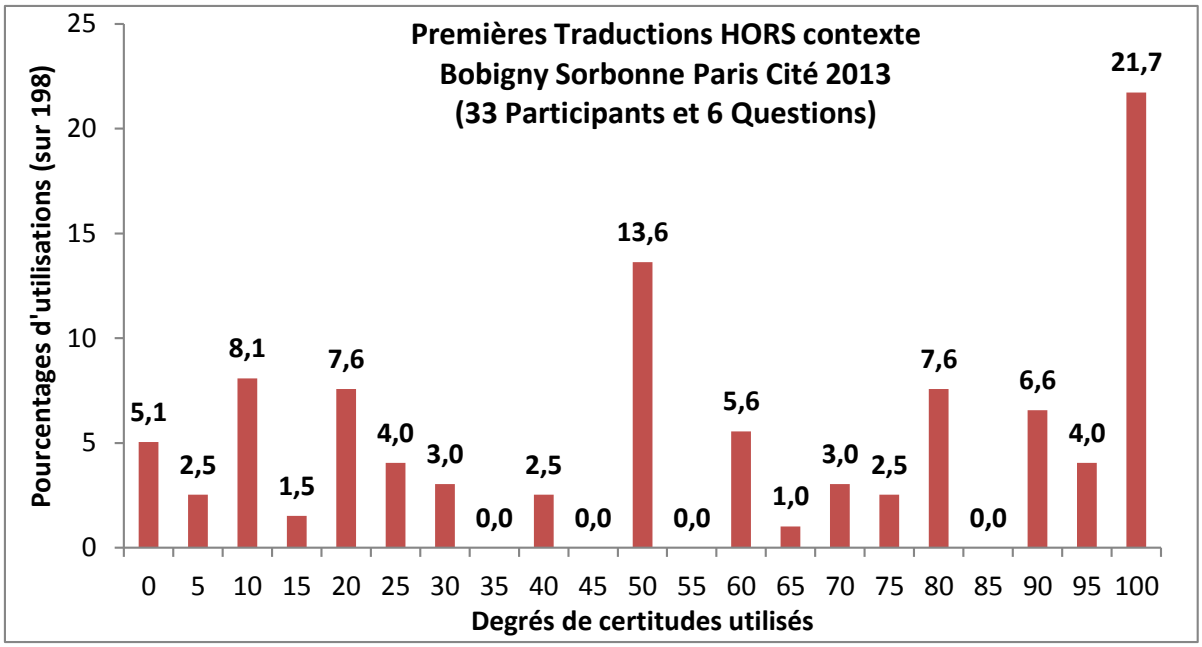


Figure 5 : Répartition (en %) des 198 pourcentages utilisés par les 33 participants pour traduire les 6 degrés de certitude exprimés en mots lors du PREtest

On constate que les valeurs multiples de 10% ont toutes été utilisées (avec des pics pour 50% et 100%), ce qui n'est pas le cas pour plusieurs valeurs intermédiaires (35, 45, 55 et 85). Les valeurs 5, 15, 25, 75 et 95 sont utilisées, mais très peu. La popularité de la réponse 50% est suspecte, étant donné la tendance, dans les échelles de jugement, à éviter les réponses extrêmes. Ce phénomène est appelé « biais de tendance centrale ».

4.3 La répartition des pourcentages dans l'expérience en contexte

La figure 6 montre les taux d'utilisation des 285 pourcentages donnés par 19 participants à 15 questions en PREtest.

Dans ce graphique, le bâton des 100% (53 observations) regroupe 50 utilisations de 100% et 3 utilisations de 99% (par la même personne).

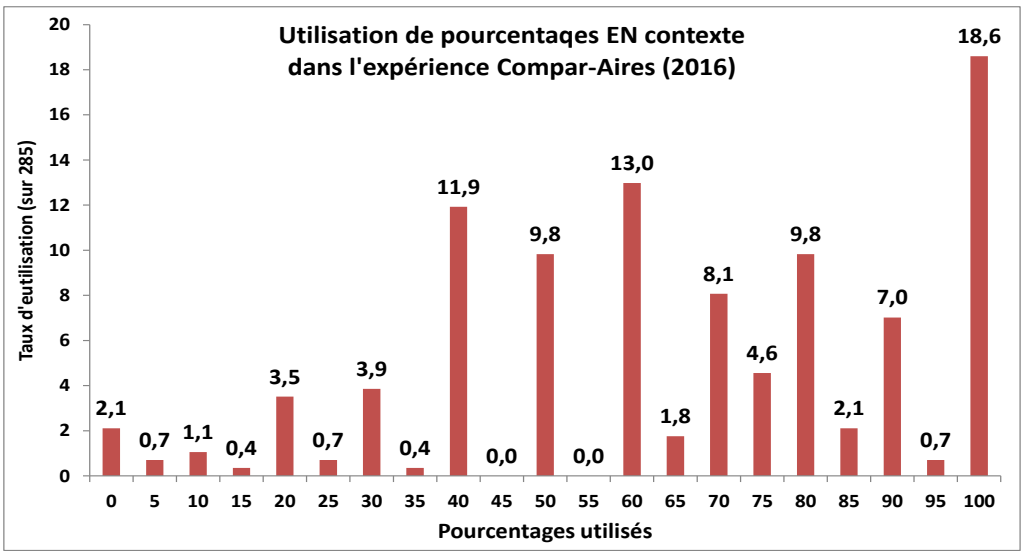


Figure 6 : Répartition des 285 pourcentages au PREtest

On constate à nouveau une concentration des réponses sur les multiples de 10%. Ce constat rejoint celui des figures 4 et 5. Ici, les pourcentages 65%, 75% et 85% font exception. Or ces trois valeurs sont situées dans la partie supérieure de l'échelle, ce qui n'a rien d'étonnant étant donné que la facilité moyenne des questions est élevée (79%). Cette observation sera discutée dans le cadre de l'hypothèse 3 (la sensibilité des répondants à la difficulté des questions). En effet, le taux d'exactitude des réponses dans l'expérience CGC (figure 4) était de 56%, et de 79% pour l'expérience Compar-Aires (figure 6). Ce qui expliquerait un « décalage » vers le haut de la distribution de la figure 6 par rapport à la distribution de la figure 5.

5. L'hypothèse 2 : La bonne répétabilité intra-individus

5.1 L'origine de l'hypothèse

En section A4, j'ai signalé n'avoir rencontré que peu de recherches (en fait 3) portant sur la répétabilité des jugements par une même personne après un certain délai. Budescu et Wallsten (1985) rapportent une bonne répétabilité (stabilité) intra-individuelle dans le temps des traductions mots-pourcentages. Les deux autres expériences sont décrites ci-après.

Preston et Colman (2000) ont utilisé le principe du test - retest, trois semaines plus tard, mais pas sur des degrés de certitude de l'exactitude d'une réponse. Ils invitent en effet les participants à porter des jugements de qualité, sur des échelles de Likert, de certains attributs (prix, vitesse, fraîcheur des aliments, etc.) de restaurants que ces participants connaissent (et qui diffèrent d'une personne à l'autre). Preston et Colman ont comparé les répétabilités selon des échelles de granularités différentes (de 2 échelons à 101 échelons). Ces répétabilités sont très élevées (corrélations supérieures à 0,9).

La dernière des trois recherches qui fonde mon hypothèse de bonne répétabilité est celle que j'ai faite avec le jeu-test CGC (voir section D1) ; j'ai pratiqué la répétition 4 semaines plus tard, pour 100 questions et 37 personnes. La valeur médiane des corrélations de répétition intra-individuelle était 0,56 pour les 100 lettres. Si l'on ne considère que les 10 premières lettres à deviner, situation où l'attention est plus soutenue et qui est plus proche d'une situation scolaire, la corrélation de répétition « monte » à 0,74.

Je fais donc l'hypothèse que les fidélités intra-personnes dans les deux expériences (2013 et 2016) seront au moins aussi élevées que 0,74 pour trois raisons :

- Dans le CGC, les participants peuvent, lors du re-test (ou posttest), faire des hypothèses différentes sur la réponse correcte de celle du test (ou PRE), hypothèses dont ils n'ont plus la trace bien que leur réponse leur ait été rendue.
- Dans le jeu Compar-Aires, qui est plus perceptif, les participants disposent de la trace de leurs processus d'élimination et de choix des solutions proposées, et, en plus dans une tâche perceptive où ils ont sous les yeux les mêmes stimuli au test (PRE) et au retest (POST).
- Le délai (2 heures) est beaucoup plus court dans le jeu-test Compar-Aires que dans le jeu-test CGC (4 semaines).

Par ces deux jeux (CGC et Compar-Aires), j'ai tenté de créer des situations où, entre les deux expressions de la certitude, l'état mental du sujet (sur la question posée) n'a pas varié. Cette condition expérimentale est facile à garantir quand la question de la répétition (de la traduction de mots vers des pourcentages) est posée hors contexte. Par contre, elle est particulièrement difficile à respecter quand la répétition se fait en contexte, plus

particulièrement sur des contenus précis. En effet, entretemps, le répondant peut avoir appris des choses ou en avoir oublié sur les contenus testés.

5.2 La répétabilité hors contexte de la traduction de mots en %, après 5 heures

Dans l'expérience hors contexte, j'ai présenté une seconde fois les 6 mêmes expressions aux mêmes 33 personnes, 5 heures après leur première « traduction ». La corrélation entre les 6 valeurs (en %) des deux traductions (la première et la deuxième, à 5h d'intervalle), est parfaite (elle vaut 1) pour 20 des 33 personnes, et la moyenne des 33 corrélations intra-individuelles vaut 0,977. Il existe cependant des personnes dont la corrélation de « répétabilité » est plus faible : 0,76 pour un participant et 0,88 pour un autre.

5.3 La répétabilité en contexte de la traduction de mots en %, après 2 heures

Deux heures après avoir répondu une première fois au jeu Compar-Aires, les participants ont reçu leur feuille de réponse dont la partie droite (avec les %) a été enlevée. Les 19 participants ont été invités à fournir à nouveau des pourcentages alors qu'ils avaient sous les yeux leur feuille où ils avaient noté (1) leurs raisonnements en barrant d'un trait les solutions qu'ils rejetaient et en entourant (bulles) les solutions entre lesquelles ils hésitaient, (2) la solution finalement choisie, en l'entourant beaucoup plus que les autres et (3) leur degré de certitude en mots. La valeur moyenne des 19 corrélations linéaires de répétition est 0,94. Leur valeur minimale observée est 0,68, chez 1 participant. Les 18 autres participants ont une corrélation comprise entre 0,87 et 1 (la valeur 1 étant observée chez 6 participants). Comme pour l'expérience hors contexte, les corrélations de répétabilité intra-individuelle de la traduction des mots vers des pourcentages sont donc en général très élevées. Contribuent à ce constat (1) le délai très court de répétition, (2) le fait que la réponse finale à chaque question soit disponible lors du « re-test » et (3) le fait que les traces du raisonnement (barres et bulles) notées à même la feuille restent elles aussi présentes au retest. Le tableau 2 et la figure 7 montrent que les moyennes se correspondent fortement entre la 1ère traduction (Test ou PRE) et la 2e traduction (retest ou POST).

Tableau 2 : Moyennes (arrondies à l'unité) en pourcentage pour chacun des 6 degrés verbaux

	Pas du tout sûr	Très peu sûr	peu sûr	sûr	très sûr	Extrêmt sûr
PRE	11	30	48	65	83	98
POST	10,5	29	47	67	84	97

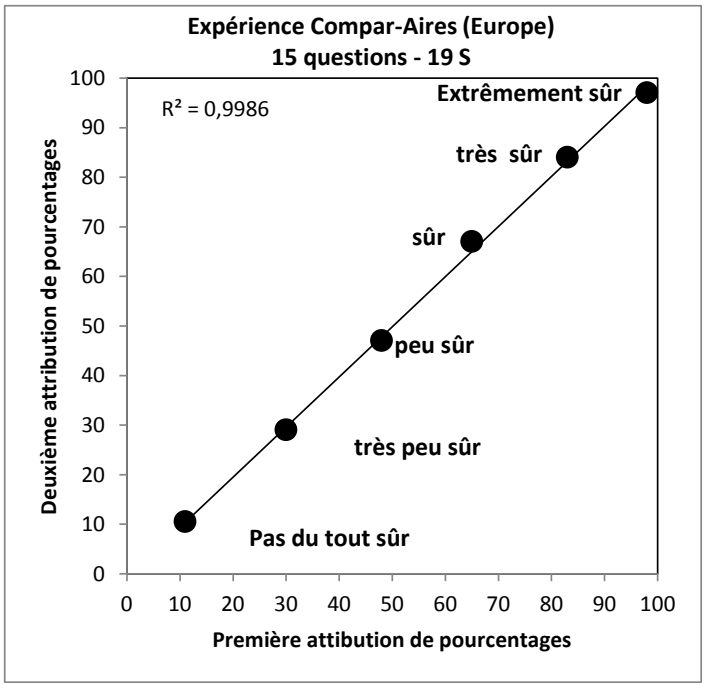


Figure 7 : Répétitions moyennes des pourcentages EN contexte

La corrélation de répétition du groupe dans son ensemble est 0,999, ce qui est plus élevé que la moyenne des corrélations individuelles (0,94).

5.4 L'ampleur des variations intra-individuelles lors de la répétition

La figure 8 montre les distributions des discordances POST-PRE (ou test-retest).

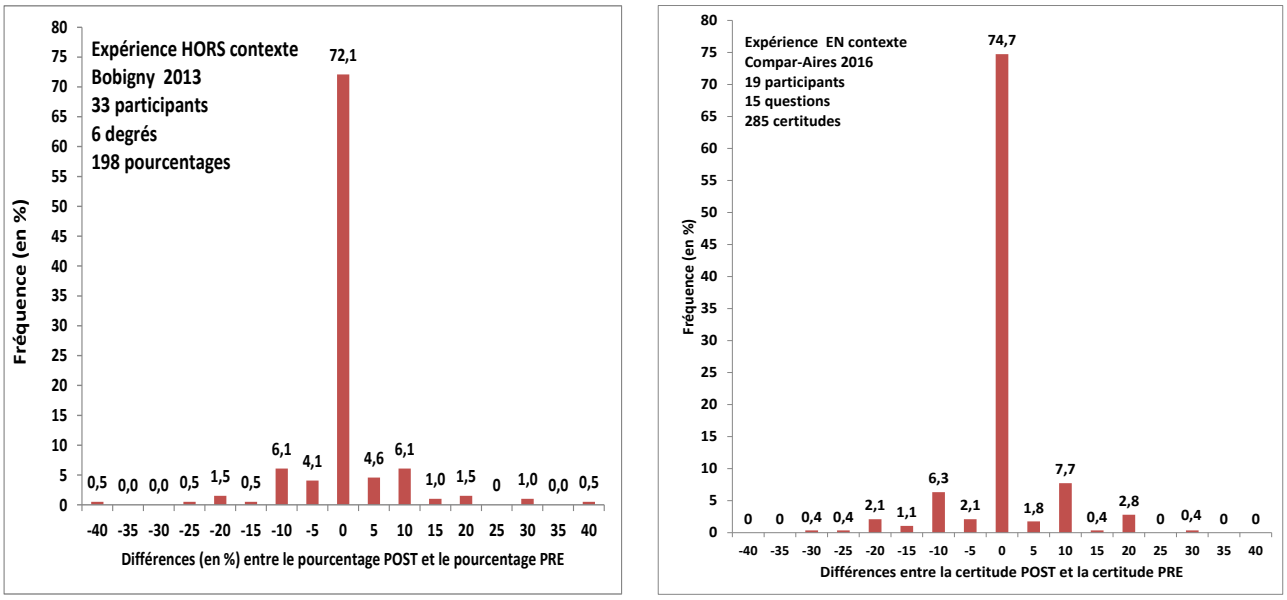


Figure 8 : Distribution des discordances calculées par la formule %POST - %PRE

La concordance parfaite (0) se produit dans plus de 70% des réponses dans les deux expériences. La symétrie de chacune des deux distributions fait penser au caractère aléatoire des (rares) discordances entre PRE (test) et POST (retest). Il faut cependant garder à l'esprit que cette bonne répétabilité n'a été observée dans ces deux expériences que pour un nombre limité (6) de degrés de certitude verbaux et ne présume pas de ce que serait la répétabilité avec des échelles comportant un nombre de degrés supérieur ou inférieur.

Cette répétabilité « globale » ne dit pas non plus pour lesquels de ces degrés la répétabilité serait la moins bonne, bref quelles seraient les valeurs optimales de l'échelle numérique de certitudes à proposer à des étudiants par exemple. Ainsi, Hamm (1991) propose une échelle en 19 degrés, mais il ne traite pas ce problème de la répétabilité, bref de la fidélité des valeurs numériques choisies.

5.5 Les Moyennes (du groupe) des traductions en % des 6 expressions verbales

Comme rappelé (entre parenthèses) en première colonne du tableau 3, les nombres de participants étaient différents dans les deux expériences. Une autre différence entre elles tient dans le nombre d'observations sur lesquelles les moyennes par degré verbal sont calculées.

Dans l'expérience hors contexte (tableau 3), c'est sur 33 valeurs en pourcentage, chacun des participants en ayant fourni une pour chaque degré verbal. Dans l'expérience en contexte, les nombres varient d'un degré verbal à l'autre (voir figure 12), puisque les participants avaient le choix de leurs degrés de certitude, ce qui est rappelé dans la dernière ligne « N (Total = 285) ».

Tableau 3. Moyennes des valeurs numériques attribuées aux 6 expressions verbales

HORS en	pas sûr du tout 1	PEU SUR très peu sûr 2	MOY SUR peu sûr 3	sûr 4	très sûr 5	extrêmemt sûr 6	r du groupe
HORS PRE *(33)*	8,7	24,2	49,5	76,5	89,2	99,1	0,988
HORS POST *(33)*	10,6	24,2	48,8	79,5	89,5	97,5	0,981
EN PRE	11	29,7	47,6	65,3	82,6	97,6	0,999
EN POST	10,5	29,1	47,4	67,2	83,5	96,7	0,998
N (Total = 285)	*11*	*26*	*73*	*69*	*44*	*62*	

Pour ces 6 degrés verbaux, les valeurs moyennes des pourcentages attribués « tournent » autour de 10%, 25%, 50%, 70%, 90% et 97,5 %. Les répondants seraient-ils (ou croiraient-ils être) plus nuancés dans les certitudes très élevées ?

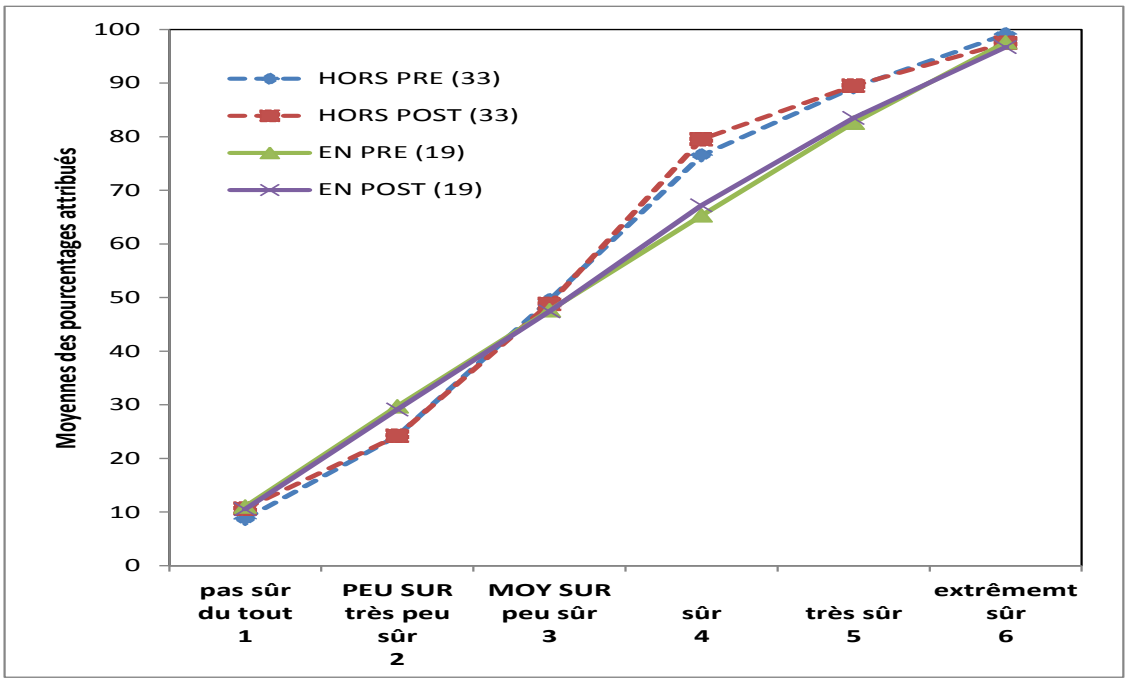


Figure 9 : Les moyennes observées dans les deux expériences pour les 6 degrés verbaux

Les courbes de l'expérience hors contexte ont une forme logistique exponentielle (en S majuscule incliné) et celles de l'expérience en contexte une forme rectiligne. Le hasard ?

5.6 La répétabilité des 19 moyennes individuelles dans l'expérience en contexte

La figure 10 (nuage de points PRE – POST ou test-retest) montre que la répétition de la certitude moyenne (en %) est aussi élevée pour chaque personne : les 19 points sont quasi tous sur la diagonale. La stabilité de groupe (figure 10) n'est donc pas due à la compensation de décalages individuels positifs par des décalages individuels négatifs, mais c'est un phénomène général (en tout cas pour ces 19 participants et dans ce contexte).

La dispersion inter-individus de ces moyennes de certitude est grande : la moyenne la plus faible est inférieure à 40% et la plus forte aux alentours de 90%, ce qui est rassurant dans la mesure où les taux de réussite sont aussi variables entre participants (le plus faible est 33%, le plus élevé 100% et le plus fréquent 87%).

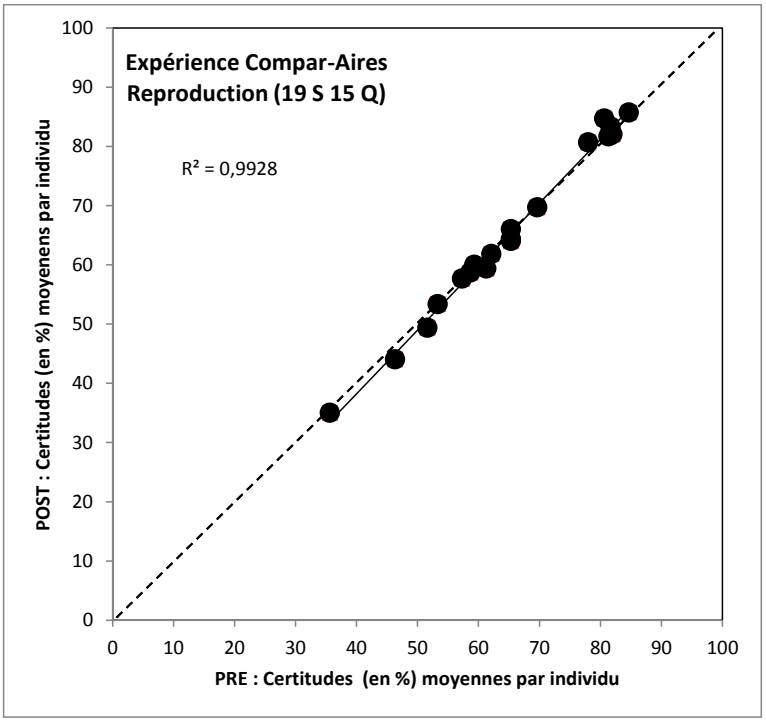


Figure 10 : Certitudes moyennes au PRE et au POST des 19 participants à l'expérience en contexte

On peut donc nourrir l'espoir (à charge de confirmations) d'une grande stabilité intra-individuelle dans l'attribution des degrés de certitude en pourcentage**,** du moins dans ce type de tâche (proche d'une QCM) où, en outre, le répondant a gardé des traces écrites non seulement de ses réponses finales, mais aussi des processus qui l'ont amené, il y a plus de deux heures, à choisir sa réponse et sa certitude.

5.7 Les erreurs de centration des moyennes

L'Erreur individuelle de centration ou EIC (Leclercq, 2003, p. 38) se calcule par la différence entre la Certitude moyenne d'une personne à l'épreuve et sa réussite moyenne (ou taux d'exactitude - TE) au total de cette épreuve, soit la formule EIC = Cmoy-TE, l'EIC idéale étant 0. Dans un graphique comme la figure 11, les EIC sont les distances verticales de chaque point à la diagonale. La figure 11 présente le graphique des 19 centrations individuelles.

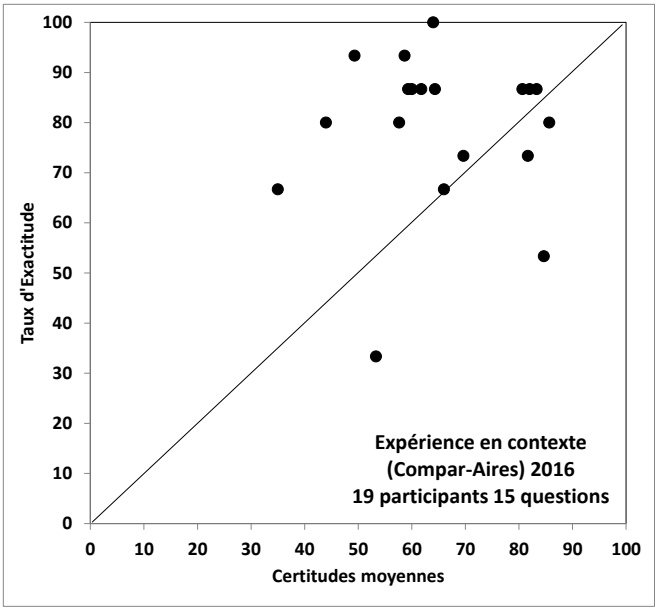


Figure 11 : Les 19 EIC de l'expérience PRE en contexte

On voit qu'un seul des participants a une centration parfaite (EIC = 0) parce que son point est sur la diagonale. Les points des (4) participants qui se sont surestimés apparaissent sous la diagonale. Pour deux d'entre eux la surestimation moyenne est faible (points proches de la diagonale) et pour les deux autres, elle est importante (points éloignés de la diagonale).

Les points de tous les autres (14) participants sont au-dessus de la diagonale : ils se sont sous-estimés (en moyenne, ils ont performé mieux que ce qu'ils avaient estimé par leurs certitudes).

Bien que ce point de la calibration (ou du réalisme) soit crucial, je ne le traiterai pas ici parce qu'il demanderait trop de développements, à commencer par se pencher sur les façons de mesurer la calibration, et l'EIC est loin d'être la meilleure de ces méthodes de mesure, car la valeur 0 pour une personne à un test peut résulter d'une compensation de ses surestimations par ses sous-estimations.

6. Hypothèse 3 : Une sensibilité à la difficulté des questions

6.1 L'origine de l'hypothèse

Avec Campbell, Lewis et Hunt (1958) et Fabre (1993a), on peut penser que le participant s'imagine que l'expérimentateur lui a fourni des questions telles que tous les niveaux de certitude devraient être utilisés de façon égale. Si l'on suit ce raisonnement, on est amené à faire l'hypothèse que si l'on déplaçait la variation de difficulté des questions vers le bas (entre 0% et 60% par exemple), au lieu de la répartition actuelle (entre 47% et 100%), la variété des degrés de certitude choisis ne subirait pas un changement d'une aussi grande ampleur, bref serait l'objet d'une certaine inertie. En toute logique, dans un examen scolaire par exemple, cette tendance à l'équi-répartition doit être contrebalancée par la difficulté des questions (ce qui, dans l'expérience en contexte, dévie la distribution vers la droite étant donné la grande facilité moyenne des questions de cette expérience). Le jeu-test Compar-Aires est à mi-chemin entre un jeu et un test, ne fut-ce que parce qu'il est administré par l'enseignant, à qui, bien que la procédure soit anonyme, les étudiants s'efforcent d'apparaître réalistes. Dans l'expérience en contexte, on devrait donc observer une assez large répartition sur les 6 degrés

verbaux avec, toutefois un « décalage » vers la droite, le taux de réussite moyen aux 15 questions étant 79%.

6.2 La répartition des 6 degrés verbaux dans l'expérience en contexte

Dans l'expérience Compar-Aires, les participants devaient, pour chacune de leurs 15 réponses, choisir un degré de certitude sur une échelle verbale à 6 degrés. La figure 12 montre les nombres totaux d'utilisation des six degrés de certitude verbaux (sur les 285 utilisations).

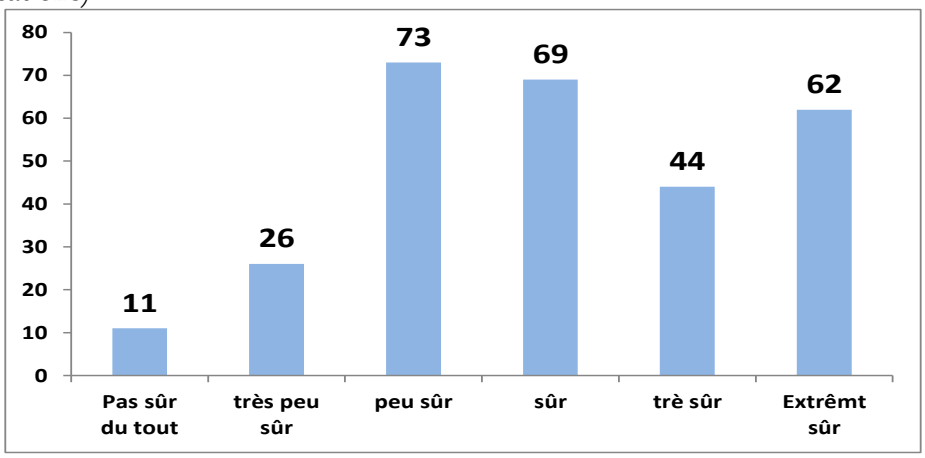


Figure 12 : Nombre de choix des 6 expressions verbales (total = 285) dans l'expérience en contexte (Compar-Aires, 2016))

Le pic ou Mode est « Peu sûr ». Etant donné la facilité du test (Moyenne 79%), (1) il n'est pas étonnant que les niveaux « Pas sûr du tout » et « Très peu sûr » aient été les moins utilisés et (2) on aurait pu s'attendre à une répartition plus concentrée encore sur les seuls degrés les plus élevés, bref à une «différenciation» moins poussée. Les deux phénomènes (tendance à l'équi-répartition et sensibilité à la difficulté des questions) ont donc probablement joué en sens contraires.

7. L'hypothèse 4 : D'importantes différences interindividuelles

7.1 L'origine de l'hypothèse

Par leurs travaux expérimentaux, de nombreux auteurs comme Johnson (1973), Beyth-Marom (1982), Lichtenstein et Newman (1987), Budescu, Weinberg et Wallsten (1988), Reagan, Mosteller et Youtz (1989), O'Brien (1989), Drudzel (1989), Clark (1990), Fabre (1993a et 1993b), Kent (1994), Bickel (2005), Fares (2006) et Bocklisch et al. (2010) ont mis en évidence une grande variété interindividuelle dans la traduction de probabilités exprimées en mots vers une expression en nombres. Cette variabilité peut aller jusqu'à une marge de variation (différence entre la plus grande valeur observée et la plus petite) de 40%. Cependant aucune de ces recherches n'a étudié le phénomène sur une échelle verbale constituée de variations autour de l'adjectif « sûr » et leur traduction vers des pourcentages. Je fais donc l'hypothèse que, dans les deux expériences que j'ai organisées, j'obtiendrai des résultats assez similaires à ceux rapportés par la littérature.

7.2 Les courbes personnelles de traduction en % des certitudes ordinales hors contexte (2013)

La figure 13 ci-après montre les réponses de 4 des 33 personnes interrogées : les personnes F, W, X et AE (noms codés). Sur l'axe vertical ont été portées les valeurs (en % de chance) que ces 4 personnes ont données aux 6 expressions verbales de l'axe horizontal. Les 6 valeurs de chaque personne ont été reliées entre elles par des segments de droite, ce qui constitue, pour chacune de ces 4 personnes, sa « courbe personnelle de certitudes ordinales» ou cpco. Cette expression est un clin d'œil aux « cci » (courbe caractéristique d'un item) dans le modèle de Rasch (Lord & Novick, 1968 ; Leclercq, 1980, 1987, et 1990). En effet, les cci ont, comme les courbes de F et de X, une forme logistique exponentielle (la forme d'une lettre S majuscule inclinée).

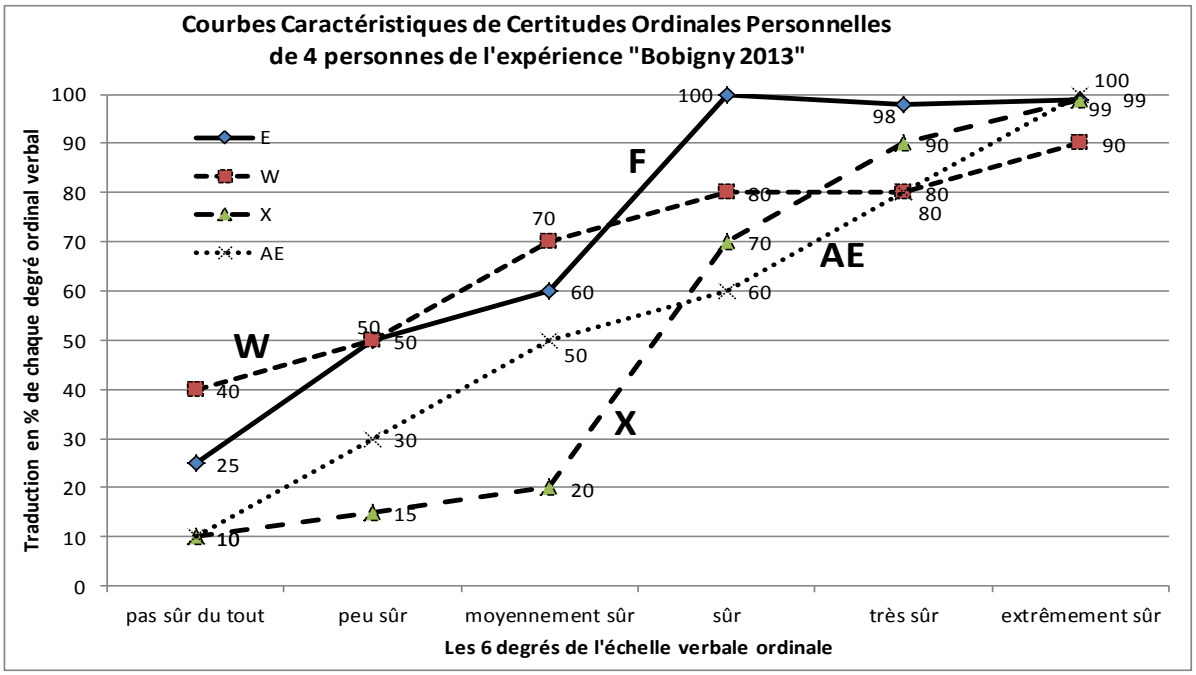


Figure 13 : CPCO de 4 des 33 participants

On peut faire les constats suivants.

1. Les courbes caractéristiques (cpco) des 4 personnes se présentent sous forme de lignes brisées monotones (à aucun moment décroissantes), à une seule exception près : pour F avec « sûr » (100%) et « très sûr » (98%). Ainsi, ces 4 personnes sont cohérentes avec elles-mêmes dans la traduction des mots en nombres (en %).
2. Les courbes (cpco) de certaines personnes (AE, et un peu moins W) sont très « progressives » : elles se présentent sous forme d'une ligne brisée proche d'une droite, donc avec une corrélation entre les certitudes en mots et les certitudes en pourcentages proche de 1, mais
   - la courbe de la personne W a une valeur minimale (40%) très élevée et une valeur maximale (90%) la moins élevée des 4 personnes. On pourrait l'appeler « peu discriminante » : la Marge de Variation (MV) étant 90 - 40 = 50%.
   - la courbe de la personne AE est beaucoup plus « discriminante » (que celle de la personne W) car son minimum est 10% et son maximum 100%, sa MV 90%.

3. Les courbes de certaines personnes (X et F) ont une forme de S majuscule, mais la personne X ne fait que très peu de distinction entre les 3 niveaux du bas (Traductions = 10%, 15%, 20%), et la personne F ne fait pas de distinction entre les 3 niveaux du haut (Traductions = 100%, 98%, 99%).

A partir de là, on devine l'impossibilité pour un évaluateur de comparer entre elles (donc d'interpréter) des données verbales dont il ignore la signification numérique que se donne *in petto* chacun des évalués.

7.3 Les courbes personnelles de certitudes ordinales en contexte (2016)

La figure 14 présente les cpco de 5 participants lors du PREtest de l'expérience en contexte.

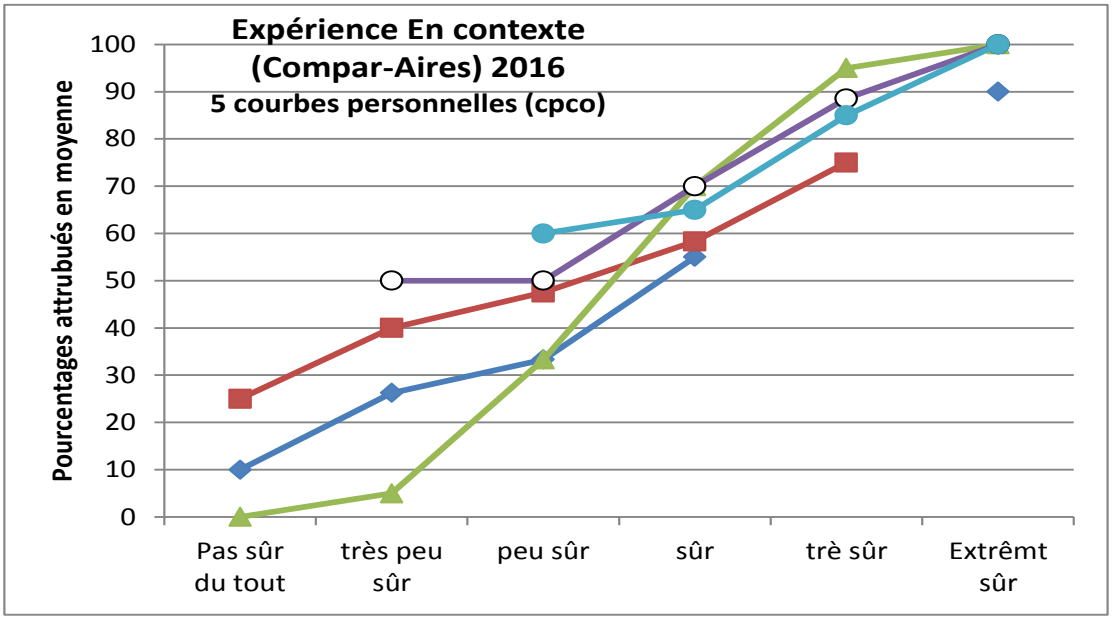


Figure 14 : CPCO de 5 des 19 participants

Certains de ces participants n'ont utilisé que 4 ou 5 des six niveaux ordinaux de certitude. Néanmoins, la ressemblance avec les cpco de l'expérience hors contexte (figure 13) est frappante. On trouvera en annexe 5 des exemples d'autres cpco pour les deux expériences.

7.4 Les Marges de Variation (MV) ou épaisseur du brouillard dans un groupe

La figure 15 ci-après montre les 33 lignes brisées individuelles (ou cpco) de l'expérience hors contexte en PRE. Attention : beaucoup de lignes obliques se superposent. Certains participants ne sont pas cohérents.

La figure 16 montre les 19 cpco de l'expérience en contexte, en PRE. Certains participants n'ont pas utilisé tous les (6) degrés de certitude verbaux.

Dans les deux graphiques, les droites verticales en pointillés montrent les Marges de Variation (MV ou écart entre l'estimation observée la plus élevée et la plus faible). Les valeurs (numériques) de ces MV sont énormes, comme on peut aussi le constater dans le tableau 4.

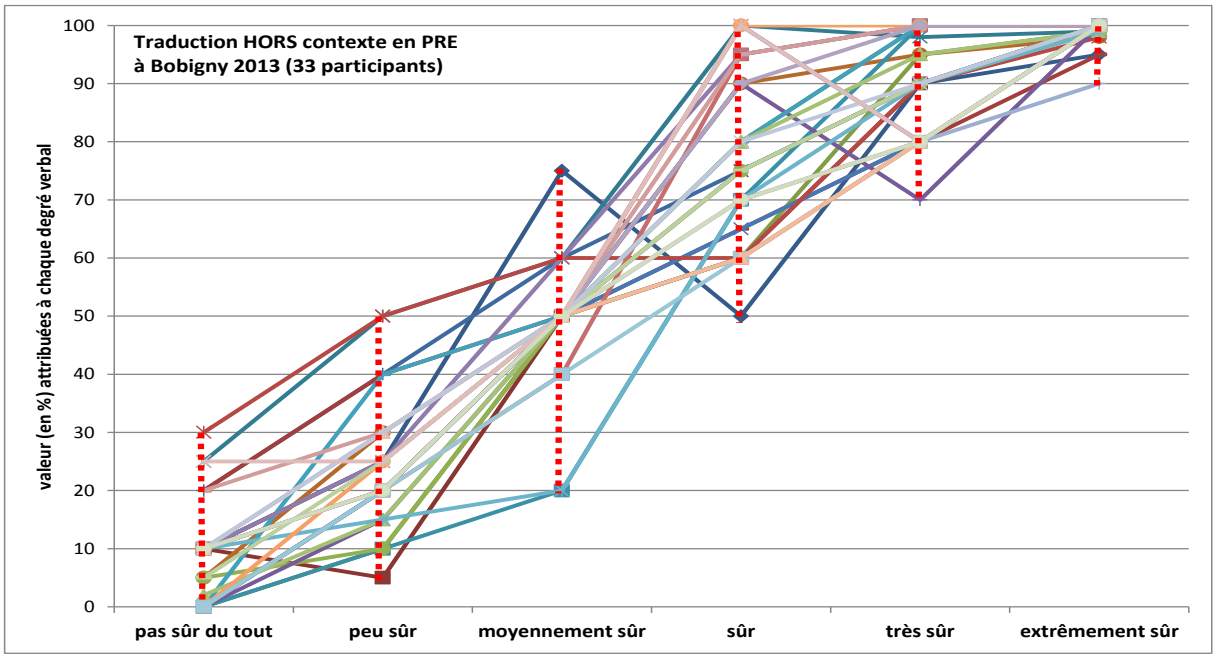


Figure 15 : Superposition des 33 cpco de l'expérience hors contexte en PRE

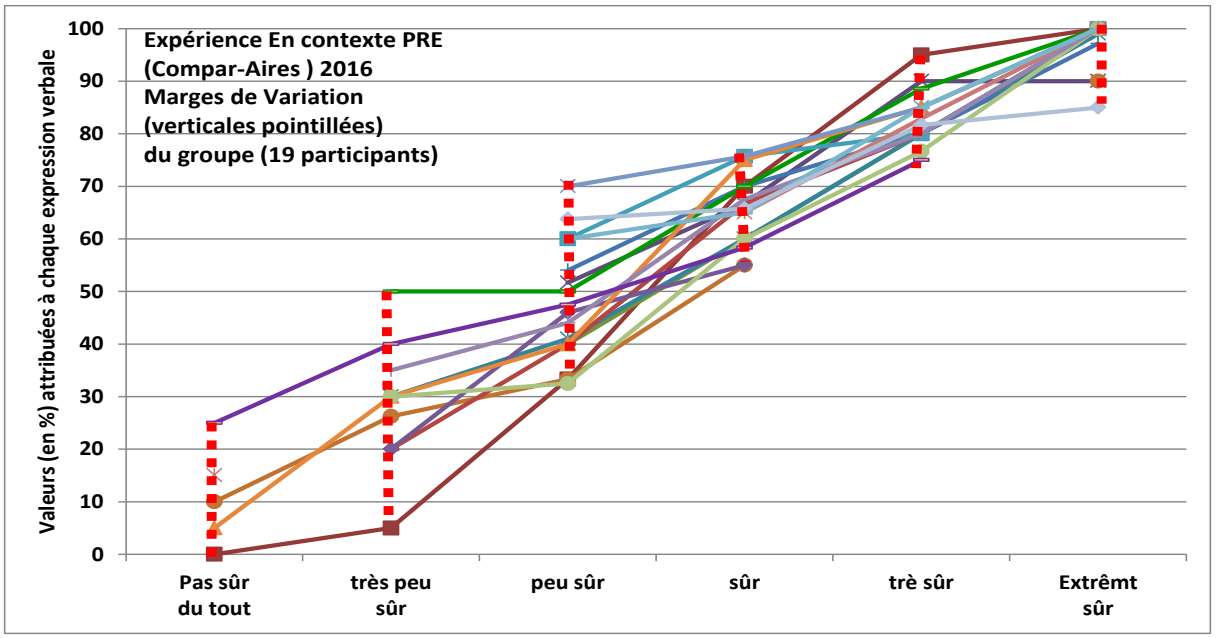


Figure 16 : Superposition des 19 cpco de l'expérience en contexte en PRE

Dans le tableau 4, le nombre d'observations est 33 pour chacun des 6 degrés de certitude verbaux dans l'expérience HORS contexte. Par contre, dans l'expérience EN contexte, le nombre d'observations diffère d'un degré de certitude à l'autre (cf. figure 12), ce qui est rappelé dans la dernière ligne (N (Total = 285).

Tableau 4 : Marges de Variation des valeurs numériques attribuées aux 6 expressions verbales

HORS contexte	pas sûr du tout	peu sûr	moyennt sûr	sûr	très sûr	extrêmemt sûr
HORS PRE *(33)*	**30**	**45**	**55**	**50**	**30**	**10**
HORS POST *(33)*	**50**	**40**	**50**	**50**	**30**	**20**
EN PRE	**25**	**45**	**35**	**20**	**20**	**15**
EN POST	**30**	**65**	**50**	**40**	**30**	**40**
N (Total = 285)	*11*	*26*	*73*	*69*	*44*	*62*
EN contexte	**pas sûr du tout**	**très peu sûr**	**peu sûr**	**sûr**	**très sûr**	**extrêmemt sûr**

On voit que, qu'il s'agisse de traduction des expressions verbales de certitude HORS contexte, ou EN contexte, sauf pour « extrêmement sûr » (et encore), ces marges de variation sont énormes.

7.5 Les Ecarts-types

Les marges de variation sont des indices extrêmes de l'ampleur des divergences. Les écarts-types en donnent une image plus pondérée.

Tableau 5 : Ecarts-types des valeurs numériques attribuées à chacune des 6 expressions verbales par des personnes différentes

HORS	pas sûr du tout	peu sûr	moyennement sûr	sûr	très sûr	extrêmemt sûr
HORS PRE *(33)*	**8,3**	**10,7**	**9,7**	**14,7**	**8,7**	**2**
HORS POST *(33)*	**10,1**	**9,6**	**8,2**	**14,6**	**8**	**5,4**
EN PRE	**9,6**	**12,1**	**10,8**	**6,5**	**5**	**4,5**
EN POST	**11,5**	**16,8**	**11,6**	**6,9**	**5**	**5,7**
N (Total = 285)	*11*	*26*	*73*	*69*	*44*	*62*
EN	**pas sûr du tout**	**très peu sûr**	**peu sûr**	**sûr**	**très sûr**	**extrêmemt sûr**

Les écarts-type des pourcentages attribués ont des valeurs proches de 10% en général. En contexte, si l'on élimine la première colonne où les valeurs sont calculées sur seulement 11 cas, on constate qu'à mesure que la certitude s'élève, les écarts-types se réduisent.

Il faut garder à l'esprit ce que signifie un écart-type par exemple de 10% (la valeur autour de laquelle tournent la plupart des écarts-types présentés dans le tableau 5) pour une expression verbale à laquelle un groupe attribue en moyenne la valeur 60% (par exemple). En théorie (pour une distribution normale),

- pour deux tiers (plus exactement 68%, c'est-à-dire situés à ± 1 écart-type de cette moyenne) du groupe, il faut considérer pour la certitude 60% un intervalle de confiance dont les valeurs limites sont 50% et 70%.
- pour que 90% du groupe (1,58 écart-types) soient inclus dans l'intervalle de confiance, celui-ci devrait aller de 44% à 76%, soit une ampleur de 32%.

Ce qui équivaut aux marges de variation du tableau 3.

Pour 15 mots traduits en pourcentage par 28 personnes, Johnson (1973, p. 7) a observé des écarts-types compris entre 12% et 21%.

7.6 Un résumé visuel des données « test » (ou PRE) des deux expériences

Les données des tableaux 3 (sur les Moyennes du groupe) et 4 (sur les marges de variations) peuvent être rassemblées en un graphique. La figure 17 rassemble ces données pour les deux expériences « test » (ou PRE), hors contexte (traits pleins) et en contexte (traits pointillés).

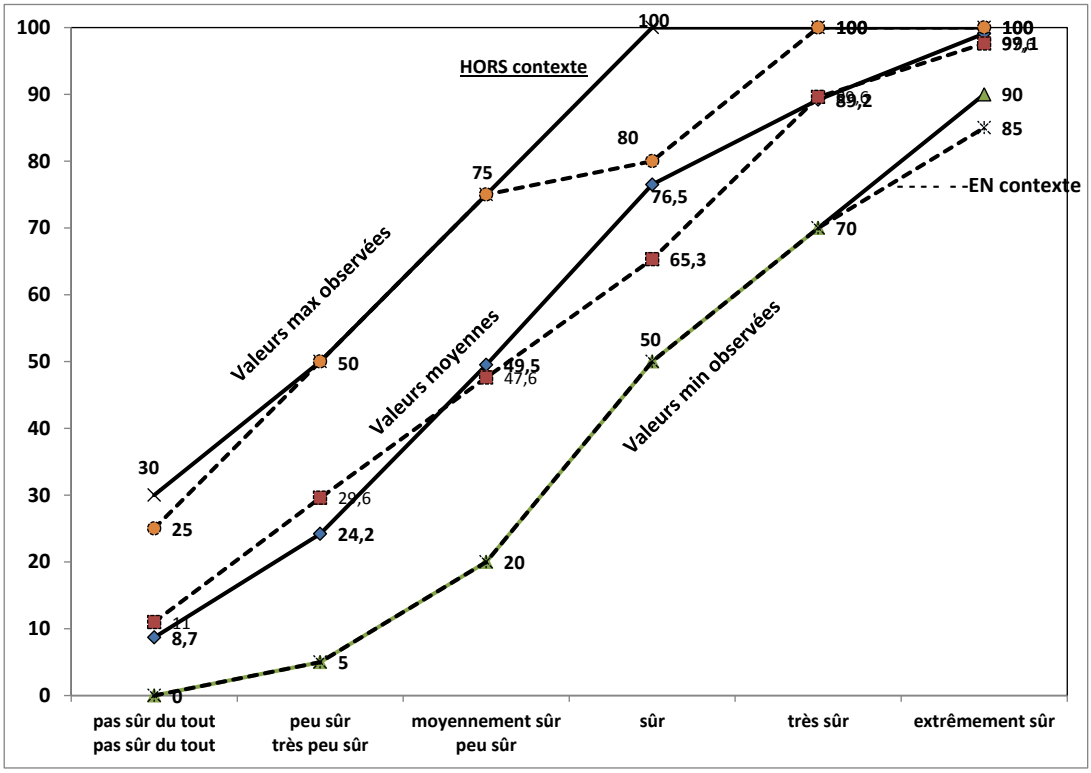


Figure 17 : Récapitulatif des valeurs des deux expériences, hors et en contexte, en PRE

Dans la figure 17 (voir tableau 6 en annexe 6), on est frappé par la proximité de la plupart des valeurs issues des deux expériences alors qu'elles différaient fortement (1) par le fait que l'une était hors contexte et l'autre en contexte et (2) par le fait qu'elles ont mobilisé des participants différents.

Simpson (1963), Hillson (1973) et Budescu et al. (1988) ont obtenu, par observation directe (et non par la méthode de l'intervalle d'acceptabilité), le même genre de données respectivement pour 25, 15 et 15 expressions verbales.

7.7 Comparaison avec les données de Fares (2006) pour d'autres contextes

Les ressemblances entre les marges de variation (tableau 4 et figures 15, 16 et 17) hors contexte et en contexte sont frappantes et rejoignent, comme on va le voir, les observations obtenues par Hamm (1991) et par Fares (2006) par la méthode de l'intervalle d'acceptabilité. Ainsi, Fares a proposé à 81 participants de déterminer hors contexte les limites inférieure (LI) et supérieure (LS) d'acceptabilité de 11 significations numériques (de 0 à 10) pour 7 mots différents. Il avait (p. 106) calculé la moyenne des LI et des LS pour chaque mot. Il avait joint ces valeurs moyennes sur un graphique et obtenu des courbes de la (même) forme logistique exponentielle (voir figure 18).

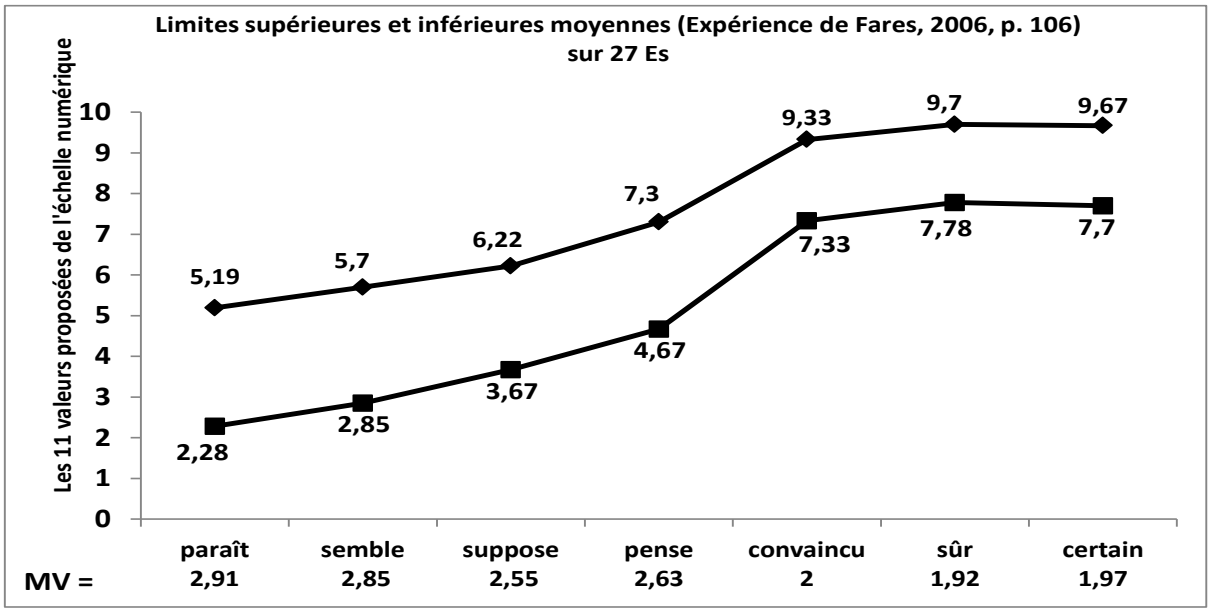


Figure 18 : Valeurs numériques minimales (Li) et maximales (Ls) moyennes d'acceptabilité (sur 27 participants) pour chacune des 7 expressions d'incertitude de l'échelle verbale observées par Fares (2006, p.106) Graphique reproduit avec l'autorisation de l'auteur.

Ce que j'appelle « épaisseur du brouillard » dans le titre G4 est à peu près le même que la zone d'acceptabilité observée par Fares, sauf que, chez ce dernier, c'est de l'aveu même des répondants.

8. L'hypothèse 5 : La similitude des résultats hors contexte et en contexte

8.1 L'origine de l'hypothèse

Fares (2006) a demandé à 27 des 81 personnes interrogées d'attribuer une valeur numérique (entre 0 et 10) pour les mêmes 7 expressions verbales (ex : Je suppose que…), mais en contexte cette fois-ci, plus exactement dans 3 situations différentes :

« Je suppose que cet arbre mourra un jour » : échelon ? .....

« Je suppose qu'au carrefour ce conducteur va tourner à droite » : échelon ? .....

« Je suppose qu'une invasion d'extra-terrestres est imminente » : échelon ? .....

Les mêmes trois fins de phrases étaient proposées, mais avec chacune des 6 autres expressions verbales (voir figure 18). La figure 19 présente ses résultats. Les lignes continues sont les valeurs moyennes des significations numériques attribuées à chacun de ces 7 mots dans les 3 situations différentes. Ces valeurs sont à l'intérieur des zones d'acceptation (entre limite inférieure et limite supérieure, en pointillés) définies par ce groupe. Ce qui montre la cohérence, chez ces étudiants, entre des valeurs données hors contexte (limites inférieure et supérieure de la zone d'acceptation) et en contexte (les 3 situations).

En outre, les courbes des trois expériences en contexte sont de forme similaire, avec des décalages verticaux dus aux différences entre contextes.

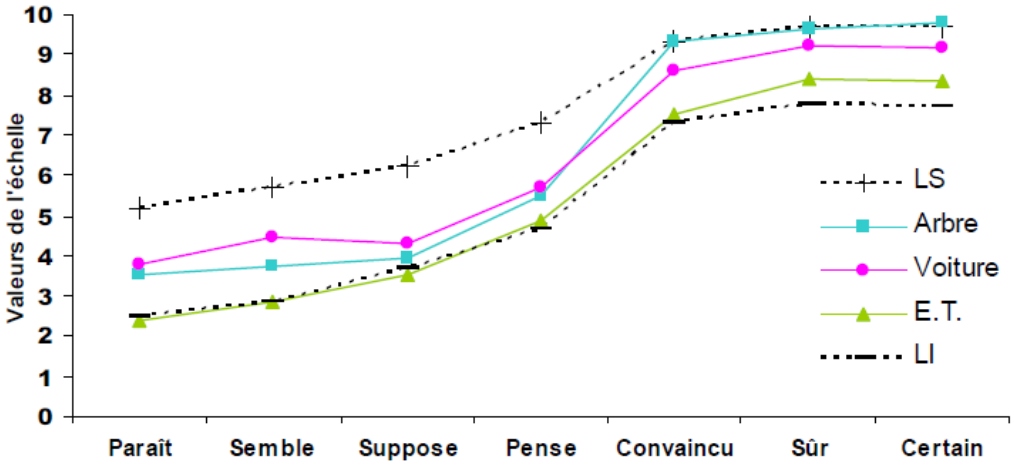


Figure 19 : Zones d'acceptation (en pointillés) et alignements moyens dans trois contextes différents (Fares, 2006, p. 108). Graphique reproduit avec la permission de l'auteur.

Ces données m'ont amené à penser que les résultats de mes deux expériences (hors et en contexte) seraient semblables entre elles.

8.2 Les résultats comparés entre les deux expériences

Pour l'hypothèse 1 (préférences des multiples de 10%), la comparaison des figures 4 (expérience de 1975), 5 (hors contexte en 2013) et 6 (en contexte en 2016) met en évidence la similitude des préférences pour des multiples de 10%, avec un décalage dû à la difficulté des tâches (hypothèse 3). Dans les traductions de certitudes exprimées verbalement en pourcentage, l'hypothèse 3 (confortée elle aussi) ne pouvait être mise à l'épreuve qu'en contexte.

Pour l'hypothèse 2 (grande fidélité de répétition intra-personnelle), les figures 9 (hors contexte) et 10 (en contexte) montrent la similitude des résultats dans les deux expériences.

Pour l'hypothèse 4 (la grande variabilité interpersonnelle), la comparaison entre les figures 13, 14, 15 et 16 montre ici aussi la similitude des résultats dans les deux expériences.

9. Conclusions et perspectives

9.1 Synthèse

Les 5 hypothèses, quant aux traductions de degrés de certitude verbaux vers des pourcentages, sont confortées par les observations :

1. les participants marquent une nette préférence pour les pourcentages multiples de 10% ;
2. la répétabilité intra-individuelle est élevée ;
3. les degrés de certitude semblent sensibles à la difficulté des questions ;
4. les différences interindividuelles d'interprétation en pourcentages des certitudes verbales sont énormes. On peut résumer grossièrement les différences individuelles d'interprétation par

   MV (Marge de Variation) = entre 30% et 50%
   ET (Ecart-Type) = entre 5% et 15% ;

5. Les résultats sont convergents, qu'il s'agisse d'une situation expérimentale hors contexte ou en contexte.

Recueillir les degrés de certitude par des expressions verbales, c'est donc introduire, dès le départ, un coefficient de « brouillard », et je dirais même de « brouillage », voire de « brouille » dans ses travaux, qu'ils soient de recherche ou à visée formative ou encore à visée docimologique.

L'expression « brouillard » est destinée à frapper les esprits. En termes plus rigoureux, il faut parler d'erreur de mesure (aléatoire) sur l'interprétation numérique des mots. Cette erreur de mesure aléatoire me paraît énorme et inutile.

9.2 Une consigne en pourcentage, à quelles conditions ?

Les résultats exposés ci-avant confortent ma conviction qu'il est préférable de demander aux étudiants d'exprimer directement leur degré de certitude en pourcentage de chance. Cependant, cela reste une conviction basée sur une série d'hypothèses, et cela le restera tant que certaines conditions ne seront pas remplies, à commencer par vérifier que les observations présentées dans cet article rejoignent celles de la littérature scientifique. C'est ce à quoi je me suis donc attelé en préparant une revue sur le sujet dans laquelle sont inclues les deux recherches présentées ici.

Bien d'autres conditions doivent cependant être réunies pour asseoir la pertinence de l'emploi des degrés de certitude dans la pratique scolaire. Je les préciserai au terme de la revue annoncée.

Le chemin à parcourir apparaît encore long, mais le plus important n'est-il pas, au départ, de ne pas se tromper de cap, de planter petit à petit des repères solides qui nous éviteront de tourner en rond ?

10. Références

Attneave, F. (1959). *Application of Information Theory to Psychology*. New-York: Holt: Rinehart & Winston.

Beyth-Marom, R. (1982). How probable is probable? A numerical translation of verbal probability expressions. *Journal of Forecasting, 1*(3), 257-269.

Bickel, J. (2005). *Probability Assessment. Strategic Decisions Group*., voir aussi https://www.coursehero.com/file/9906871/18-Probability-Assessment/

Bocklisch, F., Bocklisch, S.F., Baumann, M.R.K., Scholz, A. & Krems, J.F. (2010). The role of vagueness in the numerical translation of verbal probabilities: A fuzzy approach. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1974−1979). Austin, TX: Cognitive Science Society.

Boehm, B. (1989). *Software Risk Management*. Piscataway: NJ, USA: IEEE Computer Society Press.

Budescu, D. & Wallsten, T. (1985). Consistency in interpretation of probabilistic phrases. *Organizational Behavior and Human Decision Processes, 36*, 391-405.

Budescu, D., Weinberg, S. & Wallsten, T. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance, 14*(2), 281-294.

Campbell, D., Lewis, N. & Hunt, W. (1958). Context effects with judgmental language that is absolute, extensive, and extra-experimentally anchored. *Journal of Experimental Psychology, 55*, 220-228.

Clark, D. (1990). Verbal uncertainty expressions: A critical review of two decades of research. *Current Psychology: Research and Reviews, 9*(3), 203-235.

Druzdzel, M. (1989). Verbal uncertainty expressions: Literature Review. Technical report CMU-EPP-1990-03-02. Department of Engineering and Public Policy. Pittsburgh, PA: Carnegie Mellon University.

Fabre; J-M. (1991). The expression of uncertainty: two contextual effects. *European Journal of Cognitive Psychology, 3*, 399-412.

Fabre, J-M (1993a). *Contexte et jugement. De la psychophysique à la responsabilité*. Lille : Presses universitaires de Lille.

Fabre, J.M. (1993b). Subjective Uncertainty and the Structure of the Set of all Possible Events. In D. Leclercq & J. Bruno (1993), *Item Banking: Interactive Testing and Self-Assessment*, NATO ASI Series, F 112, Berlin: Springer Verlag, 99-113.

Fares, N. (2006). *Effet de la formulation des expressions d'incertitude, (interne ou externe) sur le choix et la prise de décision*. Thèse de doctorat Université de Provence.

Gigerenzer, G. & Hoffrage, U. (1998). Using Natural Frequencies to Improve Diagnostic Inferences. *Academic Medicine, 73*(5), 538-540.

Hamm, R. (1991). Selection of verbal probabilities: A solution for some problems of verbal probability expressions. *Organisational behaviour and Human Decision Processes, 48*, 193-223.

Hillson, D. (2005). Describing probability: The limitations of natural language. *PMI Global Congress 2005 EMEA Proceedings*. Edimburgh. http://www.risk-doctor.com/pdf-files/emeamay05.pdf.

Johnson, E. (1073). *Numerical encoding of qualitative expressions of uncertainty*. (Techn. Paper 250). Arlington, VA: US Army Research Institute for the Behavioural and Social sciences. http://www.dtic.mil/dtic/tr/fulltext/u2/780814.pdf.

Kent, S. (1994). Words of estimative probability. In Sherman Kent and the Board of National Estimates: *Collected Essays*, Washington, D.C.: ed. D. Steury. CIA. p. 130 :

Leclercq, D. (1975). *L'évaluation subjective de la probabilité d'exactitude des réponses en situation pédagogique*. Thèse de doctorat en Sciences de l'Education, Université de Liège. http://hdl.handle.net/2268/10119, 37 Mo, 520 p.

Leclercq, D. (1980). Computerised tailored testing: structured and calibrated item banks for summative and formative evaluation. *European Journal of Education*, *15*(3), 251-260. http://hdl.handle.net/2268/18555

Leclercq, D. (1982). Confidence marking, its use in testing. In N. Postlethwaite & B. Choppin (Eds), *Evaluation in Education, 6*, 161-287. http://hdl.handle.net/2268/9482

Leclercq, D. (1986). *La conception des questions à choix multiple.* Bruxelles : Labor. http://hdl.handle.net/2268/17835

Leclercq, D. (1987). *Qualités des questions et signification des scores*. Bruxelles : Labor http://hdl.handle.net/2268/17836

Leclercq, D. (1990). Intelligent Tutorial and Self Training Systems, *Proceedings of the International AI Symposium (IAIS 90)*, Nagoya, Japan, 127-135. http://hdl.handle.net/2268/18528

Leclercq, D. (Ed) (2003). *Diagnostic cognitif et métacognitif au seuil de l'université. Le projet MOHICAN mené par les 9 universités de la Communauté Française Wallonie Bruxelles.* Liège : Editions de l'université de Liège. http://hdl.handle.net/2268/28353

Leclercq, D. (2009). La connaissance partielle chez le patient : pourquoi et comment la mesurer. *Education Thérapeutique du Patient. 1*(2), 201-212. http://hdl.handle.net/2268/18728

Lichtenstein, S. & Newman, R. (1967). Empirical Scaling of Common Verbal Phrases Associated with Numerical Probabilities. *Psychon. Sci., 9*, 563-564.

Lord, F. & Novick, M. (1968). Statistical theories of Mental Tests Scores. Reading, MA: Addison-Wesley.

O'Brien, B. (1989). Words or numbers? The evaluation of probability expressions in general practice. *Journal of the Royal College of General Practitioners, 39*, 98-100.

Preston, C. & Colman, A. (2000). Optimal number of categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica. 104*, 1-15.

Reagan, R., Mosteller, F. & Youtz, C. (1989). The quantitative meaning of verbal probability expressions. *Journal of Applied Psychology, 74*, 433-442.

Simpson, R. (1944). The specific meanings of certain terms indicating degrees of frequency. *Quarterly Journal of Speech, 30*, 328-330.

Simpson, R. (1963). Stability in meanings for quantitative terms: a comparison over 20 years. *The Quarterly Journal of Speech, 49*, 146-151.

Annexe 1

Consigne de Fares (2006, p. 150), avec la permission de l'auteur

Imaginez que quelqu'un vous dise : « Il est sûr que l'événement X aura lieu ».
Selon vous sur une échelle qui va de 1 à 10 (où le 1 signifie l'incertitude totale et le 10 la certitude totale), est-ce que le chiffre « 7 » traduit bien l'incertitude exprimée par la personne (entourez votre réponse) ?

Ne traduit pas du tout A—B—C—D—E—F Traduit parfaitement

Selon vous jusqu'à quel point le chiffre « 4 » traduit bien l'expression

Ne traduit pas du tout A—B—C—D—E—F Traduit parfaitement

Selon vous jusqu'à quel point le chiffre « 8 » traduit bien l'expression

Ne traduit pas du tout A—B—C—D—E—F Traduit parfaitement

etc.

La même consigne est appliquée pour « je suis sûr que… », « il est certain que… » et « je suis certain que… ».

Annexe 2

Indices de Difficulté (taux de réponses correctes) des 15 questions de l'expérience en contexte (2016)

Cet indice est forcément le même pour chaque question au PRE comme au POST puisque lors du POST, les étudiants ne pouvaient pas changer les réponses qu'ils avaient données deux heures plus tôt. La figure 26 se lit comme suit : 1 question a été réussie par 9 étudiants, 2 questions ont été réussies par 10 étudiants…et finalement 2 questions ont été réussites par les 19 étudiants.

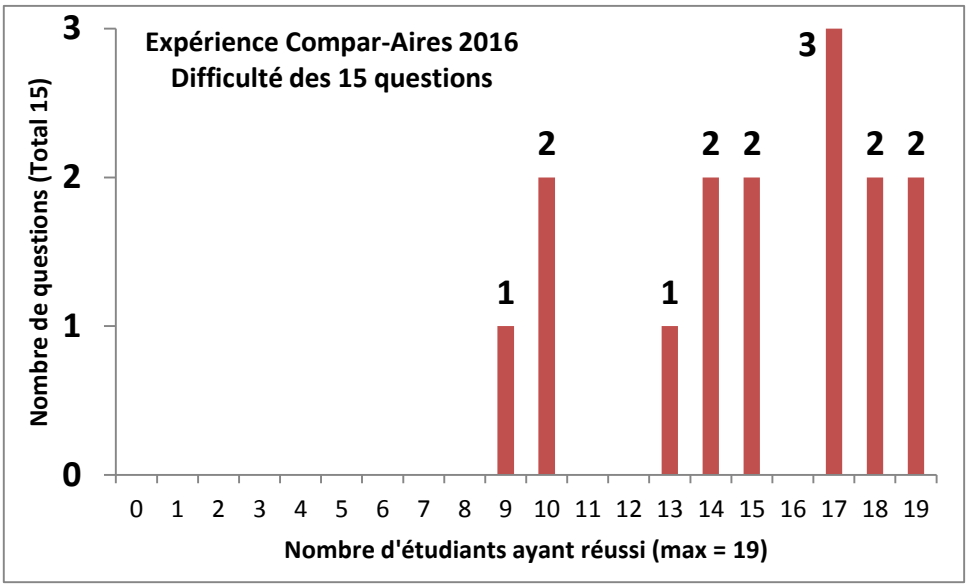


Figure 26 : Répartition des indices de difficultés (nombre de réussites) des 15 questions dans l'expérience en contexte (2016)

Annexe 3

Certitudes moyennes chez les 19 participants en contexte au test ou PRE dans l'expérience en contexte (2016)

La figure 25 montre la répartition des certitudes moyennes des 19 étudiants. Les 19 certitudes moyennes ont été regroupées en zones de 5 % en 5%.

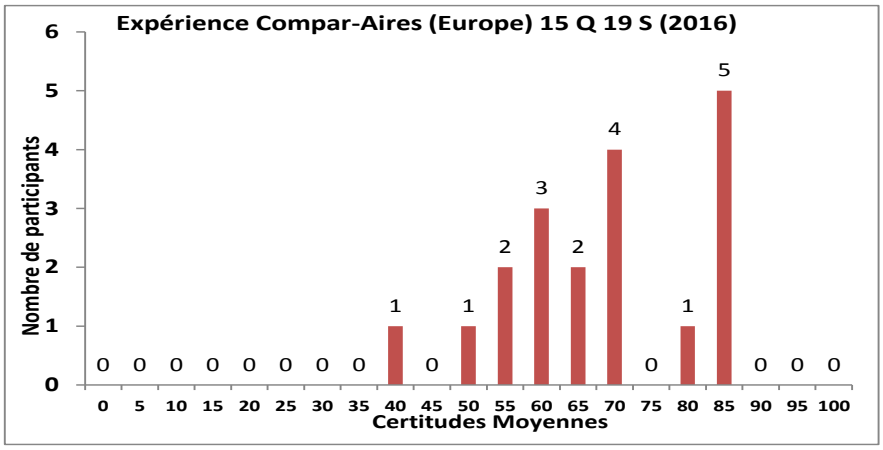


Figure 25 : Zones de « certitudes moyennes » des 19 participants à l'expérience en contexte PRE

La moyenne du groupe est 65,3% (Ecart-Type : 13,6%). La certitude moyenne minimale observée est 35,7% et la certitude moyenne maximale observée 84,6%. (Marge de variation : 48,9%).

Annexe 4

Les taux de réussite pour les 19 participants de l'expérience en contexte (2016)

Pour chacun de ces étudiants a été calculé le nombre de réponses correctes (NRC) sur un maximum de 15. La figure 24 montre la distribution des scores des 19 étudiants, avec pour maximum observé 15, pour minimum 5, pour pic ou Mode 13 (87%, ce qui est élevé), pour Moyenne 11,8 (78,6%) et Ecart-type 2,3 (15%). Les questions étaient donc trop faciles pour l'expérience (l'idéal statistique aurait été une Moyenne de 50%). Le haut degré de facilité, par contre, était favorable à la motivation des participants (il y a plus de plaisir à la réussite qu'à l'échec).

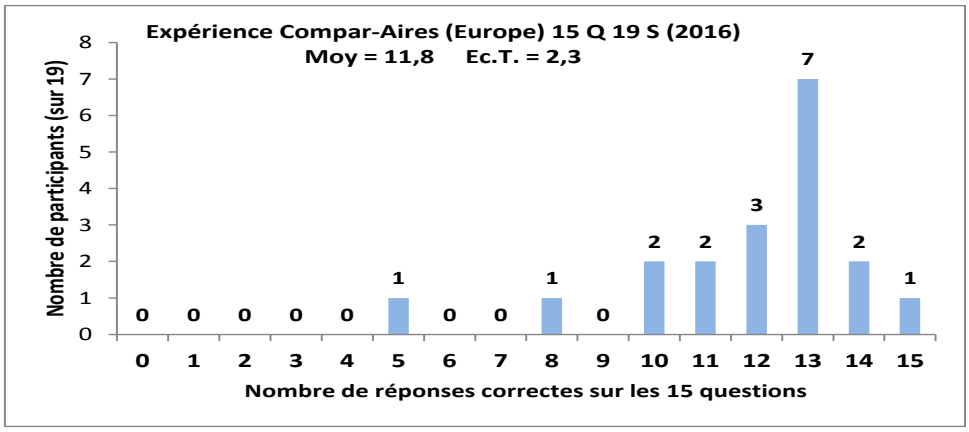


Figure 24. Distribution des 19 scores totaux (maximum possible = 15)

Annexe 5

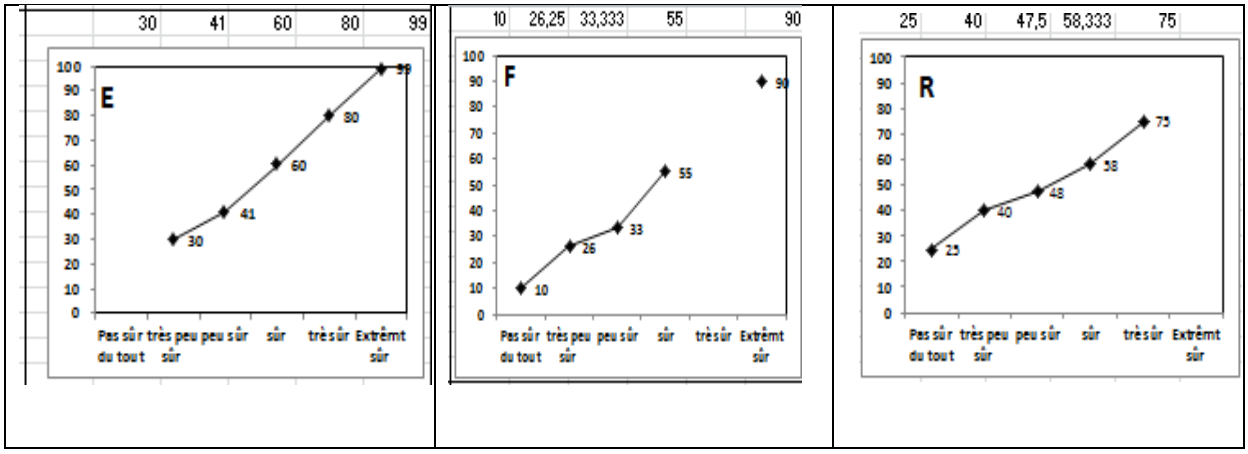
Graphiques de correspondance (alignements ou ccop) Certitudes en mots et en % de certains participants

a) Dans l'expérience Hors contexte

Les quelques graphiques de correspondance individuels entre les mots et les pourcentages qui suivent illustrent que, pour un même nombre de degrés de certitude utilisés (4 ou 5), les configurations peuvent varier d'un participant à l'autre.

Dans ces graphiques, l'axe vertical représente la valeur numérique moyenne (en %) attribuée à chaque mot par la personne. Souvent cette moyenne est une seule valeur, quand l'expression verbale n'a été choisie qu'une seule fois.

Exemples de participants ayant utilisé 5 des six degrés verbaux :



Exemples de participants ayant utilisé 4 des six degrés verbaux :

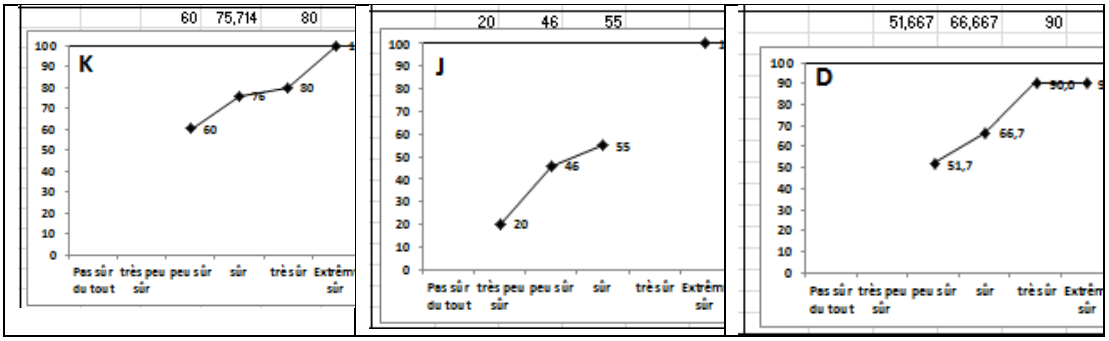


Figure 22 : Alignement (cpco) des certitudes moyennes par degré de certitude verbal pour six participants (E, F, R, K, J et D) dans l'expérience hors contexte

b) Dans l'expérience En contexte

L'alignement des valeurs numériques pour les degrés successifs peut faire l'objet d'une corrélation et être appelé la Cohérence personnelle (Leclercq, 2003, p. 36-37).

Pour les 19 participants, cet alignement (ou cohérence) est compris entre 0,95 (2 participants) et 1 (5 participants), la Moyenne étant 0,98.

Ces corrélations ont été calculées sur 15 réponses pour chaque participant. Or ils n'ont pas tous utilisé les 6 niveaux (verbaux) de certitude. Deux d'entre eux (B et L) ont utilisé les 6 degrés verbaux. Voici, figure 23 leur graphique de correspondance entre mots et %. L'indice (corrélation ordinale) de cohérence est 0,98 pour B et 0,99 pour L.

A nouveau, dans ces graphiques, l'axe vertical représente la valeur numérique moyenne (en %) attribuée à chaque expression verbale choisie par la personne.

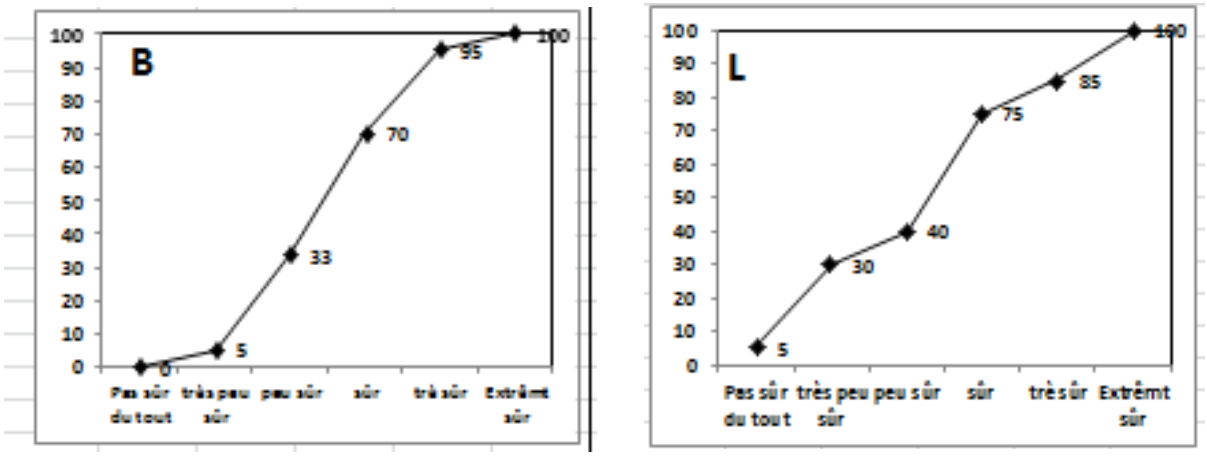


Figure 23 : Alignement (cpco) des certitudes moyennes par degré de certitude verbal pour deux participants (B et L) dans l'expérience en contexte

Annexe 6

Tableau récapitulatif des M et des MV

Le tableau 6 présente les valeurs des Moyennes et des Marges de variation, données qui ont servi à établir le graphique de la figure 17.

HORS CONTEXTE (NS = 33)	pas sûr du tout	peu sûr	moyenmt sûr	sûr	très sûr	extrêmemt sûr
Moyenne au test ou PRE	**8,7**	**24,2**	**49,5**	**76,5**	**89,2**	**99,1**
Moyenne au retest ou POST	**10,6**	**24,2**	**48,8**	**79,5**	**89,5**	**97,5**
min au test (pre)	0	5	20	50	70	90
min au retest (post)	0	10	20	50	70	80
max au test (pre)	30	50	75	100	100	100
max au retest (post)	50	50	70	100	100	100
Moyenne au test ou PRE	**11**	**29,7**	**47,6**	**65,3**	**82,6**	**97,6**
Moyenne au retest ou POST	**10,5**	**29,1**	**47,4**	**67,2**	**83,5**	**96,7**
min au test (pre)	0	5	20	50	70	85
min au retest (post)	0	5	30	50	70	60
max au test (pre)	25	50	75	80	100	100
max au restest (post)	30	70	80	90	100	100
EN CONTEXTE (NS = 19)	**pas sûr du tout**	**très peu sûr**	**peu sûr**	**sûr**	**très sûr**	**extrêmemt sûr**

Tableau 6 : Récapitulatif des Moyennes et Ecarts-types pour les deux expériences : hors contexte (2013) et en contexte (2016)