Étude expérimentale de l'applicabilité de modèles[1] d'agrégation flous à l'étude de la charge mentale[2]

Eric Raufaste, Patrice Terrier
Laboratoire Travail et Cognition (UMR 5551 du CNRS)
MDR-UTM - 5, Allées A. Machado 31058 Toulouse Cedex
raufaste@univ-tlse2.fr ; terrier@univ-tlse2.fr

Michel Grabisch
Laboratoire LIP6 - UPMC
Case 169 / 4, place Jussieu
75 252 Paris CEDEX 05
Michel.Grabisch@lip6.fr

Jérôme Lang, Henri Prade
IRIT (UMR 5505 du CNRS)
Université Paul Sabatier - 31062 Toulouse Cedex
prade@irit.fr ; lang@irit.fr

RÉSUMÉ

La mesure de charge mentale subjective est utilisée pour évaluer la difficulté ressentie face à une situation de travail. La méthode NASA-TLX calcule une valeur globale par agrégation de 21 mesures obtenues relativement à six sources de charge mentale. Le modèle d'agrégation usuel est une moyenne des estimations élémentaires, pondérée par l'importance relative des six sources. La théorie des ensembles flous ayant développé des méthodes d'agrégation plus sophistiquées, les résultats présentés sont relatifs à l'évaluation de l'apport potentiel de trois modèles d'agrégation flous pour le calcul de la note globale de charge mentale : maximum pondéré, moyenne pondérée ordonnée et intégrale de Sugeno. L'étude prenait en référence la note fournie par la NASA-TLX et une mesure directe de la charge mentale subjective globale. Les résultats suggèrent que l'intégrale de Sugeno serait le modèle d'agrégation le plus puissant et offre la perspective d'une nouvelle méthode capturant mieux la charge mentale tout en n'utilisant que 7 mesures au lieu de 21, sous réserve que suffisamment de sujets soient disponibles.

MOTS-CLÉS
Charge Mentale Subjective, NASA-TLX, Ensembles Flous, Agrégation Multi-Critères.

1 POSITIONNEMENT DU PROBLEME

La mesure de charge mentale subjective est utilisée dans l'étude des situations de travail car elle permet d'obtenir des informations spécifiques sur les difficultés ressenties par les opérateurs. En effet, des indices plus objectifs, comme l'enregistrement du rythme cardiaque, ne traduisent pas nécessairement tout ce qui est ressenti. D'autre part certaines situations ne se prêtent pas à la mise en place des dispositifs relativement lourds requis pour le recueil des mesures objectives. C'est ainsi que plusieurs méthodes d'évaluation de la charge mentale subjective ont vu le jour dans le cadre d'études relatives à la navigation aérienne (Cooper-Harper, Cooper & Harter, 1969; SWAT, Reid & Nygren, 1988; NASA-TLX, Hart & Staveland, 1988). La dernière de ces méthodes, la NASA-TLX (Task Load Index, développée à la NASA) recompose une note globale par agrégation de mesures prises sur six dimensions de la charge mentale, supposées indépendantes et envisagées comme "sources" de charge mentale. Dans le présent article, nous appellerons "cotes" les mesures prises sur chacune des sources,

[1] Nous employons ici le terme de "modèle" au lieu de "opérateur", afin de ne pas interférer avec le sens particulier que prend habituellement ce dernier terme en psychologie ergonomique.

[2] Cette étude a été subventionnée par l'INRS, convention N° 5001053 "Application des ensembles flous à l'acquisition et au traitement de données issues d'évaluations subjectives de la charge de travail à dominante mentale". Nous remercions particulièrement Daniel Liévin et Michel Neboit, de l'INRS.

et "poids" les importances relatives des sources de charge mentale dans le calcul de la note agrégée. La NASA-TLX utilise le modèle de la moyenne pondérée, calculée à partir de 21 mesures prises sur les sujets : six cotes, relatives aux six sources de charge mentale qu' envisage cette méthode[3], et 15 comparaisons binaires d' importance entre les sources de charge prises deux à deux. Ces comparaisons binaires donnent une matrice d' où dérivent, pour chaque sujet, six poids correspondant à l' importance relative des six sources dans la charge globale.

Le présent travail aborde la question des mesures individuelles minimales qu' il est nécessaire de recueillir pour composer une note globale utilisable, ainsi que celle du modèle mathématique d' agrégation qu' il est opportun de mettre en œuvre. D' un point de vue formel, il existe trois types principaux de modèles de base pour l' agrégation : les modèles qui utilisent un "et" logique et qui donnent un score élevé si tous les scores sont élevés, les modèles qui utilisent un "ou" logique et donnent un score élevé dès qu' un des scores est élevé, et les modèles du type moyenne qui sont compensatoires et donnent un score intermédiaire. Ces modèles peuvent intégrer l' importance relative des critères dans l' agrégation de scores élémentaires en un score global.

L' objectif du présent travail est d' évaluer si des modèles mathématiques d' agrégation différents, plus qualitatifs ou requérant moins de mesures, pourraient remplacer avantageusement la moyenne pondérée dans l' étude de la charge mentale subjective. L' étude présentée ici n' a pas vocation à traiter de tous les opérateurs possibles, mais simplement d' évaluer si au moins certains d' entre eux pourraient offrir des avantages comparativement à la moyenne pondérée. Dans cette perspective, nous ne considérerons ici que trois modèles d' agrégation issus de la théorie des ensembles flous : le maximum pondéré (Dubois & Prade, 1986), la moyenne pondérée ordonnée ("Ordered Weighted Average", ou OWA, Yager, 1988) et l' intégrale de Sugeno (Sugeno, 1977). Deux autres méthodes seront prises en référence : la moyenne pondérée et la mesure directe d' une évaluation subjective globale unique. Dans ce dernier cas, l' agrégation n' est pas mathématique, mais directement réalisée par le système cognitif du sujet.

2 METHODE

L' idée générale était de générer des situations susceptibles d' induire des différences objectives de charge mentale chez les participants, puis de procéder à des mesures objectives et subjectives de charge mentale. Sous l' hypothèse que la charge mentale subjective doit refléter peu ou prou les différences objectives de charge mentale, on peut s' attendre à observer des différences subjectives en présence de différences objectives. Les modèles d' agrégation qui produisent des notes globales capables de capturer ces différences peuvent ensuite être considérés comme plus puissants que les modèles qui ne détectent pas ces différences.

2.1. La situation expérimentale

La situation expérimentale retenue était une double tâche comprenant (1) Une tâche principale, en l' occurrence un jeu de réflexion sur ordinateur; Afin de pouvoir évaluer la capacité des modèles à capturer les différences de charge mentale liées à la nature de la tâche, ainsi que les différences liées au niveau de difficulté pour une même tâche, deux tâches principales différentes ont été utilisées, l' une présentant un caractère dynamique, l' autre non. Pour chacune de ces tâches, deux niveaux de difficulté ont été proposés aux sujets; (2) Une tâche secondaire indépendante consistant à appuyer sur une touche le plus vite possible lorsque l' écran se met à clignoter. Après une phase de familiarisation avec la tâche secondaire présentée seule, les sujets fournissaient une série d' estimations destinées à servir de données d' entrée pour les différents algorithmes d' agrégation testés.

Les deux tâches principales

Le premier jeu retenu était le "démineur", un jeu populaire pour lequel il est possible de sélectionner des participants présentant différents niveaux d' expertise initiale. Le jeu se présente initialement sous la forme d' un rectangle de taille paramétrable et décomposé en cases carrées, grisées et en relief. Un certain nombre de "mines" sont disposées aléatoirement dans le damier, de sorte qu' au début du jeu chaque case peut être soit neutre, soit minée, sans que le joueur ne dispose d' aucune

[3] Les 6 sources envisagées sont : activité physique; activité mentale; contraintes temporelles; performances personnelles; satisfaction-frustration; effort.

information lui permettant de savoir lesquelles. Le but du jeu est de déminer toutes les cases non minées sans exploser sur une mine, et ce dans le minimum de temps. Le niveau de difficulté (novice ou expert) dépend du nombre de cases à déminer et du nombre de cases minées.

Le second jeu était le "tetris", jeu célèbre pour lequel il est possible de sélectionner des sujets présentant différents niveaux d' expertise initiale. Il se présente sous la forme d' une colonne dans laquelle des éléments de forme géométrique tombent verticalement à des vitesses augmentant avec le niveau de difficulté. Lorsqu' une pièce rencontre le fond ou une pièce déjà posée, elle s' arrête. Pendant la chute de la pièce, le joueur peut la déplacer sur l' axe horizontal au moyen des flèches gauche et droite. Il peut aussi provoquer une rotation de la pièce dans le sens anti-horaire (flèche "haut") afin d' optimiser l' emboîtement de la nouvelle pièce avec les pièces déjà posées au fond. Le joueur peut provoquer la chute immédiate d' une pièce en appuyant sur la touche "bas". Lorsqu' une ligne est pleine, elle disparaît. Le joueur libère ainsi de l' espace. Le joueur marque des points lorsqu' une pièce s' empile, et surtout lorsqu' une ligne disparaît. Le but du jeu est d' accumuler le plus de points possible avant la fin de la partie, qui survient lorsque l' empilement des pièces atteint le sommet de la colonne.

La tâche secondaire

La bordure du terrain de jeu, normalement grise, clignotait en rouge à intervalles réguliers (30s). Cette stimulation étant très saillante, le joueur ne pouvait éviter de la remarquer. La tâche secondaire du sujet consistait à arrêter le clignotement le plus vite possible, en appuyant sur la touche "Echap", en haut à gauche du clavier. Avant l' expérience, les sujets recevaient une phase de familiarisation avec la tâche. Cette familiarisation consistait en une série de dix essais où le participant réalisait la tâche secondaire seule. Elle était réalisée avant chaque nouvelle condition de tâche principale. Les participants recevaient pour consigne de laisser le doigt sur le bouton en permanence, ceci afin de réduire les pertes de temps parasites liées au déplacement de la main et à la recherche du bouton. Du point de vue de l' indépendance des deux tâches, au plan moteur, il faut noter que la tâche principale se pilote entièrement avec la main droite (souris pour le démineur, touches "haut", "bas", "droite" et "gauche" pour le tetris), tandis que la tâche secondaire n' utilise que la main gauche (touche "Echap" à gauche du clavier).

2.2. Les participants

Les participants à l' expérience (n = 48) étaient des étudiants bénévoles recrutés sur le campus de l' Université Toulouse-II. Trente participants étaient des femmes et dix-huit des hommes. Afin d' assurer une diversité des niveaux d' expertise, certains avaient déjà une expérience du démineur et/ou du tetris, tandis que d' autres n' en avaient aucune. Toutefois, ne pouvant nous baser sur les dires des sujets pour évaluer leur niveau d' expertise, nous avons préféré opérer la classification du niveau d' expertise en prenant comme critère la médiane des meilleures performances pour le tetris, et la réussite d' au moins une partie de niveau "novice" pour le démineur.

2.3. Les mesures recueillies

Mesures subjectives de charge mentale

Les mesures subjectives étaient de trois types : (1) des estimations d' importance des sources; (2) des estimations de charge sur chacune des sources (cotes); (3) une mesure globale de charge mentale.

_ En ce qui concerne les poids, la méthode de la NASA-TLX consiste à faire procéder à 15 comparaisons : les participants doivent déterminer, pour chacun des 15 appariements possibles de sources, quelle source contribue le plus à la charge mentale (choix forcé). A cet effet, un écran a été présenté aux participants pour chaque paire de sources. L' écran donnait la consigne ainsi qu' un rappel de la signification de chaque source. Le participant disposait de deux cases à cocher, initialement vides, et devait cocher la case correspondant à la source la plus importante. Lorsqu' une case avait été cochée, un bouton OK de confirmation apparaissait. Le participant pouvait changer son choix avant de confirmer.

_ En ce qui concerne les cotes, une question était posée pour chaque source, du type :

"PERFORMANCE. Comment pensez-vous que vous avez réussi à accomplir les buts de la tâche fixés par l' expérimentateur (ou fixés par vous-même) ? Dans quelle mesure êtes-vous satisfait(e) de votre performance dans l' accomplissement de ces buts ?"

Le recueil des estimations sur chaque source était réalisé au moyen d' un curseur déplaçable par la souris. Le curseur, encadré par des marqueurs sémantiques, retournait une valeur comprise entre 0 et 100. Ces valeurs numériques n' étaient pas connues des sujets qui donnaient ainsi une réponse analogique. Outre les cotes sur les 6 sources de charge envisagées par la NASA-TLX, nous avons ajouté une septième question afin de mesurer l' estimation globale de la charge mentale des sujets :

"ESTIMATION GLOBALE: Globalement, à quel degré jugez vous que la charge mentale liée à la tâche était importante ?

La réponse à cette question était donnée au moyen du même curseur que pour les cotes.

Mesures "objectives" de charge mentale

Dans le but de fournir des critères pour les calculs de qualité des modèles d' agrégation, six indices "objectifs" de charge mentale ont été recueillis. Trois indices directs concernaient la charge mentale sur la tâche principale, tandis que trois indices indirects concernaient la dégradation de performance sur la tâche secondaire.

_ Les indices directs étaient liés au temps de réflexion entre deux actions. Le premier indice était le temps moyen entre deux actions ("moyen T1"). Le deuxième indice était le temps de réflexion maximal calculé à partir des 10 temps de réflexion les plus longs ("maximum T1"). Le dernier indice était le temps de réflexion moyen sur les 100 dernières actions ("récent T1"). La justification pour le choix des deux derniers indices provient de divers travaux qui montrent que la formation des impressions (en particulier affectives) dépend de façon privilégiée du pic de sensations et de la sensation ressentie à la fin de la période d' exposition aux stimuli (Kahneman, Frederickson, Schreiber & Redelmeier, 1993).

_ Les indices indirects ont été calculés sur la base de la dégradation de performance à la tâche secondaire. Le mode de calcul de la dégradation était le suivant. Avant chaque session de jeu, le participant était exposé à la tâche secondaire seule pendant 10 essais. Dans cette phase, le programme générait un clignotement du terrain de jeu avec un intervalle de temps entre deux clignotements choisi aléatoirement entre 5 et 15 secondes. Le sujet arrêtait le clignotement en appuyant sur la touche "Echap" et l' ordinateur enregistrait le temps mis pour réagir au stimulus. Le temps de réponse moyen des 4 meilleurs essais a été pris comme base de calcul pour la performance sur la tâche non dégradée. Pendant la phase de jeu, l' ordinateur faisait clignoter le terrain de jeu à intervalles réguliers, toutes les 30 secondes, un délai suffisamment long pour que le sujet en train de jouer ne s' y attende pas. La dégradation était calculée par différence entre le temps de réponse de base (en tâche secondaire seule) et le temps de réponse à la tâche secondaire pendant la phase de jeu. Les trois indices indirects calculés étaient la dégradation moyenne sur l' ensemble des essais ("moyen T2"), la dégradation moyenne sur les 4 derniers essais ("récent T2") et la dégradation maximale calculée à partir des 4 dégradations les plus fortes ("maximum T2").

2.4. Déroulement de l'expérience

Chaque sujet commençait par lire la consigne générale. Ensuite, le sujet effectuait successivement quatre blocs de tâches. Chaque bloc se composait d' une familiarisation à la tâche secondaire seule suivie d' une phase de jeu avec double tâche. Après 15mn de jeu, le sujet recevait la consigne pour les mesures de charge mentale. Ensuite, le sujet répondait aux questions : les 21 questions classiques de la NASA-TLX, la mesure directe de charge mentale subjective globale, et enfin un écran de comparaison directe des importances des critères (non traité dans le présent article). Les quatre blocs correspondaient aux conditions expérimentales (tetris / démineur x facile / difficile). L' ordre de ces conditions était contrebalancé d' un sujet à l' autre, sachant toutefois qu' un même jeu était toujours présenté dans sa condition facile avant d' être présenté dans sa condition difficile.

2.5. Analyse

La note globale produite par la NASA-TLX a été évaluée par la méthode classique (moyenne pondérée). Les poids obtenus par cette méthode ont été transformés en rangs, ce qui a permis de calculer le max pondéré. Le calcul des intégrales de Sugeno a été réalisé sur la base des synergies entre les différentes coalitions de sources ({1}, {2}, …, {6}, {1, 2}, {1, 3}, … {2,3,4,5,6}). Les jeux de synergies étaient calculés, pour chaque condition, en fonction de sept indices : la note subjective globale, et les six indices objectifs. Il a donc été obtenu 4 x 7 jeux de synergies, donnant ensuite 7

notes globales différentes par condition. L' intégration de la contribution respective des sources à chaque coalition constitue une autre méthode de calcul de l' importance des critères, importance donnée par sa "valeur de Shapley" (Denneberg & Grabisch, 1999). Une moyenne pondérée a été calculée en prenant comme critère d' importance de chaque source sa valeur de Shapley, calculée avec la note subjective globale comme critère d' ajustement. Ce même jeu de pondération a aussi été utilisé pour calculer la moyenne pondérée ordonnée (OWA). Afin de tester la capacité des différents modèles à capturer les différences de tâches et de difficulté, nous avons réalisé des ANOVAs univariées à deux facteurs, Tâche x Difficulté, en prenant successivement comme variables dépendantes chacun des modèles d' agrégation considérés pour évaluation.

3 RÉSULTATS

Modèle d'agrégation	Démineur	Tetris	F	Sig.
Note subjective globale	61.12	59.48	0.38	NS
Moyenne pondérée (méthode NASA-TLX)	58.73	60.01	0.19	NS
Moyenne pondérée (par valeurs de Shapley)	53.0	55.5	0.79	NS
Max pondéré	55.40	60.64	6.49	p = .012
OWA (par valeurs de Shapley)	30.4	42.6	27.44	p < .001
Sugeno (d' après valeur subjective Globale)	36.1	48.3	27.92	p < .001
Sugeno (d' après indice moyen de T1)	52.8	68.0	37.00	p < .001
Sugeno (d' après indice maximum T1)	38.3	62.5	116.76	p < .001
Sugeno (d' après indice récent T1)	53.7	67.7	35.34	p < .001
Sugeno (d' après indice moyen de T2)	69.5	59.2	17.94	p < .001
Sugeno (d' après indice maximum T2)	65.6	49.9	42.89	p < .001
Sugeno (d' après indice récent T2)	69.7	62.7	7.75	p = .006

Tableau 1 : Sensibilité aux différences de tâches

Ce premier tableau montre que tous les opérateurs flous testés, et eux seuls capturent les différences entre tâches. Les seuils de signification sont calculés en bilatéral.

Modèle d'agrégation	Facile	Difficile	F	Sig.
Note subjective globale	57.14	63.39	4.01	p=.048
Moyenne pondérée (méthode NASA-TLX)	56.22	62.51	7.32	p=.008
Moyenne pondérée (par valeurs de Shapley)	52.2	56.4	2.62	NS
Max pondéré	56.74	59.38	1.72	NS
OWA (par valeurs de Shapley)	36.8	36.4	0.03	NS
Sugeno (d' après valeur subjective Globale)	43.2	41.2	0.71	NS
Sugeno (d' après indice moyen de T1)	56.2	64.7	12.49	p= .001
Sugeno (d' après indice maximum T1)	47.8	53.3	6.10	p=.014
Sugeno (d' après indice récent T1)	57.0	64.6	10.86	p < .001
Sugeno (d' après indice moyen de T2)	63.6	65.1	0.38	NS
Sugeno (d' après indice maximum T2)	56.4	59.0	1.13	NS
Sugeno (d' après indice récent T2)	65.4	66.9	0.36	NS

Tableau 2 : Sensibilité à la difficulté des tâches

Les degrés de signification du tableau précédent sont calculés en bilatéral. Ce tableau montre que si la NASA-TLX produit une bonne estimation, la note subjective globale et les intégrales de Sugeno fondées sur les indices objectifs de charge primaire détectent aussi les différences de difficulté.

Jusqu' à 39% de la variance sur l' intégrale de Sugeno sont expliqués par les conditions de tâche et de difficulté (10% pour l' intégrale basée sur la note subjective globale). C' est une excellente performance pour ce type d' analyse, performance que l' on peut rapporter à moins de 3% de variance de la NASA-TLX et de la note subjective globale expliqués par les différences inter- et intra-tâches.

Modèle d' agrégation	Variance expliquée (R^2 ajusté)
Note subjective globale	1.0%
Moyenne pondérée (méthode NASA-TLX)	2.4%
Moyenne pondérée (par valeurs de Shapley)	< 1%
Max pondéré	3.3%
OWA (par valeurs de Shapley)	13.5%
Sugeno (Global)	12.0%
Sugeno (Mean T1)	22.1%
Sugeno (Max T1)	41.0%
Sugeno (Recent T1)	20.4%
Sugeno (Mean T2)	7.5%
Sugeno (Max T2)	17.8%
Sugeno (Recent T2)	3.3%

Tableau 3 : Pourcentage de la variance des notes globales expliquées par les facteurs Tâche x difficulté

4 DISCUSSION

En première analyse, les résultats sont très encourageants dans la perspective d' appliquer les modèles d' agrégation flous à l' étude de la charge mentale subjective : seuls les modèles flous capturent les différences entre tâches. Le max pondéré, en particulier, est intéressant à cet égard car il ne requiert pas plus de sujets que la NASA-TLX. Nous pouvons donc d' ores et déjà suggérer que les études utilisant la NASA-TLX complètent l' utilisation de la moyenne pondérée par l' utilisation d' un max pondéré. L' intégrale de Sugeno possède une propriété qui pourrait s' avérer intéressante dans certaines études : elle permet de combiner des mesures subjectives avec des indices objectifs, ce qui donne, dans notre étude, les meilleurs résultats. Tout ceci soulève une question d' ordre plus théorique : si des modèles mathématiques qualitatifs capturent mieux les différences de charge mentale qu' un modèle d' agrégation quantitatif comme la moyenne pondérée, il se pourrait que les mécanismes cognitifs sous-jacents à la production des estimations subjectives soient eux-mêmes qualitatifs.

5 BIBLIOGRAPHIE

Cooper, R.P., & Harper, Jr (1969). *The use of pilot rating in the evaluation of aircraft handling qualities* (Report NASA TN-D-5153), Moffett Field, CA, Ames Research Center, National Aeronautics and Space Administration.

Denneberg D., & Grabisch, M. (1999). Interaction transform of set functions over a finite set. *Information Sciences, 121*, 149-170.

Dubois, D., & Prade, H. (1986) Weighted minimum and maximum operations in fuzzy set theory. *Information Sciences, 39*, 205-210.

Hart, S.G. & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P.A. Hancock and N. Meshkati (Eds.), *Human Mental Workload* (p. 139-185). Amsterdam, The Netherlands: North Holland Press.

Kahneman, D., Fredrickson, B.L., Schreiber, C.A., & Redelmeier, D.A. (1993). When more pain is preferred to less: adding a better end. *Psychological Science*, *4*, 6, 401-405.

Reid, G.B., & Nygren, T.E. (1988). The Subjective Workload Assessment Technique: A Scaling Procedure for Measuring Mental Workload. In P. A. Hancock and N. Meshkati (Eds.), *Human Mental Workload* (p. 185-218). Amsterdam, The Netherlands: North Holland Press.

Sugeno, M. (1977). Fuzzy measures and Fuzzy Integrals: A survey. In M.M. Gupta, G.N. Saridis, and B.R. Gaines (Eds.), *Fuzzy automata and decision processes* (pp 89-102), North-Holland.

Yager, R.R. (1988). On ordered weighted averaging aggregation operators in multicriteria decision-making. *IEEE Trans, Systems, Man & Cybernetics*, 18, 183-190.