Une mata-analyse des degrés de certitude exprimés en mots

A mata-analysis of confidence degrees expressed in words

Dieudonné LECLERCQ
Université de Liège, Département Education et Formation
mailto:d.leclercq@ulg.ac.be

Résumé.
Demander aux étudiants d'accompagner chacune de leurs réponses à un test par un degré de certitude est une pratique encore assez rare. La majorité des enseignants qui la découvrent utilisent spontanément une échelle verbale. L'hypothèse de l'auteur est que cela revient à introduire, dès le départ, une énorme erreur aléatoire de mesure car il existe de grandes variations interindividuelles dans l'interprétation (la traduction) en pourcentage des mots de l'échelle verbale, comme l'a montré Leclercq (2016) par deux expériences, l'une hors contexte et l'autre en contexte. Une grande convergence entre ces deux études apparaît quant à l'épaisseur du brouillard communicationnel introduit par des mots pour exprimer des valeurs qui pourraient / devraient l'être en pourcentage (de 0% à 100%). Les marges de variation (MV) des « traductions » de mots (tels que « peu sûr », « sûr », « très sûr », etc.) en % ont une valeur modale de 40% et les écarts-types, une valeur comprise entre 10% et 15%. L'hypothèse de l'auteur est donc confortée par ces deux expériences, mais qu'en est-il dans la littérature ? Dans une première partie du présent article, une révision de la littérature « à charge », appelée mata-analyse, accumule des données allant dans le sens de l'hypothèse. Dans une deuxième partie, sont mises en évidence les valeurs numériques les plus fréquemment utilisées par les répondants dans différentes recherches de traduction mots – pourcentages, ainsi que les valeurs qui seraient optimales non seulement en termes de préférences des répondants, mais surtout en termes de la précision (ou granularité) maximale qui reste fiable (ou « fidèle » en termes de répétabilité dans un court intervalle de temps).

Mots-clés
Degrés de certitude, variations interindividuelles, expressions verbales, mata-analyse, méta-analyse, granularité, précision, fiabilité

Abstract
Asking to students to add a confidence degree to each of their responses to a test is rather rare, and the large majority of those who practice that use verbal scales such as "weakly sure", "sure", "strongly sure", etc. instead of probabilities or percentages of chances. My hypothesis is that consists in introducing, from the beginning, an enormous random error of measurement since there exist large differences in the interpretation (the translation) into percentages of the word used in verbal scales. I demonstrated this in two experiences (Leclercq, 2016), one in context and one context free. A strong convergence appear in the results of the two experiences in terms of communicational fog produced by words in place of percentages (from 0% to 100%). Variation Ranges (VR) of translations of words into percentages have a modal value of 40% and standard deviations (SD) from 10% to 15%. Therefore my hypothesis is confirmed by these two experiences, but what does the specialized literature say? In a first part of the present article I have browsed many reviews and books with the purpose to find data that contribute to fight (and kill?) this habit of using words instead of percentages to express confidence degrees. Therefore I have named my method a "*mata*" analysis (*matar* meaning

"to kill" in Spanish), that is distinct from meta-analysis, as will be shown. In the second part of the article, I underline data that help approach what could be the optimal number of numerical degrees (and their exact values) not only in terms of students' preferences, but mainly in terms of the reliability (measured but the repeatability criterion, i.e.; stability in a short period of time) of the declared confidence,

Keywords
Confidence degrees, inter-individual variations, verbal expressions, mata-analysis, meta-analysis, reliability, precision, granularity

Pour citer cet article : Leclercq, D. (2017). Une mata-analyse des degrés de certitude exprimés en mots. *Evaluer. Journal international de Recherche en Education et Formation, 2*(3), pp. 69-105.

1. Le concept de degré de certitude

1.1 Définition

Une définition approfondie du concept de « degré de certitude en évaluation (et en auto-évaluation) pédagogique » nécessiterait une page entière, car, pour asseoir une telle définition, bien des concepts préalables devraient être explicités, à commencer par ceux qui vont être présentés dans cet article. C'est pourquoi on se limitera ici à la définition simplifiée suivante :

> *« un degré de certitude est l'expression par une personne (un étudiant par exemple), pour chacune de ses réponses (à un test par exemple) de sa probabilité subjective (allant de nulle à totale) que sa réponse sera jugée correcte par le correcteur (l'enseignant par exemple) ».*

On remarquera que cette définition minimale ne précise pas comment se fait cette « expression ». Il est possible en effet de recourir à diverses modalités, non seulement numériques (ex : « 2 chances sur 10 » ou « 3 chances sur 4 » ou « 60% ») mais aussi par des expressions ordinales verbales (avec des mots). Voici, proposée par Farès (2006, p. 50), une typologie de ces expressions verbales (et les auteurs qui y ont eu recours) :

> *«*
> - *probabilistes (ex : « probable », « presque certain », etc.),*
> - *de fréquences (ex : « souvent », « la plupart du temps », « jamais », « rarement », « fréquemment », « presque jamais ») chez Bryant & Norman, 1980 ; Budescu & Wallsten, 1985 ; Kong, Barnett, Mosteller et Youtz, 1986 ; Lichtenstein & Newman, 1967 ;.*
> - *comparatives (ex : « le plus vraisemblable » chez Beyth-Marom, 1982 ;*
> - *logiques (ex : …ne peut être exclu »), chez Bryant & Norman, 1980 ;*
> - *autres (ex : « imprévisible », « courant » chez Budescu & Wallsten, 1985 ou encore « nécessairement », « faisable », « concevable », chez Reyna, 1981. ».*

Or ce choix est un problème majeur. En effet, un trop grand nombre des utilisations qui ont été faites des degrés de certitude en éducation ont été décevantes. Ainsi, des revues pessimistes ont été publiées par Koehler (1971), Echternacht (1972a et b), Echternach et al. (1976) et par Frary (1989). Ce dernier auteur conclut (p. 92) sa revue par ces mots : « Ce qui caractérise le plus les recherches empiriques étudiées ici c'est la fréquence de résultats contradictoires. Ce qui semble bien fonctionner dans un cas ne fonctionne pas dans un autre cas ou lors d'une réplication ». Le recours à des degrés de certitude verbaux n'est qu'une des causes (mais la première) de cette déception. Le présent article n'aborde pas les autres causes (par exemple le système de notation ou *proper scoring rule*). Il n'aborde pas non plus les raisons des préférences des utilisateurs pour les certitudes verbales ni les conséquences d'un tel choix. Tous ces thèmes ne seront traités dans des articles ultérieurs qu'une fois démontré, par la présente revue de la littérature, que le brouillard communicationnel engendré par les échelles de mots est ample et général, c'est-à-dire quelles que soient ces échelles et quels que soient les contextes.

1.2 Un besoin de recherche fondamentale sur le sujet

Bien qu'un certain nombre d'utilisations scolaires soient rapportées dans la littérature, la capacité d'adultes de formuler des degrés de certitude a été trop peu étudiée en elle-même au préalable. Koriat (2012) souligne ce que Vickers (2001, p. 148) exprimait comme suit :

> *« Malgré son importance pratique et son omniprésence, la variable « certitude » semble n'avoir joué qu'un rôle de Cendrillon dans la psychologie de la cognition ; elle est reconnue pour son utilité, mais oubliée en tant que variable à étudier intéressante en soi ».*

En outre, les recherches sur le sujet (voir ci-après) ont été publiées dans des revues variées, allant de la psychophysique à l'économie en passant par la statistique, la médecine et le renseignement (ou services secrets – en anglais *intelligence*). Cette dispersion contribue au manque de visibilité de ces recherches.

Devant l'ampleur des différences interpersonnelles, de nombreux auteurs tels que Beyth-Marom (1982), Nakao et Axelrod (1983) ou Budescu et Wallsten (1985) recommandent d'utiliser des nombres et non des mots pour exprimer les degrés de certitude. L'essentiel de cet article vise à présenter un certain nombre de recherches (leurs méthodes et leurs résultats) qui ont amené l'auteur de cet article à la même conviction.

1.3 Les deux questions de recherche du présent article

La revue de la littérature qui suit porte sur deux questions de recherche :

Question de recherche 1 :

Quelle est l'ampleur des différences interindividuelles dans les traductions de degrés de certitude en mots vers des expressions numériques (spécialement des pourcentages de chances) ?

Ce sera le fil conducteur de l'article et les expériences seront présentées selon les différentes méthodes de recueil des données, autrement dit les différentes consignes aux répondants. On verra que les travaux cités confirment l'hypothèse de la grande variation interindividuelle des significations en pourcentages d'une même expression verbale. Comme le suggèrent d'autres auteurs, il nous faut demander aux étudiants d'exprimer directement leurs degrés de certitude en pourcentages. Se pose alors la question de la consigne, c'est-à-dire doit-on inviter les répondants à fournir n'importe quelle valeur comprise entre 0% et 100% (*continuous confidence marking*) ou doit-on leur proposer un nombre de valeurs entre lesquelles choisir ? Et, dans ce dernier cas, combien de valeurs et lesquelles ?

Question de recherche 2 :

Quelles sont les valeurs numériques (entre 0% et 100%) majoritairement choisies par les utilisateurs (notamment les enseignants) ?

Cette question vise à dégager ce que pourrait être une échelle optimale de pourcentages pour exprimer les degrés de certitude.

Dès que je me suis intéressé aux degrés de certitude, j'ai opté pour un nombre limité de degrés, à partir de 1971 pour 4 zones, puis à partir de 1990 pour 7 zones, puis enfin, à partir de 1999, pour 6 valeurs-repères (voir détails en annexe 1). Ce dernier choix (0%, 20%, 40%, 60%, 80%, 100%) rejoint le « *five stars system* » de De Finetti (1965) qui se base sur les multiples de 20% (5 étoiles = 100%, 4 = 80%, etc… et aucune étoile = 0%). Cette intuition est-elle confortée par la revue des recherches sur les traductions des mots vers les pourcentages ? La plupart des expériences décrites ci-après apportent des réponses aussi à cette deuxième question de recherche. Les résultats sont exprimés tantôt par les valeurs dans leur ordre de popularité (la valeur la plus populaire étant appelé le mode), tantôt par la médiane de ces valeurs, tantôt par leur moyenne… mais jamais les trois à la fois.

Le mode est évidemment l'indice le plus révélateur des préférences des répondants pour certaines des valeurs numériques possibles, celles qui leur viennent le plus souvent à l'esprit. Dans les recherches publiées, on trouve les valeurs moyennes, médianes ou modales des pourcentages utilisés pour traduire un degré de certitude exprimé en mots. Malheureusement, c'est la valeur modale qui, des trois, est la plus souvent absente des rapports de recherche.

La suite de cet article se structure en trois grandes parties, chacun comprenant une série de sections. La première partie présente le concept de mata-analyse. Les deux parties suivantes correspondent chacune à une des deux questions de recherche. La deuxième partie traite des différences inter-individuelles et la troisième partie aborde le problème du nombre et des valeurs optimales proposées aux sujets lorsque les degrés de certitude sont utilisés sous la forme de pourcentages.

Première partie

2. Une mata-analyse

2.1 Une méthode conçue pour vacciner

Dans cet article, il s'agit de vacciner les utilisateurs et futurs utilisateurs contre la tentation forte (on verra par la suite pourquoi) d'utiliser des mots (et non des pourcentages) pour exprimer les degrés de certitude. Le mot « vacciner » indique que, dans mon esprit, il s'agit de combattre l'équivalent d'une maladie et que l'idéal est de la prévenir avant qu'elle se déclenche. Sinon le risque est grand de rencontrer les arguments classiques de la persévération aveugle : « On a toujours fait comme cela » ou « Maintenant les étudiants sont habitués à cette modalité-là », même quand la consigne verbale n'a été utilisée qu'une seule fois. Plus profondément existe la crainte chez l'enseignant de devoir expliquer à ses étudiants pourquoi il change, et ainsi leur avouer que lui aussi apprend, qu'il n'est pas, comme ses étudiants auraient pu le penser, omniscient et parfait. Ou encore ce sentiment (erroné) que le bond en avant qui consiste à recourir aux degrés de certitude est déjà tellement gigantesque que la consigne apparaît comme un « détail négligeable ». Un peu comme l'idée de la transfusion sanguine. Était-ce vraiment un « détail » de tenir compte des groupes sanguins et du facteur rhésus ? Au nombre de morts par réaction immunitaire lors de transfusions avant la prise en compte de ce « détail », on comprend que certains y soient opposés ou, pour le moins, doutaient de son intérêt.

Puisque cet article est le premier à utiliser le concept de mata-analyse, j'utiliserai ci-après l'expression « **cette** mata-analyse » ou « **ma** mata-analyse » car je manque de recul pour apprécier la généralité du concept.

2.2 Rappels sur les méta-analyses

On sait que les méta-analyses les plus courantes en éducation consistent à repérer dans la littérature des publications qui présentent, sur un sujet précis, des données expérimentales dans des valeurs réductibles à un indice de mesure commun : l'ampleur de l'effet (Cohen, 1969 et 1986, Glass, 1976, Hedges & Olkin, 1985). On sait qu'en psychologie et en éducation il existe des « synthèses des méta-analyses ». C'est d'ailleurs le titre du chapitre 2 d'une des plus connues de ces synthèses en éducation, celle de Hattie (2009) dans son livre « *Visible learning* ». Une méta-analyse est une sorte d'instruction (dans le sens judiciaire du terme) « à charge et à décharge » sur une question du genre « La méthode A est-elle plus efficace que la B ? », dont on ne préjuge, a priori, ni de la réponse (positive ou négative) ni de l'ampleur de l'effet (AE) du phénomène. En ce sens, les résultats d'une méta-analyse ne sont pas « bornés » : ils peuvent aller (en théorie évidemment) de $-\infty$ à $+\infty$, ou plus exactement les bornes exactes ne sont pas connues à l'avance. Dans la présente mata-analyse des variations des traductions en pourcentage des degrés de certitude exprimés en mots, par contre, ces bornes sont bien connues à l'avance : 0% pour la borne inférieure et 100% pour la borne supérieure.

Pour Glass (1976, p. 3), une méta-analyse est l'analyse systématique d'un ensemble (*collection*) de résultats d'analyses issus d'études individuelles dans le but d'intégrer les observations *(findings*). Comme on va le voir, une mata-analyse n'a pas pour but d'intégrer les observations, mais d'accumuler celles que le mata-analyste a (subjectivement) jugées les plus convaincantes. Le tableau 1 présente (colonne de gauche) les 4 volets de la définition de Glass de ce que fait un méta-analyste et (colonne de droite) leurs correspondants quand à ce que j'ai fait dans ma mata-analyse.

Tableau 1 : Contrastes entre méta-analyse (selon Glass) et la présente mata-analyse (Leclercq, 2017)

Pour Glass, un méta-analyste….	**Dans ma mata-analyse,**
(a) utilise des méthodes objectives pour trouver les études à réviser	l'approche a été plus subjective : je m'en tiens à quelques publications que je juge les plus démonstratives
(b) décrit les données des études en termes quantitatifs ou semi-quantitatifs	c'est aussi ce que j'ai fait
(c) exprime les effets du traitement sur une échelle commune d'ampleur d'effet	j'ai recouru à des indices et représentations graphiques variés, souvent ceux et celles des études originales
(d) utilise des techniques statistiques pour mettre en relation les caractéristiques de l'étude avec les résultats qu'elle produit	ces mises en relation ont été faites qualitativement, par l'analyse de chaque consigne verbale

Le tableau 2 ci-après présentera à nouveau ces oppositions mais de façon plus détaillée, après que chaque volet a été examiné en détails.

2.3 Le but de cette mata-analyse : une instruction à charge

Ma mata-analyse est une instruction totalement à charge, d'où le préfixe « mata » (en espagnol, *matar* signifie « tuer »). Elle se distingue d'une méta-analyse par son but et par sa méthode. Le but de cette mata-analyse est de montrer le caractère inapproprié, voire nuisible, de certaines conceptions et de certaines pratiques…fort répandues et récurrentes (chaque génération de chercheurs et de praticiens la « redécouvre » et l'applique). Or, comme le disait Richard Clark (USC) à Rolland Viau (communication personnelle) : **«** En science, il est aussi utile de fermer des portes que d'en ouvrir **».**

Je pars du principe que les variations interindividuelles de « traductions » de mots (exprimant le degré de certitude) vers les pourcentages est une faiblesse méthodologique et que les valeurs de ces variations ~~résultats~~ (dispersions des pourcentages pour traduire un même mot) seront supérieures à 0 (l'idéal), ou, au mieux, nulles. Même s'il arrivait que cette dispersion soit faible, elle resterait un défaut. A moins que cette dispersion soit négligeable ! C'est la seule incertitude avant d'entreprendre cette mata-analyse. On se trouve un peu dans la situation du rejet de l'hypothèse nulle. A partir de quand une telle dispersion (la valeur de l'écart-type des pourcentages fournis) est- elle négligeable ? J'en discuterai à la fin de l'article. Chaque domaine de l'activité humaine rencontre cette question. Un crash aérien entraînant des morts est un crash aérien de trop et pourtant on continue de prendre l'avion. Un mort par vaccination est un mort de trop, et pourtant on continue à vacciner, etc. Une instabilité de 5% dans la répétition à une heure d'intervalle chez une même personne de son degré de

certitude sur un savoir doit-il faire renoncer à l'emploi de cette technique ? J'ai répondu à cette question dans un précédent article (Leclercq, 2016).

2.4 Les reproches faits aux méta-analyses s'appliquent-ils à ma mata-analyse ?

Glass et al. (1981, chap. 7) ont eux-mêmes énuméré les reproches qui ont été faits aux méta-analyses.

Une première critique (qu'ils réfutent) est que les méta-analyses donneraient trop d'importance à des études de faible qualité.

> *Dans ma mata-analyse, les études produites se caractérisent par leur simplicité et leur reproductibilité, si bien que le lecteur qui voudrait prouver le contraire des résultats avancés dans les études retenues peut à volonté reproduire les expériences et/ou en améliorer les conditions.*

Une autre critique est que les méta-analyses mêleraient des pommes et des poires.

> *Dans ma mata-analyse, c'est plutôt une force de rendre compte d'études ayant utilisé des méthodologies différentes. En ce qui concerne la critique portant sur le mot « mêler », elle ne s'applique pas à ma mata-analyse ci-après car les résultats des études choisies ne sont à aucun moment mélangés. Ils sont utilisés comme des coups de marteaux (différents) tapant sur le même clou (la thèse de départ). Le lecteur reste libre de juger que certains marteaux sont abusivement utilisés. Les arguments en ce sens seront les bienvenus : la science est un débat d'idées…argumentées.*

Une autre critique encore est qu'il arrive que certaines méta-analyses surexploitent une même étude de base, la faisant « peser » indument sur le résultat final.

> *Dans ma mata-analyse, il arrive que les données d'une étude donnent lieu à plusieurs « coups de marteaux », ce qui est justifié quand les données produites par cette étude sont indépendantes. Au lecteur d'en juger dans chaque cas.*

2.5 La méthode de ma mata-analyse

La méthode présentée ci-après diffère de la méthode des méta-analyses par trois aspects :

2.5.1 La triangulation par convergence des preuves convaincantes

Il ne s'agit pas de récolter le plus grand nombre possible de publications, mais un nombre suffisant et si possible les plus connues (voir justification en Annexe 1), pour emporter la conviction par la qualité de leurs données ou de leurs raisonnements. En effet, la présente mata-analyse ne débouche pas sur des valeurs numériques comme l'ampleur de l'effet. J'ai visé à rassembler un nombre suffisant d'études, qui, on le verra, se confirment l'une l'autre. Ma démarche correspond à un principe mis en évidence par Granhag & Strömwall (2000) et énoncé comme suit par Koriat (2012) : « Dans le système judiciaire relatif aux crimes, l'indice le plus fréquemment signalé de crédibilité d'un témoin est la cohérence des témoignages à travers des interrogations répétées. ». Ici ce sont des recherches différentes qui se confortent l'une l'autre. C'est donc un principe de triangulation des approches visant à vérifier la convergence des données.

2.5.2 La facilité de reproduction

Dans ma mata-analyse, sont privilégiées les études dont les méthodes (et par conséquent les données) sont facilement reproductibles, donc vérifiables par chaque enseignant… à condition qu'il bénéficie de la collaboration d'une trentaine de ses étudiants (comme sujets). Pour moi, tout enseignant dans de telles conditions est potentiellement un praticien-

chercheur…mais sur des thèmes qui exploitent les données que produit sa pratique au bénéfice de ses étudiants.

2.5.3 Le détail des descriptions

Dans une méta-analyse, pour chaque recherche retenue, sont mentionnées (1) les valeurs de descripteurs communs (par exemple le nombre de sujets et/ou le nombre de questions) ; (2) quelques données complémentaires (par exemple : le domaine de l'épreuve comme les mathématiques ou la géographie, etc.), mais sans entrer dans les détails, que le lecteur trouvera en se reportant à chacune des publications mentionnées.

Dans la mata-analyse ci-après, des extraits significatifs des instruments et résultats sont fournis avec suffisamment de détails pour permettre au lecteur, sans devoir retourner aux publications de base, d'imaginer une réplication (avec ou sans variante) des expériences mentionnées.

2.6 Un tableau schématique des différences entre cette mata-analyse et une méta-analyse

Le tableau 2 présente de façon plus détaillée que le tableau 1 synthétique les différences principales entre une méta-analyse classique et la mata-analyse proposée dans cet article.

Tableau 2 : Comparaison méthodologique ente les méta-analyses classiques et la présente mata-analyse

	En méta-analyse	**Dans cette mata-analyse**
Les recherches retenues	Toutes celles qui, répondant aux critères méthodologiques (contenu des hypothèses et design expérimental), apparaissent dans un ensemble de documents (le plus souvent des articles), ensemble défini, soit par des dates, soit par des listes de revues, soit par une combinaison des deux.	Il n'y a pas de souci d'exhaustivité. Les recherches retenues sont « de convenance » pour le propos. Les designs expérimentaux diffèrent.
La facilité de reproduction des données	La plupart du temps, les recherches citées ne sont pas décrites dans suffisamment de détails (la liste des questions posées par exemple) pour que l'expérience puisse être reproduite par le lecteur, qui doit, pour ce faire, lire les articles originaux, qui, eux-mêmes, ne donnent pas toujours les informations nécessaires à la reproduction.	Pour toutes les recherches citées, la présente mata-analyse est suffisamment précise pour que le lecteur puisse reproduire les expériences.
Les bornes dans les résultats	Quand l'unité de comparaison est l'Ampleur de l'Effet, les valeurs de cet indice peuvent aller dans les deux sens : négatif ou positif sur un continuum non borné (du moins en théorie) de $-\infty$ à $+\infty$.	L'unité de mesure de la présente recherche est **l'ampleur de la dispersion (AD)**, dont la borne inférieure est 0%, qu'elle soit exprimée par l'écart-type ou l'écart interquartile (EIQ). En outre, elle a aussi une borne supérieure : 100%.

2.7 Les modes, moyennes et médianes des valeurs choisies et des indices de dispersion

Quand elles traduisent des mots en pourcentage, les personnes interrogées fournissent très généralement des valeurs « arrondies » aux multiples de 10% (ex : 10%, 20%, 30%, etc.) ou de 5% (ex : 25%, 35%, etc.). Les rares exceptions sont aux extrêmes (ex : 1%, 2%, 98%, 99%) ou aux valeurs correspondant à 1/3 et 2/3 soit 33% et 66%. Qu'en dehors de ces exceptions, ces personnes ne fournissent jamais des valeurs plus nuancées en dit déjà long sur la granularité (la limite de la précision) avec laquelle un humain peut exprimer les degrés de certitude. Et nous allons voir que cette autolimitation spontanée dans la gamme des **valeurs préférées** est pleine de sagesse.

Dans la revue de certaines recherches, il m'a paru aussi intéressant de mettre en évidence, quand ils sont disponibles, (1) des indices de popularité des pourcentages chez (les choix préférés par) les répondants, (2) les valeurs les plus populaires chez (les plus spontanément choisies par) les expérimentateurs pour construire leurs échelles et (3) les indices de dispersion des significations numériques données par les répondants à des expressions verbales des divers degrés de certitude proposées. Pour ce faire, les données les plus utiles sont l'ensemble des valeurs (en %) et la popularité (en taux) de chacun de ces choix. Or les recherches publiées fournissent rarement ces données complètes. Elles s'en tiennent souvent à leur mode, à leur médiane ou à leur moyenne. C'est ce qui explique que, dans certains tableaux qui suivent, on ne retrouve pas toujours ces trois indices.

Le mode (la valeur en % ou la catégorie qui est la plus populaire), apparaît dès lors légitime comme un « résumé » possible de la popularité des degrés de certitude exprimés en %. Le nombre de valeurs choisies peut être, lui, un résumé de la dispersion de ces valeurs choisies spontanément. L'inconvénient de la moyenne de ces valeurs est qu'elle pourrait prendre elle-même une valeur (en %) qui n'a été le choix d'aucune des personnes qui ont répondu. La médiane ne présente pas ce défaut de la moyenne, mais elle ne peut prétendre à remplacer le mode, tout spécialement quand la distribution des valeurs choisies est multimodale et asymétrique.

2.8 Mata-analyse et méta-analyses qualitatives

En recherche médicale, des principes de méta-analyses qualitatives ont été proposés (voir détails en annexe 3), pour procéder à des synthèses de recherches n'appliquant pas le même schéma expérimental (par exemple permettant la mesure de gains entre pré et posttest), ou ne produisant pas le même type de résultats (transformables en ampleur d'effet). Par exemple, Bland, Meurer et Maldonado (1995) sélectionnent les publications sur base de critères de validité (pondérés) tels que ceux définis par Campbell et Stanley (1966). La démarche de ces auteurs est cependant plus proche d'une méta-analyse classique que de ma mata-analyse.

La mata-analyse qui suit recourt à une procédure de sélection qualitative (voir tableaux 1 et 2) et débouche, pour certains aspects, sur des valeurs quantitatives et, pour d'autres, sur des conclusions qualitatives.

Deuxième partie

3. Les causes de variabilité dans la traduction de « mots en % »

3.1 Variabilité d'une langue à l'autre

La traduction, d'une langue naturelle à une autre, d'une expression verbale des certitudes pose souvent problème. Essayons, par exemple, de traduire en français les expressions anglaises *scarecely* ou *likely* ou *hardly* ou *cocksure*, de manière à refléter exactement le même sens que l'original. *Cocksure*, par exemple, selon les contextes…et les dictionnaires traductifs peut devenir *outrecuidant, prétentieux, trop sûr de soi, péremptoire, condescendant, suffisant, arrogant,* ou encore *excessivement confiant.*

Bien sûr, on ne parle pas ici des congénères interlinguaux (ou cognates) qui, par essence, véhiculent des significations différentes sous des formes proches parce qu'ayant la même origine étymologique, mais avec « bifurcation ». En latin, *cognatus* signifie « lié par le sang ». Un bel exemple est le terme anglais *eventually* qui a un tout autre sens (*finalement*) que le terme français *éventuellement.*

Par contre, il y a mathématiquement équivalence entre des expressions comme 1/5 et 20%, entre 4/5 et 80%, entre 6/12 et 50%, etc., ce qui ne pose aucun problème de traduction d'une langue à l'autre.

3.2 Variabilité selon les cultures

Chee (2006) a présenté 23 expressions verbales en anglais à 3 groupes de personnes : 13 Britanniques, 9 Grecs et 10 Malais. Ces 32 personnes ont été invitées à traduire ces 23 expressions sur une échelle de 11 valeurs numériques (pas des %) allant de 0 à 10. Les comparaisons des moyennes dans les trois groupes montrent de grandes variabilités intergroupes, tout spécialement pour les valeurs extrêmes. Je n'en retiens, dans le tableau 3, que quelques exemples :

Tableau 3 : Moyennes des valeurs numériques (en %) traduisant des expressions verbales par des personnes de 3 cultures différentes

Expression en anglais	Britanniques (N = 13)	Grecs (N = 9)	Malais (N = 10)
certain	96,5	86	86
almost certain	87	79	76
not probable	20	37	31
not possible	6	30	16
impossible	3	11	4

Quant aux variabilités intra-groupes (non présentes dans le tableau 1), les deux termes qui offrent le moins de variabilité d'interprétation numérique chez ces 9 Grecs sont :

- « *définite* » (M = 89%). C'est également le terme qui fait le plus l'unanimité chez les 13 Britanniques (M = 97%)
- « *perhaps* » (peut-être) dont M = 56% chez ces Grecs, 50% chez ces Britanniques et 56% chez ces 10 Malais. Cependant, chez ces 10 Malais, « *perhaps* » est un des 2 mots (sur les 23) qui a la plus grande variabilité de significations intra-groupe.

Bien entendu, les nombres de sujets de l'expérience de Chee sont très petits et des vérifications avec de plus grands nombres seraient nécessaires. C'est plus pour sa méthode que pour ses résultats que j'ai présenté cette recherche.

3.3 Variabilité selon le contexte et les taux de base

On appelle « taux de base » (*base rate*s) les probabilités a priori (avant plus ample information) d'un événement. Les traductions de mots vers des pourcentages sont sensibles à certains contextes, et tout particulièrement aux probabilités a priori ou aux enjeux. Ainsi, dans le domaine médical, 10% peut apparaître comme défavorable quand la probabilité a priori est plus élevée (que 10%) et que l'événement est désirable. Par contre, 10% peut apparaître comme « une bonne chance » quand la probabilité est plus élevée au départ (à la base) et que l'événement est défavorable.

Plusieurs chercheurs ont étudié l'effet de ces taux de base sur la traduction de certitudes en mots vers des certitudes numériques (en %).

Wallsten, Fillenbaum et Cox (1986) ont demandé à des sujets de traduire des termes comme *sure, unlikely, improbable, frequently, unusual, seldom,* etc. dans des contextes différents. Ces contextes pouvaient être le moment de l'année (exemple : Les chutes de neige à Montréal en septembre / en novembre / en mars) ; idem pour des pics de pollution à divers moments de l'année, etc. Dans tous les cas, le contexte (le taux de base) a pesé sur les résultats.

Pepper et Prytulak (1974) ont observé que les pourcentages attribués à *almost never, often, frequent, rarely*, dépendaient de l'événement à quantifier : selon qu'il s'agissait, par exemple, de la présence de coups de feu dans un western hollywoodien ou de tremblements de terre (d'une certaine magnitude) en Californie pendant une période donnée.

O'Brien (1989) donne un exemple d'effet de contexte dans le domaine médical (la connaissance préalable des statistiques concernant un événement). Pour ce faire, il cite les résultats de Mapes (1979). Ce dernier a demandé à des médecins généralistes d'évaluer, par des mots et par des nombres, la probabilité d'effets secondaires de divers médicaments. Le mot « rare » a été traduit par « moins d'une chance sur 1000 » par 59,4% des médecins pour les bêtabloqueurs et par 20,7% pour les antihistaminiques.

Druzdzel (1989, p. 5 et 6) donne un exemple humoristique d'une telle combinaison du contenu (par exemple la météo) avec le contexte (par exemple la nationalité). Ainsi, en Angleterre, le mot « probablement » dans l'expression « Le parti travailliste devrait probablement gagner » (*The Labor Party is likely to win*) peut avoir une traduction numérique plus élevée que dans « Nous allons probablement avoir un bel été ». En effet, en Angleterre, on est loin des 50% en temps normal, alors qu'à l'époque, l'alternance des deux grands partis au pouvoir les met souvent au coude à coude (donc près de 50% chacun).

Beyth-Marom (1982) observe non seulement une grande variabilité des significations interindividuelles des traductions des degrés de certitude en mots vers des pourcentages, mais aussi que cette variabilité s'accroît encore quand les certitudes verbales sont utilisées en contexte. Ce qui est d'autant plus grave puisque c'est cette utilisation en contexte (pédagogique) qui intéresse les utilisateurs des degrés de certitude en pédagogie.

A la suite des travaux de Fabre (1993), Fares (2006) a fait grosso modo les mêmes observations (grande variabilité interindividuelle de traduction) et a étudié en particulier les effets du contexte.

C'est pourquoi j'ai tenté (Leclercq, 2016) d'éclairer un peu cette question de la variabilité interindividuelles dans deux expériences, l'une hors contexte (en 2013 avec 33 participants) et l'autre en contexte (en 2016 avec 19 participants). Dans les deux, j'ai invité les participants à traduire en pourcentage des expressions verbales de certitude telles que « pas sûr du tout », « très peu sûr », « peu sûr », « sûr », « très sûr » et « extrêmement sûr ».

Les résultats de ces deux expériences « de traduction » sont convergents : la répétabilité intra-individus est élevée, mais le « brouillard traductif inter-individus » est épais (voir ci-après).

3.4 Variabilité d'une personne à l'autre (interpersonnelle) et profusion des mots de la langue

Prolongeant leurs travaux sur le sujet (Budescu & Wallsten, 1985 et 1987), dans une de leurs expériences, Budescu, Weinberg et Wallsten (1988) ont présenté 11 probabilités sous forme numérique (0%, 10%, 20%, etc.) à 20 personnes en leur demandant de les traduire en mots. Ils ont obtenu 111 phrases (expressions) distinctes, soit en moyenne 10 façons différentes d'exprimer un même nombre en mots. Ce ne serait pas gênant si tous donnaient à chaque mot le même sens en termes de nombres. Or toute la suite va montrer que c'est loin d'être le cas.

4. Une revue critique pionnière des traductions « Mots vers % »

Il y a un quart de siècle, Clark (1990) a publié une « Revue critique de deux décennies de recherches sur les expressions verbales d'incertitude ».

4.1 Les enseignements (de la revue de Clark) sur les méthodes

La revue de Clark met en lumière certains aspects méthodologiques auxquels il faut être attentif.

(1) Plusieurs facteurs sont confondus dans ces recherches, comme la calibration (conformité à la réalité), la familiarisation avec l'usage de degrés de certitude, la granularité (le nombre d'échelons) de l'échelle, etc. Ces problèmes, importants, ne seront pas abordés ici, mais je serai vigilant à ne pas les confondre avec les deux objets de la présente étude : les différences interindividuelles de traduction et les valeurs numériques les plus fréquemment utilisées (valeurs modales).

(2) Dans plusieurs recherches, les expressions verbales ne portent pas vraiment sur la probabilité de survenue d'un événement, bref sur la certitude ou l'incertitude. Par exemple, parmi les 44 expressions de Lichtenstein et Newman (1980), on trouve les mots *inconclusive* et *rather* (sans autre mot). Chez Bryant et Norman (1980), on trouve *hopefully*, *classic*, *normally*, *consistent with*, et *sometimes*.

Ces observations m'ont amené à m'en tenir, dans mes expériences de traduction, à l'adjectif « sûr » et à ses modérateurs (pas du tout, très peu, peu), ainsi qu'à ses amplificateurs (très, extrêmement).

4.2 Les enseignements (de la revue de Clark) sur les résultats

(1) Les différences inter-individus de traduction de mots vers des pourcentages sont énormes et s'observent dans toutes les recherches recensées.

(2) Plusieurs distributions sont bimodales et certains résultats laissent dubitatif. Ainsi, Bryant et Norman (1980) rapportent un pourcentage moyen attribué à *never* plus élevé qu'à *doubtful* et *unlikely*.

(3) La moindre nuance verbale a un impact sur la traduction. Ainsi, Zadeh (1975) a observé que *not likely* et *unlikely* ne donnaient pas les mêmes traductions en pourcentage de chances.

(4) La répétabilité des certitudes chez une même personne (ou stabilité intra-personnelle) est élevée. En particulier, Budescu et Wallsten (1985) ont demandé à 32 participants de traduire 19 expressions verbales en pourcentage à 3 reprises, distantes de 3 semaines à chaque fois. Ils ont observé une forte répétabilité intra-individuelle en termes de classement des phrases entre elles. Bryant et Norman (1980), qui ont demandé de traduire 30 expressions à deux reprises, ont observé une corrélation de répétition de 0,96.

4.3 Les enjeux pour la pratique éducative

Ainsi, dès le départ, l'utilisateur de telles consignes verbales introduit dans son dispositif, qu'il soit à visée de recherche ou de notation (docimologique), un coefficient de confusion (ou erreur aléatoire de mesure) qui contribuera grandement à rendre ses résultats ininterprétables. Autrement dit, une même situation mentale (par exemple hésiter également entre trois possibilités, donc attribuer à chacune des solutions une probabilité de 0,33 ou un pourcentage de chances de 33%) est traduite en mots de façon différente d'une personne à l'autre. Par exemple, l'un dira «je n'en suis pas sûr », un autre dira « j'en suis très peu sûr » et un troisième dira « j'en suis peu sûr ».

A l'inverse, si, dans son for intérieur, un sujet A attribue à « sûr » une valeur de 50%, un sujet B une valeur de 60% et un troisième sujet C une valeur de 70%, le chercheur qui ne dispose que des déclarations verbales, les tiendra pour équivalentes et risque de conclure que la certitude de ces trois personnes est élevée, alors qu'elle n'est que de 60% en moyenne. Si 3 autres personnes (D, E et F) attribuent (dans leur for intérieur) à « peu sûr » des valeurs respectivement de 60%, 70% et 80%, le chercheur aura tendance à conclure que ce deuxième groupe est moins sûr que le premier alors que sa probabilité subjective moyenne (pensée mais non déclarée) est de 70%. Etant donné les effets de contexte du genre de ceux décrits par Mapes (voir section 3.3 ci-avant), qui eux, portent sur l'instabilité INTRA-individuelle de la signification numérique de mots, les erreurs d'interprétation deviennent gigantesques, et expliquent les déboires constatés en fin de section 1.1 par des auteurs comme Frary.

Que la répétabilité intra-individuelle (ou stabilité dans le temps chez une même personne) soit souvent bonne (dans une situation donnée) signifie que chaque fois qu'il se retrouve dans le même état mental, et dans des circonstances semblables, le sujet utilise grosso modo les mêmes mots pour exprimer son degré de certitude.

En situation scolaire, pour inférer l'état mental (de connaissance) de chaque personne en termes de pourcentage, un enseignant serait dès lors obligé d'étudier l'échelle personnelle de chacun de ses étudiants et, en quoi elle produit des erreurs systématiques par rapport à

l'interprétation moyenne de la population (de la classe par exemple). Travail colossal et sans fin car cette échelle peut varier, chez un même individu, selon les contenus ou les circonstances, ce qui rend les erreurs non plus systématiques (donc relativement « corrigibles ») mais aléatoires.

Or, c'est cette pratique (des consignes en mots) que l'on retrouve majoritairement chez les enseignants qui découvrent le principe des degrés de certitude accompagnant chaque réponse d'une épreuve constituée de QVF (Questions Vrai-Faux) ou QCM (Questions à Choix Multiple) ou encore de QROC (Questions à réponses Ouvertes Courtes).

5. L'ambiguïté interpersonnelle des mots présentée en tableaux

La présente section est structurée selon les différentes méthodes de recueil des données (les consignes).

5.1 Méthode 1 : Positionner le mot sur une échelle numérique

Fabre (1993b, p. 108-109) a posé à 143 participants la question suivante :

> *« Voici 5 expressions courantes qui expriment le degré de confiance qu'un événement particulier arrivera. Donnez un nombre compris entre 0 et 10 à chacune, 0 signifiant « incertitude totale quant au fait que cet événement arrivera » et 10 signifiant « certitude totale ».*

Voici ces résultats, synthétisés dans le tableau 4 où les valeurs présentées dans les colonnes sont les suivantes :

M = moyenne ; Ec-Type = écart-type ; MV1 = marge de variation ; MV2 = marge de variation après élimination d'une valeur extrême ; Mé = médiane ; Q3-Q1 = écart interquartile (EIQ) ou différence entre le quartile Q3 et le quartile Q1).

Tableau 4 : Distribution des valeurs chiffrées attribuées par 143 personnes

Degrés de certitude --->		**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	M	Ec-T	**MV1**	MV2	**Mé**	Q3-Q1
J'affirme que	**143**			1			2	2		11	25	102	9,49	1,12	**9**	6	10	1
Je suis sûr que *	**143**					1	1		3	17	44	77	9,31	0,97	**7**	6	10	1
Je pense que *	**143**		1	2	4	3	31	39	30	21	11	1	6,33	1,57	**10**	9	6	2
Il me semble que	**143**		7	19	17	25	39	23	9	3	1		4,37	9	**9**	8	6	3
Je suppose que *	**143**	3	7	17	26	32	26	16	11	4	1		4,13	1,84	**9**	8	4	1

Ces données montrent qu'il existe une assez grande dispersion inter-personnes (colonne écart-type) quant à la quantification d'une même expression verbale. Les marges de variation (MV), exprimées ici en nombre de catégories utilisées, sont, quant à elles, gigantesques (colonnes MV_1 et MV_2). Ainsi, « Je pense que » correspond à la valeur 1 pour une personne et à la valeur 10 pour une autre.

Les trois termes marqués du signe * sont repris d'une expérience de Foley (1959) qui avait invité 38 personnes à attribuer une valeur numérique (comprise entre 1 et 10) à 5 termes exprimant la probabilité qu'un événement se produise. Ses deux autres termes étaient *positif* et *certain*. Il avait obtenu des résultats semblables à ceux de Fabre.

5.2 Méthode 2 : Donner un pourcentage pour chaque mot

5.2.1 Les marges de variation (MV)

Reagan, Mosteller et Youtz (1989) ont observé que l'expression « *Likely* » donne lieu à des traductions numériques allant de 5% à 95%. Avant eux, Lichtenstein et Newman (1967) avaient demandé à 180 personnes une traduction « Mot → Nombre » de 41 expressions verbales. Les étendues ou marges de variation (MV), c'est-à-dire l'écart entre la traduction la plus élevée et la plus basse, laissent rêveur. Le tableau 5 présente, en colonne de droite, ces étendues (MV) pour 6 de ces expressions :

Tableau 5 : Moyennes et étendues des traductions de 6 des 41 mots de Lichtenstein et Newman (1967)

	Moyenne (en %)	**Médiane (en %)**	**Ec. Type (en %)**	**Etendue ou Marge de Variation (en %)**
Very likely	87	**90**	6	45% - 99%
Probable	71	**75**	17	1%-99%
Rather likely	69	**70**	9	15% - 99%
Uncertain	40	**50**	14	8%-90%
Highly improbable	6	**5**	5	1%-30%

Ce tableau 5 m'inspire les commentaires suivants :

a) Ce sont des Moyennes qui sont calculées. Or les distributions sont rarement symétriques, ce qu'illustrent, en section 5.1, les données de Fabre (1993b). Dès lors, ce sont plutôt les Médianes qu'il faut considérer (voir ci-après), ce qui ne changerait rien à l'étendue (ou marge de variation – MV).

b) Les étendues sont impressionnantes, mais elles pourraient être dues à un individu isolé. C'est ce que montrent les différences entre les colonnes MV1 et MV2 dans le tableau 3. C'est donc l'écart semi-interquartile (ESI) qu'il faut prendre en considération pour juger de la dispersion. Rappelons que cet écart est obtenu en repérant la valeur de Q1 (quartile 1 ou la valeur du 25° sur 100) et de Q3 (quartile 3 la valeur du 75° sur 100), les estimations numériques ayant été au préalable ordonnées entre elles. L'ESI est obtenu par la formule (Q3-Q1)/2. C'est ce qu'a fait O'Brien (voir section 5.2.2 ci-après).

S'appuyant sur le travail de Lichtenstein et Newman (1967), Bickel (2005, p. 3) conclut « Les phrases habituelles ne sont pas fiables. Elles amènent des désaccords et une perte de temps ».

5.2.2 Les variations mesurées par l'Ecart Interquartile (EIQ)

O'Brien (1989) a demandé à 52 médecins généralistes de donner une valeur numérique à 23 expressions verbales. Le tableau 6 présente les données pour 12 de ces expressions que j'ai choisies en fonction des valeurs numériques de leur médiane (Mé), de leur ampleur de Q3-Q1 (l'écart interquartile ou EIQ) et de leur moyenne (Moy). Les médianes sont aussi reprises en figure 1 (section 6.1).

Les médianes constituent une approximation des modes (non fournis par O'Brien). Sur les 23 médianes, 14 sont des multiples de 10 et 5 des multiples de 5. Ce qui conforte l'hypothèse que les personnes ont tendance à utiliser des nombres « ronds » (en l'occurrence des multiples de 10).

Tableau 6 : Valeurs numériques observées dans la traduction de 12 expressions verbales (mots)

	Mé	*Q3-Q1*	Moy
certain	**95**	*10*	84
fort probable	**80**	*19*	72
probable	75	*20*	70
vraisemblablement (likely)	70	*20*	69
chance significative	**60**	*47*	49
chance raisonnable	50	*27*	49
risque modéré	**40**	*20*	39
possible	25	*20*	30
parfois	**23**	*40*	28
il y a une chance	**15**	*20*	23
peu vraisemblable (unlikely)	13	*10*	19
jamais	**0**	*0,5*	6

Dans les distributions habituelles (normales par exemple), l'EIQ (ou Q3-Q1) prend une valeur moins élevée que la marge de variation (MV), comme je l'illustre en section 6.2.

On constate néanmoins que cet EIQ est, pour 9 des 12 expressions, égal ou supérieur à 20%, alors que, par définition, 50% des réponses sont en dehors de cet intervalle qui va de Q1 à Q3.

5.3 Méthode 3 : Juger de l'ambiguïté de chaque expression verbale

O'Brien (1989) a aussi demandé aux 52 médecins (de l'expérience ci-dessus) de noter les 23 expressions verbales (en mots) sur une échelle de 1 (sans ambiguïté) à 3 (très ambigu) pour mesurer ce qu'il appelle « l'ambiguïté ressentie ou prédite ». Celle-ci s'avère corrélée à 0,86 (rho) avec l'ambiguïté observée (EIQ). Ces médecins sont donc conscients des différences de niveaux d'ambiguïté entre les différents termes.

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 et les résultats de cette expérience ont été rapportées ailleurs (Leclercq, 2016).

5.4 Méthode 4 : Puissance d'extrémisation d'adverbes modificateurs (ex : « *very* »)

On peut modifier l'interprétation numérique d'expressions verbales (comme « peu probable ») ou d'adjectifs comme « certain » ou « sûr » en leur adjoignant des adverbes comme « très », « extrêmement » ou « totalement ». Ce faisant, on rend ces expressions plus « extrêmes ». Quelle est la « puissance d'extrémisation » de tels adverbes ajoutés ? Johnson (1973) a étudié l'effet (en réduction de l'ambiguïté interpersonnelle) de termes comme *very*, *quite* et *fairly*. On pourrait calculer divers coefficients d'extrémisation d'une expression verbale comme *very* accolé à une autre expression comme *likely* ou *infrequent*, etc. La formule que je

propose consiste à diviser la valeur moyenne attribuée par un groupe de personnes à une expression (*infrequent* par exemple) par la valeur moyenne attribuée par ce même groupe à la même expression précédée de *very*. J'ai calculé de tels coefficients pour des données fournies par Mosteller et Youtz (1990, p. 9). Ceux-ci ont procédé à une méta-analyse de 20 études, au terme de laquelle ils fournissent les données que je présente dans les trois premières colonnes du tableau 7. Ce qui m'a permis (dans la 4° colonne) de calculer le coefficient d'extrémisation (ici d'environ 2 fois la valeur moyenne).

Tableau 7 : Tableau (légèrement) modifié de Mosteller & Youtz (1990, p.9) et calcul du coefficient d'extrémisation suite à l'ajout du mot « *very* »

Mosteller & Young, 1990, p. 9			**Auteur**
	Moyenne (en %) **sans** « *very* »	Moyenne (en %- **avec** « *very* »	Coefficient d' extrémisation
Infrequent	17	7	2,43
Low probability	16	6	2,66
Unlikely	17	11	1,54
Improbable	16	7	2,28
Seldom	12	7	1,71
Rarely	9	4	2,25
			Moyenne = 2,14

On voit que l'adverbe « *very* » réduit considérablement l'ambiguïté. Mosteller et Youtz (1990) estimaient cette puissance de réduction de « *very* » d'un facteur 2 (dans ces 6 cas). Cependant, la dispersion inter-individus des interprétations numériques pour une même expression « extrémisée » reste très importante. Ainsi, dans un tableau présentant les médianes de ces valeurs, Mosteller et Youtz (1990) observent, pour « *unlikely* » (Mé = 17%), un EIQ de 13% et pour « *very unlikely* » (Mé = 5%), un EIQ de 7%. Pour « *seldom* » (Mé = 10%) ils observent un EIQ de 10% et pour « *very seldom* » (Mé = 5%) un EIQ de 4,5%.

5.5 Méthode 5 : Adverbes modérateurs et amplificateurs qui réduisent l'ambiguïté

Diverses expressions sont la combinaison d'un terme de base (par exemple « *probable* ») et de modificateurs « réducteurs » (ex : peu, faiblement) ou amplificateurs (ex : très, extrêmement). Lichtenstein et Newman (1967) ont étudié leur effet sur la réduction de l'ambiguïté. Le tableau 6 présente ce que, dans leurs études, 5 variations autour du mot « *probable* » (en anglais !) donnent comme modifications de signification (de traduction) en termes de pourcentage. Nous savions déjà (par le tableau 3) que « *probable* » avait une dispersion maximale, et donc qu'il ne peut pas servir de point d'ancrage. Evidemment, les modificateurs réduisent cette ambiguïté. La preuve : les marges de variation (étendues entre les valeurs minimale et maximale) observées par ces deux auteurs chez 188 sujets adultes sont réduites par rapport à la marge de variation pour l'expression « *probable* » (dont la MV va de 1 à 99). C'est ce que montre la représentation graphique que je donne, dans la partie de droite du tableau 8 des valeurs numériques de ce tableau. Dans les zones grisées, j'ai placé à chaque fois la valeur minimale (à gauche) et la valeur maximale (à droite). Ces zones visualisent les marges de variation (MV) des valeurs rencontrées par Lichtenstein et Newman (1967) pour ces 6 expressions verbales.

Tableau 8 : Réduction de l'ambiguïté de *probable* (Lichtenstein et Newman (1967)

	Moy	Min	Max	1	10	20	30	40	50	60	70	80	90	99
higly probable	89%	60%	99%							60				99
very probable	87%	60%	99%							60				99
probable	71%	1%	99%	1										99
not very probable	20%	1%	60%	1						60				
improbable	11%	1%	50%	1					50					
highly improbable	6%	1%	30%	1			30							

5.6 Méthode 6 : Indiquer « quel % » en plus ou en moins

Le problème du flou des degrés de certitude verbaux se retrouve à divers endroits de l'activité humaine. Ainsi, en aviation militaire, Rigby et Swain (1971) ont étudié les variations inter-individus d'estimation de nombres représentés par des expressions verbales telles que *a bunch of* ou *many* ou encore *a lot of*.

De même, dans le cadre de l'explosion de la navette spatiale Challenger en janvier 1986, Marshall (1988, p. 1233) critique la pratique d'alors d'évaluer la probabilité de défaut de chaque composant par des expressions verbales (et non par des fréquences numériques). Pourtant, la sonnette d'alarme avait été tirée plusieurs années auparavant par Kent.

Sherman Kent, spécialiste de la CIA, est considéré comme le père du « renseignement » (Intelligence Service) aux USA. En 1951, pendant la guerre froide, à propos de l'invasion de la Yougoslavie par l'URSS, la CIA avait utilisé l'expression (verbale) « *serious possibility* ». Or cette expression signifiait alors pour la CIA une probabilité comprise entre 0,2 et 0,8. Commentant 43 ans plus tard cette utilisation de mots et non de nombres, Kent (1994) écrit : "*We did not use the numbers, …and we misused the words.*" Ce qui l'avait amené (dès 1964) à adopter des pourcentages pour exprimer les probabilités.

Dans une de ses plus récentes grilles (*chart*s), Kent (1994) recommande la correspondance suivante (tableau 8) entre mots et valeurs numériques, avec une « marge d'erreur en plus **et** en moins ». J'écris « et » bien que Kent écrive « ou ». En effet, encadrer une estimation (ex : 50%) de 10% en plus **ou** en moins, c'est se donner une marge d'erreur de 20%.

C'est la dernière colonne du tableau 9 qui nous intéresse le plus ici : **l'incertitude (de la certitude).**

Tableau 9 : Les mots de Kent et leurs probabilités correspondantes

	Traduction recommandée par Kent	**(ajouter ou soustraire)**
Certain	100%	*Give or take 0%*
Almost certain	93%	*Give or take 6%*
Probable	75%	*Give or take 12%*
Chances about even	50%	*Give or take 10%*
Probably not	30%	*Give or take 10%*
Almost certain not	7%	*Give or take 5%*
Impossible	0%	*Give or take 0%*

L'imprécision estimée (subjective) par cet auteur est plus grande (environ 10%) au centre de l'échelle qu'aux extrémités (de 0% à 5%). Ces valeurs intuitives de Kent se vérifient dans diverses études : celles d'Edwards (1967) ou de Leclercq (1982, p. 250-251) ou, plus récemment, de Mosteller et Youtz (1990).

Notons que ce tableau de Kent ne précise pas « avec quel niveau de risque ($p<.05$ par exemple) ». On en est réduit à faire l'hypothèse que ces « ajouts » ou « retraits » (*Give or take*) sont à comprendre comme l'écart-type d'une distribution normale (gaussienne) des pourcentages.

5.7 L'évolution de l'élasticité (ou l'ambiguïté) sur 20 ans

Simpson (1963) appelle « Indice d'élasticité » d'une expression verbale l'écart interquartile (EIQ = Q3-Q1) des valeurs numériques (en %) attribuées par différentes personnes à cette expression. A 20 ans d'intervalle (1942 et 1962), sur des populations différentes, il a présenté les mêmes 20 expressions. Voici (tableau 10), pour 3 de ces expressions, les indices d'élasticité (ou EIQ) qu'il a observés (pour les hommes de ses deux populations) :

Tableau 10 : Stabilité 20 ans plus tard de l'ambiguïté (l'élasticité) ou Ecart Interquartile (EIQ) de la traduction de 3 expressions verbale sen pourcentages

	1942	**1962**
Often	20	30
Now and Then	25	28
Infrequently	19	19

Cette « élasticité » (ou variation inter-individus) n'avait donc pas changé en 20 ans pour la traduction de ces trois termes. Il serait facile de vérifier ce qu'il en est un demi-siècle plus tard. Toutes les données accumulées dans le présent article permettent de faire l'hypothèse d'une grande stabilité (hélas) sur ce point.

6. L'ambiguïté interpersonnelle des mots visualisée par graphiques

Les graphiques sont évidemment des données numériques visualisées. Et ce, afin de diminuer la charge mentale exigée du lecteur, afin de faciliter à la fois sa traduction des observations dans une langue naturelle, sa réflexion et son imagination pour aller au-delà de l'observé (en produisant des hypothèses explicatives par exemple). Or, dans ce cas-ci, quelle est la meilleure représentation graphique pour ce faire ? J'ai encore trop peu de certitude sur ce point. C'est pourquoi je recourrai à des modalités variées.

6.1 Visualisation par des boîtes d'EIQ (Q3-Q1)

Revenons à l'expérience d'O'Brien (1989) qui a demandé à 52 médecins généralistes de traduire 23 expressions verbales en valeurs numériques (voir sections 5.2.2 et 5.3). L'intérêt de ses résultats est que, pour chaque expression, en plus de la moyenne, il présente le médian et l'écart interquartiles (EIQ ou Q3-Q1), mais, hélas, pas la valeur modale. Dans la figure 1, j'ai visualisé ces valeurs pour 12 des 23 expressions verbales que j'ai choisies sur base des valeurs de leurs médianes et de leur EIQ. Pour chaque expression verbale, la valeur médiane apparaît à l'intérieur d'une boîte rectangulaire, dont le bord gauche représente Q1 et le bord droit représente Q3. Chaque mot est suivi (entre parenthèses) de son EIQ.

On constate

(1) Pour une majorité des expressions verbales (laissées en anglais), l'EIQ atteint ou dépasse 20%. Une telle dispersion est énorme. En effet, il faut se rappeler que, par définition, chaque boîte n'inclut que 50% des observations. On s'attendait à une dispersion aussi grande, sur base de démonstrations précédentes.

(2) des valeurs médianes correspondant à des « nombres ronds » : pour 5 des 12 expressions verbales, les pourcentages sont des multiples de 10 et pour 4 autres des multiples de 5.

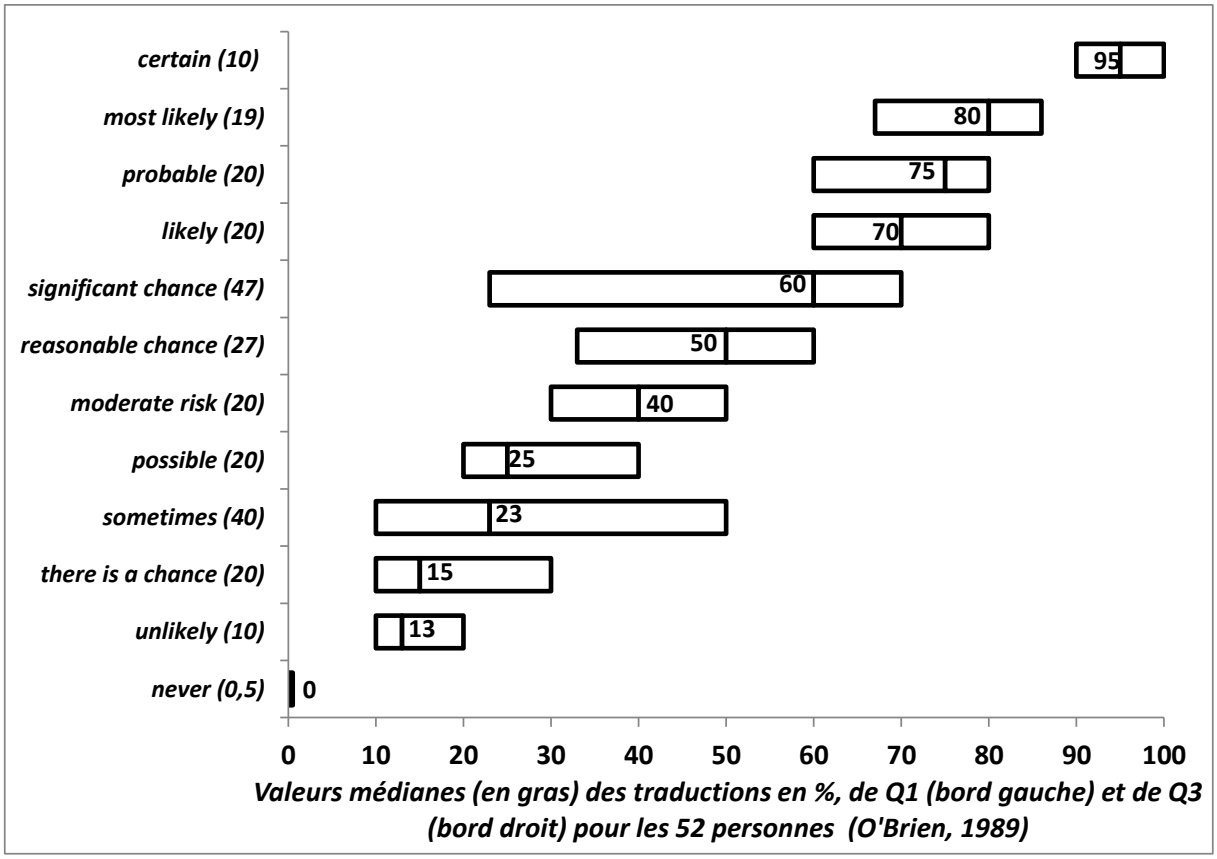


Figure 1 : Valeurs numériques observées dans la traduction de 12 expressions verbales (adapté de O'Brien, 1989)

6.2 Visualisation par des boîtes des marges de variation (MV)

Dans un précédent article (Leclercq, 2016), j'ai présenté les résultats de deux expériences où les participants étaient invités à traduire par des pourcentages 6 expressions construites autour de « sûr » (allant de « Pas du tout sûr » à « Extrêmement sûr »). La première expérience (à Bobigny Sorbonne Paris Cité, en 2013, avec 33 participants) était hors contexte. La deuxième (à Liège en 2016 avec 19 participants) était en contexte, avec grosso modo les mêmes mots. C'est cette dernière expérience qui est la plus proche d'une situation scolaire (en contexte). En effet, elle demandait aux participants, pour chacune des 15 questions d'un test, de fournir une réponse (choisie parmi 6 solutions), une certitude verbale et une certitude en un pourcentage compris entre 0% et 100%. Ces attributions de valeurs numériques (en %) ont été demandées aux participants à deux reprises (pour tester leur répétabilité ou stabilité dans le temps) : à 5 heures d'intervalle (en 2013) et à 2 heures d'intervalle (en 2016).

La figure 2 présente, pour chacune de ces 6 expressions verbales (quasi les mêmes dans les deux expériences), la marge de variation (MV) des « traductions » des mots en pourcentage effectuées par les participants ainsi que leur Moyenne.

Puisqu'elle prend en compte les deux valeurs extrêmes (la valeur minimale observée et la valeur maximale observée), la marge de variation (MV) est l'indice qui maximise l'ampleur des différences. L'écart interquartile (EIQ), lui, ne prend en compte que la 25e et la 75e des valeurs ordonnées. On est néanmoins frappé par la ressemblance entre les résultats d'O'Brien (figure 1) et ceux (présentés en figure 2) de mes deux expériences (Leclercq, 2016) de 2013 et 2016, bien que les divers pourcentages aient été obtenus dans des situations différentes (hors contexte et en contexte), avec expressions verbales différentes (et même dans des langues différentes) et avec des participants différents. Les valeurs modales, elles, sont présentées dans les figures 5 et 6.

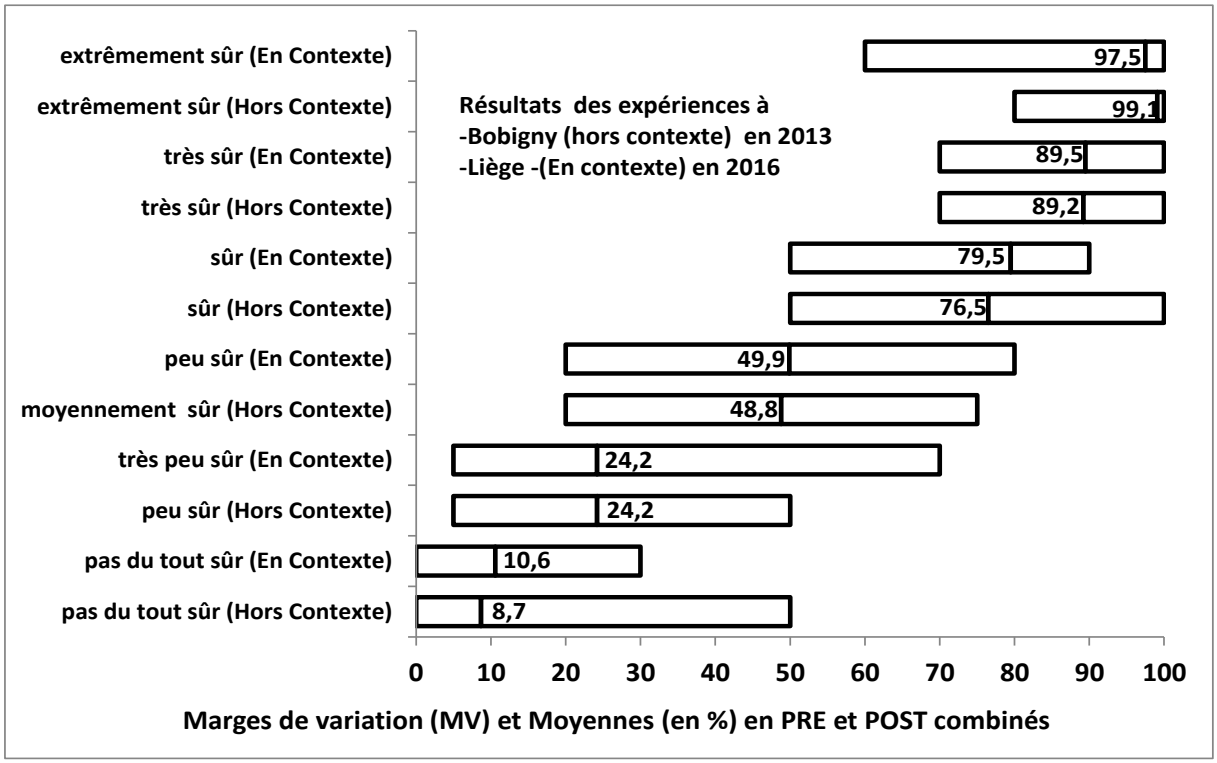


Figure 2 : Marges de variation et moyennes pour les 12 expressions autour de « sûr » dans deux expériences décrites dans Leclercq (2016)

Ces marges de variation sont du même ordre de grandeur que celles rapportées dans les expériences de Fabre (1993) dans le tableau 2, de Lichtenstein & Newman (1967) dans le tableau 3, de O'Brien (1989) dans le tableau 4 et la figure 1 et de Bocklisch, Baumann, Scholz, et Krems (2010) dans les figures 3 et 4 ci-après.

6.3 Visualisation par fonctions d'appartenance

Gigerenzer & Hoffrage (1998) considèrent que demander à une personne d'estimer un pourcentage peut être avantageusement remplacé (je dirais plutôt "complété") par la question « Dans combien de cas sur 100 ? ». C'est cette consigne « fréquentiste » que Bocklisch, Bocklisch et al. (2010, p. 1977) ont utilisée à propos de diagnostics médicaux pour 13 expressions verbales traduites numériquement par 121 participants avec la consigne :

« Dans combien de cas sur 100 un diagnostic est-il correct quand il est jugé ***doubtful****".*

Leur liste de 13 mots était : « 1. *Impossible ; 2. Very improbable ; 3. Quite improbable ; 4. Improbable ; 5. Hardly probable; 6. Sparsely probable ; 7. Doubtful ; 8. Thinkable ; 9. Possible ; 10. Probable ; 11. Quite probable ; 12. Very probable ; 13. Certain.* ».

Pour ces auteurs, le nombre optimal de niveaux est celui qui minimise les recouvrements des significations numériques. Deux de leurs représentations graphiques sont reproduites en figures 3 et 4, avec la permission du premier auteur (Bocklisch). L'axe vertical y représente des probabilités d'appartenance d'une valeur numérique donnée (en abscisse) au mot, qui est caractérisé par une « Fonction d'appartenance » (*fuzzy Membership Function ou MFs,* dont ces auteurs fournissent la formule mathématique) tenant compte de 4 paramètres de la distribution des valeurs observées : leur moyenne, leur écart-type, leur asymétrie (*skewness*) et leur aplatissement (*kurtosis*). En outre, pour chaque mot (ou expression verbale), la valeur modale de la distribution des traductions (des mots vers les nombres) correspond à 1 et les autres valeurs ont été « réduites » en proportion de cette valeur modale, selon la formule de la fonction d'appartenance. Une telle représentation graphique aide à visualiser non seulement les valeurs modales, mais aussi les recouvrements entre les traductions des 13 mots dont les numéros apparaissent au sommet de leur distribution.

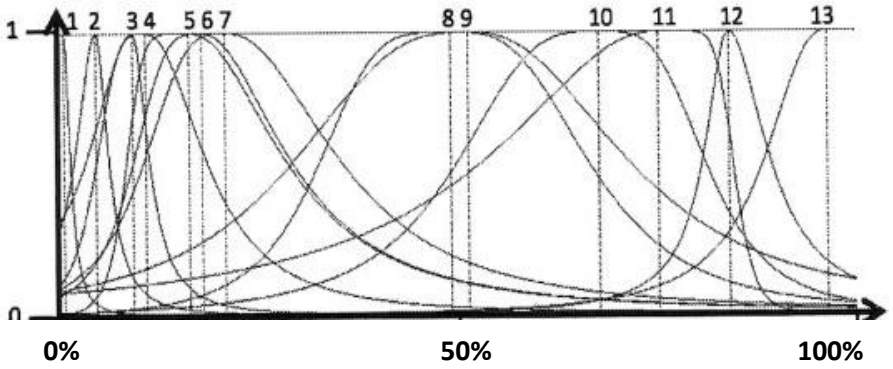


Figure 3 : Distributions lissées des proportions relatives de traductions numériques de 13 mots (Bocklisch et al (2010)

Sur base de la figure 3, en vue de minimiser les recouvrements de significations, Bocklisch et al (2010, p. 1978) suggèrent de travailler avec 5 niveaux (figure 4) : 0–20-50–70–100.

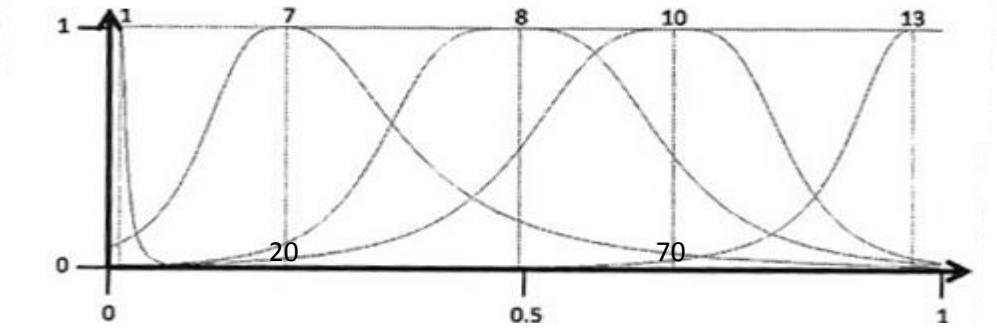


Figure 4 : La proposition de Bocklisch et al. (2010) de se limiter à 5 niveaux (asymétriques) de certitude : 0% 20% 50% 70% 100%

Malgré cette limitation à 5 niveaux seulement, les recouvrements de signification interindividuels restent énormes. Par exemple entre les mots 8 (*thinkable*) et 10 (*probable*).

6.4 Visualisation par histogrammes lissés

Les figures 5 et 6 présentent (en fréquences absolues de chaque valeur numérique) des données des expériences évoquées en section 6.2 (figure 2). La figure 5 montre que les distributions de « peu sûr » et « « très sûr » sont « bimodales » (plus exactement ont 2 pics) et que celle de « sûr » est « tri-modale » (plus exactement a 3 pics).

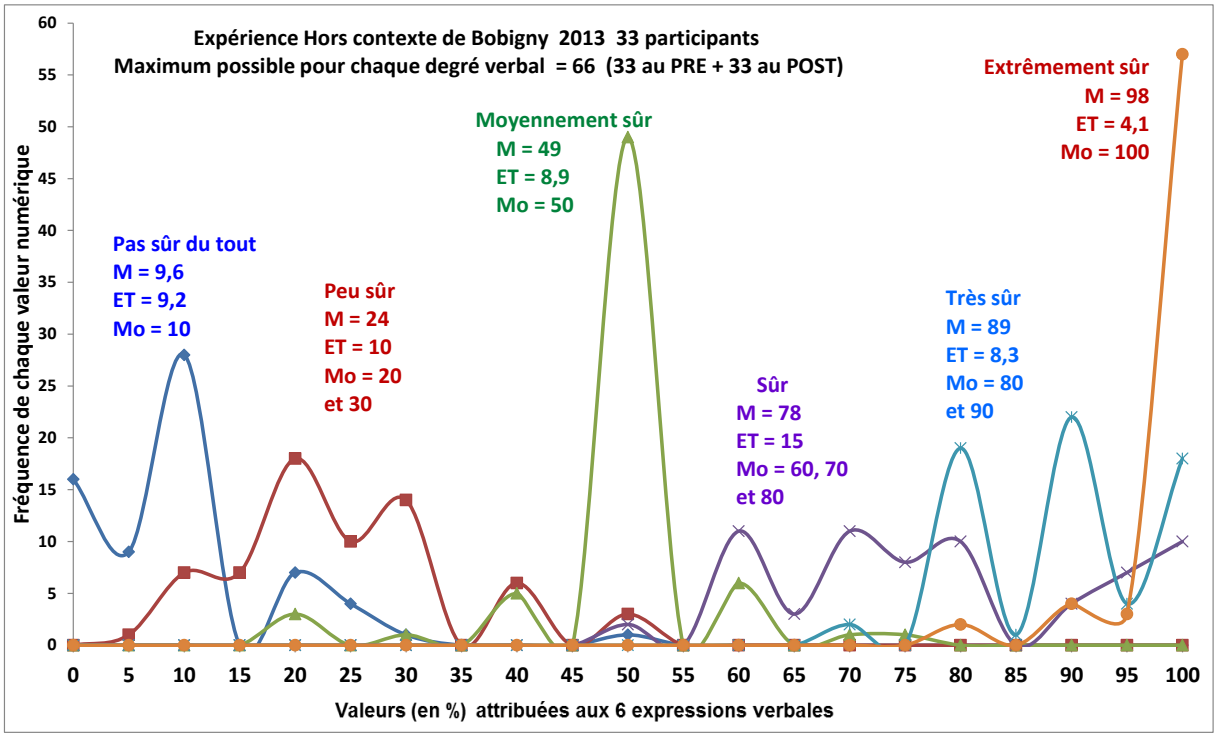


Figure 5 : Les 6 distributions lissées de l'expérience hors contexte

La figure 6 montre les distributions des valeurs (en %) données à 2 reprises par les 19 participants au pré- et au post- test dans l'expérience en contexte (Le jeu Compar-Aires) qui comptait 15 questions.

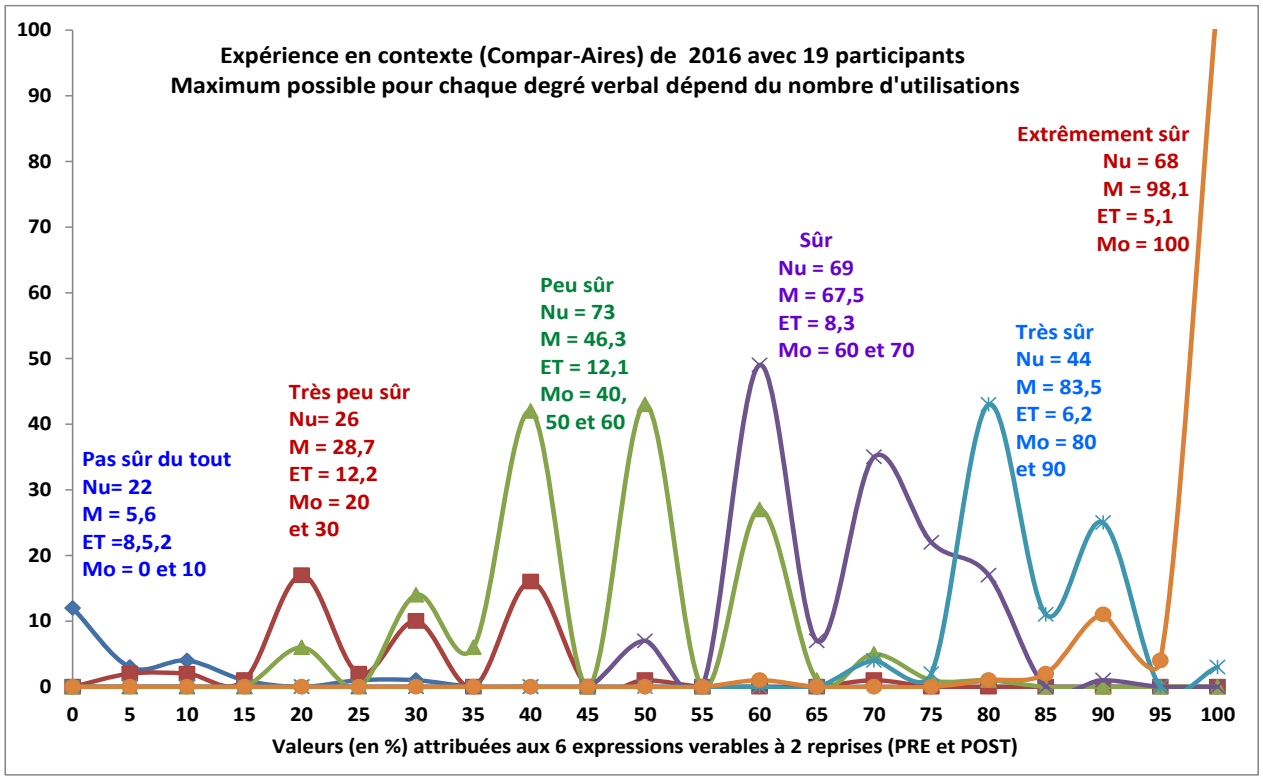


Figure 6 : Les 6 distributions lissées de l'expérience en contexte

Ces 570 pourcentages ne sont pas distribués de façon égale pour les 6 expressions verbales. C'est pourquoi ce nombre d'utilisations (Nu) est mentionné dans la figure 6.

Ces faibles nombres (19 participants, 15 questions avec 2 certitudes pour chacune) sont équivalents à bien des situations de « classes ». Ces données sont cependant suffisantes pour faire apparaître (en figure 6) le flou interprétatif, le brouillard communicationnel engendrés par le recours à des expressions verbales pour exprimer les degrés de certitude. Ainsi, si l'on se réfère à la figure 6, un enseignant qui reçoit la certitude « peu sûr » (courbe verte) doit-il l'interpréter comme 40%, 50% ou 60% ? Puisque parmi les 19 participants toutes ces valeurs numériques ont été associées dans des proportions assez semblables à cette expression verbale. Sans parler des (plus rares) fois où les valeurs étaient 30% et 70%.

7. Conclusions sur les différences interindividuelles

7.1 **Conclusion 1** (Bonne nouvelle pour les degrés de certitude, même en mots) :

Des adultes d'un bon niveau de scolarité (le supérieur) sont capables d'utiliser 6 degrés de certitude verbaux en leur attribuant une **signification** en % qui est **stable** dans le temps, à 5 heures d'intervalle (voir 6.2 et Leclercq, 2016). Cela rejoint les observations de Clark (1990) et de Druzdzel (1989).

7.2 **Conclusion 2** (Mauvaise nouvelle pour les degrés de certitude en mots) :

Ces mêmes adultes n'attribuent pas les mêmes significations numériques (en %) aux degrés verbaux que leurs collègues. La marge de variation (MV) peut aller jusqu'à 50% (tableaux 4, 5 et 7). Cela rejoint les conclusions de Bickel (voir section 5.2.1), de Beyth-Marom (voir ci-après), de Clark (1990) et de Druzdzel (1989).

Sur ces bases, il faudrait donc utiliser des certitudes exprimées non pas verbalement mais en pourcentage. Ce que Nakao et Axelrod (1983) annoncent dans le titre même de leur publication: « Les nombres valent mieux que les mots : les significations verbales de fréquence n'ont pas leur place en médecine. ». Ma conviction est qu'il en va de même en éducation, mais je reconnais que, pour convaincre les enseignants, certaines conditions devraient être réunies. Elles sont évoquées dans les perspectives ci-après (en section 10 : « Les critères de qualité attendus des degrés de certitude en pourcentage »).

Une façon très simplifiée de présenter les résultats présentés jusqu'ici consiste à dire que recourir aux expressions verbales pour exprimer les degrés de certitude, c'est introduire dès le départ dans ses modalités d'évaluation des erreurs de mesure aléatoires (dans l'interprétation de ces mots en nombres) se caractérisant, pour la majorité des 6 degrés de l'échelle, par

-une marge de variation (MV)	de 40%
-un écart interquartile (EIQ)	de 20%
-un écart-type (Ec. T.)	de 10% à 15%.

L'expression « erreur aléatoire de mesure » me paraît se justifier dans les situations (scolaires par exemple) où l'évaluateur, et souvent l'évalué lui-même, ne connaît pas les erreurs systématiques (s'il y en a) de chaque personne (de chaque étudiant).

De ces mêmes données qui ont été présentées ci-avant ressortent de bonnes et de mauvaises nouvelles… mais pas pour les mêmes idées. Je préciserai donc pour lesquelles.

7.3 **Conclusion 3** (Bonne nouvelle pour la recherche)

Les expériences décrites dans les sections précédentes peuvent être reproduites (avec des variations le cas échéant) par un enseignant avec ses propres étudiants ou, pour augmenter le nombre d'observations, en collaboration avec des collègues.

Troisième partie

8. A la recherche des valeurs optimales

8.1 Résumé des Modes numériques OBSERVES

Le tableau 11 présente les différentes valeurs des modes dans les diverses situations (voir figures 5 et 6) lorsque les consignes sont données hors contexte ou en contexte, lors du pré- et du post- test. La seule différence observée apparaît pour les colonnes « PEU SUR » et « MOYENNEMENT SUR (HORS contexte) et « très peu sûr » et « peu sûr » (en contexte).

Tableau 11 : Modes 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	N
HORS PRE (33)	10	20	50	60	80-90	100	33
HORS POST (33)	10	20	50	70-80	90	100	33
EN PRE (19)	0	20-40	40	60-70	80	100	19
EN POST (19)	0	20	40	60	80	100	19

Boehm (1989, rapporté par Hillson, 2005, p. 3) a étudié la traduction de 8 mots vers des %. Pour 5 de ces 8 mots, la traduction modale est un multiple de 10 et 2 autres mots sont des multiples de 5.

Hamm (1991, p. 217) propose 19 valeurs repères (qui sont, à très peu de choses près, des pas de 5%) ; 10 d'entre elles sont des multiples de 10 ; 8 sont des multiples de 5 et une est 0,33.

Tous ces résultats plaident en faveur de degrés de certitude exprimés en multiples de 10. Mais lesquels au juste ? Voyons ce que certains auteurs recommandent.

8.2 Valeurs recommandées pour les pourcentages

Les modes permettent de se faire une idée de ce que pourrait être une échelle de 5 à 7 degrés numériques. C'est ce que certains chercheurs ont recommandé :

Beyth-Marom (1982), après avoir observé un écart interquartile (Q3-Q1) de 14% en moyenne, recommande d'utiliser des valeurs numériques et propose une échelle à 7 degrés, mais d'intervalles.

Bocklisch et al. (2010) proposent (voir figure 4) 5 valeurs repères.

Zimmer (1983, 1986) propose une échelle à 5 valeurs repères.

Dans le tableau 12, j'ai rassemblé ces diverses propositions. Le tableau comporte 7 colonnes, pour prendre en compte les propositions allant de 5 à 6 degrés. J'ai intercalé ma proposition juste en-dessous de celle de Beyth-Marom pour faire remarquer que je recommande, en fait, les valeurs centrales de ses intervalles (une consigne que j'ai moi-même utilisée pendant des années). J'ai choisi (Leclercq, 2003) de travailler avec les centres de ces intervalles.

Tableau 12 : Valeurs recommandées (en %) par divers auteu**rs**

	1	2	3	4	5	6	7
Beyth-Marom (1982)	0-10	10-30	30-50	50	50-70	70-90	90-100
Leclercq (1998)	5	20	40		60	80	95
Bocklisch et al. (2010)	0	20		50	70		100
Zimmer (1983, 1986)		15	37	50		67	86
soit, en rapports		1/6	1/3	1/2		2/3	5/6

Zimmer propose des valeurs qui sont basées sur des rapports de chances (comme au tiercé).

L'échelle de Bocklisch et al. n'est pas symétrique.

8.3 Le cas des 50%

L'échelle que je recommande se distingue des trois autres notamment par l'absence du repère 50%, présent dans les trois autres échelles.

Les données de mes expériences autour de « sûr » n'indiquent pas qu'il soit incontournable d'y recourir. Hors contexte, en PRE (figure 5), la valeur 50% a été choisie bien plus (dans 13% des cas) que ses valeurs rondes voisines 40% (2,5% des cas) et 60% (5,6% des cas).

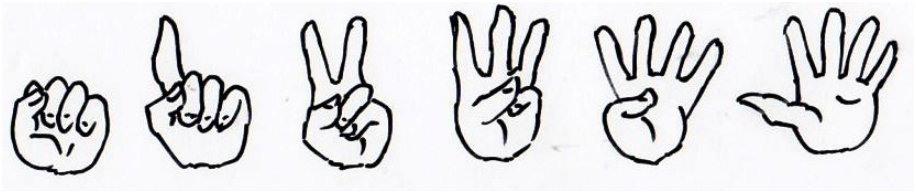
Par contre, en contexte, (figure 6), c'est l'inverse : 40% est choisi dans 11,9% des cas ; 50% est choisi dans 9,8% des cas et 60% est choisi dans 13% des cas.

Donc, abandonner 50% ne serait pas contradictoire avec les résultats observés ci-dessus, si l'on privilégie les observations « en contexte ». En outre, on connaît l'effet de refuge dans la valeur centrale. Ne pas proposer 50%, c'est empêcher cette réponse de facilité.

On peut faire une exception pour les questions Vrai-Faux, et réintroduire 50% qui signifie alors « je réponds au hasard ». L'échelle devient : 50 60 80 95.

8.4 Des arguments de validités écologique et d'acceptabilité

La consigne que je propose (les six premiers multiples de 20%, allant de 0 fois 20 à 5 fois 20) présente des avantages. L'encodage par ce multiple (0 1 2 3 4 ou 5) serait facilité, tant pour les étudiants sur leur feuille de test que pour les enseignants qui les introduisent au clavier dans un logiciel ad hoc. En outre, ces six multiples de 20 correspondent à une façon naturelle d'exprimer les probabilités avec les 5 doigts d'une main (figure 7).



0 **20** **40** **60** **80** **100**

Figure 7 : une façon « naturelle » d'exprimer ses degrés de certitude, sur 100% (les 5 doigts dressés), par paliers de 20%

Se baser sur les doigts d'une main présente une validité écologique, dans le sens donné par la littérature en évaluation.

Il faudrait cependant compter avec la réticence (bien compréhensible) de beaucoup de personnes à fournir les valeurs extrêmes (0% et 100%). La consigne dès lors pourrait être

	0 à 5%	20%	40%	60%	80%	95 à 100%
ou même	5%	20%	40%	60%	80%	95%.

8.5 Les mots peuvent-ils accompagner les pourcentages ?

Le but du présent article est de vacciner les enseignants contre l'utilisation, en contexte scolaire, pour caractériser les degrés de certitude, de mots au lieu de pourcentages de chance. Entendons-nous bien : si des pourcentages sont utilisés, on PEUT y ajouter ce que l'on veut (des mots, des codes alphabétiques ou numériques, des couleurs, et même des dessins comme des émoticônes). On court cependant le risque que les répondants accordent plus d'importance à ces décors ou à ces codes qu'aux pourcentages…et on n'aura rien gagné. Bref, seules les valeurs en pourcentage sont indispensables ; tout le reste est superflu, voire confusionnel.

9. Limites de certaines situations expérimentales

Avant de conclure, une mise au point plus théorique s'impose quant aux situations expérimentales créées et quant aux significations que leur donnent les participants.

9.1 Les auto-instructions à communiquer la différenciation

Pour Hamm (1991, p. 218), « présenter un ensemble de phrases qui ont des significations symétriques peut amener celui qui répond à les juger comme ayant des étendues de signification égales ».

Les réflexions de Fabre (1993a) permettent d'approfondir cette question en plaçant plusieurs des situations expérimentales évoquées ci-avant dans le cadre plus large de la communication entre les testés et l'expérimentateur.

Fabre (1993a, p. 99) cite Mellers et Birnbaum (1983, p. 158) qui font l'hypothèse que les sujets ont « tendance à utiliser des portions égales de l'étendue de réponses avec une fréquence égale […] ». Il donne une explication à ce phénomène, explication dans laquelle il appelle « cohérence » le fait de ne pas classer une même situation dans plusieurs catégories de certitude (ce que l'on cherche à vérifier, dans les expériences de répétition). Voici l'explication qu'il donne (p. 104) :

> *« …l'activité du sujet » [ici répondre à 15 questions avec des degrés de certitude] « est insérée dans un processus de communication. Le sujet doit ainsi :*
>
> *-d'une part concilier la cohérence (référenciation actuelle et antérieure des réponses) et la différenciation (expression des différences perçues) ;*
>
> *-d'autre part, répondre aux attentes supposées d'un interlocuteur qui a l'initiative de la situation, et s'adapter aux caractéristiques que cet interlocuteur est censé avoir introduites dans le matériel ».*

A l'appui de cette réflexion, Fabre (p. 104 toujours) invoque les travaux de Campbell, Lewis et Hunt (1958, p. 221) qui traitent des « auto-instructions à communiquer la différenciation ». Il cite un extrait de leur page 221 :

> *«…le sujet suppose que l'expérimentateur souhaite le voir utiliser toutes les catégories de réponses données au départ et qu'il lui présente une étendue symétrique de stimulation. ».*

Etant donné les considérations de Fabre (1993a) développées ci-avant, je fais l'hypothèse que, quand ce sont 6 modalités (verbales) de certitude qui leur ont été proposées, les modes des valeurs en pourcentage choisies par les participants sont quasi équidistants les uns des autres, c'est-à-dire ici à peu près tous les 20% sur l'étendue allant de 0% à 100%.

9.2 Le choix des valeurs multiples de 10

Rien dans la consigne (donnez une valeur comprise entre 0% et 100%) ne force les répondants à fournir des multiples de 10%. Sur base d'autres travaux expérimentaux (Leclercq, 1975 et 1982), je fais l'hypothèse que les répondants (adultes instruits) sont limités dans leur capacité de discriminer fiablement entre plus de 6 à 7 zones sur le continuum allant de 0% à 100%...et qu'ils en sont conscients. Dès lors, il leur paraît inutile d'entrer dans une précision illusoire.

Avec quelques exceptions ; par exemple, quand ils jugent en termes de rapport, du genre 3 chances sur 4, soit 75%, ou 1 chance sur 3, soit 33%. Mais le dénominateur de ce genre de rapport ne dépasse pas 5. Dans l'expérience en contexte appelée Compar-Aires (Leclercq, 2016), sur les 19 participants, 18 n'ont utilisé que des pourcentages multiples de 10 et de 5. Un seul participant a utilisé à 6 reprises (3 fois au PRE et 3 fois au POST) 99%. Ces 6 valeurs ont été assimilées à 100% dans le schéma de la figure 6.

10. Les critères de qualité attendus des degrés de certitude en pourcentage

10.1 Un postulat

Il serait naïf de considérer que les critiques formulées ci-avant à l'encontre des degrés de certitude exprimés en mots constituent un passeport de validité pour leur expression en pourcentage.

Je postule que les degrés de certitude numériques (en pourcentage de chances) ne comportent pas d'ambiguïté inter-individus, sauf avec des personnes pour qui, par exemple, 3 chances sur 5 ne signifieraient pas la même chose que 60%.

Néanmoins, ces degrés de certitude numériques (en pourcentage) ne devraient être largement adoptés dans la pratique scolaire que s'ils satisfont eux aussi à des critères de qualité.

Dans une autre publication (Leclercq, 2006), j'ai proposé 8 critères de qualité attendus d'un dispositif d'évaluation des apprentissages (DEA), que j'ai appliqués à l'évolution des QCM au cours des 100 dernières années. Ils sont identifiés par l'acronyme ETICPRAD, les premières lettres de chacun d'entre eux, dans un ordre choisi pour que l'acronyme soit prononçable comme un mot en soi. C'est le fil conducteur des perspectives que j'envisage dans la suite de cette section. Il n'y a aucun autre ordre (logique ou d'importances relatives) dans la liste qui suit.

10.2 Caractéristiques, qualités ou validités ?

Dans mon article de 2006, j'ai utilisé le terme « **validité** », que j'ai décliné en lui adjoignant à chaque fois un adjectif décrivant une des huit « qualités ». Le concept de « validité » fait l'objet de débats au point que périodiquement des numéros spéciaux de revues y sont consacrés[1]. Newton et Shaw (2016) distinguent trois « courants » sur le sujet :

(1) Ceux qu'ils appellent les « ultra-conservateurs » (ex : Borsboom et Markus, 2013, Borsboom et Wijsen, 2016), qui limitent le concept à sa définition la plus dépouillée (la propriété de la mesure de permettre d'inférer un trait – une capacité- latent(e) » ;

(2) les « conservateurs » qui s'en tiennent à une définition un peu plus large : « Le degré par lequel les données et la théorie concordent avec les interprétations des scores au test générés par un usage donné » (AERA, 1999). Cette même définition est souvent traduite en 3 types de validité : de *construct* (ou de théorie), de contenu et prédictive » et

(3) les « néo-libéraux » qui incorporent sous le terme « validité » **entre autres** le concept de validité conséquentielle (liée aux usages qui sont faits des mesures), notion introduite par Messick (1965, 1980). Ainsi, alors qu'il est principalement destiné à prédire les chances de réussite dans les études supérieures aux Etats-Unis, l'outil ACES[2] distingue 9 types de validité.

Puisqu'il faut se positionner dans ce débat qui est loin d'être clos, j'ai pris les options suivantes.

1 Educational Measurement en 1997, Journal of Applied Testing Psychology en 2005, Assessment in Education en 2013.

2 http://www.collegeboard.com/highered/apr/aces/vhandbook/evidence.html

Option 1 : Considérer le terme validité comme l'équivalent de « qualité » (ou force selon son origine latine, *validus* signifiant « fort »).

Option 2 : Parler non pas seulement de la validité du score d'un test mais d'un **système** ou **dispositif** d'évaluation, qui englobe non seulement le contenu, mais aussi l'annonce (et donc son impact sur la préparation), les séquelles (par exemple l'impact des résultats sur les méthodes d'étude des étudiants), la faisabilité pratique, etc.

Option 3 : Ajouter au terme validité une série d'adjectifs qui en précisent la facette. On pourra me dire « Plutôt que d'affubler ainsi le terme « validité » d'adjectifs divers, il eut été préférable de parler de « propriétés » ou de « qualités » d'un Dispositif d'Evaluation des Apprentissages. Cette objection est fondée, et d'ailleurs je mentionnerai ces synonymes (dans la section 11 ci-après). Néanmoins, j'ai trouvé plus efficace de m'en tenir à 8 adjectifs, car ma « taxonomie » est destinée à des enseignants qui doivent en permanence avoir à l'esprit ces multiples critères de qualité qui, bien que parfois corrélés, sont, en principe, disjoints.

11. Conclusions : 8 directions pour les de recherches ultérieures

Le présent article est loin d'avoir résolu tous les problèmes liés à l'utilisation des degrés de certitude, et n'en est même que la première étape. Je pense cependant qu'il permet de définir une des bases (l'expression numérique en % et la granularité 6 sur cette échelle) pour approfondir, théoriquement et expérimentalement, ces problèmes.

Ces approfondissements utiles devraient se développer dans les 8 directions ETICPRAD décrites par ailleurs (Leclercq, 2006) et que je détaille ci-après. Chacune des « validités » explicitées ci-après est un programme de recherche en soi. Certaines de ces recherches déjà publiées, ont été signalées. Elles utilisaient une consigne numérique (en %) avec les consignes décrites en section 8.4. Sans le présent article, ce choix serait apparu comme arbitraire. D'autres articles décrivant des recherches terminées sont en préparation. Ils ne peuvent cependant prétendre épuiser le sujet.

11.1 La validité écologique (ou authenticité)

Les valeurs-repères, c'est-à-dire des valeurs proposées (ex : 20%, 40%, 50%, 75%, etc. parmi lesquelles l'évalué est invité à choisir) sont-elles jugées normales ou justifiées par les répondants ? On pourrait aussi parler d'un critère d'acceptabilité pour l'étudiant. Ceci devrait être vérifié par des observations de terrain et sur plusieurs terrains (de l'enseignement primaire au supérieur), et non plus, comme ci-avant, par des expériences de laboratoire.

Ce critère apparaît évident. Or il n'est pas acquis d'avance car les personnes préfèrent exprimer leur degré de certitude par des mots plutôt que par des pourcentages, pour des raisons que je développe dans un article en préparation.

11.2 La validité théorique (ou « de *construct* » et « de contenu »)

Le recours aux degrés de certitude en pratique scolaire devrait s'appuyer sur un modèle théorique de l'activité mentale d'une personne qui répond à une question.

Un tel modèle devrait :

- être diversifié selon les natures différentes des questions (par exemple des questions ouvertes ou à choix multiple, ou vrai-faux) et donc des processus mentaux requis par la résolution de problèmes différents. Choppin, entre autres (Leclercq, 1982, p. 165-183) a proposé des modèles de tels processus mentaux.
- proposer une définition de ce qu'est « maîtriser » ou « connaître ». En plus de la connaissance ou de la maîtrise objective devrait se généraliser la mesure des connaissances et de la maîtrise subjectives, en particulier via le concept de connaissance partielle (De Finetti, 1965, p. 101 et Leclercq, 1982, p.257-270).
- traiter du réalisme (ou calibration) c'est-à-dire établir une bonne correspondance (*fit*) entre les pourcentages de réussite annoncés et les pourcentages observés des utilisateurs (étudiants, patients, enfants, etc.). Ainsi, le taux d'exactitude (TE) des réponses données par une personne avec la certitude 60% devrait avoisiner 60%. Encore faut-il définir le réalisme d'une façon scientifiquement fondée et en proposer une (des) mesure(s) qui soi(en)t opérationnelle(s) (Schraw et al., 2013 ; Leclercq, 1982, p.227-236).

11.3 La validité Informative (ou diagnostique et utilisable)

On parle aussi de validité diagnostique. L'acuité ou granularité ou précision fiable des échelles en pourcentage est un constituant important de l'enjeu. Autrement dit, avec quel degré de « grossissement » les degrés de certitude numériques permettent-ils d'observer les phénomènes ? Quelle est leur valeur ajoutée en termes de sensibilité ou de précision ? Cela aussi devrait faire l'objet d'une revue (seulement ébauchée ici) de la littérature. Se confirmera-t-il qu'une échelle de pourcentages à six degrés comme celle envisagée en section 8.4 (les doigts de la main ou les 6 multiples de 20) est appropriée, voire optimale ? Par « appropriée » il faut comprendre « qui respecte la capacité humaine de distinguer de manière fiable entre plusieurs degrés de certitude différents sur l'échelle qui va de 0% à 100% ». On sait que ces degrés de certitude numériques fournissent des indices se rapportant à la métacognition, comme les indices dits « de résolution » que sont

- la confiance ou certitude moyenne des réponses correctes,
- l'imprudence ou certitude moyenne des réponses incorrectes (Leclercq, 2009, p. 209),
- la nuance ou discriminance ou différence entre confiance et imprudence. (Leclercq, 2003, p. 42).

11.4 La validité Conséquentielle (ou d'impacts)

Cette expression, due à Green (1998), désigne les effets positifs et l'absence ou la minimisation d'effets négatifs qu'a la pratique de degrés de certitude numériques.

D'abord, les conséquences néfastes (critère de non nocivité) de ce « brouillard communicationnel » (les certitudes en mots) pour la pratique scolaire ou pour l'apprentissage en général, devraient être identifiées et documentées : par exemple, dans des tâches de communication ou de vérification en cas de doute.

Ensuite, les avantages (critère de fécondité) de recourir à des valeurs numériques devraient être démontrés et illustrés dans l'enseignement collectif, dans l'apprentissage individuel, dans le pilotage de systèmes scolaires et dans la recherche.

11.5 La validité prédictive (ou concourante ou corrélationnelle avec un critère extérieur)

Les certitudes ajoutées aux réponses devraient permettre d'expliquer (de prédire) mieux que les réponses seules une série de phénomènes. Ainsi, Leclercq et Detroz (2003) ont mis en évidence que des indices incorporant les degrés de certitude dans des tests à l'entrée à l'université étaient plus prédictifs de la réussite en fin d'année académique que les indices qui ne tenaient pas compte de ces données.

11.6 La validité de répétition (ou de reproductibilité ou de fidélité)

Il s'agit de la fidélité de la mesure (en anglais *reliability*). On peut tenter d'apprécier cette fidélité par des procédures classiques (comme la *half-split* ou l'alpha de Cronbach), mais la méthode la plus pertinente me paraît ici le test de répétabilité ou de stabilité dans le temps. Cette répétabilité devrait être satisfaisante. Autrement dit, un degré de certitude sous forme de pourcentage donné par une personne à une réponse donnée à une question précise est-il le même (ou approximativement le même) si on demande à cette personne de le fournir à nouveau quelque temps plus tard si son état mental n'a pas changé ? De telles procédures expérimentales ont été appliquées avec des degrés de certitude verbaux et numériques (Leclercq, 2016). Il est apparu que ce critère est lié au précédent (la granularité) : moins l'échelle comporte de degrés, plus il est facile d'être stable, c'est-à-dire de répéter les degrés de certitude à l'identique lors d'un test et d'un retest, dans les mêmes conditions.

11.7 La validité d'acceptabilité (ou de faisabilité pratique)

Les postulats, les enjeux et les procédures doivent convenir, être acceptables pour les enseignants comme pour les étudiants. On peut imaginer qu'ils soient même désirés par les uns et les autres. Par exemple, les degrés de certitude numériques pourraient diminuer le temps de réponse lors d'un test, ou diminuer le stress des étudiants lors du passage d'une épreuve ou améliorer les scores des étudiants au test. Ces trois dernières hypothèses sont volontairement provocantes car je sais que bon nombre de personnes sont persuadées que, sur ces trois points, c'est l'inverse qui se passe. Il importerait de dépasser le niveau des opinions et de documenter ce genre de débats par des données concrètes, et d'étudier les conditions qui les façonnent.

11.8 La validité éthique (ou de valeur morale ou déontologique)

Les usages docimologiques (dans le sens d'attribution de notes, de points, qui tiennent compte des degrés de certitude) les plus appropriés doivent être identifiés ainsi que ceux qui amènent les étudiants, légitimement intéressés à maximiser leur score total, à déformer (biaiser) leur certitude estimée au moment de l'exprimer (Leclercq, 1982). Ce qui n'affecte pas seulement la validité théorique de l'approche, mais débouche sur des injustices.

De nombreux indices concernant ces critères de qualité ont déjà fait l'objet de publications. Cependant ces publications sont dispersées et ne répondent pas à toutes les questions.

Les 8 critères de qualité énoncés ci-dessus sont liés entre eux. S'il s'avérait qu'ils ne sont pas suffisamment satisfaisants, alors il faudrait sans doute abandonner l'évaluation par degrés de certitude.

L'absence de synthèse de recherches, voire d'études tout court, sur certains de ces divers critères explique la méfiance quant à l'utilisation des degrés de certitude en contexte scolaire.

Défiance qui en freine le développement. C'est à une telle tâche de légitimation des degrés de certitude que participe le présent exposé ainsi que plusieurs articles en préparation.

L'édifice à construire est (sera) élevé et complexe.
Ma préoccupation est que les fondations en soient solides.

11. Références

AERA. American Educational Research Association, Psychological Association, & National Council on Measurement in Education. (1999). Washington, DC: American Educational Research Association. Standards for Educational and Psychological Testing

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*. https://www.coursehero.com/file/9906871/18-Probability-Assessment/

Bland CJ, Meurer L.N. & Maldonado G. (1995). A systematic approach to conducting a non-statistical meta-analysis of research literature. [Review]. *Academic Medicine*; *70*, 642-53.

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.

Bocklisch, F; (2011). The vagueness of verbal probability and frequency expressions. *International Journal of Advanced Computer Science, 1*(2), 52-57.

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

Borsboom, D. & Markus, (2013). Truth and Evidence in Validity Theory. *Journal of Educational Measurement, 50,* 110-114.

Borsboom, D. & Wijsen, L. (2016). Frankeinstein's validity monster. The value of keeping politics and science separated. *Assessment in Education. Principles Policy and Practice, 23*(2), 281-283.

Bryant, G. and Norman, G. (1980). Expressions of probability: Words and Numbers. *New England Journal of Medicine, 302*(7), 441a.

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

Budescu, D. & Wallsten, T. (1987). Subjective estimation of precise and vague uncertainties. In G. Wright & P. Ayton (Eds). *Judgmental Forcasting*. NY: Wiley & Sons.

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.

Campbell, D. & Stanley, J. (1963, 2nd ed. 1966). *Experimental and quasi experimental designs for research*. Chicago: Rand Mc Nally & Co.

Chee, C-S. (2006). "In-Isolation" study on verbal uncertainty expressions. *Proceedings of the 2nd IMT-GT regional Conf. on Mathematics, Statistics & Applications*. University Sains. Malaysia. Penang.

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

Cohen J. (1969). *Statistical Power Analysis for Behavioral Sciences*. New York: Academic Press

Cohen, B. L. (1986). *The effect of outcome desirability on comparisons of numerical and linguistic probabilities*. Unpublished M. A. thesis, University of North Carolina at Chapel Hill.

De Finetti, B. (1965), Methods for discriminating levels of partial knowledge concerning a test item. *British Journal of Mathematical and Statistical Psychology*, *18*, 87-123.

Delvenne (1999). Meta-analyse qualitative. http://www.ebm.lib.ulg.ac.be/prostate/metaqual.htm

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

Echternacht, G. (1972a). The use of confidence testing in objective testing. *Review of Educational Research*, *42,* 217-236.

Echternacht, G., Boldt, R. & Sellman, W. (1972b). Personality influences on confidence test scores. *Journal of Educational Measurement, 9*, 235-241.

Echternacht, G. (1976). Reliability and validity of option weighting schemes. *Educational and Psychological Measurement*, *36,* 301-309.

Edwards, W. (1967), Probabilistic information processing by men and man-machine systems, Actes du symposium OTAN "*La simulation du comportement humain"*, Paris : Dunod.

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

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

Fares, I. (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.

Foley, J. (1959). The expression of certainty. *American Journal of Psychology, 72*, 614-615.

Frary R.B. (1989). Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education*, *2*(1), 79-96.

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

Glass, G.V. (1976). Primary, secondary, and meta-analysis of research, *Educational Researcher*, *5*, 3-8.

Glass, G., Mc Gaw, B. & Smith, M. (1981). *Meta-analysis in social research*, Beverly Hills CA: Sage Publications.

Granhag, P. & Strömwall, L. (2000). Deception detection: Examining the consistency heuristic. In C. Breuer, M. Kommer, J. Nijboer & J. Reintjes (Eds). *New trends in criminal investigation and evidence, 2*, (pp. 309-321). Antwerp: Intersentia.

Green, D. R. (1998). Consequential aspects of the validity of achievement tests: A publisher's point of view. *Educational Measurement*, *17*, 16-19, 34.

Grevisse, M. (2011). *Le Bon Usage*, quinzième édition par André Goosse, de Boeck Duculot.

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.

Hattie, J. (2009). *Visible learning*. London: Routledge.

Hedges LV & Olkin I. (1985). *Statistical methods for meta-analysis*. London: Academic Press.

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

Johnson, E. (1973). 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 D. Steury (Ed.). *Sherman Kent and the Board of National Estimates: Collected Essays*. Washington, D.C.: CIA.

Koehler, R. (1971). A comparison of the validities of conventional choice testing and various confidence marking procedures, *Journal of Educational Measurement, 8*, 297-303.

Koehler, R. (1974). Overconfidence on probabilistic tests. *Journal of Educational Measurement*, *11*, 101-108.

Kong, A., Barnett, G., Mosteller, F. & Youtz, C. (1986). How medical professionals evaluate expressions of probability. *New England Journal of medicine,* 740-744.

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review*, *119*(1), 80-113.

Leclercq, D. (1982). Confidence marking, its use in testing. In Postlethwaite & Choppin, *Evaluation in Education*, *6*, 161-287, Oxford: Pergamon Press. 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. (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., Boxus, E., De Brogniez, Ph., Wuidar, H. & Lambert, F. (1993). The TASTE approach: General implicit solutions in MCQs, open books exams and interactive testing and self-assessment. In D. Leclercq & J. Bruno (Eds), *Item Banking: Interactive Testing and Self-Assessment*, NATO ASI Series F112, Berlin : Springer Verlag, 210-232. http://hdl.handle.net/2268/22610

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. & Detroz, P. (2003). Liens entre caractéristiques de départ (dont les résultats aux check-up) et les réussites en première candidature. In 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/2268/27520

Leclercq D. (2006). L'évolution des QCM. In Figari G & Mottier-Lopez, L. (Eds.) *Recherches sur l'évaluation en éducation*. Paris : L'Harmattan,: 139-46. http://hdl.handle.net/2268/10124.

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

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 Deux études sur la variabilité inter-individus des significations métriques données aux degrés de certitude verbaux. *Evaluer : Journal International de Recherche en Education et Formation*, 2(1), p. 89-125. http://hdl.handle.net/2268/202730

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

Mapes, R. (1979). Verbal and numerical estimates of probability in therapeutic contexts. *Social Science and Medicine. 13A*, 277-282.

Marshall, E. (1988). Academy panel defaults NASA's safety analysis. *Science, 239*, 1233.

Mellers, B. & Birnbaum, M. (1983). Contextual effects in social judgment. *Journal of Experimental Social Psychology, 19*, 157-171.

Messick, S. (1965). Personality measurement and the ethics of assessment. *American Psychologist, 20*, 146-162.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012-1027.

Mosteller, F. & Youtz, C. (1990). Quantifying probabilistic expressions. *Statistical Science, 5*(1), 2-34.

Nakao, M. & Axelrod, S. (1983). Numbers are better than words: verbal significations of frequency have no place in medicine. *American Journal of Medicine, 74*, 1061-1065.

Newton, P. & Shaw, S. (2016). Disagreement over the best way to use the word "validity" and options for reaching consensus. *Assessment in Education. Principles Policy and Practice, 23*(2), 178-197.

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.

Parent, N. & Vissandjée, B. (2008) Evidence-based cardiovascular nursing practice: Why? For whom? Where and how? *Canadian Journal of Cardiovascular nursing, 18*(3), 26-30 (EN) & 32-36 (FR). http://www.medsp.umontreal.ca/IRSPUM_DB/pdf/25158.pdf

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

Pepper, S. & Prytulak, L. (1974). Sometimes frequently means seldom: Context effects in the interpretation of quantitative expressions. *Journal of Research in Personality, 8*, 95-101.

Pugh, R. & Bunza, J. (1975). Effects of a confidence weighted scoring system on measures of test reliability and validity. *Educational and Psychological Measurement, 35*, 73-78.

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

Reyna, V. (1981). The language of possibility and probability: Effects of negation in meaning. *Memory & Cognition, 9*, 642-650.

Rigby, L. & Swain, A. (1971). In-flight target reporting. How much is "A bunch"? *Human Factors, 13*(2), 177-181.

Schraw, G., Kuch, F. & Gutierrez, A. (2013). Measure for measure: Calibrating ten commonly used calibration scores. *Learning and Instruction, 24*, 48-57.

Shuford, E., Albert, A. & Massengil, N.E. (1966). Admissible probability measurement procedures, *Psychometrika, 31*(2), 125-145.

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.

Van Naerssen, R.F (1962). A scale for the measurement of subjective probability, *Acta Psychologica*, *20*(2), 159-166.

Vickers, D. (2001). Where does the balance of evidence lie with respect to confidence? In E. Sommerfield, R. Kompass, & T. Lachmann (Eds.). *Proceedings of the 7th annual meeting of the International society for Psychometrics,* Lengerich Germany, 148-153.

Wallsten, T., Fillenbaum, S. & Cox, J. (1986). Base rate effects in the interpretation of probability and frequency expressions. *Journal of Memory and Language, 25*, 571-587.

Zadeh, L. (1974). The concept of a linguistic variable and its application to approximate reasoning. In K. Fu & J. Tow (Eds.), *Learning systems and intelligent robots* (pp.1-10). New York: Plenum Press.

Zimmer, A. (1984). A model for the interpretation of verbal predictions. *International Journal of Man-Machine Studies, 20*, 121-134.

Annexe 1

Voici les consignes que j'ai utilisées successivement pour recueillir les degrés de certitude en contexte scolaire et académique :

- 4 zones (Leclercq, 1975 et 1982) :

0-25%	25- 50%	50-75%	75-100%

- 6 zones : voir Leclercq et al. (1993, p. 214, 217).

0-25%	25-50%	50-70%	70-85%	85-95%	95-100%

- - 6 valeurs repères (Leclercq, 2003) en % :

0 20 40 60 80 100

- - 6 valeurs repères « adoucies » (vers 2015) en % :

0-5 20 40 60 80 95-100

Annexe 2

En section 2.5.1, je dis « Il ne s'agit pas de récolter le plus grand nombre possible de publications, mais un nombre suffisant et si possible les plus connues ». Cette définition de la base de données m'est inspirée par celle que donne Maurice Grevisse (2011) pour justifier sa propre base de données sur laquelle il fonde son célèbre livre « Le bon usage » (de la langue française) : « La meilleure partie des bons auteurs ». Qui dit plus subjectif ?

Annexe 3

Un exemple de méta-analyse qualitative est développé par Delvenne (1999) dans http://www.ebm.lib.ulg.ac.be/prostate/metaqual.htm

Par ailleurs, Parent et Vissandjée (2008) fournissent trois illustrations dans des domaines médicaux différents de l'application de méta-analyses qualitatives.
En ce qui concerne les pratiques post-opératoires en chirurgie cardiaque, une de leurs synthèses débouche sur des valeurs moyennes (et des intervalles autour de cette moyenne) de la réduction de la durée (en jours) d'hospitalisation et (en heures) de séjour en soins intensifs.
Une autre méta-analyse qualitative, qui pourtant ne regroupe que trois études, permet de conclure à la supériorité de certaines procédures sur d'autres.
Une troisième méta-analyse qualitative (sur le sevrage tabagique) retient 42 études (totalisant 15000 participants) qui permettent de conclure à la supériorité de certains types d'interventions par rapport à d'autres.