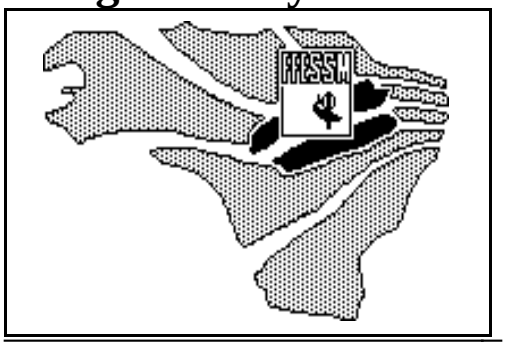
Fédération Française d'Études
et de Sports Sous-Marins

Commission Technique Régionale
Bretagne & Pays de Loire



Evaluation et Plongée

Application aux épreuves théoriques

Mémoire d'Instructeur Régional
Marc Gogny
MF2 n°654
Juin 1996

Table des matières

Avertissement....4

Introduction....5
Les objectifs de ce fascicule.... 5
Qu'est-ce que l'évaluation ?.... 5
Quelques remarques préliminaires.... 5
Une définition de l'évaluation.... 6
La perception de l'évaluation.... 6
Le plan du fascicule.... 7

Evaluer pourquoi ? - Les types d'évaluation....9
La place de l'évaluation dans le système.... 9
Les trois types d'évaluation....12
L'évaluation prédictive....12
L'évaluation formative....14
L'évaluation sommative....16
L'évaluation au deuxième degré....17
En résumé....18

Evaluer quoi ? - Les objectifs....19
La liste des objectifs....20
Une méthode : l'analyse descendante....20
Les composantes de l'apprentissage....21
Les objectifs du domaine cognitif....23
La construction de l'évaluation....25
Comment définir un objectif ?....26
L'influence des systèmes de notation....28
Le risque du système des moyennes....28
Vers la définition d'un "Niveau Acceptable de Performance"....29
Unités capitalisables ou examen ponctuel ?....31
En résumé....32

Evaluer avec quoi ? - Les outils....33
La correspondance objectifs-outils....33
Les limites de l'évaluation....34
Les risques d'évaluer....34
Les limites extrinsèques....35
Les limites intrinsèques....35
Les effets-parasites....36
Les outils pour épreuves théoriques....38
Les items de sélection....39
Les items de production....44
Après la rédaction....46
En résumé....47

L'expression des résultats.....48
La notation des items.....48
Notation quantitative classique.....49
Notation quantitative pondérée.....49
Notation qualitative.....49
Exemple.....50
La notation des questionnaires.....51
Notation quantitative.....51
Notation qualitative.....53
La notation sur l'ensemble des épreuves.....53
Le Niveau Acceptable de performance et la médiane.....54
Les coefficients (moyenne pondérée).....54
La moyenne simple.....55
L'importance de la délibération du jury.....55
L'analyse de l'examen.....55
L'analyse des épreuves.....56
L'analyse des items.....58
En résumé.....60

Bibliographie.....61

Annexe I : Résumé des points principaux.....62

Annexe II : Simulation de résultats d'examen Niveau 4.....64
Pratique.....64
Pierre.....64
Paul.....64
Jacques.....65
Théorie.....66

Annexe III : Analyse d'un examen théorique Niveau 4.....67
Rappel des résultats.....67
Epreuve de Tables.....67
Epreuve d'Accidents.....68
Epreuve de Physique.....68
Epreuve de Physiologie.....69
Epreuve de Matériel.....70
Epreuve de Législation.....70

Avertissement

Ce document est un mémoire pour l'obtention du titre d'Instructeur Régional du Comité Bretagne & Pays de Loire. Comme tel, il n'engage que la responsabilité de l'auteur et ne représente pas la position officielle du Comité, ni de la Fédération, en matière de brevets ou d'évaluation.

J'espère modestement que cette réflexion ajoutera à la somme des travaux qui sont actuellement menés dans le but de revoir les épreuves qui constituent les brevets attestant des niveaux des plongeurs, donc de leurs compétences en fonction de leurs prérogatives, et surtout d'harmoniser les pratiques d'évaluation.

Ce mémoire est plus particulièrement destiné à l'évaluation des "connaissances" théoriques (les "savoir"). Néanmoins, la plupart des éléments de réflexion sont transposables à l'évaluation des aptitudes dans l'eau (les "savoir-faire"), et, dans une moindre mesure, à la composante affective de l'apprentissage (le "savoir-être").

Dans la mesure du possible, les termes "savants" utilisés en Sciences de l'Education sont évités. Dans le cas contraire, ils sont clairement définis. Pour simplifier le style, le plongeur en formation, "l'apprenant", le stagiaire, est appelé dans ce texte "l'élève". L'enseignant est appelé "le moniteur". Enfin, le genre masculin est employé, sans pour autant traduire une quelconque discrimination vis-à-vis de nos charmantes plongeuses et monitrices…

Introduction

Les objectifs de ce fascicule

Ce fascicule a deux objectifs clairement identifiés : apporter aux moniteurs un certain nombre de conseils concernant la façon d'évaluer la progression de leurs stagiaires, ainsi que leur prestation aux examens, notamment dans le domaine de la théorie de la plongée, et élargir la réflexion, plus principalement destinée aux instructeurs, sur les pratiques d'évaluation au sein du Comité, et sur la façon d'aboutir à une meilleure harmonisation.

Qu'est-ce que l'évaluation ?

Quelques remarques préliminaires

Etymologiquement, le terme d'évaluation revêt deux acceptions différentes.

Evaluer, c'est mesurer…

L'évaluation est souvent une notion simplement quantitative ; à l'aide d'un instrument de mesure approprié (par exemple une montre avec trotteuse ou chronomètre), on traduit en chiffres l'activité évaluée (ex. : une durée d'apnée). Le mesurable paraît facile ; encore faut-il utiliser des unités comparables, et préciser les conditions de la mesure.

Utiliser des unités comparables…

Si l'on mesure la durée d'une apnée statique, d'une apnée dynamique ou encore le temps passé à parcourir une distance donnée, on emploie toujours une unité de temps, mais les résultats obtenus ne sont pas comparables, puisque les activités évaluées sont différentes. Si la compétence à évaluer fait intervenir les trois activités, il faudra utiliser un barème de notation, en attribuant éventuellement un coefficient différent à chaque activité si l'on veut leur donner des poids différents.

La correction d'un questionnaire de 20 questions à 1 point ne semble pas poser de problème… mais ces 20 questions mesurent-elles des connaissances ayant la même importance ? Que conclure si un élève obtient la note de 8 ?.

Préciser les conditions de la mesure…

Un 800 m P.M.T. en piscine, sans combinaison, et la même épreuve en mer, avec combinaison et lestage, par force 4 avec 2 noeuds de courant… donnent des résultats différents pour un même plongeur ayant la même condition physique. Il faut en tenir compte.

Evaluer, c'est apprécier…

Mais l'évaluation amène dans la plupart des cas à poser un jugement de nature qualitative. La performance ne se mesure pas directement, elle est estimée, appréciée par le moniteur (par exemple la qualité d'une immersion en canard). Sur quoi repose une telle évaluation ? Sur l'écart observé entre la prestation de l'élève et la prestation idéale attendue.

Certains types d'épreuves théoriques relèvent aussi de l'appréciable plus que du mesurable. C'est le cas notamment des questions ouvertes ou de synthèse.

… dans un cadre de références.

Finalement, dans les deux cas, on est amenés à définir un système de références, avec un niveau d'exigences qui s'y rapporte.

Une définition de l'évaluation

En tenant compte de ces remarques préliminaires, la définition de l'évaluation qui paraît la plus appropriée à l'activité plongée est la suivante :

> **L'évaluation est un jugement *outillé* en vue d'une prise de décision *éclairée* correspondant à des *buts* fixés *a priori*.**

Outillé …

Un adage dit : "il n'y a pas de bon ouvrier sans de bons outils". En matière d'évaluation aussi, le choix des outils, des instruments de mesure est déterminant. Dans certains cas, ils sont fixés par un texte et la marge d'interprétation de l'évaluateur est limitée (ex. : épreuve du 800 m P.M.T. au niveau 4). Dans d'autres cas, le choix du test ou de la façon dont il est mis en oeuvre dépend entièrement du jury (ex. : épreuve de "Code à 40 m" du niveau 4 ; épreuves théoriques).

Eclairée…

La décision n'est pas arbitraire. Elle repose sur un ensemble de critères précis, qu'un autre évaluateur pourrait utiliser avec le même résultat. La subjectivité ne peut être éliminée. Elle doit cependant être limitée.

Buts…

La prestation n'est évaluée que par son écart avec une prestation idéale espérée, reflet direct de la compétence voulue pour le plongeur. Autrement dit, il n'y a pas d'évaluation possible sans une définition des objectifs de l'enseignement.

A priori…

Ces objectifs sont clairement identifiés avant la séance d'évaluation. Ils ont été annoncés et, dans l'idéal, les élèves se les sont appropriés et savent ainsi très exactement ce que l'on attend d'eux.

Dans tous les cas, il ne faut pas confondre évaluation-*sanction* et évaluation-*motivation* ; si les élèves se sentent dévalorisés, leur échec est perçu comme une faute, et peut devenir un blocage. S'ils sont impliqués dans leur évaluation, et que l'échec est bien présenté, il n'est qu'une erreur comprise et devient un élément de motivation pour progresser.

La perception de l'évaluation

En plongée, comme dans le domaine scolaire, l'évaluation est souvent perçue comme l'acte simple qui consiste à attribuer une note à une prestation. Dès que l'on demande selon quel(s) critère(s) cette note a été fixée, la réponse est déjà très floue. Dans le meilleur des cas, elle correspond à une échelle arbitraire que se fixe le moniteur en toute bonne foi, à travers le prisme de ses propres connaissances et de l'idée

qu'il se fait de la plongée. Dans le pire des cas, la notation est totalement aléatoire.

Enfin, dans tous les cas, c'est encore dans le domaine des connaissances théoriques que les divergences sont les plus importantes et que l'effort d'harmonisation doit être le plus soutenu.

Trop souvent, les efforts pédagogiques développés par les moniteurs ne sont axés que sur le transfert des connaissances ; l'évaluation n'est mise en œuvre qu'au moment de l'examen, sans qu'y soit consacrés le temps et l'attention nécessaires.

"Prépare-moi vite quelques questions de physique, moi je m'occupe des accidents…"

L'expérience des examens théoriques régionaux anticipés pour le niveau 4 a montré l'intérêt d'une préparation minutieuse des épreuves, et que même avec vigilance, il arrive qu'on laisse échapper des sujets qui donnent une mesure faussée des aptitudes réelles des candidats.

Le plan du fascicule

Les informations qui circulent entre les différents niveaux d'un système pédagogique ne sont en fait que des formes différentes d'évaluation. Cela veut dire qu'il existe plusieurs types d'évaluation dont le rôle est différent, ce qui peut conduire à des pratiques d'évaluation différentes. C'est ce qui sera envisagé en premier lieu *(évaluer pourquoi ?)*.

En matière d'enseignement, il est devenu classique de parler d'objectifs pédagogiques, sur la base de l'adage "si l'on ne sait pas où l'on veut aller, on a toutes chances de se retrouver ailleurs…". Il en va de même pour l'évaluation : comment être certain de la mesure si le système de référence n'est pas clairement étalonné ? Ces éléments font l'objet du deuxième chapitre *(évaluer quoi ?)*.

Mais l'un des points essentiels de ce fascicule reste de montrer comment les différents outils de mesure dont nous disposons sont adaptés à certains types d'évaluation, mais pas à d'autres, et de réfléchir sur l'utilisation de ces outils *(évaluer avec quoi ?)*.

Enfin, la nécessité d'évaluer l'enseignement lui-même, de le remettre en question et de le faire évaluer doit amener à chercher de l'information dans les résultats obtenus par les plongeurs aux différentes épreuves ; le dernier chapitre est donc consacré à l'*expression des résultats.*

Ce plan reprend en fait l'ordre chronologique selon lequel toute démarche d'évaluation doit être réfléchie (fig. 1).

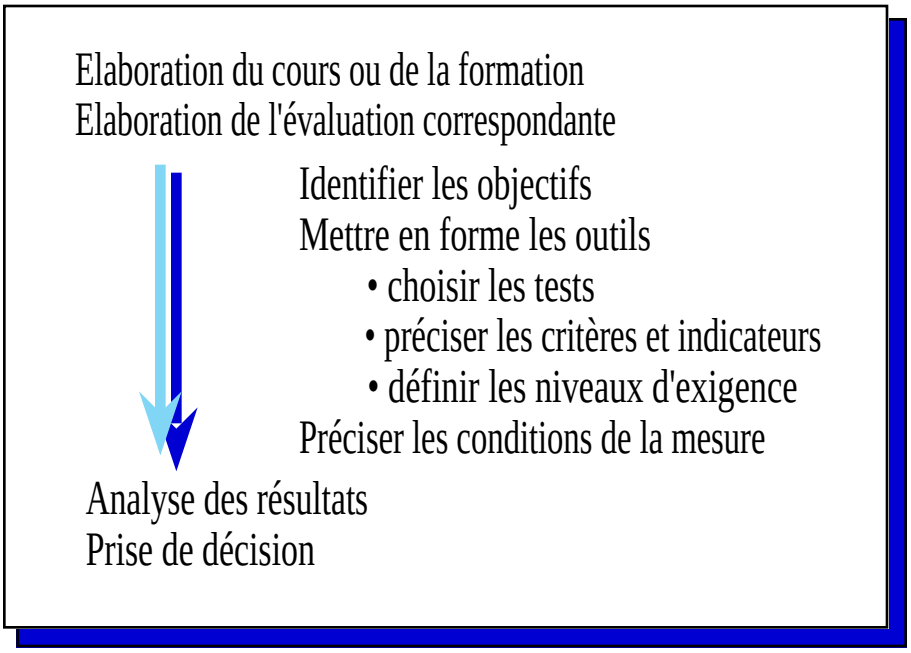


Figure 1 : Les étapes de la démarche d'évaluation. La réflexion doit être menée avant d'enseigner, juste après l'élaboration du cours. Les objectifs ont donc en principe déjà été définis. Par rapport à ces objectifs, on choisit les outils d'évaluation et la manière dont on va traiter les résultats (critères et niveaux d'exigence, conditions de déroulement). L'ensemble du processus a donc bien été fixé *a priori* et annoncé aux élèves.

Evaluer pourquoi ? - Les types d'évaluation

La place de l'évaluation dans le système

Le triangle pédagogique

Dans un système pédagogique, l'évaluation (fig. 2) est le centre de la relation entre :

- l'élève (le stagiaire), qui suit la formation,
- l'enseignant (le moniteur), ou l'équipe pédagogique qui la dispense,
- et l'institution, qui délivre le diplôme (FFESSM), ou qui en fixe les prérogatives (état).

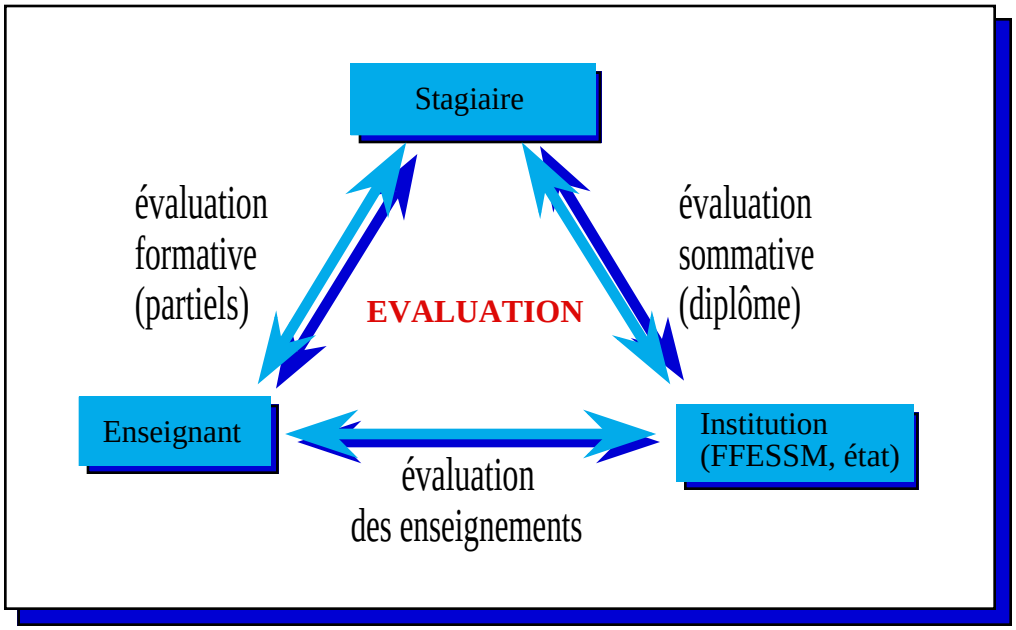


Figure 2 : Le triangle pédagogique et la place de l'évaluation.

En termes cybernétiques, l'évaluation n'est qu'un transfert d'informations entre les différents niveaux d'organisation du système pédagogique. C'est un circuit de régulation indispensable au fonctionnement de l'ensemble :

- entre le moniteur et l'élève, deux flux d'informations coexistent : le transfert de compétence, qu'il s'agisse de connaissances ou de savoir-faire (démonstration), et l'évaluation, appelée alors *formative*, permettant de mesurer la progression vers l'objectif à atteindre. Souvent, l'élève est évalué en début de cursus, afin de savoir s'il est apte à le suivre : c'est l'évaluation *prédictive.*

- entre l'élève ou le moniteur, d'une part, et l'institution, d'autre part, l'évaluation représente le seul transfert significatif d'information. L'élève, s'il satisfait aux critères établis par l'institution, reçoit son diplôme (évaluation

sommative). Le moniteur est, ou devrait être, évalué dans son enseignement (évaluation *prescriptive*), et sa progression est elle-même objet de diplômes (MF2, instructeur régional, instructeur national). Les centres de plongée, ou les équipes pédagogiques, sont de plus en plus évalués, même indirectement sous forme d'audits. C'est une étape nécessaire à l'harmonisation des formations et à la validité d'un brevet par rapport à des prérogatives d'évolution en plongée.

Ce "triangle" pédagogique se transforme même en "tétraèdre" si l'on prend en compte l'activité elle-même, et son évolution avec la société et les avancées techniques (fig. 3).

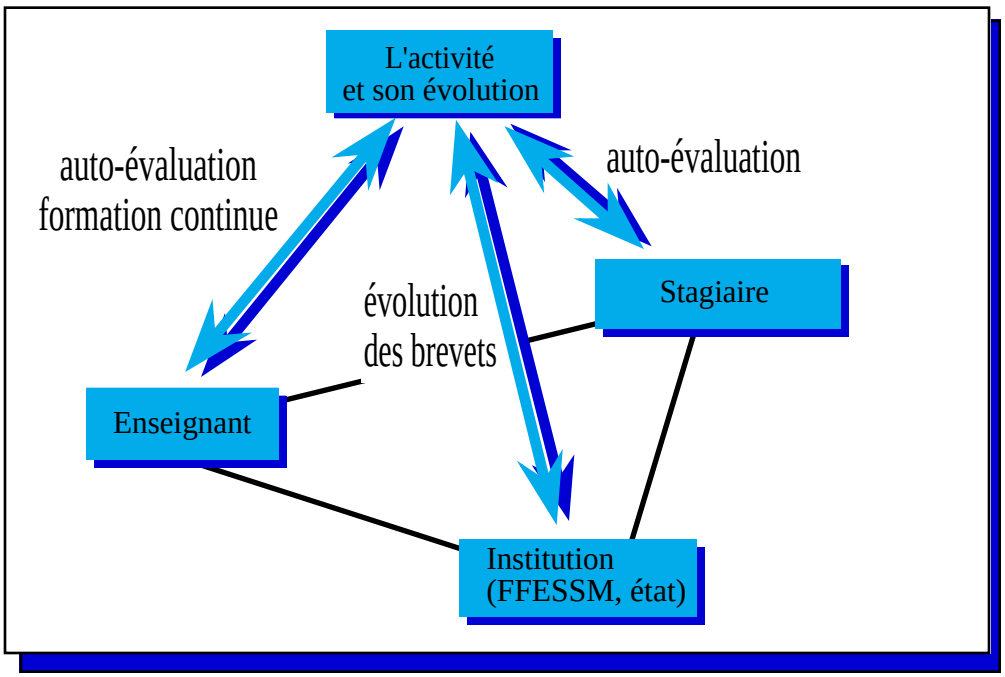


Figure 3 : L'évaluation, flux d'informations vital pour le système dans son ensemble.

L'élève exprime constamment le besoin de s'évaluer par rapport à l'activité (*auto-évaluation*), au moins au cours de sa formation. après obtention d'un brevet, il devrait être incité à se remettre fréquemment en question, pour savoir si avec le temps il n'a pas oublié telle notion ou tel geste technique. Lorsque l'activité évolue, le décalage peut devenir important s'il ne le fait pas.

Cette nécessité est encore plus nette pour le moniteur, qui continue d'enseigner, qui assume la responsabilité de son enseignement et de la sécurité de ses stagiaires. Il est impératif de lui donner des moyens de s'auto-évaluer, et de pallier ses carences par la formation continue. Ce n'est pas par hasard que les Colloques de Moniteurs se développent actuellement, et que la demande grandit.

Enfin, l'institution devrait être soucieuse de mettre en adéquation les brevets qu'elle délivre avec l'activité telle qu'elle évolue, que cette évolution soit souhaitable ou imposée par des faits de société.

L'évaluation est donc bien une des activités-pivot sur laquelle repose tout l'édifice pédagogique, et revêt des formes très différentes selon les niveaux d'organisation concernés. Paradoxalement, les réflexions la concernant sont beaucoup moins fréquentes que celles qui ont trait aux méthodes de transfert de compétence. **Si l'on prend le simple exemple des formations de moniteurs (MF1 ou MF2), quelle est la part de temps consacrée à apprendre à évaluer, par rapport à apprendre à enseigner ?**

L'insertion dans le temps

A mesure que se développe la formation, les types d'évaluation mis en place sont différents (fig. 4).

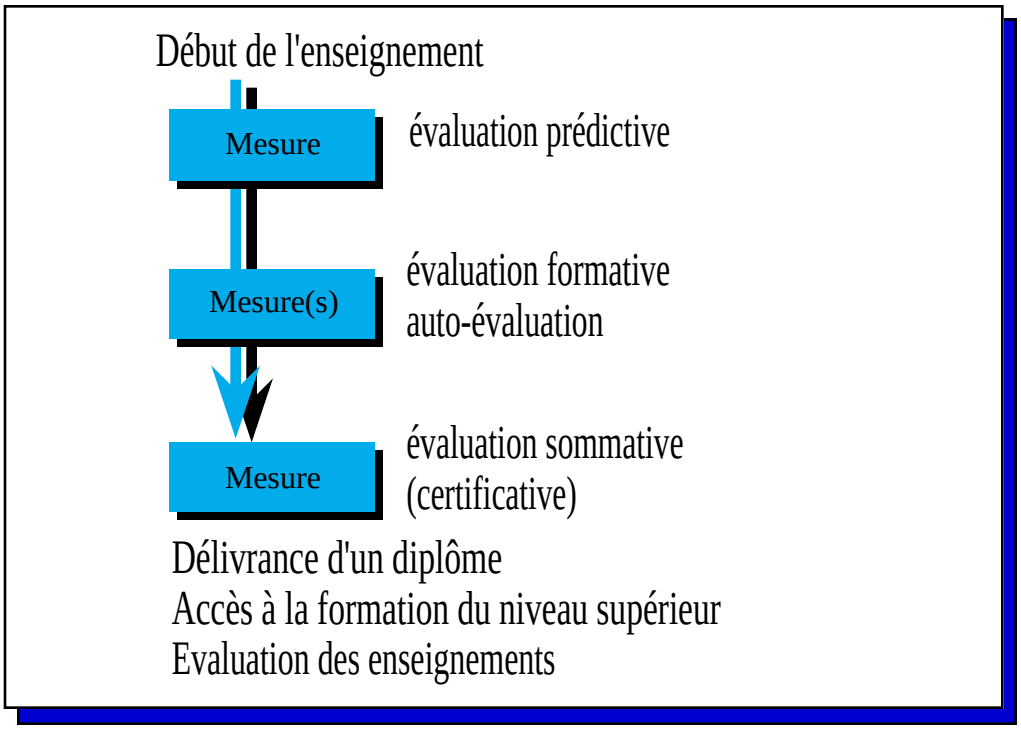


Figure 4 : Les différents types d'évaluation et leur insertion dans le processus de formation.

Les trois types d'évaluation

Il existe donc trois principales formes d'évaluation ; chacune d'entre elles s'insère à un moment particulier du cursus et joue un rôle différent (fig. 5).

	Evaluation Prédictive	*Evaluation Formative*	*Evaluation Sommative*
L'élève	*Eviter d'aller à l'échec*	*Auto-évaluation*	*Faire reconnaître sa compétence*
Le moniteur	*Orienter Constituer des groupes homogènes*	*Adapter son enseignement*	*Evaluer son enseignement*
L'institution	*Eviter les surcoûts*	*Surveiller l'enseignement*	*Délivrer le diplôme Harmoniser les examens*

Figure 5 : Apports des trois types d'évaluation aux trois composantes du système pédagogique.

L'évaluation prédictive

…est-il (suis-je) capable d'atteindre l'objectif ?…

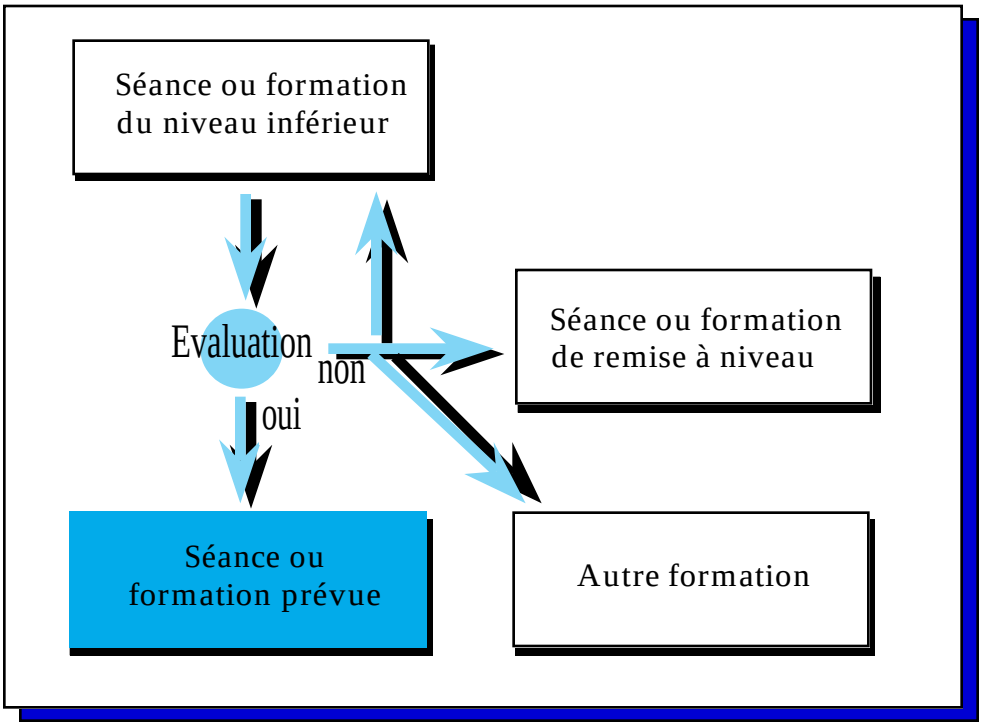


Figure 6 : Domaine d'intervention de l'évaluation prédictive.

L'évaluation prédictive, en début de séance ou de formation, correspond à la vérification des acquis, notion bien connue des moniteurs de plongée. Son importance est cependant souvent négligée. Elle joue pourtant plusieurs rôles intéressants pour tous les partenaires du système. Ses caractéristiques sont les suivantes (fig. 5 et 6) :

- elle peut être *formelle* (l'élève effectue des tests bien définis), ou *informelle* (on le regarde évoluer, on le questionne,…).
- elle a une valeur *diagnostique* : elle identifie les points forts et les lacunes de l'élève.
- elle permet un *pronostic* : l'élève pourra, ou ne pourra pas, suivre le cours ou la formation.
- elle ouvre vers une *orientation* adaptée : compte-tenu du profil observé, telle autre formation serait préférable, ou une remise à niveau est nécessaire.

Pour l'élève…

Identifier ses propres lacunes est un facteur supplémentaire de motivation pour l'élève présentant déjà une motivation forte envers la formation souhaitée. Les autres sont au contraire démotivés, et abandonnent leur prétention à un brevet qu'ils ne peuvent pas ou pas encore décrocher.

L'exemple type est le plongeur niveau 4 souhaitant préparer le MF1 alors même que ses connaissances théoriques ont fortement régressé depuis le brevet…Une évaluation prédictive correctement menée permet d'éviter de le laisser courir à l'échec inévitable.

Pour le moniteur…

L'évaluation prédictive, en permettant la réorientation de tous ceux qui n'ont pas le niveau minimal pour suivre la formation, et le regroupement de ceux dont le niveau est comparable, permet d'espérer la création de groupes d'élèves homogènes.

Le gain d'efficacité pédagogique est alors considérable. C'est évident pour une séance pratique ; pour les cours théoriques, il en va de même, pour pouvoir s'occuper au mieux des auditeurs, en limitant les lenteurs liées aux élèves qui ne peuvent pas suivre.

Pour l'institution…

Dans les Clubs associatifs, le problème du coût d'une formation, dès l'instant qu'il n'est pas supporté entièrement par les élèves, ou qu'on souhaite le limiter, amène souvent à établir un *numerus clausus* pour les formations.

Dans ce cas, seule une évaluation prédictive permet une juste sélection des stagiaires. Les autres systèmes (ancienneté d'inscription, âge, nombre de plongées, voire favoritisme,…) ne sont pas équitables.

Constat actuel

L'évaluation prédictive est la plupart du temps réalisée dans de bonnes conditions pour la formation "dans l'eau". En revanche, elle est beaucoup moins souvent pratiquée pour la formation théorique, sauf peut-être avant les formations au monitorat (MF1, MF2).

L'évaluation formative

…est-il (suis-je) en voie d'atteindre l'objectif ?…

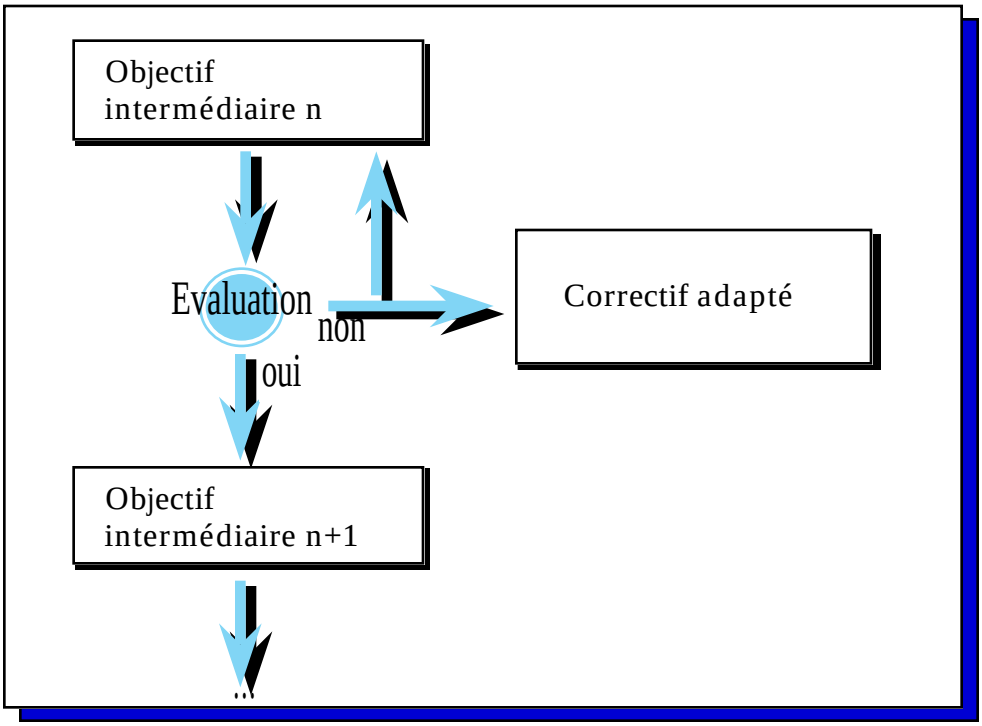


Figure 7 : Domaine d'intervention de l'évaluation formative.

L'évaluation formative est celle qui est pratiquée pendant le processus de formation, et dont les résultats permettent d'améliorer la progression des élèves et de les amener, tous si possible, à l'objectif.

Les caractéristiques de l'évaluation formative sont les suivantes (fig. 5 et 7) :

- elle peut être *informelle* ou *formelle* . Informelle, elle résulte d'une simple d'observation de l'élève par le moniteur, ou d'une discussion. Mais en plongée, elle est très souvent formalisée par le biais de l'évaluation de la réussite à des exercices plus ou moins codifiés.
- elle a une valeur *diagnostique*, différente de l'évaluation prédictive : au cours de la formation, elle identifie les points forts et les lacunes, mais doit permettre de déceler les causes d'un échec éventuel, donc d'y remédier.
- correctement présentée à l'élève, elle est *facteur de motivation.*

- tout comme le feed-back de la communication, elle est indispensable au moniteur pour réguler, *adapter l'enseignement*.

Pour l'élève…

Au sein d'un groupe de niveau, et face à un objectif clairement identifié, l'élève peut mettre en place sa propre évaluation, et suivre, voire prendre en charge lui-même sa progression. Sa motivation s'en trouve alors renforcée.

S'il est laissé dans le flou, au contraire, il risque de se démotiver, faute de pouvoir se situer par rapport aux attentes du moniteur et au reste du groupe.

Ce phénomène est encore plus flagrant pour les épreuves théoriques, pour lesquelles il y a toujours une forte demande d'évaluation, ou d'auto-évaluation, à l'aide d'exercices du même type que ceux qui risquent d'être proposés à l'examen.

Pour le moniteur…

L'évaluation formative est le seul moyen qui permette au moniteur d'estimer la progression de l'élève et d'adapter son enseignement en conséquence. C'est d'ailleurs à ce stade que l'évaluation revêt sa forme la plus complète :

- dans un premier temps, l'élève est évalué, par rapport à l'objectif à atteindre,
- puis, en cas d'insuccès, le moniteur procède à une analyse des causes de cet échec, afin de pouvoir mieux y remédier, par exemple en imaginant un correctif adapté (fig. 7). Ce temps n'existe pas dans les autres formes d'évaluation.

La force de l'évaluation formative, c'est que le moniteur, ou l'équipe pédagogique, disposent d'une très grande liberté dans le choix des outils d'évaluation et des moyens à mettre en oeuvre. Le seul facteur limitant est en fait le temps disponible, et donc le coût.

Pour l'institution…

L'institution intervient peu, en principe, dans le processus d'évaluation formative, qui ne concerne que le moniteur et ses élèves. Elle peut pourtant avoir son rôle à jouer dans le cas où le moniteur est lui-même en formation. Les MF2 ou instructeurs, représentant l'institution (le Club, la CTD ou la CTR), se font alors les garants de l'efficacité de la formation des plongeurs, même si la plupart du temps, ils laissent agir les moniteurs en formation.

Constat actuel

Dans l'eau, l'évaluation formative est conduite, de façon consciente ou intuitive, par la plupart des moniteurs. Ils apprennent vite à détecter l'erreur, et plus difficilement à en analyser les causes et à y remédier. Faute de temps ou de disponibilité souvent, mais aussi parce qu'on ne leur a pas appris à le faire, l'évaluation formative est pratiquée de façon beaucoup plus irrégulière pour la partie théorique et les élèves sont souvent livrés à eux-mêmes.

L'évaluation sommative

…a-t-il atteint l'objectif ?…

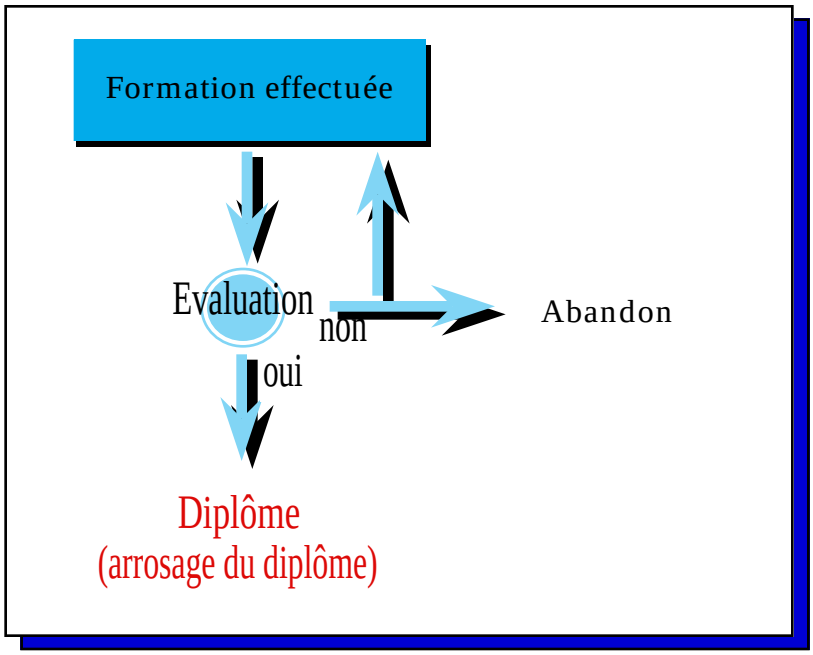


Figure 8 : Domaine d'intervention de l'évaluation sommative.

L'évaluation sommative intervient à la fin du processus de formation (examen), ou de ses phases principales (contrôle continu, brevet obtenu par unités de valeur capitalisables).

Les caractéristiques de l'évaluation sommative sont les suivantes (fig. 5 et 8) :

- elle est toujours *formelle* : ses épreuves sont codifiées à l'avance par le moniteur ou par l'institution.
- elle permet de délivrer ou non le diplôme : elle a valeur *certificative*.
- elle ouvre sur la voie de l'*évaluation des enseignements.*
- elle permet, plus ou moins facilement selon le type d'examen, de tendre vers une *harmonisation* régionale ou nationale des enseignements et des pratiques d'évaluation.

Pour l'élève…

Bien qu'en rapport avec une activité de loisir, sans enjeu aussi important que dans le domaine scolaire ou professionnel, elle est presque toujours génératrice d'angoisse et mérite, rien qu'à ce titre, une attention et un soin particuliers.

En cas de succès, elle permet au plongeur de faire valoir ses prérogatives, de faire reconnaître (de façon d'ailleurs excessive chez certains) leur niveau. Ce type d'évaluation a donc une dimension sociale très nettement exprimée.

Pour le moniteur… L'évaluation sommative devrait inciter le moniteur à remettre en question son enseignement, dans le cas où les résultats de ses élèves seraient décevants (plus de 50 % d'échec est la barre couramment admise).

Pour l'institution… C'est enfin par le biais d'examens réfléchis et travaillés à l'échelle départementale ou régionale que l'on peut tendre vers une harmonisation des enseignements et surtout des exigences lors des épreuves. L'exemple des épreuves théoriques anticipées pour le niveau 4 est riche d'enseignement.

Constat actuel Le constat reste le même que pour les autres types d'évaluation : si l'évaluation dans l'eau a fait l'objet de réflexions nombreuses et évolutives depuis bien longtemps, les épreuves théoriques ont été un peu laissées pour compte, et l'on apprend pas, ou peu, aux futurs moniteurs à les concevoir et à les noter.

Il y a sans doute une raison simple à cela : même les objectifs à atteindre, pour les différents niveaux, n'ont pas encore été clairement définis !

L'évaluation au deuxième degré

L'objet n'est pas ici de développer l'évaluation au deuxième degré, qui concerne l'évaluation d'un enseignement, et l'évaluation des pratiques d'évaluation. On se bornera simplement à effectuer les remarques suivantes :

- cette évaluation est *impérative* : comment penser s'adapter, évoluer, si l'on ne dispose pas de mesures les plus réalistes possibles de notre efficacité pédagogique et de la précision de nos évaluations ? Des tentatives allant dans ce sens sont pourtant très souvent mal perçues par les moniteurs, qui n'aiment pas qu'on remette en question la pertinence de leur jugement.

- elle est *réalisable* : tout ce qui est valable au premier degré est transposable au second. Tous les chapitres de ce fascicule pourraient donc s'y appliquer sans problème. On en donnera quelques exemples succincts dans le chapitre "l'expression des résultats".

En résumé...

1. L'évaluation est un flux d'informations qui circule en permanence au sein du système pédagogique et qui en permet la régulation. Pourtant, les cadres sont peu ou mal préparés à évaluer.

2. Il y a trois principaux types d'évaluation : prédictive (pourra t-il suivre ?), formative (progresse-t-il ?), sommative (a-t-il atteint le niveau ?). En plongée, dans les trois cas, l'évaluation a progressé pour les épreuves dans l'eau, mais reste très imparfaite pour les épreuves théoriques.

3. A l'heure où les coûts d'une formation ne sont plus négligeables, l'évaluation des enseignements devient indispensable.

Evaluer quoi ? - Les objectifs

Les plus graves erreurs en matière d'évaluation ne viennent pas de la méconnaissance des outils et de la façon de s'en servir, mais de son objet même. Paradoxalement, en plongée comme dans de très nombreux autres domaines, l'enseignant est confronté à un problème qu'il surmonte tant bien que mal : il sait ce qu'il doit enseigner, mais pas bien ce que ses élèves doivent savoir, ou savoir faire.

Autrement dit, il existe une confusion ancestrale entre un *programme d'enseignement* et les *objectifs pédagogiques* qui s'y rapportent. L'enseignant manque d'un *cadre de référence* (ou système de références) sur lequel baser son évaluation.

Cela se traduit inconsciemment par des pratiques d'évaluation oscillant entre deux systèmes nettement différents : l'évaluation *normative* et l'évaluation *critériée* (fig. 9) :

- l'évaluation normative est basée sur l'hypothèse que la production de l'ensemble des élèves se distribue selon une courbe en cloche (dite courbe de Gauss ou loi normale). La plus grande partie des élèves présente un résultat moyen, et aux extrémités on trouve peu de très bons et de très mauvais. Autrement dit, dans ce système, on évalue les élèves par rapport à la production moyenne de l'ensemble des autres. *A la limite, ce n'est pas l'enseignant qui définit ce qu'il attend de ses élèves.*

- dans l'évaluation critériée, c'est bien l'écart par rapport à la compétence recherchée qui est mesuré. *L'objectif doit donc avoir été clairement défini par l'enseignant,* ainsi que l'ensemble des critères qui lui permettront de savoir si cet objectif a été atteint. La notion de moyenne n'intervient pas.

En plongée, la recherche d'une compétence en fonction des prérogatives des plongeurs a alimenté, depuis plusieurs années, nos réflexions quant aux objectifs et à leur critères d'évaluation, pour les épreuves dans l'eau. En revanche, les aptitudes théoriques, voire "intellectuelles" qui s'y rapportent restent encore mal définies. Tant que les différents groupes de travail qui se sont attelés à ce travail n'auront pas terminé, le même flou persistera quant à l'évaluation de la théorie.

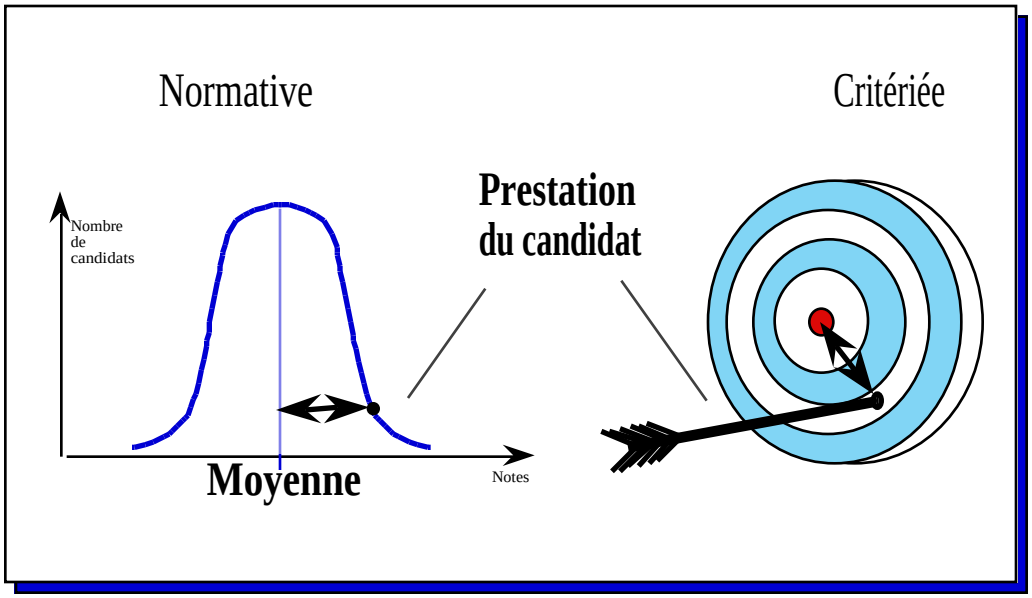


Figure 9 : Les cadres de référence. Dans une évaluation normative, le candidat est évalué par rapport à une norme (la moyenne, le plus souvent) ; dans l'évaluation critériée, la référence est l'objectif à atteindre (la compétence recherchée).

Ce fascicule n'a pas pour but de se substituer à ces groupes de travail, et à une démarche nécessairement collégiale, mais d'attirer l'attention sur trois points principaux :

- il faut établir la liste des objectifs à atteindre pour obtenir chaque niveau de plongée,
- un objectif n'est recevable que dès lors qu'il est parfaitement défini, et en aucun cas un simple programme d'enseignement ne peut suffire,
- nos barèmes et systèmes de notation devraient être remaniés (et non nécessairement abandonnés), dans une optique d'évaluation critériée.

La liste des objectifs

Une méthode : l'analyse descendante

L'analyse descendante est une méthode puissante qui permet de définir les composantes d'une formation, en étant sûr de n'en oublier aucune facette. Comme son nom l'indique, elle est menée en identifiant en premier lieu le niveau final à atteindre[1]. On tente ensuite de préciser les grands domaines de compétence, les capacités qui sont nécessaires. Puis, pour chaque compétence, viennent les objectifs sous-jacents, bien spécifiques, que doit acquérir l'élève depuis son niveau de départ.

[1] Pour simplifier, les termes "officiels" tels que finalité, objectif final, intermédiaire ou opérationnel ont été omis.

La figure 10 schématise les phases de cette analyse. Les trois grandes compétences proposées ne sont qu'indicatives, d'autres découpages peuvent être trouvés.

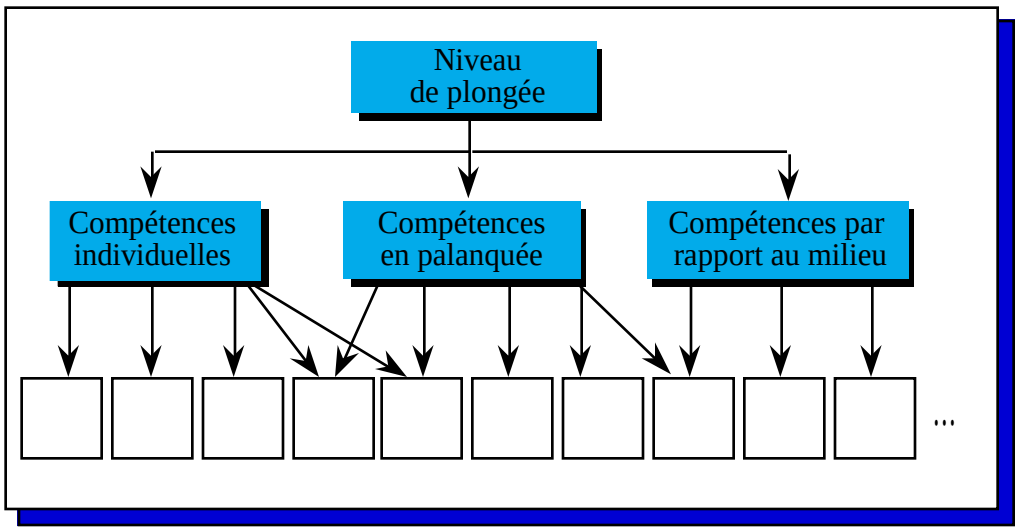


Figure 10 : Méthode de l'analyse descendante permettant de définir les objectifs de formation. En fonction du niveau de plongée clairement identifié, et de ses prérogatives, on définit les grands domaines de compétence à faire acquérir, ce qui permet ensuite de faire la liste des objectifs d'apprentissage nécessaires.

Il faut faire à propos de cette méthode deux remarques :

- le nombre d'objectifs ainsi définis doit être le minimum nécessaire et suffisant. Si l'on retire un objectif, le niveau ne peut pas être atteint. Mais il n'y a pas de découpage excessif, de "saucissonnage" de la formation. C'est la principale erreur à ne pas faire, au risque de se perdre dans un dédale de micro-objectifs ne correspondant plus à rien, en perdant de vue l'objectif final.

- ce travail facilite grandement le processus d'évaluation. On ne risque pas d'être amené à évaluer, dans le flou, une composante qui n'aurait pas été définie. A l'opposé, on vérifie que tous les objectifs identifiés sont atteints, et pas seulement les seules épreuves des brevets.

Les composantes de l'apprentissage

L'analyse est facilitée lorsque l'on sait que tout apprentissage, quel qu'il soit, résulte de la combinaison plus ou moins complète de trois composantes (fig. 11) :

- le "savoir", donc les connaissances et la façon de s'en servir,
- le "savoir-être", c'est-à-dire la personnalité de l'élève, qui retentit sur son comportement,

- et pas seulement le "savoir-faire", la simple réalisation d'un acte moteur.

On remarquera d'ailleurs que les épreuves des brevets prennent déjà en compte ces trois dimensions, puisqu'il y a des épreuves pratiques, alliant des gestes techniques et des comportements (ex. conduite de palanquée), et des épreuves théoriques.

Ce découpage est également une aide précieuse en évaluation formative, dans l'analyse des causes d'échec.

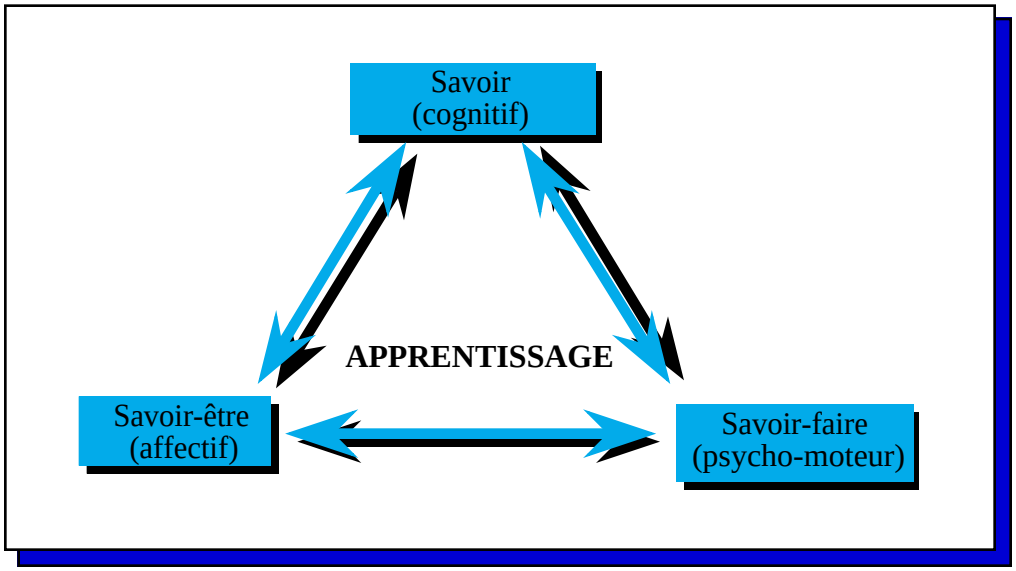


Figure 11 : Les trois composantes de tout apprentissage.

Avec plus ou moins de bonheur selon les niveaux de formation en plongée, les objectifs correspondant aux actes moteurs, aux savoir-faire sont à peu près définis.

Les objectifs correspondant aux savoir-être, à l'influence de l'affectif sur le comportement, n'ont jamais été identifiés. Sans doute la tâche est-elle trop ardue et nos moniteurs fédéraux pas du tout préparés à la psychopédagogie. L'exemple-type de cette distorsion reste l'évaluation du comportement à 40m du candidat au niveau 4. L'épreuve de "Code à 40m" fluctue entre une stricte application du texte et la notation d'une réponse à quelques signes, souvent stéréotypés, et l'évaluation d'un véritable comportement de guide de palanquée, impliquant un comportement responsable, vigilance et prévention tout autant qu'efficacité dans l'intervention.

Qu'en est-il de la définition des objectifs ayant trait aux connaissances théoriques ?

Les objectifs du domaine cognitif

En matière de théorie de la plongée, l'état des lieux n'est pas reluisant. Les contenus ne sont pas clairement définis pour chacun des niveaux de plongée. Seules quelques têtes de chapitre sont mentionnées, ouvrant la voie à de larges interprétations (exemples les plus frappants : physiologie, matériel). Par conséquent, les objectifs d'apprentissage n'ont jamais été clairement identifiés, de façon harmonisée au plan national.

Faute des bases de l'analyse, une grande majorité de moniteurs se réfugient, s'ils en ont eux-mêmes les compétences, dans l'apport pléthorique d'informations : ils font du "surniveau". Ils veulent "bien faire", dans la carence où ils sont d'objectifs pédagogiques clairement identifiés. Ils transposent ainsi vers l'élève la responsabilité du tri des connaissances et de leur application dans le cadre du brevet préparé.

Quant à ceux qui n'ont pas - ou plus - eux-mêmes le niveau, l'efficacité de leur enseignement est aléatoire…

Selon Bloom (1964), les objectifs permettant de définir ce que l'on attend d'une performance "intellectuelle" peuvent être classés en six niveaux de difficulté croissante (fig. 12).

Connaissance	Restituer un contenu.
Compréhension	Résoudre un problème, toutes les informations nécessaires étant données dans l'exposé.
Application	Résoudre un problème, en apportant ses propres connaissances et ses informations.
Analyse	Examiner une situation selon ses connaissances et des critères explicites.
Synthèse	Assembler les éléments de l'analyse de façon à comparer les différentes solutions possibles. Conclure.
Evaluation	Utiliser ses facultés d'analyse et de synthèse au recueil d'informations et à la mesure.

Figure 12 : La hiérarchie des objectifs cognitifs, classés par ordre de complexité croissante.

Cette classification peut, dans le contexte de la plongée, être simplifiée. Si l'on admet que *compréhension* et *application* n'ont pas lieu d'être dissociés, et si l'on se réfère aux besoins de formation théorique en fonction du brevet, on aboutit à une réduction à deux ou trois grands types d'objectifs (fig. 13 et 14).

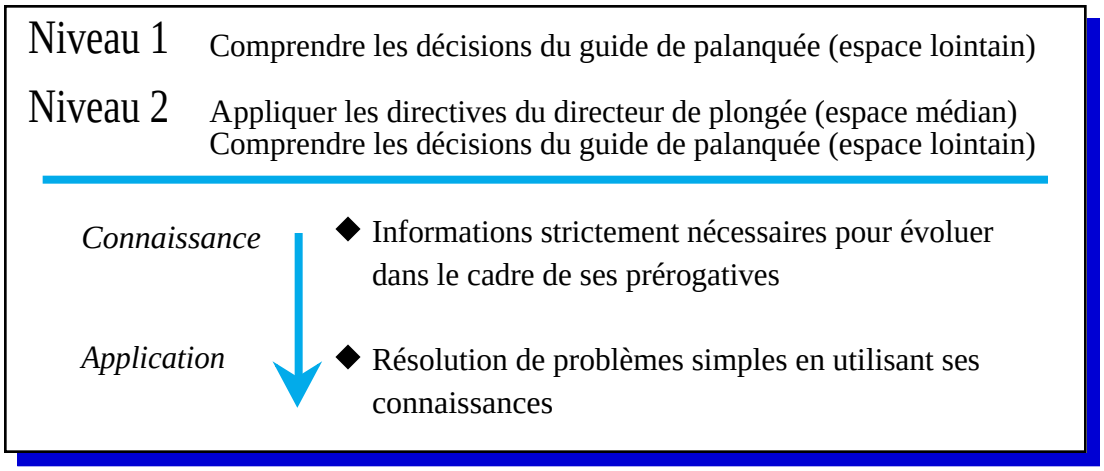


Figure 13 : Les besoins en théorie des niveaux 1 et 2 et les deux grands types d'objectifs correspondants.

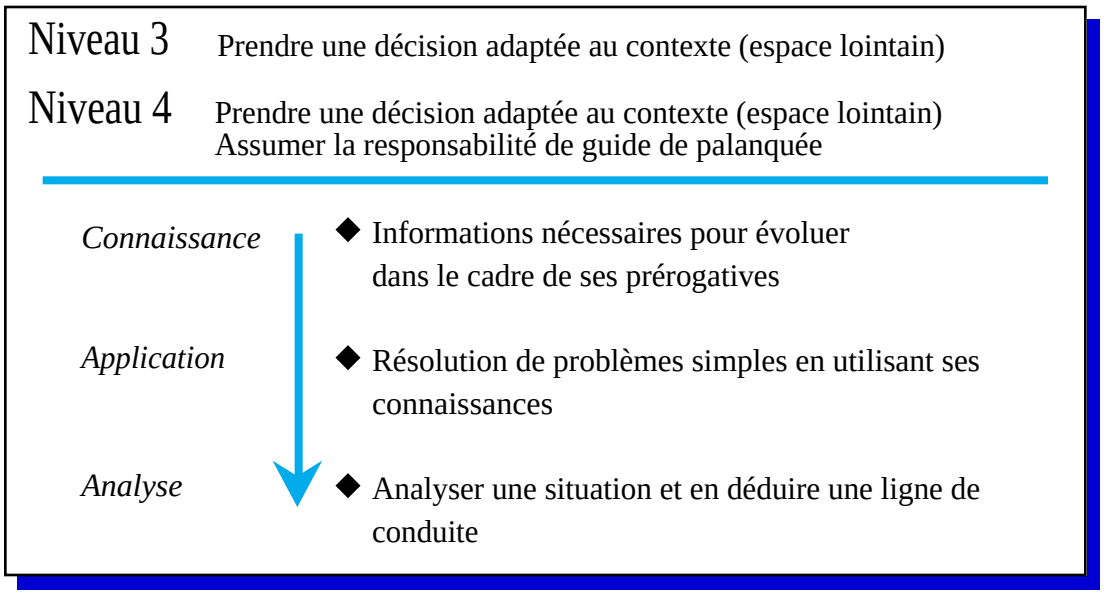


Figure 14 : Les besoins en théorie des niveaux 3 et 4 et les trois grands types d'objectifs correspondants.

Cette hiérarchisation des objectifs en fonction des niveaux ne doit pas rester une simple construction intellectuelle. Trop souvent réduit à la simple vérification de connaissances peu ou mal digérées, l'examen théorique n'a que peu de valeur s'il néglige la prise en compte de ces niveaux d'aptitude.

Exemple 1, concernant les tables de plongée. Pour le Niveau 1, évoluant encadré, quelques *connaissances* suffisent : la courbe de sécurité. Le Niveau 2 a besoin d'en *connaître* un peu plus sur la décompression (pourquoi des tables ?), mais il a surtout besoin d'*appliquer* ces connaissances, donc de résoudre des exercices simples sur l'utilisation des tables. Pour le Niveau 4, guide de palanquée, on sera plus exigeant sur les *connaissances*, sur leur *application* à la manipulation des tables, mais on y ajoutera l'évaluation de ses facultés d'*analyse*, face à des situations plus complexes, plus "bâtardes". Face à ce genre de situation non nécessairement prévue à l'avance, que décide-t-il ?

Exemple 2 : accidents de décompression. Les Niveaux 1 et 2 n'ont besoin que de quelques *connaissances* de base, à des degrés différents, concernant surtout les causes et la prévention. Le Niveau 4 doit faire la preuve de *connaissances* plus

poussées, incluant les mécanismes, les symptômes et la conduite à tenir. Mais il doit surtout être capable d'*analyser* une situation, de faire le diagnostic, et de décider de la conduite à tenir. Lui faire simplement réciter des symptômes, c'est tester ses connaissances. Lui faire poser un diagnostic, par le biais d'un problème ou d'un cas concret, c'est évaluer ses facultés d'analyse. Le type d'outil d'évaluation utilisé dépend donc bien de l'identification de ces niveaux d'objectifs.

La construction de l'évaluation

Les spécialistes en sciences de l'éducation recommandent d'avoir une idée précise, dans une évaluation, du nombre de questions à réserver respectivement à chacun des niveaux ainsi définis. On obtient ainsi un tableau à double entrée, dit *tableau de spécification*, permettant de cerner l'ensemble des questions sur lesquelles portera l'examen, de ne pas se limiter simplement au niveau le plus bas ("connaissance"), et de préparer le barème de notation (fig. 15 et 16).

Surpression pulmonaire Niveau 2	*Objectif 1 (mécanismes et causes)*	*Objectif 2 (symptômes et conduite à tenir)*	*Objectif 3 (prévention)*
Connaissance	30	20	30
Application	10	_	10

Figure 15 : Exemple de tableau de spécification pour une évaluation portant sur la surpression pulmonaire au Niveau 2. Les chiffres représentent, en pourcentage, le poids respectif des différents objectifs.

Comparons les figures 15 et 16, par exemple à la ligne 1, colonne 1 (connaissance des mécanismes et causes de la surpression pulmonaire) : dans les deux cas, l'élève doit acquérir un certain nombre de connaissances. Est-ce pour autant que le niveau d'exigences, que la somme d'informations mémorisées, seront les mêmes au niveau 2 et au niveau 4 ? Il est clair pour tous que la réponse est non.

Surpression pulmonaire Niveau 4	*Objectif 1 (mécanismes et causes)*	*Objectif 2 (symptômes et conduite à tenir)*	*Objectif 3 (prévention)*
Connaissance	15	15	15
Application	_	_	15
Analyse	_	25	15

Figure 16 : Exemple de tableau de spécification pour une évaluation portant sur la surpression pulmonaire au Niveau 4. Les chiffres représentent, en pourcentage, le poids respectif des différents objectifs.

> Cela signifie donc qu'en aucun cas, une tête de chapitre (ex. "mécanismes et causes"), ne peut constituer un objectif suffisamment défini pour permettre une évaluation univoque et harmonisée entre moniteurs.

De la même façon, dans la colonne 3 (prévention), la formation du niveau 2 sera axée sur la prévention individuelle, alors que le niveau 4 sera préparé à intervenir sur l'ensemble de la palanquée dont il est responsable.

Comment définir un objectif ?

Il est donc nécessaire de définir avec précision chaque objectif, faute de quoi les exigences de l'évaluation peuvent ne pas correspondre au niveau préparé : celle-ci ne sera pas **pertinente**.

Il n'y a donc pas d'évaluation possible sans définition complète de l'objectif à atteindre. Depuis les débuts de l'utilisation des objectifs dans l'apprentissage, que l'on peut attribuer à Ralph Tyler (1934), un objectif n'est utilisable, n'est "opérationnel", que s'il revêt trois aspects :

- la référence à un **comportement** observable, clairement et précisément défini,
- la référence aux **conditions** dans lesquelles ce comportement est attendu,
- la référence à l'**évaluation**, c'est à dire aux critères qui permettront de savoir si l'objectif est atteint.

Dans la pratique, il est commode de retenir que la définition d'un objectif se décompose en 6 points (fig. 17).

Les quatre premiers cernent l'objectif lui-même :

- à quelle phase de la formation doit-il être atteint ?
- qui est concerné ?
- et surtout quel comportement, exprimé en des termes précis, directement évaluable, est attendu ?
- dans quelles conditions (de lieu, de temps, de météo,…) ?

Les deux autres éléments concernent directement les conditions dans lesquelles se fera l'évaluation :

- quels sont les critères d'évaluation, quel est le niveau d'exigence, de "perfection", souhaité ?
- combien de fois l'objectif doit-il être réalisé pour que l'apprentissage puisse être considéré comme ancré ?

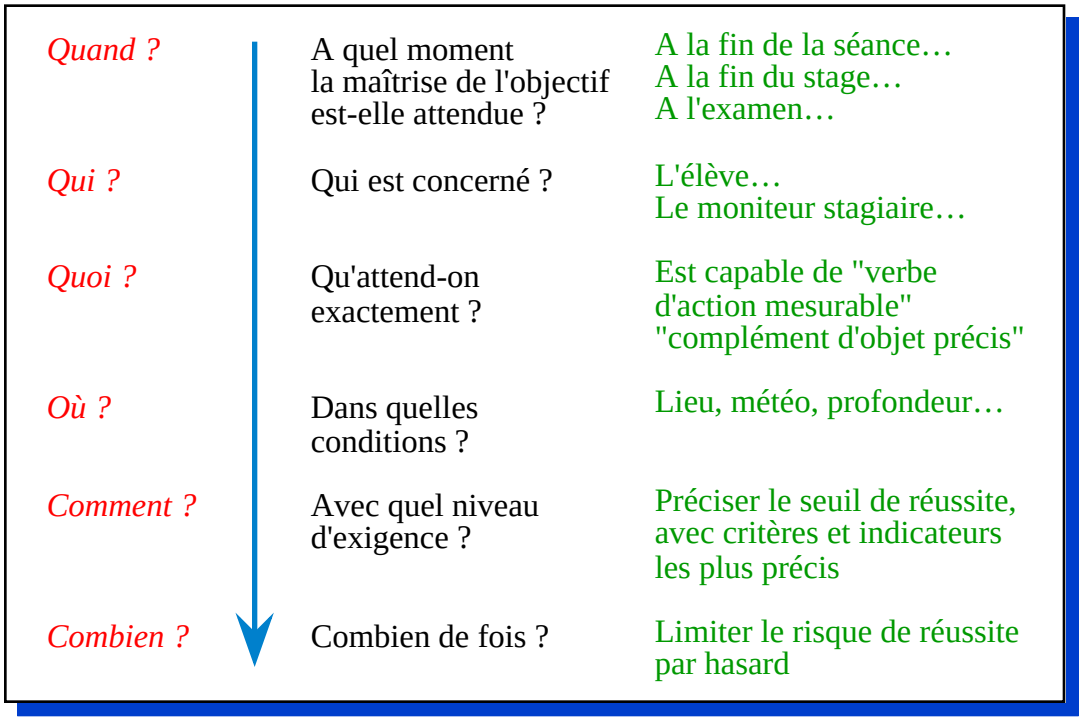


Figure 17 : Les 6 points de la définition d'un objectif.

Deux exemples, l'un concernant la pratique, l'autre la théorie, permettront de bien comprendre l'intérêt de ces précisions et l'avantage qu'on en tire en terme d'évaluation.

Exemple 1 Objectif concernant l'utilisation individuelle du gilet de stabilisation dans le cadre de la préparation au niveau 2 :

"A l'issue de la première moitié de la formation, le stagiaire est capable de remonter individuellement, à vitesse normale, d'une profondeur de 20 m, à l'aide de son gilet stabilisateur. L'objectif sera atteint si trois remontées au moins ont été effectuées dans les conditions suivantes : après une stabilisation rapide (<20 s), il est remonté régulièrement à une vitesse comprise entre 12 et 17 m/min. Un ou deux arrêts seront tolérés, mais il ne doit jamais redescendre. Le réglage de la vitesse se

fait aux repères naturels et non aux instruments. L'usage accessoire des palmes pendant des périodes de moins de 10 s est toléré. L'arrêt à 3 m est impératif et marque la fin de l'exercice."

Exemple 2 Objectif concernant l'intervention courante sur problème de matériel dans le cadre de la préparation au niveau 4 :

" A l'examen, le candidat est capable de trouver l'origine d'une panne concernant une pièce souple (joint, siège) sur un 1er étage de détendeur non compensé, et d'y remédier en moins de 30 minutes. Le détendeur est d'un modèle très courant, type MKII Scubapro, et aucune pièce n'est grippée. Outils et pièces de rechange sont disponibles."

Il est important que les élèves aient connaissance de la liste des objectifs à atteindre pour le brevet convoité, et qu'ils en comprennent la justification. Cette acceptation de l'objectif est un facteur de motivation primordial.

L'influence des systèmes de notation

Les principes de la notation des épreuves, et surtout des questions composant une épreuve, seront envisagés plus loin (cf. § "L'expression des résultats"). On se bornera ici à montrer que, même si les objectifs ont été définis, un système de notation inapproprié, mal évalué, peut entraîner une distorsion telle que l'évaluation perd également sa pertinence.

Le risque du système des moyennes

Le système de notation classique, utilisé en plongée comme ailleurs, fait appel à une échelle de notes de 0 à 20. Le brevet est acquis lorsque le candidat a obtenu la moyenne (10) sur l'ensemble des épreuves ou sur chaque groupe d'épreuves. Cette moyenne est souvent pondérée par des coefficients. Enfin, le plus souvent, une note inférieure à 5 est éliminatoire.

Ce système de notation présente le risque d'entretenir la confusion entre évaluation normative et critériée. La C.T.N. en a pris conscience il y a quelques années, puisque dans certaines épreuves, le seuil d'élimination a été ramené à toute note inférieure à 10. Car il y a deux façons d'utiliser une telle échelle.

Il suffit de considérer la note maximale, soit la note de 20 : *cette note correspond-elle à la prestation totalement parfaite, ou montre-t-elle simplement que l'objectif est atteint ?* Dans les épreuves théoriques rédactionnelles, aussi bien que dans l'eau, cette note est très rarement attribuée : le moniteur, inconsciemment, procède de l'évaluation normative.

- Si effectivement la note de 20 correspond à une prestation très au-delà de l'objectif à atteindre, alors la moyenne (10) est attribuée lorsque l'objectif est tout juste atteint. Dans ce

cas pourtant, la note éliminatoire (objectif non atteint) devrait être systématiquement à 9 et non à 4.

- Si la note de 20 est délivrée lorsque l'objectif est juste atteint, l'échelle de 21 points ne sert à rien et il n'y a pas de moyen de différencier les bons des très bons.

L'inadéquation de l'échelle 0-20 pour les brevets de plongée est d'ailleurs perçue, même inconsciemment, par de nombreux moniteurs en jury. Il n'est qu'à écouter les délibérations à propos de la délivrance de notes éliminatoires. On s'aperçoit bien que de nombreux jurys raisonnent désormais en termes de profil global du candidat (évaluation critériée). Malgré tout, au Niveau 4 et au MF1 notamment, le jeu des moyennes et des coefficients laisse encore filtrer de nombreux candidats qui n'ont visiblement pas atteint un ou plusieurs objectifs importants au regard de leurs prérogatives futures.

Vers la définition d'un "Niveau Acceptable de Performance"...

Le malaise est donc issu de l'utilisation de l'échelle 0 - 20, traditionnelle en France. Un tel système ne pose aucun problème dans les formations de type éveil, ou culture générale (études primaires et secondaires), dans lesquelles l'évaluation normative reste utilisable. Face à un ensemble d'épreuves, le fait d'atteindre 100% des objectifs n'est pas prépondérant. On s'inscrit dans la durée (plusieurs années). Culture d'ensemble, imprégnation, éveil, , voire "teinture" peuvent suffire, et il est intéressant que les élèves puissent se situer par rapport au niveau moyen de la classe.

Dans les formations spécifiques, impliquant l'acquisition d'une compétence en rapport avec des prérogatives, il en va tout autrement. Seule l'évaluation critériée permet de mesurer si l'ensemble des objectifs est atteint ou non. Dans ce cadre, l'échelle 0 - 20 est malcommode et elle entretient la confusion.

Si un des objectifs n'est pas atteint, le niveau ne l'est pas non plus : toute note comprise entre 0 et 9 est superflue. Lorsque l'objectif est atteint, il peut être utile de faire la différence entre les "passables", les "bons" et les "très bons". En ce cas, il n'est pas nécessaire de disposer de 11 degrés (notes de 10 à 20).

Notation d'une épreuve

On peut donc proposer, à l'instar de certains pays (ex. : Canada, Etats-Unis), une notation à quatre degrés, notés de 0 à 3 ou de D à A[1] :

- 0 ou D : objectif non atteint, éliminatoire,
- 1 ou C : niveau acceptable de performance atteint,

[1] Attention, cette notation n'a rien à voir avec l'échelle ordinale A - E proposée en France dans les années 1970, où E correspondait à la plage 0-3, D à 4-7, C à 8-11, B à 12-15 et A à 16-20. Superposé à l'échelle traditionnelle, ce système n'avait donc pas changé grand-chose sur le fond.

- 2 ou B : bien,
- 3 ou A : très bien.

Il suffit d'un 0 (ou D) dans l'une des épreuves pour que le candidat ne soit pas admis. Lorsqu'au contraire, il a atteint partout le niveau acceptable de performance (N.A.P.), il est reçu.

Il est possible d'ailleurs d'utiliser un tel système, malgré la persistance de l'échelle 0-20 : il suffit de n'utiliser systématiquement que quatre notes : 0, 10, 12 et 16 par exemple (il suffit de se mettre d'accord en réunion préalable du jury).

Moyenne d'une série d'épreuves

On remarque que dans un tel système, la performance minimale à atteindre (la note 1) n'est pas située à la moyenne de l'échelle des notes (0 à 3). Dans une série d'épreuves, évaluer la performance globale en calculant la moyenne arithmétique des notes ne signifie donc rien.

La prestation globale peut alors être évaluée en utilisant comme indice la **médiane** des notes obtenues, c'est à dire la note séparant l'ensemble des notes en deux parties égales par le nombre.

Soit une série de neuf épreuves auxquelles l'élève obtient :
1 3 2 2 1 1 1 2 3
soit quatre fois la note 1 (passable), trois fois la note 2, et deux fois 3.

Les notes, ordonnées dans l'ordre croissant, donnent :
1 1 1 1 2 2 2 3 3

Puisqu'il y a neuf épreuves, la médiane sera 2 (quatre autres notes de chaque côté).

Pour mieux situer l'élève dans sa marge de progression, on peut aussi utiliser la *moyenne harmonique*, employée dans les démarches qualitatives. Cette moyenne se rapproche plus des valeurs basses que des valeurs hautes, et son calcul est impossible si une des épreuves est notée 0 (l'élève est éliminé).

Pour un échantillon de n mesures individuelles notées x_i (x_1 à x_n) :

Moyenne harmonique : $\overline{H} = \dfrac{n}{\sum \frac{1}{x_i}}$

Moyenne arithmétique : $\overline{M} = \dfrac{\sum x_i}{n}$

Soit une série de huit épreuves auxquelles l'élève obtient :
1 3 2 2 1 1 1 2
soit quatre fois la note 1 (passable), trois fois la note 2, et une fois 3.

$\overline{H} = \dfrac{8}{(4 \,.\, 1)+(3 \,.\, 0{,}5) + 0{,}333} = 1{,}37$ $\qquad \overline{M} = \dfrac{(4 \,.\, 1)+(3 \,.\, 2) + 3}{8} = 1{,}63$

Sur la même série d'épreuves, un autre élève obtient :
3 3 2 2 3 3 1 2
soit quatre fois la note 3 (très bien), trois fois la note 2, et une fois 1.

$\overline{H} = \dfrac{8}{(4 \,.\, 0{,}333)+(3 \,.\, 0{,}5) + 1} = 2{,}09$ $\qquad \overline{M} = \dfrac{(4 \,.\, 3)+(3 \,.\, 2) + 1}{8} = 2{,}38$

Unités capitalisables ou examen ponctuel ?

En évaluation sommative, pour délivrer le brevet, se pose immanquablement la question : faut-il diviser le brevet en groupes d'épreuves, qu'il est possible d'acquérir progressivement, ou bien faut-il organiser un examen ponctuel, dans lequel l'ensemble des épreuves seront organisées ?

En fait, cette question échappe presque à la problématique de l'évaluation. D'autres considérations entrent en ligne de compte. Elles sont d'ordre économique (coût des formations et des examens), pratique (sites, bateaux, disponibilité en moniteurs et instructeurs), humain (risque de dérive par délivrance d'un groupe "à l'usure", ou par complaisance), ou encore institutionnel (surveillance des conditions de délivrance de certains brevets, etc…).

Ces aspects ne seront pas abordés ici. On se contentera de trois remarques :

- c'est dans l'optique d'une évaluation critériée que le système des Unités Capitalisables (ou Unités de Valeur) a été mis en place : dès que l'élève a atteint les objectifs correspondant à un domaine de compétence, on lui valide cette compétence[1]. On remarquera, d'ailleurs, que pour les Niveaux 1, 2 et 3 de plongée, pour lesquels des groupes d'épreuve ont été créés, de nombreux moniteurs ne s'en tiennent pas à la lettre aux épreuves des brevets, mais ont mis en place une évaluation des compétences.

- mais le principe des groupes d'épreuve est risqué, au contraire, lorsque les objectifs de formation ne sont pas clairement définis et acceptés par tous. Trop de dérives pourraient en découler.

- groupes d'épreuves ne signifie en aucun cas déconnexion entre les épreuves théoriques et les épreuves pratiques. Chaque domaine de compétences fait appel à la fois à des savoir, des savoir-faire et des savoir-être...

[1] Attention, cette notion n'a rien à voir avec le *contrôle continu*, dans lequel l'élève est évalué plusieurs fois, de façon répétée, au cours de sa progression dans un même domaine de compétences.

En résumé…

1. Les épreuves théoriques font partie d'un tout. C'est une erreur de trop les déconnecter de l'objectif final de formation. Vers une remise en question des examens théoriques anticipés ?..

2. Il n'y a donc aucune raison de ne pas apporter autant de soin à la définition des objectifs du domaine cognitif.

3. Il ne faut pas se limiter à la simple restitution de connaissances. Selon les domaines et les niveaux préparés, des objectifs d'application et d'analyse doivent être définis. Même des savoir-faire sont à évaluer (cas des épreuves de matériel).

4. Il est préférable de définir, avant de rédiger les épreuves, la répartition des questions, en pourcentage, en fonction des différents types d'objectifs à évaluer.

5. Tous les objectifs doivent avoir été définis avec soin et précision, faute de quoi les élèves ne pourront pas être évalués dans de bonnes conditions. Les élèves doivent connaître ces objectifs et, de préférence, les accepter et se les approprier.

6. L'échelle de notation traditionnelle 0-20 ne s'applique pas bien à une pédagogie de maîtrise basée sur des objectifs à atteindre clairement définis et acceptés. Il faut lui préférer une échelle critérielle à quatre degrés : 0 (objectif non atteint), 1 (objectif juste atteint), 2 (bien) et 3 (très bien).

Evaluer avec quoi ? - Les outils

La correspondance objectifs-outils

Il est classique d'affirmer qu'il existe une correspondance entre le type d'objectifs à évaluer et l'outil d'évaluation qui permet de le faire. Ainsi, les questions plutôt fermées (type Q.C.M. [on reviendra plus loin sur ce sigle]) correspondraient mieux à l'évaluation d'objectifs "de bas niveau", comme la restitution de connaissances, tandis que les questions les plus ouvertes (comme les sujets de synthèse) permettraient d'évaluer des objectifs plus complexes (fig. 18).

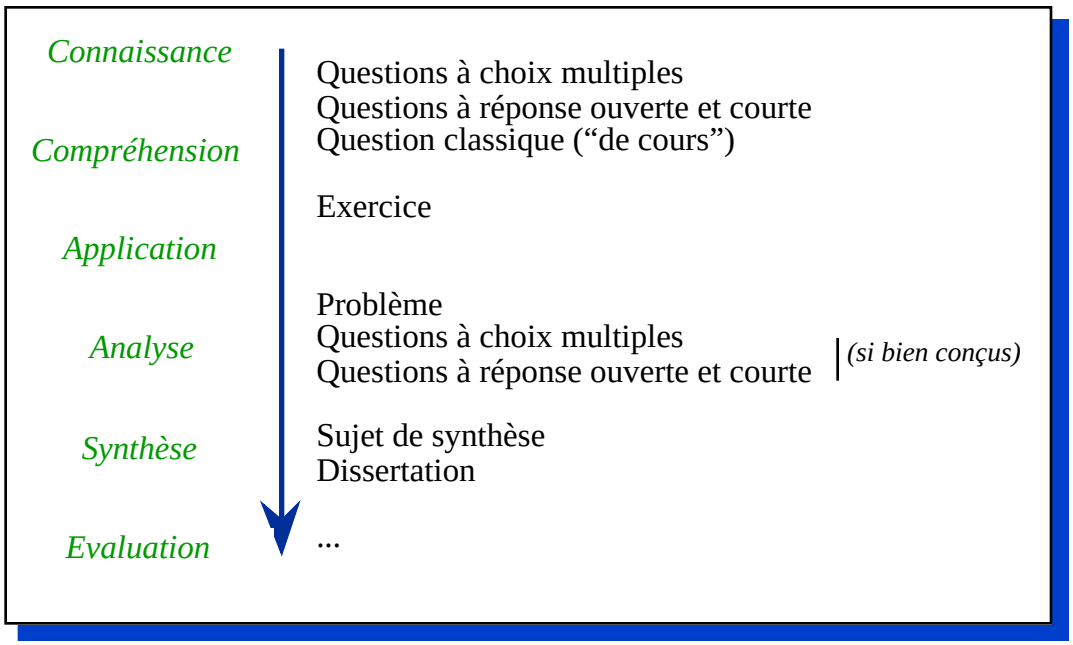


Figure 18 : La correspondance objectifs-outils d'évaluation. Schématiquement, les questions fermées servent plutôt l'évaluation des objectifs de bas niveau, tandis que les outils ouverts correspondent mieux aux objectifs de haut niveau. En fait, tout dépend surtout de la manière dont ils sont utilisés.

En fait, cette notion n'a qu'une importance mineure dans notre activité, pour trois raisons :

- nous avons vu que, finalement, l'apprentissage de la théorie de la plongée ne fait intervenir que deux, voire trois niveaux d'objectifs. Les sujets de synthèse, les dissertations, les sujets faisant appel à l'imagination ou à la création n'entrent pas dans le cadre de notre activité.

- ces objectifs ne sont, à l'heure actuelle, toujours pas clairement définis.

- la manière d'utiliser les outils d'évaluation compte beaucoup plus que les outils eux-mêmes. Nos cadres ne sont pas du tout formés à l'utilisation des outils d'évaluation ; les erreurs dans la conception et la rédaction des sujets d'examens sont donc finalement plus fréquentes et plus gênantes que dans la correspondance entre les objectifs et les outils.

Les limites de l'évaluation

L'évaluation n'est qu'une forme de mesure, ou d'appréciation, pour laquelle on utilise un ou des instruments. Or, toute mesure est obligatoirement entachée d'une erreur plus ou moins grande, qui tient :

- à l'instrument de mesure lui-même, ou à la façon dont il est utilisé (limites intrinsèques),
- au contexte de cette utilisation (limites extrinsèques).

Avant d'étudier ces limites à l'évaluation, il faut se poser une question préliminaire : que risque-t-on lorsque l'on évalue ?

Les risques d'évaluer

Reprenons la définition de l'évaluation (p. 6) :

L'évaluation est un jugement *outillé* en vue d'une prise de décision *éclairée* correspondant à des *buts* fixés *a priori*..

Imaginons que l'on porte un **jugement**, en n'étant pas sûr de l'**outil**, ou de l'utiliser de façon **éclairée**, ou bien que l'on n'ait pas identifié clairement le **but** poursuivi, l'objectif à atteindre ?

Première possibilité

La mesure effectuée est proche de la compétence réelle du candidat. C'est un coup de chance, mais c'est possible. Tout va pour le mieux.

Deuxième possibilité

Le jugement aboutit à reconnaître une compétence qui, en réalité, n'existe pas. Ce risque est probablement le plus grave, étant donné l'évolution actuelle des textes réglementaires : il équivaut à donner à un plongeur des prérogatives qu'il n'est pas capable d'assumer... On pourrait assimiler cette erreur au risque de *première espèce*, en statistiques, qui consiste à affirmer qu'un résultat est significatif alors qu'il ne l'est pas.

Les conséquences sont lourdes, à la fois pour le candidat, content de son résultat, mais que l'on laisse évoluer au-delà de ses moyens, et pour le cadre (ou le jury), qui pourrait (pourquoi pas ?) être poursuivi pour "mise en danger"...

Troisième possibilité

Le jugement aboutit à refuser une compétence qui, en réalité, est acquise. Ce risque a moins de conséquences pour le moniteur, mais pas pour le candidat ! On peut assimiler cette erreur au risque de *deuxième espèce*, en statistiques, qui consiste à affirmer qu'un résultat n'est pas significatif alors qu'il l'est.

Les limites extrinsèques

Les limites extrinsèques de l'évaluation sont les facteurs qui échappent au moniteur ou au jury, obligés de "faire avec", et qui peuvent entraîner une erreur de mesure.

Parmi celles-ci, les contraintes de terrain sont les plus importantes. Même avec des objectifs clairement définis, des outils bien affûtés, on se trouve très souvent dans la situation où l'on est amené à évaluer dans des conditions éloignées de ce que l'on aurait souhaité. C'est le problème classique de l'adéquation fonctionnalisme-opportunisme, du meilleur compromis entre idéal d'évaluation et contraintes de terrain.

Exemple concernant la pratique : l'épreuve dite "code à 40m" du niveau 4. Souvent, pour gagner du temps, parce qu'il y a peu de moniteurs dans le jury, selon le site,… ou pour cause de météo, on est amenés à descendre dans des conditions très variables. L'évaluation se fait en quelques minutes, sur la base d'une simulation de narcose assez stéréotypée et d'un signe amenant à un échange d'embout. Est-ce ainsi que l'on peut vraiment évaluer l'aptitude d'un Niveau 4 à "assurer" à 40 mètres ? Mais il y a malheureusement peu d'alternatives...

Exemple concernant la théorie : le moniteur souhaiterait une évaluation par oral, mais il y a 42 candidats et un créneau de deux heures maximum...

La préparation des élèves est un autre élément risquant de fausser l'appréciation. Le "bachotage" les amène parfois à réussir certains types de tests alors que la compétence réelle n'y est pas.

Les limites intrinsèques

Un outil d'évaluation idéal doit posséder quatre qualités cardinales, sans lesquelles le risque d'erreur est largement augmenté : **pertinence, validité, sensibilité, fidélité**.

Pertinence

La **pertinence** de l'évaluation a déjà été étudiée à propos des objectifs (p. 26). Elle se rapporte à l'objet même de l'évaluation. Une épreuve d'examen n'est pertinente que si elle correspond exactement aux objectifs d'apprentissage tels qu'ils ont été définis. Savoir si son évaluation est pertinente, c'est répondre à la question : la compétence recherchée et les objectifs correspondants sont-ils clairement identifiés ?

Exemple "extrême" d'absence de pertinence : évaluer un candidat au Niveau 1 en testant son aptitude à répondre au signe "je suis essoufflé", en piscine, à genoux sur un fond de 3 m, en faisant le signe tout en respirant le plus calmement du monde...

Si l'on admet que le Niveau 1 n'a pas à intervenir sur le signe "je suis essoufflé", mais simplement savoir le faire, cette évaluation n'est pas pertinente.

Validité

La **validité** (ou justesse) est le degré de correspondance entre ce qui doit être évalué (la compétence, les objectifs), et ce qui est effectivement évalué (la performance à un test, à une épreuve).

Ce que je mesure correspond-il à ce que je veux mesurer ? Le résultat (la performance) est-il le plus proche possible de l'objectif à atteindre (composante de la compétence) ?

Les défauts de validité sont de loin les plus fréquents, pour des raisons différentes ne s'excluant pas les unes des autres :

- exercices, outils d'évaluation non adaptés à l'objectif à évaluer,
- outils d'évaluation non maîtrisés, mal conçus,
- barème de notation mal adapté,
- présence d'effets-parasites faussant la mesure.

Exemple "extrême" d'absence de validité : évaluer un candidat au Niveau 2 en testant son aptitude, dans le cadre de ses futures prérogatives, à répondre au signe "je suis essoufflé" en piscine, à genoux sur un fond de 3 m, en faisant le signe tout en respirant le plus calmement du monde...

Les conditions dans lesquelles le candidat est évalué n'ont rien avoir avec une situation réelle dans laquelle un essoufflement peut se produire. L'évaluation n'est pas valide.

Sensibilité

La **sensibilité** se rapporte à la précision de la mesure. Suis-je capable de différencier deux performances voisines mais différentes ? La sensibilité ne pose pas de problème particulier en évaluation critériée si les objectifs sont bien établis : le problème n'est pas de comparer deux élèves, mais de savoir si chacun d'entre eux à atteint ou non son objectif.

Fidélité

La **fidélité**, ou reproductibilité, est en revanche une limite grave de l'évaluation, retrouvée très souvent en plongée, tant pour les épreuves théoriques que pratiques. Le jugement porté est-il stable dans le temps ? La même prestation évaluée à différents moments par un même notateur aboutit-elle au même résultat ? Deux notateurs différents donnent-ils la même note ?

Exemple : dans une épreuve à 40m, le candidat qui passe le troisième est-il toujours évalué rigoureusement avec autant d'attention que le premier, à prestation identique ?

Les effets-parasites

Les effets-parasites sont des facteurs qui affectent la note même si l'outil d'évaluation est adapté à l'objectif ou à la compétence à évaluer. Chaque enseignant est contaminé par au moins l'un d'entre eux ; ils sont inévitables. Chaque moniteur, que ce soit en jury ou dans son for intérieur, doit l'avouer. L'essentiel est donc à la fois de *les* connaître et de *se* connaître.

Les effets-parasites sont :

- la fatigue,
- l'effet de halo ou l'effet Narcisse,
- l'effet d'ordre ou de contraste,
- l'effet de contamination,
- l'effet de stéréotypie,
- l'effet choc ou l'effet débordement.

Fatigue

La fatigue est une cause évidente d'erreur de mesure, entraînant un défaut de validité et/ou de fidélité.

Effet de halo

L'effet de halo correspond au processus, conscient ou inconscient, qui amène à surévaluer celui ou celle qui paraît sympathique, avenant, qui correspond au même cadre social de référence, qui est du même club, etc… Le processus de séduction relève du même mécanisme.

L'effet de halo poussé à l'extrême, c'est le favoritisme, bien connu, ou son inverse envers un élève mal-aimé ou détesté.

Parmi les critères qui peuvent intervenir figure ce qu'on pourrait appeler l'effet Narcisse : le fait de retrouver, dans un élève, une image de soi au même stade de l'apprentissage, ou plus simplement le fait que l'élève restitue un de ses "dadas" amène le moniteur à noter plus favorablement.

Effet d'ordre et de contraste

La note d'un candidat, ou d'une copie, dépend souvent de celle du précédent. Il y a confrontation, et l'on entend souvent des réflexions comme : "on a mis 11 au précédent, donc celui-ci ne vaut pas plus de 9...". Cette apparente préoccupation de justice, d'une évaluation à l'autre, cache souvent, en fait, une notation relative, normative, éloignée de la prise en compte de l'objectif à atteindre.

De plus, il est montré que dans une série longue de notations, la dispersion des notes augmente : l'enseignant attribue plus facilement de bonnes ou de très mauvaises notes au fur et à mesure que le temps passe.

Effet de contamination

L'évaluation d'un candidat peut aussi être influencée par l'avis d'autres cadres, par des appréciations entendues avant l'épreuve. Certains moniteurs sont peu influençables, d'autres beaucoup trop...

Effet de stéréotypie

Un moniteur qui suit régulièrement un élève peut être tenté de le "cataloguer", de le cantonner dans une échelle de notes dont il aura du mal à sortir, même si l'élève a progressé.

On échappe à ce biais en confiant de temps en temps ses élèves à un autre moniteur pour faire le point.

Effet de choc et de débordement

Enfin, l'évaluation d'une prestation dans sa globalité peut être faussée parce que le candidat a dit ou fait quelque chose qui aura un effet tout particulier sur le moniteur : c'est l'effet de choc. Une énormité, ou au contraire une notion très intéressante, aboutissent ainsi à une très mauvaise note, ou une très bonne, alors que le reste est moyen. L'effet de débordement, c'est "la goutte qui fait déborder le vase". L'élève réitère quelque chose qui en soit n'est pas grave, mais qui exaspère le moniteur à partir d'un certain seuil, même si sa prestation est correcte.

Les outils pour épreuves théoriques

Il n'existe pas de terminologie internationale normalisée pour désigner les différentes formes de questions et questionnaires utilisés. Par exemple, le terme bien connu de "QCM" est traduit, selon les auteurs, par "question à choix multiples", ou par "questionnaire à choix multiples". De même, certains considèrent que le "choix multiple" s'applique aux différentes propositions de réponse, mais que l'élève ne doit en choisir qu'une et une seule, tandis que d'autres estiment que dans un vrai "choix multiple", cet élève doit pouvoir cocher plusieurs bonnes réponses parmi les propositions.

Pour lever toute ambiguïté, j'utiliserai ici la terminologie suivante :

- ➜ dans les questions *ouvertes* (exemple : QROC, question à réponse ouverte et courte), le contenu pédagogique est situé dans la question elle-même ; c'est l'élève qui répond ensuite. On peut donc conserver le terme de *question.*

- ➜ dans les questions *fermées*, le contenu pédagogique principal n'est pas dans la question elle-même, mais dans les différentes propositions de réponse offertes à l'élève. On appelle *item* l'ensemble [question/propositions de réponse].

- ➜ un *questionnaire* est un ensemble de questions, quel qu'en soit le type.

Les différents types de questions (ou d'items) peuvent de plus être rassemblés en deux groupes :

- les items *de sélection* sont des questions fermées dans lesquelles l'élève doit choisir (ou trier, classer) parmi des propositions de réponse qui lui sont faites.

- les items *de production* sont des questions ouvertes auxquelles l'élève doit produire lui-même une réponse.

Les exemples donnés ci-dessous sont volontairement choisis hors du domaine de la plongée afin de ne pas perturber le lecteur avec un contenu trop signifiant.

Les items de sélection

Les items de sélection sont très utilisés, à juste titre, en vertu de trois qualités principales :

- ils se caractérisent avant tout par leur *objectivité* : il n'y a pas de marge d'interprétation sur la réponse de l'élève. Leur fidélité est donc très bonne également. Si le barème est correctement établi, ils peuvent être corrigés par différentes personnes, même non moniteurs, en arrivant au même résultat.

- regroupés en questionnaires, ils permettent de tester l'élève sur la plus *grande partie du programme* et des objectifs à évaluer. Ils s'adaptent très bien au concept de tableau de spécification : chaque grand groupe d'objectifs sera évalué par un nombre donné d'items, correspondant à un nombre de points précis.

- ils sont de *correction très rapide.* Avec des masques de correction bien conçus, on peut corriger une copie en quelques secondes. Il est même possible d'utiliser des outils informatisés de correction automatique.

Les items de sélection ont été largement critiqués. On leur a reproché :

- de ne vérifier que l'acquisition de connaissances factuelles trop pointues,
- d'induire les élèves en erreur en leur offrant à retenir trop de propositions fausses assénées sur un mode affirmatif,
- de laisser, pour certains de ces items, une trop grande part au hasard,
- de négliger la "connaissance partielle".

La grande majorité de ces critiques vient en fait d'une mauvaise formulation des items, de difficultés non résolues dans la notation, et d'un mauvais agencement des différents items au sein d'un questionnaire.

Leur principal inconvénient, donc, est qu'ils nécessitent une certaine expérience, une pratique, et ne tolèrent pas d'être bouclés trop rapidement, juste avant l'examen.

Le second inconvénient, qu'il faut connaître si l'on veut y remédier, c'est que les items de sélection facilitent la fraude par coup d'oeil sur les copies voisines. On y remédie en isolant les élèves ou en distribuant, soit des questionnaires différents, soit, pour éviter toute contestation, des questionnaires dans lesquels l'ordre des items est changé une rangée sur deux (ou un élève sur deux).

Les différents types sont présentés par fréquence d'utilisation décroissante.

Item à choix unique

Une question est posée. Plusieurs propositions de réponse sont données (cinq en général). Une seule est à sélectionner. Les autres sont appelés leurres ou distracteurs.

Exemple :
Quelle est la couleur du soleil ?
- *Vert*
- *Bleu*
✔ - *Jaune*
- *Marron*
- *Rouge vermillon*

C'est la forme la plus répandue d'items de sélection. Le fait qu'il n'y a qu'un seul choix possible leur confère des avantages :

- ils sont robustes, c'est à dire utilisables même s'ils ne sont pas parfaits : les élèves choisissent la réponse la plus plausible.
- parmi les items de sélection, ce sont les plus appréciés des élèves car ils savent qu'il n'y a qu'un seul choix possible.

Item Vrai/Faux ou Oui/Non

Une question est posée, soit sur le mode affirmatif, soit sur le mode interrogatif. L'élève doit simplement déterminer si cette proposition est vraie ou fausse, ou répondre par oui ou non.

Ce type d'item doit être évité en évaluation sommative (en examen). En effet, l'élève qui répond au hasard a une chance sur deux de bien répondre. Il est donc encouragé à répondre même s'il ne sait pas, ce qui est à éviter, et l'évaluation sera de toute façon biaisée.

En revanche, ces items sont très intéressants en cours de progression, lorsque l'élève veut se situer (évaluation formative et surtout auto-évaluation).

Item à choix multiple

Une question est posée. Plusieurs propositions de réponse sont données (cinq en général). Parmi elles, plusieurs peuvent être sélectionnées.

Exemple :
Parmi les teintes suivantes, quelles sont celles que peut prendre le soleil ?
- ✔ - *Jaune soleil*
- - *Vert pomme*
- ✔ - *Orange*
- ✔ - *Rougeâtre*
- - *Bleu pétrole*

Ces items sont mal aimés par les enseignants, qui ne savent pas bien comment établir un barème de notation, ni par les étudiants, qui se sentent frustrés, n'étant jamais sûrs d'avoir trouvé la bonne combinaison de réponses.

Pourtant, bien utilisés, et surtout bien notés (cf. plus loin), ces items sont très intéressants, car ils permettent une véritable pondération de l'évaluation, en dissociant les élèves qui appréhendent la totalité de la question, de ceux qui n'en ont qu'une connaissance partielle.

Item à réponses groupées

Une question est posée. Plusieurs propositions de réponse sont données (cinq en général). Ensuite, plusieurs combinaisons sont données parmi lesquelles une seul est à sélectionner.

Exemple :
Voici cinq mots censés être des noms de nuage. Quelle combinaison donne la liste de ceux qui en sont vraiment?

A : cumulus	- *A,B,D,E*
B : nimbus	- *A,B*
C : aureolus	- *B,D,E*
D : pluvius	✔ - *A,B,E*
E : stratus	- *A,B,C,D,E*

Ce type d'item a été introduit pour contourner les problèmes de notation des items à choix multiples. Plusieurs bonnes propositions sont à identifier, mais la liste des combinaisons ne propose qu'une seule bonne combinaison. Il s'agit donc d'un cas particulier d'item à choix unique.

Les items à réponses groupées sont peu utilisés.

Item d'appariement

Une question est posée. Plusieurs propositions de réponse sont données, qu'il faut associer deux par deux en général. Encore appelé item associatif.

Exemple d'appariement dit parfait :
Associez une seule lettre à un seul chiffre :

A : pluie	*1 : flocon*
B : vent	*2 : goutte*
C : cyclone	*3 : éclair*
D : neige	*4 : oeil*
E : orage	*5 : brise*

Réponse : A2, B5, C4, D1, E3.

Exemple d'appariement dit imparfait :
Associez une lettre à zéro, un ou plusieurs chiffres :

A : eau sous forme gazeuse	*1 : verglas*
B : eau liquide	*2 : nuage*
C : eau sous forme solide	*3 : pluie*
	4 : givre
	5 : brouillard

Réponse : A2 : rien. B : 2, 3, 5. C : 1, 4.

Ce type d'item, peu utilisé, est pourtant très utile pour évaluer l'aptitude à relier des faits ou des idées entre eux.

L'appariement imparfait est préférable à l'appariement parfait, qui permet des déductions trop faciles (le fait d'être sûr d'une ou deux paires diminue très nettement le nombre de possibilités sur les autres).

Item de classement

Une question est posée. Plusieurs propositions de réponse sont données, qu'il faut classer selon un ordre donné.

Exemple :
Classez ces étendues d'eau de la plus petite à la plus grande :
A : lac
B : étang
C : mer
D : flaque
E : mare

Réponse : D,E,B,A,C

Les items de classement posent surtout un problème de notation, lorsque la combinaison n'est que partiellement trouvée. Habituellement, les points ne sont accordés que lorsque la bonne combinaison est entièrement trouvée.

Conseils de rédaction

Quel que soit le type d'item de sélection, quelques conseils peuvent permettre d'en améliorer la qualité.

1. *Chaque item porte sur un objectif bien identifié, et sur ses aspects essentiels.*
   On ne se lance pas à rédiger des items sans savoir précisément ce que l'on veut évaluer.

2. *L'énoncé utilise un langage simple et clair, compréhensible par tous.*
   Eviter le jargon technique, les doubles négations, les phrases trop longues, etc…

3. *L'énoncé ne laisse pas de part à l'appréciation ou l'interprétation individuelle.*
   Les mots ambigus sont évités, l'énoncé comporte tout les mots essentiels à sa compréhension univoque. Les items ambigus sont très fréquents quand le moniteur, plongé dans sa logique personnelle, oublie qu'un élève pourrait sortir l'item de ce contexte en lui trouvant une autre interprétation possible.

Eviter les adverbes ou locutions laissant la place à l'appréciation ("en général", "plutôt", "relativement", …).

4. *Il n'y a qu'une seule solution à la question ; elle n'est pas contestable.*

Il faut se méfier de rédiger des items lorsque l'on n'a pas une connaissance étendue du sujet : en voulant rédiger des leurres, on écrit parfois des propositions tout aussi valables que la bonne réponse.

Eviter les bonnes réponses portant sur des sujets contestés.

5. *Il faut toujours laisser à l'élève la possibilité de répondre "Je ne sais pas".*

L'élève peut toujours ne pas répondre du tout, mais il est psychologiquement préférable d'inciter les élèves à affirmer leur ignorance, plutôt que de répondre au hasard.

6. *En examen, éviter le trop faible nombre de propositions de réponse.*

Un choix parmi deux ou trois laisse trop de place au hasard. L'idéal, c'est cinq propositions de réponse.

7. *Eviter les indices qui permettent de trouver indirectement la bonne réponse , ou d'exclure des mauvaises réponses.*

Eviter que la ou les bonnes réponses soient plus longues, plus courtes, plus précises que les autres, etc… Veiller à l'homogénéité grammaticale entre l'énoncé et les différentes propositions de réponse. Toute différence peut mettre l'élève sur la voie. Les propositions de réponse ont un contenu et une forme homogènes.

8. *La(es) bonne(s) réponses ne sont pas toujours au même endroit.*

Les élèves sont attentifs à la position des bonnes réponses dans la série.

9. *Les leurres sont incontestablement faux, mais plausibles.*

Des leurres trop loufoques, souvent utilisés, facilitent le travail déductif de l'élève, puisqu'ils neutralisent une possibilité.

A l'opposé, il ne doit y avoir aucun doute possible sur le fait que cocher chaque leurre est une mauvaise réponse.

10. *Il n'y a pas de propositions de réponse connexes, ou s'excluant.*

Les propositions connexes facilitent le choix de l'élève. De même, dans les items à choix unique, si deux propositions s'excluent l'une l'autre, les autres ne servent à rien : la bonne réponse fait obligatoirement partie de ces deux-là...

11. *La réponse à une question ne dépend pas de la réponse à une autre question.*

Eviter qu'une mauvaise réponse à un premier item entraîne obligatoirement une mauvaise réponse à d'autres items.

Eviter que, par recoupements sur des items connexes, on puisse deviner une réponse qu'on ne connaît pas.

13. *En examen, toujours préciser aux candidats, pour chaque item, le type d'item utilisé, et le barème de notation choisi.*
Sinon, les contestations ne sont pas rares.

12. *En cours de formation, éviter d'engendrer la mémorisation d'affirmations fausses.*
Les questionnaires d'auto-évaluation distribués en cours de formation, devraient contenir plus d'affirmations vraies que fausses. A l'examen, ce problème ne se pose pas.

Les items de production

Question à réponse ouverte et courte

Une question est posée. L'élève doit répondre en un, deux, ou quelques mots précis. La réponse peut être du texte, un nombre, une date, une heure, etc… Les exercices et problèmes peuvent, indirectement, être rattachés à ce groupe lorsque seule la réponse importe, indépendamment du cheminement pour y arriver.

Exemple :
Quelle est la couleur généralement attribuée au soleil (répondre uniquement par l'adjectif) ?

Réponse : jaune.

Les célèbres Q.R.O.C. représentent une charnière entre les items de sélection et les items de production. La question, si elle est bien rédigée, est en fait "semi-ouverte" : les possibilités de bonne réponse sont limitées et la correction peut être effectuée avec le même résultat par divers correcteurs. Le test reste donc objectif et fidèle. De plus, il n'est pas possible de trouver la bonne réponse en répondant au hasard.

Les conseils 1 à 5 ci-dessus s'y appliquent également : la question ne vaut que si elle est formulée de façon incontestable.

Texte lacunaire

Une phrase est proposée, avec des blancs à compléter selon une règle définie en énoncé.

Exemple :
Compléter en conjuguant le verbe "aller" correctement.

Demain, nous ..... visiter Belle-Ile.
Réponse : irons.

Sauf en langues (lorsqu'il s'agit de grammaire, ou de syntaxe, exemple ci-dessus), le texte lacunaire peut être également formulé sous la forme d'un Q.R.O.C., dont il n'est que la forme affirmative.

La forme la plus classiquement utilisée en plongée est le schéma dont il faut trouver les légendes.

Exercice

Un énoncé donne tous les éléments nécessaires, qu'il suffit d'appliquer pour trouver le résultat.

Exemple :
Convertissez 20°C en degrés Kelvin, en considérant que le zéro absolu (en °K) vaut -273°C.

Bien adapté à la physique, ou aux tables, l'exercice ne pose en général pas de problème. Il faut bien veiller à ne pas oublier de données indispensables à la réalisation correcte de l'exercice.

Si le cheminement qui permet de trouver le résultat est jugé aussi important que le résultat lui-même, il faut veiller à donner des consignes de rédaction ; par exemple : écrire la formule ou la loi utilisée, écrire les résultats intermédiaires, faire un schéma, etc…

Problème

L'énoncé ne donne pas systématiquement tous les éléments nécessaires ; certaines lois, formules ou valeurs doivent être connues de l'élève, et utilisées à bon escient pour trouver le(s) résultats(s).

Exemple :
Un avion vole à 450 km/h et consomme 50 litres de carburant à l'heure. Il part de Paris avec le plein (son réservoir fait 200 l). Combien de temps mettra-t-il pour aller à Marseille ? Doit-il faire escale et, si oui, où ?

Dans un problème, il est possible de ne pas fournir certaines données, dès l'instant qu'elles sont univoques et supposées être connues. S'il y a la moindre ambiguïté pouvant amener à des résultats différents mais corrects, il vaut mieux fournir les valeurs. En revanche, il est classique de fournir des données inutiles. Il ne faut pas en abuser, cela déstabilise souvent les élèves peu sûrs d'eux.

Question classique

A l'écrit comme à l'oral, l'épreuve peut enfin être constituée d'une ou plusieurs questions, amenant un développement plus ou moins important selon l'ampleur du sujet, ou selon qu'elle fait appel plutôt à la connaissance ou plutôt à l'esprit de synthèse..

Exemple de question dite "de cours":
Exposez les principales caractéristiques de la notion de responsabilité au sens juridique.

Exemple de sujet de synthèse :
Le fonctionnement des fédérations sportives en France est-il fondé sur n système démocratique ? Argumentez votre analyse.

Plus la question est ouverte et nécessite une réflexion de la part de l'élève, plus le temps accordé doit être long. Pour un oral, il faut un temps de préparation suffisant et sensiblement égal pour tous les candidats.

A l'oral, le temps nécessaire à la réalisation de l'épreuve peut donc être très long. A l'écrit, c'est la correction qui prend beaucoup de temps. c'est le principal écueil des sujets vastes.

Ce type de question, pourtant très employé, pose de gros problèmes de correction, tant de validité que de fidélité : effets-parasites, exigences différentes des correcteurs, difficultés d'expression de certains candidats, nombreux sont les écueils.

Dans tous les cas, la double correction devrait être imposée. Même le "découpage en tranches" de la réponse souhaitée, avec un barème diffusé aux correcteurs, ne permet pas d'homogénéiser les corrections.

En outre, à l'écrit, ce type d'épreuve ne permet pas de balayer la totalité des objectifs à évaluer ; les "impasses" sont possibles.

A l'oral, une grande part de ces inconvénients tombent puisqu'on peut :

- balayer tout le programme par de nombreuses questions variées,
- tenter de cerner si une mauvaise prestation du candidat est liée à un manque de connaissances, ou à un "oubli" momentané (problème de restitution), lui donner une chance de prouver sa connaissance partielle du sujet,
- siéger au moins à deux dans le jury.

En revanche, certains élèves, impressionnés à l'oral (en général ou devant tel moniteur en particulier), peuvent perdre leurs moyens. Il faut impérativement se comporter de façon à les mettre en confiance. On impute aussi à l'examen oral un manque de validité lié au risque d'effet-parasite comme l'effet de halo : les connaissances ne seraient pas les seules à être évaluées, mais aussi la manière d'être du candidat.

Après la rédaction…

Le travail de préparation d'une évaluation ne s'arrête pas à la rédaction. Pour un examen écrit, il faut encore :

- construire le correctif et le barème de notation aussitôt, tant que la rédaction des questions est fraîche à la mémoire du rédacteur. Dans une évaluation critériée, basée sur l'acquisition d'objectifs, la notation se prévoit *toujours a priori*.

- évaluer, le plus précisément possible, la durée globale accordée aux élèves pour chaque épreuve. Des ajustements en fin d'épreuve sont toujours mal vécus par les candidats.

- faire relire et vérifier l'ensemble par au moins un autre moniteur, afin de déceler toute imperfection éventuelle.

- veiller à rédiger les consignes générales permettant aux élèves de répondre correctement. Même des élèves habitués à un type de test donné peuvent encore se tromper. Dans les items de sélection notamment, il est important que l'on sache, pour chaque item, quel type de réponse est attendu. La durée de l'examen doit figurer, ou être annoncée. enfin, le barème de notation doit être précisé. La place pour écrire le nom et le prénom du candidat doit être prévue sur chaque feuille séparée à rendre.

En résumé…

1. L'évaluation est une mesure de laquelle on doit au maximum limiter l'erreur.

2. Une bonne évaluation est pertinente (son objet est clairement identifié), valide (ce qu'elle mesure correspond bien à cet objet), sensible (de faibles écarts de performance sont détectés), et fidèle (la même prestation, répétée, aboutit toujours à la même note).

3. Les effets-parasites, qui affectent la validité et la fidélité, sont nombreux et nul n'y échappe totalement. Chaque moniteur doit bien se connaître.

4. Les outils d'évaluation sont nombreux. Ils présentent tous des avantages et des inconvénients. Les items de sélection, bien conçus, sont des outils performants, à panacher avec des questions plus ouvertes.

5. Il faut apporter le plus grand soin à la rédaction. Une épreuve d'examen ne se bâcle pas. Tant le fond que la forme ont une influence considérable sur le résultat, indépendamment de la valeur du candidat.

6. Toujours faire relire les sujets par un autre moniteur. Exprimer clairement les consignes. Mentionner le barème de notation et la durée de l'épreuve.

L'expression des résultats

L'expression des résultats, en matière d'évaluation, se fait en quatre étapes différentes :

- la notation, le cas échéant, de chacun des items composant l'épreuve,
- le calcul de la note globale obtenue à l'épreuve,
- l'analyse de la performance du candidat sur l'ensemble des épreuves composant l'examen,
- l'analyse de l'examen lui-même (évaluation de l'évaluation).

La notation des items

Il y a trois manières de noter un item, une question au sein d'une épreuve :

➜ la notation quantitative classique,
➜ la notation quantitative pondérée,
➜ la notation qualitative.

Cette gamme permet, au gré du moniteur, de jouer sur le poids qu'il souhaite donner à l'item au sein de l'épreuve. Il peut vraiment procéder à une évaluation critériée, c'est-à-dire de subordonner la réussite à l'atteinte effective de l'objectif.

Quel que soit l'item et le type de notation, c'est chaque proposition de réponse (items de sélection), chaque champ de réponse (items de production), chaque partie du développement attendu (exercices, problèmes, questions classiques), qui fait l'objet d'une notation.

➜ Dans les items où une seule réponse est à sélectionner ou à produire (choix unique, vrai/faux, oui/non, réponses groupées, QROC), l'élève obtient la note affectée à sa réponse.

Dans les items à choix multiple ou dans les textes lacunaires, il obtient la somme des notes affectées à chaque élément de réponse.

Dans les questions classiques, il obtient, pour les différentes composantes de son développement, la somme des notes prévues dans le barème de correction.

Notation quantitative classique

C'est la notation la plus répandue. Une bonne réponse est notée en points positifs, une mauvaise réponse ou une absence de réponse ("je ne sais pas") ne rapporte pas de points.

Notation quantitative pondérée

Il est possible, dans un item de sélection, de donner un poids différent aux divers choix de réponse. Même dans des questions ouvertes, les différentes réponses, bonnes ou mauvaises, que peut donner l'élève peuvent être prévues et notées différemment. Une bonne réponse, ou composante de la réponse (selon le type de questions) est notée en points positifs, une mauvaise réponse en points négatifs.

La notation en points négatifs est en général mal vécue par les élèves et fait l'objet de nombreuses controverses. Son intérêt dépend en fait du type d'évaluation considéré. En évaluation normative (culture générale, par exemple), il n'y a pas lieu de pénaliser certaines réponses. On ne crédite que ce qui est su et compris. Par contre, en évaluation critériée, pour laquelle une performance est attendue, on peut être amené à considérer que telle mauvaise réponse est plus grave que telle autre.

Exemple : imaginons une question à choix multiples dans laquelle on demande à un futur Niveau 4 quelle est la conduite prioritaire à tenir face à un accident de décompression. On notera différemment une réponse type "ré-immerger la victime" (-2 points), "alerter les secours" (0 point), "faire boire de l'eau" (1 point), "mettre sous oxygène" (2 points).

Notation qualitative

Dans ce type de notation, une bonne réponse est notée en points positifs, une mauvaise 0 point ou en points négatifs. Mais de plus, certaines réponses peuvent être qualifiées d'**obligatoire** ou d'**éliminatoire**.

Si une bonne réponse est **obligatoire**, le fait de ne pas l'avoir sélectionnée entraîne *l'élimination pour tout le questionnaire* : le niveau n'est pas atteint, cet élément de l'item étant considéré comme fondamental.

De la même façon, si une réponse fausse est **éliminatoire**, cela revient à dire que le fait de l'avoir sélectionnée est impardonnable et entraîne *l'élimination à tout le questionnaire.*

Dans un tel système, le fait de ne pas répondre (ou de répondre "je ne sais pas"), entraîne logiquement l'élimination aussi.

Ce système est très puissant mais à manier avec précautions, strictement en rapport avec les objectifs minimaux à atteindre pour tel niveau de plongeur.

Exemple

Soit l'item à choix multiple suivant, avec les réponses considérées comme bonnes :

Parmi les teintes suivantes, quelles sont celles que peut prendre le soleil ?

✔ - *Jaune soleil*
- - *Vert pomme*
✔ - *Orange*
✔ - *Rougeâtre*
- - *Bleu pétrole*

En notation quantitative classique, on aurait par exemple :

1 - *Jaune soleil*
0 - *Vert pomme*
1 - *Orange*
1 - *Rougeâtre*
0 - *Bleu pétrole*

En notation quantitative pondérée, on aurait :

2 - *Jaune soleil*
0 - *Vert pomme*
1 - *Orange*
1 - *Rougeâtre*
-1 - *Bleu pétrole*

Mais, supposons que l'on décide qu'il est impardonnable, pour atteindre l'objectif, d'oublier "*jaune soleil*", tout autant que de cocher "*bleu pétrole*". On passe alors en notation qualitative, et l'on obtient :

Obligatoire (3) - *Jaune soleil*
0 - *Vert pomme*
1 - *Orange*
1 - *Rougeâtre*
Eliminatoire - *Bleu pétrole*

Cet exemple permet de bien comprendre les différences entre ces systèmes de notation. Pour simplifier, comparons simplement la notation pondérée et la notation qualitative.

Premier cas : l'élève répond :

✔ - *Jaune soleil*
- - *Vert pomme*
✔ - *Orange*
- - *Rougeâtre*
- - *Bleu pétrole*

Quant. : 3 sur 4 Qual. : 4 sur 5

L'élève a oublié de cocher "*rougeâtre*", mais le moniteur a estimé que ce n'était pas dramatique. Pas de différence notable entre les deux systèmes.

Deuxième cas : l'élève répond :

- *Jaune soleil*
- *Vert pomme*
- ✔ - *Orange*
- ✔ - *Rougeâtre*
- *Bleu pétrole*

Quant. : 2 sur 4 Qual. : Eliminé

L'élève a oublié de cocher "*jaune soleil*", et le moniteur a estimé que c'était impardonnable. Il est éliminé au lieu d'avoir la même note que dans le cas précédent.

Troisième cas : l'élève répond :

- ✔ - *Jaune soleil*
- *Vert pomme*
- ✔ - *Orange*
- ✔ - *Rougeâtre*
- ✔ - *Bleu pétrole*

Quant. : 3 sur 4 Qual. : Eliminé

L'élève a coché "*bleu pétrole*", et le moniteur a estimé que c'était impardonnable. Il est éliminé au lieu d'avoir la même note que dans le premier cas.

Quatrième cas : l'élève répond :

- ✔ - *Jaune soleil*
- ✔ - *Vert pomme*
- ✔ - *Orange*
- ✔ - *Rougeâtre*
- *Bleu pétrole*

Quant. : 4 sur 4 Qual. : 5 sur 5

L'élève a, en trop, coché "*vert pomme*", et le moniteur a estimé que ce n'était pas dramatique. Pas de différence notable entre les deux systèmes.

La notation des questionnaires

Notation quantitative

Trois principaux types de barèmes sont utilisés :

- le barème additif,
- le barème soustractif,
- le barème progressif.

Barème additif

En notation quantitative classique, la note finale est simplement la somme des notes obtenues pour chaque item ou composante de la question, le tout ramené à 20 points si nécessaire.

> Le résultat n'est donc pas directement corrélé à l'atteinte des objectifs de base dans la matière considérée. Le moniteur suppose implicitement que si l'élève a obtenu la "moyenne", c'est qu'il a le niveau requis.

Ce système est de très loin le plus répandu. Il offre cependant la possibilité de laisser passer des carences dans certains domaines : il a tout les inconvénients de l'échelle 0-20 avec succès à partir de la moyenne (10).

Exemple : une épreuve est composée de trois parties, jugées d'importance comparable pour atteindre le niveau. Deux parties sont notée sur 6 points, une sur 8 points. L'élève répond parfaitement aux deux questions à 6 points, et ne répond rien du tout à la troisième. Avec 12 sur 20, il est admis...

Certains jurys, dans les Niveaux 1 à 3, conscients du problème, placent la barre un peu plus haut. Ils exigent, soit une note de 12 ou 14, soit, sans calculer de note globale, une bonne réponse à un pourcentage d'items très supérieur à 50% (souvent 75 ou 80%). C'est une solution, mais qui n'est pas totalement satisfaisante : si, parmi les 20% de questions auxquelles l'élève a mal répondu, figurent des concepts fondamentaux ?

L'alternative est alors souvent la notation pondérée : le barème reste additif, mais certaines réponses apportent des points négatifs. Il se peut d'ailleurs que la note finale soit négative, ce qui n'a pas grande signification. Dans ce cas, on la ramène à zéro.

On peut regretter qu'au Niveau 4, dont la délivrance a pourtant des conséquences en terme de responsabilité, le système classique, non pondéré, avec moyenne à 10 soit encore en vigueur dans la grande majorité des cas.

Le problème des effets du hasard a souvent préoccupé les enseignants, dans d'autres domaines que la plongée. Il existe une formule qui tente d'en pondérer les effets. Si B est le nombre de bonnes réponses, F le nombres de réponses fausses, et T le nombre total d'items, la note se calcule ainsi :

$$\text{Note} = B - \frac{F}{T - 1}$$

Exemple : 20 bonnes réponses. L'élève est bon, le hasard a sûrement une part modérée dans sa réussite. Sa note est de 20. S'il a 10 bonnes réponses, il obtient 9,5. S'il n'a qu'une seule bonne réponse, il a 0.

Barème soustractif

Le barème soustractif est parfois employé : l'élève a, a priori, la note maximale, et on lui déduit les points correspondant à ses erreurs. S'il tombe en-dessous d'un certain seuil, il n'est pas admis.

En fait, ce barème n'est qu'un autre mode de calcul du barème additif (on procède simplement à l'envers). Il en a les mêmes défauts.

Barème progressif

Le barème progressif, réservé aux questionnaires à items de sélection, est très rarement employé. Les points affectés à chaque item augmentent en fonction du nombre de bonnes réponses, par tranches.

Exemple : Sur 20 items, bien répondre à 9 items ne rapporte rien. A partir de 10 bonnes réponses, 1 point par réponse. A partir de 15, deux points par réponse.

Ainsi, un élève ayant 7 bonnes réponse obtient un 0. Un élève ayant 12 bonnes réponses conserve la note de 12. Un élève ayant 17 bonnes réponses et au-delà obtient la note maximale (20).

Ce système présente l'avantage de réduire l'échelle de notation. Dans l'exemple proposé, seules les notes 0, 10, 11, 12, 13, 14, 16, 18 et 20 sont attribuées, soit 9 degrés. Mais il est un peu bâtard, entre système normatif et système critérié.

Côté critérié, il partage avec l'échelle à 4 degrés (cf. p. 29) la caractéristique de n'avoir pas de note en dessous de la moyenne sauf 0. Mais son défaut est de n'être pas sélectif sur le ou les items auxquels il faut impérativement répondre correctement. Ce barème est bien quantitatif : à moins de 50% de bonnes réponses, échec... Il n'est donc utilisable que si tous les items ont un poids, une importance comparable dans le questionnaire.

Côté normatif, il avantage les élèves qui répondent bien à plus de 15 questions. Quel intérêt ?

Notation qualitative

A condition d'avoir défini précisément les objectifs et compétences à acquérir, la notation qualitative est une excellente solution. Dans ce cas, le questionnaire doit être agencé en deux parties :

- la première ne contient que des items éliminatoires. Une mauvaise réponse, ou une absence de réponse à un item seulement entraîne l'élimination du candidat (note 0). Si l'élève a satisfait à cette première partie, c'est qu'il a atteint le **Niveau acceptable de performance**. Il est en prévenu à l'avance, ainsi que dans les consignes générales qui figurent sur son questionnaire.

- la seconde partie du questionnaire est notée de façon quantitative. Cela permet d'aller plus loin dans l'évaluation des connaissances et de délivrer des notes allant du Niveau acceptable, à **bien** et **très bien** (1, 2 ou 3).

Ce type de questionnaire est tout à fait réalisable dans le contexte actuel des brevets, et avec l'échelle de notation 0-20 : on ne se sert que de 4 notes (0, 10, 12, 16 par exemple).

Mais, au risque d'insister, cette manière de procéder n'est applicable que si les objectifs sont clairs et admis de tous. Nous n'en sommes pas là...

La notation sur l'ensemble des épreuves

Le problème de la décision, sur la base de la prestation à l'ensemble des épreuves, est toujours le même : peut-on tolérer des faiblesses dans une des disciplines formant l'examen et, si oui, dans quelle mesure ?

Le Niveau Acceptable de performance et la médiane

En notation qualitative, le problème ne se pose pas. Il s'agit véritablement d'évaluation critériée, et il faut que l'élève ait le Niveau Acceptable de Performance dans toutes les épreuves.

Si c'est le cas, pour attribuer ensuite une note globale à l'élève, on cherche la médiane (cf. p. 30). Pas de multiplications à faire, pas de calculs complexes (une médiane ne se calcule pas), simplification des bordereaux.

Attention, la médiane n'est pas utilisable dans les systèmes plus classiques de notation (cf. simulations en Annexe II).

Les coefficients (moyenne pondérée)

La solution "ancestrale" consiste à pondérer les différentes épreuves en leur attribuant un coefficient. Telle épreuve, de coefficient 4, a quatre fois plus d'importance dans la note finale que telle autre, de coefficient 1. La note éliminatoire, dans la majorité des cas, est une note inférieure ou égale à 4.

Faute d'être capable de procéder à une réelle évaluation critériée, qui rend ce système de pondération inutile, cette solution est la moins mauvaise.

Il faut impérativement que tous les moniteurs soient bien conscients des limites de ce système, ce qui est loin d'être le cas :

- les limites de l'échelle 0 - 20 (cf. p. 28) sont amplifiées par les coefficients.
- le caractère tout relatif d'une note doit amené à mûrir sa décision avant de la multiplier par 4. L'erreur de mesure est également multipliée...
- même en limitant l'erreur de mesure, un élève n'ayant pas du tout satisfait à certains objectifs fondamentaux pour un niveau de plongeur donné peut très bien décrocher son brevet par le jeu des coefficients.

Les simulations de résultats présentés en Annexe II sont significatives à cet égard...

En fait, la pondération des épreuves par des coefficients n'est qu'une parodie d'évaluation critériée, qui permet de faire l'économie de la définition des objectifs de la formation.

La moyenne simple

Le simple calcul de la moyenne, à partir des notes de chaque épreuve, est la pire des solutions.

Faire la moyenne d'une épreuve de physiologie, de physique, d'accidents et de réglementation, par exemple, revient à faire la moyenne entre des kilomètres et des kilogrammes : l'échelle est conçue de la même façon, avec des subdivisions analogues (kilo, micro, etc…), mais l'unité de mesure, au départ, n'est pas la même.

Si l'on prend en compte, en plus, les problèmes de fidélité de certains tests (la même copie pouvant être notée de façon *très* différente), on imagine bien comment la réussite ou l'échec peuvent être dénués de toute signification.

Certains moniteurs, dans certains jurys, en sont parfaitement conscients et se forgent, forts de leur longue expérience, une idée globale de la valeur du candidat ("çà passe", ou "çà ne passe pas"). En fonction de cette appréciation d'ensemble, ils ajustent les notes. Mais certains candidats peuvent passer à travers les mailles du filet.

L'importance de la délibération du jury

Dans tous les cas de figure, il est fondamental que la décision soit prise par une équipe pédagogique, un jury, et non pas par un individu isolé. C'est un niveau global que l'on évalue. En conséquence, chaque note n'est qu'une facette de la compétence à obtenir. Seul le jury peut donc en faire la synthèse et gommer les imperfections.

Il convient donc de dénoncer les pratiques, pourtant courante, de certains moniteurs :

- ceux qui, se posant en censeurs infaillibles, ne reviennent jamais, par principe, sur une note, quel que soit le profil global du candidat.

- ceux qui se permettent de ne pas assister à la délibération finale du jury, se contentant de donner leur note, assortie parfois d'un bref historique commenté.

L'analyse de l'examen

La dernière étape, l'analyse de l'examen, n'est que trop rarement effectuée. C'est dommage, car elle ouvre immédiatement vers deux évaluations indispensables, en complément de celle des candidats :

- l'évaluation de l'examen lui-même. Que tous les candidats, par exemple, même les meilleurs, échouent ou réussissent à une épreuve

doit amener à se poser une question : est-ce mérité ou y a-t-il un biais dans la mesure ?

• l'évaluation des enseignements : le résultat d'un candidat ne dépend pas seulement de sa valeur et de son travail, mais aussi de sa formation. Il est difficile cependant d'analyser cela à l'échelon individuel. Mais l'analyse des résultats de tous les candidats à un même examen permet de tirer des tendances globales (points forts et points faibles), sur la formation reçue.

De tels renseignements ne peuvent pas être tirés d'impressions plus ou moins subjectives. Seuls des chiffres permettent de statuer, et encore faut-il les interpréter correctement. Il s'agit donc de déterminer des *indices*, des paramètres statistiques interprétables. Il y a deux types d'indices, portant sur :

- l'analyse des résultats obtenus pour l'ensemble de l'examen,
- l'analyse de chacun des items le composant.

L'analyse des épreuves

Trois principaux types d'indices sont utilisés :

- les pourcentages de réussite,
- les indices de tendance centrale,
- les indices de dispersion.

Pourcentage de réussite

Le pourcentage de candidats reçus (en global ou par épreuve) est un bon indice d'analyse des épreuves, à condition d'être interprété correctement.

Il faut d'abord considérer le *nombre de candidats total* (en statistiques, on parle de taille d'échantillon). 100% de réussite pour trois candidats et 100% de réussite pour 300 candidats n'ont pas la même valeur informative. Il faut se méfier des pourcentages lorsque la taille de l'échantillon est faible (pour simplifier, inférieure à 30). Or, mis à part les examens théoriques anticipés du Niveau 4, les échantillons dans les examens de plongée sont toujours petits… Le pourcentage de réussite, dans ces conditions, ne signifie pas grand-chose.

Il faut aussi prendre en compte le *nombre de candidats ayant passé l'examen*, par rapport au nombre de candidats inscrits. 100% de réussite après un stage où 50% des stagiaires ont abandonné ou n'ont pas été autorisés à passer l'examen, cela ne veut rien dire non plus, sinon que la sélection opérée a bien fonctionné !

Tendance centrale

Le plus utilisé reste la moyenne arithmétique des notes obtenues par tous les candidats de l'examen (moyenne par épreuve, moyenne des moyennes générales).

Cet indice ne s'applique pas bien aux outils de l'évaluation critériée. Il est beaucoup plus utile en mode normatif, où il faut situer un candidat par rapport à la production moyenne de tous.

Il est tellement facile à calculer qu'il ne faut pas s'en priver. La moyenne peut, dans certains cas, permettre de déceler une carence globale des élèves dans un domaine donné, donc de remettre en question l'enseignement.

La moyenne harmonique et surtout la médiane peuvent être utilisés aussi. La médiane est beaucoup moins sensible que la moyenne aux scores extrêmes. Si moyenne et médiane sont basses, et que l'analyse des items révèle que ceux-ci n'étaient pas trop difficiles, c'est que l'enseignement est à revoir, l'objectif n'est pas atteint pour la majorité des élèves.

Dispersion

Les indices de dispersion permettent de regarder comment se répartissent les notes autour de la tendance centrale. Le plus utilisé est l'écart-type, noté σ, indice complémentaire de la moyenne arithmétique.

$$\sigma = \sqrt{\frac{\sum (x_i - \overline{M})^2}{n}}$$

On peut aussi calculer l'écart à la médiane.

Ces indices ont en réalité peu d'intérêt dans une optique d'évaluation critériée. Néanmoins, ils sont un complément indispensable de la moyenne si l'on veut s'en servir pour évaluer les enseignements.

Exemple : Une épreuve d'examen où 30 candidats ont obtenu une note moyenne de 7,2 ± 1,2 donne plus d'informations sur une carence de l'enseignement qu'une autre notée 5,6 ± 6,1.

Autres indices

D'autres indices ont été développés, pour mieux cerner les limites d'un examen, en tentant d'évaluer notamment la validité et la fidélité d'un examen. Ils sont difficilement exploitables dans le domaine de la plongée.

La validité d'un test n'est pas mesurable. Il s'agit d'une notion qualitative, en référence à l'objectif à mesurer. Un coefficient de validité prédictive a été proposé. Basé sur le calcul d'un coefficient de corrélation entre deux tests, il permet de mettre au point des tests ou examens blancs adaptés à mesurer la probabilité de réussite à tel examen institutionnel. Ce coefficient n'a pas grand intérêt en plongée.

Le calcul d'un coefficient de fidélité n'est possible, là encore, qu'en calculant une corrélation entre deux séries de notes obtenues à plusieurs épreuves identiques, à condition que la valeur intrinsèque des candidats n'ait pas varié entre temps ! Un coefficient, appelé KR (pour Kuder-Richardson), est utilisé Outre-Atlantique ; il peut se calculer à partir des résultats d'un seul examen. Sa valeur est cependant contestée.

L'analyse des items

Trois études peuvent être faites :

- l'analyse des choix de réponse, dans les items à choix unique ou à choix multiple,
- l'indice de difficulté,
- l'indice de discrimination.

Ces études ne sont jamais faites, ou presque. C'est dommage, parce qu'a posteriori, on peut détecter des failles dans un item ou dans l'une de ses composantes. Si l'on souhaite constituer une banque d'items d'examen, au niveau de la C.T.R. ou au niveau national, cette phase sera incontournable.

Analyse des choix de réponse

Dans un item à choix unique ou à choix multiple, il est utile de mesurer, sur l'ensemble des candidats, le pourcentage de choix de chacune des propositions de réponse de chaque item.

Pour se fixer un ordre d'idée, les leurres devraient attirer en moyenne 10% des élèves, et la(les) bonne(s) réponse(s) 50 à 75% d'entre eux.

Ainsi, un leurre n'ayant attiré aucun des candidats n'est pas un bon leurre. A l'opposé, un leurre ayant attiré un pourcentage élevé d'élèves peut être trop proche de la bonne réponse, au moins dans l'esprit des élèves. Cela ne veut pas dire pour autant que l'item soit à écarter. Il faudra simplement l'éviter comme item éliminatoire. S'il est manifestement faux, c'est que l'enseignement qui doit être remis en cause.

Si la bonne réponse est trouvée par moins de 25% des élèves, on estime que l'item est trop difficile. Là encore, il peut être conservé s'il est inclus dans un questionnaire pour différencier le "bien" du "très bien".

Indice de difficulté

L'indice de difficulté p d'un item correspond au pourcentage d'étudiants qui l'ont réussi complètement.

Plus il est proche de zéro, plus l'item est donc difficile.

L'indice de difficulté doit être interprété en tenant compte des indices de tendance générale (c'est à dire de la valeur du groupe), et de la formulation de l'item. Il permet donc de détecter, selon les cas, un problème de formulation, de construction de l'item, ou une carence de l'enseignement sur l'objectif se rapportant à l'item.

Lorsque de nombreux items ont été omis, par exemple en fin de questionnaire par manque de temps, il faut les déduire du calcul de l'indice de difficulté : ce n'est bien sûr pas, dans ce cas, parce qu'ils étaient difficiles que les élèves n'ont pas répondu.

L'influence du hasard peut être pondérée en utilisant une formule appropriée. Si n est le nombre d'options de l'item, on a :

$$p \text{ corrigé} = \left(p \cdot \frac{n}{n-1} \right) - \frac{1}{n-1}$$

Par exemple, un item de 5 choix ayant un p égal à 0,43 (43% de réussite), obtient un p corrigé de 0,29. On estime donc que plus de 10% des élèves ont bien répondu par l'effet du hasard dans un item à 5 choix.

Indice de discrimination

En évaluation normative, on calcule l'indice de discrimination d'un item, qui correspond à son aptitude à faire la différence entre les bons élèves, les élèves moyens et les mauvais élèves. Cet indice n'a que peu d'intérêt en évaluation critériée.

Ainsi, par exemple, un item auquel 75% des élèves répondent correctement, même les mauvais élèves qui ont une mauvaise note à l'ensemble de l'examen, discrimine peu. On évalue cet indice en calculant le coefficient de corrélation entre les résultats au test et les résultats à l'item.

En résumé…

1. Bien manipulés, en référence à des objectifs clairs, la notation pondérée, incluant notamment des points négatifs, ou bien la notation qualitative (un candidat peut être éliminé à tout le questionnaire en fonction de sa réponse à un item fondamental), sont très intéressantes.

2. Les items à choix multiple, qui gênent souvent les moniteurs, sont pourtant simples à noter : il suffit d'affecter une note à chacune des propositions de réponse. La note obtenue est égale à la somme des notes affectées aux réponses de l'élève.

3. La note finale est égale à la somme des notes obtenues pour chaque item. En notation qualitative, le fait de placer en première partie de test des items éliminatoires permet de savoir si le Niveau Acceptable de Performance est atteint, le reste des items sert à faire la différence entre 'l'acceptable", le "bien" et le "très bien".

4. Sur l'ensemble des épreuves, en notation qualitative, il suffit que l'élève ait atteint le N.A.P. dans toutes les épreuves.

5. En notation classique, la moyenne des notes obtenues à toutes les épreuves n'est pas une bonne solution. La pondération par des coefficients peut atténuer les imperfections, mais en aucun cas elle ne résout le problème.

6. La délibération du jury joue un rôle majeur dans l'évaluation globale d'un candidat, dans l'appréciation de son profil, et ne doit pas être sous-estimée.

7. L'analyse de l'examen est une phase souvent négligée qui apporte des informations intéressantes pour évaluer l'enseignement et les tests eux-mêmes. Il faudrait au moins, à chaque fois, connaître pour chaque épreuve le pourcentage de réussite, la moyenne et l'écart-type.

8. L'analyse des items est importante, notamment dans la phase de constitution d'une banque d'items d'examen. On peut analyser les choix de réponse par option au sein d'un item, et pour chaque item son indice de difficulté.

Bibliographie

Albernot Y : Les méthodes d'évaluation scolaire. Bordas, Paris, 1988.

Aubégny J : Les pièges de l'évaluation. Evaluer pour (se) former. Editions Universitaires, Paris, 1987.

Bernard H, Fontaine F : Les questions à choix multiple ; guide pratique pour la rédaction, l'analyse et la correction. Service Pédagogique, Université de Montréal, 1982.

Blais JG, Laurier M, Lévesque M, Pelletier G, Van der Maren JM : Guide pédagogique. G.R.I.P.U., Université de Montréal, 1993.

Bloom D : A taxonomy for educational objectives.

Cardinet J : Evaluation scolaire et mesure. De Boeck, Bruxelles, 1986.

Cardinet J : Pour apprécier le travail des élèves. De Boeck, Bruxelles, 1988.

Daigneault A : Guide docimologique. A.I.E.S., Montréal, 1973.

De Lansheere G : Dictionnaire de l'évaluation et de la recherche en éducation. P.U.F., Paris, 1979.

De Peretti A : Recueil d'instruments et de processus d'évaluation formative. I.N.R.P., Paris, 1980.

Fontaine F : Dossier sur l'évaluation. Service Pédagogique, Université de Montréal, 1989.

Houssaye J : La pédagogie, une encyclopédie pour aujourd'hui. E.S.F., Paris, 1993.

Mager RF : Comment définir des objectifs pédagogiques. Bordas, Paris, 1990.

Mager RF : Comment mesurer les résultats de l'enseignement. Bordas, Paris, 1986.

Monestiez P : Evaluation et niveaux de plongée. Mémoire d'Instructeur National F.F.E.S.S.M., Octobre 1993.

Payne DA : The specification and measurement of learning outcomes. John Wiley, New York, 1968.

Piéron H : Examens et docimologie. P.U.F., Paris, 1969.

Popham JW : Criterion-referenced measurement. Prentice Hall, New Jersey, 1978.

Annexe I : Résumé des points principaux

1. L'évaluation est un flux d'informations qui circule en permanence au sein du système pédagogique et qui en permet la régulation. Pourtant, les cadres sont peu ou mal préparés à évaluer.

2. Il y a trois principaux types d'évaluation : prédictive (pourra t-il suivre ?), formative (progresse-t-il ?), sommative (a-t-il atteint le niveau ?). En plongée, dans les trois cas, l'évaluation a progressé pour les épreuves dans l'eau, mais reste très imparfaite pour les épreuves théoriques.

3. A l'heure où les coûts d'une formation ne sont plus négligeables, l'évaluation des enseignements devient indispensable.

4. Les épreuves théoriques font partie d'un tout. C'est une erreur de trop les déconnecter de l'objectif final de formation. Vers une remise en question des examens théoriques anticipés ?..

5. Il n'y a donc aucune raison de ne pas apporter autant de soin à la définition des objectifs du domaine cognitif.

6. Il ne faut pas se limiter à la simple restitution de connaissances. Selon les domaines et les niveaux préparés, des objectifs d'application et d'analyse doivent être définis. Même des savoir-faire sont à évaluer (cas des épreuves de matériel).

7. Il est préférable de définir, avant de rédiger les épreuves, la répartition des questions, en pourcentage, en fonction des différents types d'objectifs à évaluer.

8. Tous les objectifs doivent avoir été définis avec soin et précision, faute de quoi les élèves ne pourront pas être évalués dans de bonnes conditions. Les élèves doivent connaître ces objectifs et, de préférence, les accepter et se les approprier.

9. L'échelle de notation traditionnelle 0-20 ne s'applique pas bien à une pédagogie de maîtrise basée sur des objectifs à atteindre clairement définis et acceptés. Il faut lui préférer une échelle critérielle à quatre degrés : 0 (objectif non atteint), 1 (objectif juste atteint), 2 (bien) et 3 (très bien).

10. L'évaluation est une mesure de laquelle on doit au maximum limiter l'erreur.

11. Une bonne évaluation est pertinente (son objet est clairement identifié), valide (ce qu'elle mesure correspond bien à cet objet), sensible (de faibles écarts de performance sont détectés), et fidèle (la même prestation, répétée, aboutit toujours à la même note).

12. Les effets-parasites, qui affectent la validité et la fidélité, sont nombreux et nul n'y échappe totalement. Chaque moniteur doit bien se connaître.

13. Les outils d'évaluation sont nombreux. Ils présentent tous des avantages et des inconvénients. Les items de sélection, bien conçus, sont des outils performants, à panacher avec des questions plus ouvertes.

14. Il faut apporter le plus grand soin à la rédaction. Une épreuve d'examen ne se bâcle pas. Tant le fond que la forme ont une influence considérable sur le résultat, indépendamment de la valeur du candidat.

15. Toujours faire relire les sujets par un autre moniteur. Exprimer clairement les consignes. Mentionner le barème de notation et la durée de l'épreuve.

16. Bien manipulés, en référence à des objectifs clairs, la notation pondérée, incluant notamment des points négatifs, ou bien la notation qualitative (un candidat peut être éliminé à tout le questionnaire en fonction de sa réponse à un item fondamental), sont très intéressantes.

17. Les items à choix multiple, qui gênent souvent les moniteurs, sont pourtant simples à noter : il suffit d'affecter une note à chacune des propositions de réponse. La note obtenue est égale à la somme des notes affectées aux réponses de l'élève.

18. La note finale est égale à la somme des notes obtenues pour chaque item. En notation qualitative, le fait de placer en première partie de test des items éliminatoires permet de savoir si le Niveau Acceptable de Performance est atteint, le reste des items sert à faire la différence entre 'l'acceptable", le "bien" et le "très bien".

19. Sur l'ensemble des épreuves, en notation qualitative, il suffit que l'élève ait atteint le N.A.P. dans toutes les épreuves.

20. En notation classique, la moyenne des notes obtenues à toutes les épreuves n'est pas une bonne solution. La pondération par des coefficients peut atténuer les imperfections, mais en aucun cas elle ne résout le problème.

21. La délibération du jury joue un rôle majeur dans l'évaluation globale d'un candidat, dans l'appréciation de son profil, et ne doit pas être sous-estimée.

22. L'analyse de l'examen est une phase souvent négligée qui apporte des informations intéressantes pour évaluer l'enseignement et les tests eux-mêmes. Il faudrait au moins, à chaque fois, connaître pour chaque épreuve le pourcentage de réussite, la moyenne et l'écart-type.

23. L'analyse des items est importante, notamment dans la phase de constitution d'une banque d'items d'examen. On peut analyser les choix de réponse par option au sein d'un item, et pour chaque item son indice de difficulté.

Annexe II : Simulation de résultats d'examen Niveau 4

Afin d'alimenter la réflexion, voici quelques simulations de résultats individuels à un examen Niveau 4. Pour plus de commodité, on séparera les résultats à la partie Pratique et les résultats en Théorie.

Pratique

Imaginons trois candidats, Pierre Paul et Jacques. D'après leurs résultats ci-dessous, Pierre et Paul sont reçus à la Pratique, tandis que Jacques est ajourné.

Epreuve	Coef.	Pierre		Paul		Jacques	
Mannequin	2	10	20	18	36	2	4
800 P.M.T.	1	10	10	18	18	14	14
Apnée	1	10	10	16	16	16	16
Sauvetage Palmes	3	10	30	6	18	14	42
Bleu à 40 m	1	10	10	10	10	13	13
Code à 40 m	4	10	40	6	24	15	60
R.S.E.	2	10	20	15	30	17	34
Bouée 30 m	2	10	20	6	12	13	26
Matelotage	2	10	20	16	32	12	24
Total pondéré (>180 ?)			**180**		**196**		**233**
Moyenne pondérée (total/18)			10		10,89		12,94
Moyenne arithmétique		10,00		12,33		12,89	
Moyenne harmonique		10,00		9,97		8,43	
Médiane		10,00		15,00		14,00	

Etudions respectivement leurs profils, afin d'en tirer quelques leçons.

Pierre

La simulation des notes de Pierre est provocatrice. Il est peu probable qu'une telle situation se produise : il a 10 partout ! Elle est cependant intéressante, car, même en faisant confiance au système d'évaluation actuel, et à son échelle 0-20, on sent bien, intuitivement, que "quelque chose cloche"...

Pierre a-t-il bien le niveau de compétence pratique correspondant aux prérogatives du Niveau 4 ? Compte-tenu des aléas de l'évaluation, où se situe-t-il réellement ? Ne prend-on pas un risque à lui confier des palanquées (cf. risque de première espèce, p. 36) ?

Paul

Dans le système tel qu'il est actuellement, Paul a acquis la Pratique du Niveau 4. Il a 196 points (pour 180), pas de note éliminatoire. Observons ses notes. Paul est "physique". Il est jeune, il a "la pêche"… Il est bon en nage, en apnée, en matelotage. En revanche, il

manque cruellement de maturité technique, et sa compétence de guide de palanquée n'est pas certaine : 6 à 40m, et 6 à chacune des épreuves de sauvetage en scaphandre...

Ce type de configuration est parfois rencontrée. Le jeu des coefficients, et le nombre d'épreuves "physiques" par rapport aux épreuves "techniques" sont à l'origine d'une distorsion. Quels sont les moyens d'y remédier ?

Si l'on changeait les textes, le meilleur moyen serait de passer à l'évaluation critériée. Dans ce contexte, Paul serait ajourné, il n'a pas le niveau.

A défaut, une solution pourrait être de changer les coefficients. En passant l'épreuve de bouée à un coefficient de 3, et de ramener les épreuves de mannequin et de R.S.E. à un coefficient 1, Paul est ajourné (169 pts pour 170). Cependant, pondérer davantage les épreuves qui paraissent incontournables n'est pas la meilleure solution : les élèves réussissant bien les "grosses" épreuves, maix ayant moins travaillé les autres pourraient réussir. On admettrait simplement que c'est moins grave que l'inverse (cas de Paul).

On remarquera :

- que la moyenne arithmétique des notes (12,33 pour Paul), non coefficientées, n'est pas du tout un bon reflet de la performance.
- que la médiane (ici, 15 !) n'est pas applicable dans ce type d'évaluation.
- que la moyenne harmonique (9,97) situe mieux l'élève dans cet exemple.

Jacques

Jacques est un bon plongeur. Tous les moniteurs qui l'ont suivi pendant le stage s'accordent pour dire qu'il a le profil d'un bon Niveau 4. Ses notes l'attestent. Pourtant, une crampe pendant le remorquage du mannequin, en fin d'épreuve, l'a contraint à arrêter avant la fin du parcours de 100 m. Il est donc éliminé. Il n'est pas rattrapable, il y a de nombreux témoins, ce serait un précédent de lui refaire passer l'épreuve, et il n'est pas possible de remonter sa note à 5.

Pourtant, tout le monde sait qu'il a le niveau. Où est l'erreur ?

Remarque : tant la moyenne harmonique que la médiane reflètent mal, dans ce type d'évaluation, la valeur du candidat. C'est bien le profil global du candidat qui doit donc compter, plus que quelques indices chiffrés qui ont tous leurs limites...

Théorie

Soient les résultats suivants :

Epreuve	Coef.	Pierre		Paul		Jacques	
Tables	3	10	30	7	21	16	48
Accidents	3	10	30	7	21	14	42
Physique	1	10	10	16	16	3	3
Physiologie	1	10	10	17	17	16	16
Réglementation	2	10	20	18	36	13	26
Matériel	2	10	20	7	14	15	30
Total pondéré (>120 ?)			120		125		165
Moyenne pondérée (total/12)			10		10,42		13,75
Moyenne arithmétique		10,00		12,00		12,83	
Moyenne harmonique		10,00		9,91		8,91	
Médiane		10,00		11,50		14,50	

Les profils sont les mêmes que dans l'exemple concernant la pratique. Les commentaires seraient donc identiques : Pierre, admis avec 10 partout, Paul, admis avec trois notes de 7 en tables, accident et matériel, et Jacques, éliminé parce qu'il a rendu copie quasi blanche en Physique (il a des problèmes avec les règles de trois…).

Annexe III : Analyse d'un examen théorique Niveau 4

L'analyse portera, à titre d'exemple, sur les sujets posés lors de la session anticipée, régionale, de l'examen théorique du Niveau 4, du 13 Mai 1995. Les épreuves figurent ci-joint.

Rappel des résultats

Centre d'examen	Angers	Chateaulin	Rennes	Moyenne
Nb candidats (total : 93)	23	34	36	
Tables	10,78	11,75	13,16	12,06
Accidents	8,90	10,81	10,79	10,33
Physique	14,30	15,21	15,03	14,92
Physiologie	9,57	8,54	12,03	10,15
Matériel	10,57	12,82	12,36	12,09
Réglementation	14,22	15,68	15,78	15,36
Moyenne Théorie	11,04	12,37	12,93	12,26

Epreuve de Tables

Pertinence Rien à signaler. Les 2 problèmes et les 10 exercices sont en accord avec les prérogatives du Niveau 4.

Validité Rien à signaler. L'association d'exercices simples à exécuter en temps limité (objectif : application) et de deux problèmes faisant intervenir la réflexion (objectif : analyse) a fait ses preuves désormais. Les deux problèmes sont très voisins. C'est dommage : autant évaluer des facultés d'analyse portant sur des situations très différentes.

Rédaction On peut regretter l'utilisation abusive des symboles T, P, T1, T2 (malgré la présence d'une légende). Il paraît préférable d'utiliser des abréviations plus directement "parlantes" pour tous : Temps, Prof., … En outre, les consignes n'étaient pas suffisamment claires : absence d'indication sur le temps maximal accordé, pas de barême de notation, même indicatif.

Notation Le total était noté sur 60, dont 30 points pour les deux problèmes et 30 pour les 10 exercices. Ceci signifie que, dans l'esprit du concepteur, les objectifs d'application et d'analyse pèsent le même poids : à discuter.

Les 10 exercices, de difficulté inégale, font l'objet d'une notation différente : trois à 2 points, quatre à 3 pts, trois à 4 pts (sur 30 pts).

Résultats : Moyenne régionale 12,06. A noter une répartition inégale dans les trois centres organisateurs : Rennes fait une performance nettement supérieure (13,16 de moyenne, contre 11,75 pour Chateaulin et 10,78 pour Angers). Meilleur niveau général,

meilleure préparation des candidats à ce type d'examen, indulgence des correcteurs, … fuites ?

Epreuve d'Accidents

3 exercices, valant respectivement 9, 5 et 6 points.

Pertinence R.A.S. Ce type de sujet atteste véritablement des réflexions sur les objectifs de formation en fonction des prérogatives de niveau.

Validité L'épreuve est très intéressante, car elle évalue tout à la fois les connaissances, leur application et les facultés d'analyse.

Rédaction R.A.S. Libellé clair, consignes données, barême annoncé.

Notation Le seul défaut qu'on pourrait voir dans ce sujet est le "saucissonnage" un peu excessif du barême. Mais il est rendu nécessaire par le souci d'harmoniser la correction sur les trois centres. Il aurait peut-être fallu ajouter un "note d'impression globale", entrant dans le calcul de la note finale.

Résultats : 10,33 de moyenne régionale, avec un moins bon résultat sur Angers (8,90 et 2 notes éliminatoires). Ce résultat appelle deux remarques :

- dès l'instant que l'on évalue des objectifs d'*analyse,* de niveau taxonomique plus élevé, le résultat est toujours moins bon, surtout si les candidats n'ont pas été préparé à réfléchir et analyser, mais simplement à digérer des *connaissances.*
- vue la qualité de l'épreuve, le résultat reflète donc en toute probabilité le niveau des N4 actuels en ce qui concerne les accidents : des connaissances, mais quelques difficultés d'analyse et de prise de décision.

Epreuve de Physique

20 items à 1 point (7 Vrai/Faux, 13 Choix Unique).

Pertinence Les objectifs de formation Niveau 4, en ce qui concerne la Physique, mériteraient d'être précisés. Le sujet est dans la lignée de ce qui se fait actuellement.

Validité L'épreuve est contestable, en revanche, au plan de la validité : 7 questions de type "Vrai/Faux" (à proscrire en évaluation sommative), et les 13 autres n'ont que trois propositions de réponse (33,33% de chances au hasard). Pas de point négatif. On peut donc prévoir que le hasard et les coups d'oeil périscopiques associés ont entraîné une nette surestimation du niveau réel des candidats.

3 items (items 3, 8 et 13) posaient problème : le corrigé était faux, ou l'item avait deux solutions (item 13) !

Ce sujet est, à nos yeux, l'exemple-type de l'évaluation telle qu'elle est pratiquée actuellement, dans laquelle on ne sait pas bien ce que la note obtenue signifie. Que vaut un candidat qui a la moyenne ?

Rédaction L'intitulé de l'épreuve ne figure pas sur la feuille ! Les consignes sont claires, mais les abréviations ne sont pas légendées.

Notation Les résultats illustrent parfaitement les problèmes de validité : moyenne régionale 14,92, pas de note éliminatoire.

Les candidats étaient donc tous excellents en Physique ?

Epreuve de Physiologie

2 questions classiques (l'ensemble sur 12 points), 2 items à choix multiples (respectivement 5, 2 points), et 1 item à choix unique (1 point).

Pertinence Quelques éléments, dans les items à choix multiples, paraissent bien au-delà de ce qui est exigible pour un Niveau 4. Même remarque que pour la Physique : les objectifs n'ont jamais été clairement définis pour l'Anatomie/Physiologie.

Validité Le principe de l'épreuve est intéressant : le niveau de base est évalué par deux questions classiques notées sur 12 points. Ensuite, la distinction entre les candidats moyens et très bons s'effectuent par des items à choix multiples, dont certaines affirmations sont un peu plus complexes.

Dans l'ensemble, seules les connaissances sont évaluées ici. Est-il nécessaire d'aller au-delà dans une épreuve de Physiologie ? A définir…

Rédaction R.A.S.

Notation Un aspect intéressant, pas souvent rencontré : dans les items à choix multiples, chaque proposition de réponse est notée. L'élève obtient 0,5 point s'il a coché à bon escient une réponse, et idem s'il ne l'a pas cochée et qu'il ne le fallait pas. Il n'y a donc aucune ambiguïté de correction. Comme cela a été montré dans le mémoire, les items à choix multiples sont très faciles à corriger s'ils sont conçus ainsi.

Résultats : moyenne régionale à 10,15 ; six notes éliminatoires. Ce nombres de notes éliminatoires n'est pas surprenant. L'équilibre de conception de l'épreuve implique une évaluation davantage critériée, et ceux qui n'ont pas atteint le niveau minimal sont détectés. Cependant, il y a peut-être eu un manque de fidélité, avec une correction plus

sévère à Chateaulin (5 des 6 notes éliminatoires) et plus favorable à Rennes (prestation très nettement meilleure que les autres).

Epreuve de Matériel

12 items de type oui/non (50% de hasard). 8 items orientés "mécanismes, fonctionnement", et 4 orientés "pratique courante".

Les résultats à ce questionnaire n'ont pas grande signification pour ceux qui y ont répondu correctement (effet du hasard peut-être très important), ni pour les mauvais scores (effet du hasard, orientation du questionnaire). Il n'y a pas lieu de pousser l'analyse plus loin, car l'essentiel de l'évaluation s'est fait à l'oral. L'écrit n'avait qu'une valeur indicative.

Résultats globaux : 12,09 de moyenne générale, avec une bonne homogénéité des trois centres. Les tentatives d'harmonisation de l'évaluation du matériel commencent à porter leurs fruits...

Epreuve de Législation

20 items à 1 point, tous à Choix Unique parmi cinq propositions. Mis à part les remarques concernant l'évaluation sur 20 questions de même poids et la signification d'une moyenne dans ces conditions, le sujet était bien conçu. Les candidats ont obtenu une moyenne régionale de 15,36, répartie de façon homogène dans les trois centres. Pas de note éliminatoire.

La formation à la Réglementation s'est améliorée et les candidats se préparent bien à cette épreuve.