APPLICATION OF THE CHOQUET INTEGRAL TO
SUBJECTIVE MENTAL WORKLOAD EVALUATION

M. Grabisch\textsuperscript{1} H. Prade\textsuperscript{2} É. Raufaste\textsuperscript{3} P. Terrier\textsuperscript{3}

\textsuperscript{1} Université Paris I Panthéon-Sorbonne ; LIP6
\textsuperscript{2} Université Toulouse-III Paul Sabatier ; IRIT
\textsuperscript{3} Université Toulouse-II Le Mirail, LTC

Abstract: We describe a methodology based on Choquet integral to build general models of subjective evaluation. The model is compared to the NASA-TLX, which is of current use in assessment of workload, in an experiment based on two classical computer games. Copyright ©2005 IFAC.

Keywords: subjective evaluation, Choquet integral, importance index, interaction, workload assessment, NASA-TLX

1 Introduction: the NASA-TLX method

A very important topic for ergonomics and human factors is measuring the mental workload associated with the situations under study. For example, if one is to compare two human-computer interfaces, the device that produces a lower level of subjective mental workload is generally preferred. Among the most widely used methods are the National Aeronautics and Space Administration-Task Load Index (NASA-TLX, Hart and Staveland, 1988) and the Subjective Workload Assessment Technique (SWAT, Reid and Nygren, 1988). Contrary to the Cooper-Harper scale (Cooper and Harper, 1969), where operators are asked a single workload estimate, the NASA-TLX and SWAT both assume that workload is a multidimensional concept, with six and three workload sources respectively.

Supposedly, a multidimensional approach provides a richer and less biased picture of workload. In the general case, a multidimensional approach makes difficult to directly compare various settings (work situations, interfaces, and so on).

The NASA-TLX rating procedure provides an overall workload score based on a weighted average of the ratings on six subscales: Mental Demands, Physical Demands, Temporal Demands, Performance, Effort, and Frustration. Depending on situations, the various sources may differently contribute to the operator’s subjective workload. Taking into account the relative weights of the sources first requires obtaining a measure of their relative importance. For example, during the standard NASA-TLX procedure participants provide the 15 possible pair-wise comparisons of the six subscales. In each comparison, subjects select the source that contributed to the workload more than the other. Each source receives one point for each comparison where it was deemed to contribute more. The relative weight of a source is then given by the sum of those points, divided by 15 for normalization purposes. In order to avoid confusion, in this paper we will call "rating" the value provided for each source, and "weight" the relative importance of a source. The "score" will denote the global score provided by an aggregation method.

After information about ratings and weights is collected, the question is to choose the aggregation method. The NASA-TLX makes use of a classical weighted mean, which simply sums the products of ratings by their normalized weights ($\Sigma w_i = 1$). Thus, noting $x_i$ the rating about the $i^{th}$ source and $a_i$ the relative importance of the same source, the subjective workload $SW$ in the NASA-TLX method is provided by

$$SW = \sum_{i=1}^{6} w_i a_i$$

where $w_i$ and $a_i$ respectively denote the weight and rating associated with the $i^{th}$ workload source.

Although some previous studies applied fuzzy sets theory to workload measurement by means of linguistic terms (e.g., Chen, Jung, and Peacock, 1994; Liou and Wang, 1994), another question is addressed here: what is the best way of aggregating data about the six NASA-TLX workload sources into a single workload value that enables direct comparison of different settings?

Is weighted average the best model of aggregation? Indeed, it is easy to compute and familiar to most researchers. On the other hand, despite its apparent simplicity, the weighted average model is built upon several strong mathematical assumptions that are not
necessarily verified in workload assessment. For example, it requires independence between ratings and weights. This condition could be attained, for example, by having operators providing the ratings and external experts providing the weights. Unfortunately, in the standard NASA-TLX procedure, each operator provides both the ratings and weights. Second, weighted average does not allow taking into account interactions between sources. Is it a reasonable choice to neglect dependencies and interactions between sources of workload? By neglecting such interaction effects, a weighted average model might induce measurement biases.

As soon as a measure can be used as a learning criterion, the Choquet integral provides a potential solution to those problems because it enables computing weights from an adjustment criterion in a mathematically sound manner.

2 The Choquet integral for multi-criteria decision making

(For a detailed presentation, see Grabisch, Duchêne, Lino, and Perny, 2002; Grabisch, 2003; Grabisch and Labreuche, 2004). We present here the necessary material for introducing our model based on Choquet integral. Let $N = \{1, \ldots, n\}$ be the index set of criteria.

**Definition 1** A capacity (Choquet, 1953) or fuzzy measure (Sugeno, 1974) $\mu$ on $N$ is a function $\mu : \mathcal{P}(N) \rightarrow [0,1]$, satisfying the following axioms.

(i) $\mu(\emptyset) = 0$.

(ii) $A \subseteq B \subseteq N$ implies $\mu(A) \leq \mu(B)$.

We will assume in addition $\mu(N) = 1$ (normalized capacity). For any $A \subseteq N$, $\mu(A)$ represents the importance of the coalition $A$ of criteria for making decision. The capacity is additive if it is a probability measure.

**Definition 2** Let $\mu$ be a capacity on $N$. The Choquet integral of a function $f : N \rightarrow \mathbb{R}_+$ with respect to $\mu$ is defined by

$$C_{\mu}(f) := \sum_{i=1}^{n} (f^{(i)} - f^{(i|-1)}) \mu(A(i)),$$  \hspace{1cm} (1)

where we have written for simplicity $f^{(i)} := f(i)$, and $A(i)$ indicates that the indices have been permuted so that $0 \leq f^{(1)} \leq \cdots \leq f^{(n)}$, and $A(i) := \{i\}, \ldots, \{n\}$, and $f^{(0)} = 0$.

The Choquet integral of $f$, considered as an alternative to be evaluated, is the overall score of the alternative considering the importance of coalitions of criteria. An important property of the Choquet integral is that

$$C_{\mu}(1_A) = \mu(A)$$

where $1_A$ is the characteristic function of $A$, $A \subseteq N$. This gives a clear interpretation of the quantity $\mu(A)$. Another property worth to be mentioned is that when $\mu$ is additive, then the Choquet integral reduces to a weighted average $\sum_i w_i f_i$, with $w_i = \mu(\{i\})$.

Since the definition of $\mu$ involves $2^n$ values, which may cause some interpretation problem in terms of the importance of criteria, a convenient concept is the one of Shapley index (Shapley, 1953), coming from cooperative game theory. For any criterion $i \in N$, the Shapley index of $i$ is defined by:

$$\phi_i := \sum_{K \subseteq N \setminus i} \frac{(n - |K| - 1)!|K|!}{n!} [\mu(K \cup \{i\}) - \mu(K)].$$  \hspace{1cm} (2)

Roughly speaking, the Shapley index $\phi_i$ computes the average contribution of criterion $i$ in all coalitions, the average being weighted by a coefficient taking into account the cardinality of the coalition. In this sense, it can be taken as definition of the average importance or average contribution of a single criterion for the decision process. The Shapley index satisfies $\sum_{i=1}^{n} \phi_i = \mu(N) = 1$, so that the sum of importance degrees is a constant. The idea to use the Shapley index for multicriteria decision making is due to Murofushi (1992).

Another important topic is the notion of interaction among two criteria, as proposed originally by Murofushi and Soneda (1993).

$$I_{ij} := \sum_{K \subseteq N \setminus \{i, j\}} \frac{(n - |K| - 2)!|K|!}{(n - 1)!} [\mu(K \cup \{i, j\}) - \mu(K \cup \{i\}) - \mu(K \cup \{j\}) + \mu(K)].$$  \hspace{1cm} (3)

A positive interaction $I_{ij}$ occurs whenever criteria $i, j$ are complementary, i.e., the satisfaction of both is necessary to get overall satisfaction (the score of $i$ and $j$ are aggregated conjunctively). On the contrary, if it is sufficient to satisfy only $i$ or $j$, then $i$ and $j$ are substitutive, and $I_{ij} < 0$ (the score of $i$ and $j$ are aggregated disjunctively).

Later, Grabisch has generalized this notion to any number of criteria, leading to the following definition of the interaction index (Grabisch, 1997), defined for all coalitions (including the empty one), which is, $\forall A \subseteq N$:

$$I(A) := \sum_{K \subseteq N \setminus A} \frac{(n - k - a)!k!}{(n - a + 1)!} \sum_{B \subseteq A} (-1)^{|A| - |B|} \mu(K \cup B),$$  \hspace{1cm} (4)

where $k := |K|$ and $a := |A|$. Note that $I(\{i\}) = \phi_i$, and $I(\{i, j\}) = I_{ij}$. Also, it is easy to show that for an additive measure, $I(A) = 0$ whenever $|A| > 1$, and $\phi_i = \mu(\{i\})$. It is interesting to note that giving $I(A)$ for all $A \subseteq N$ allows for recovery of the capacity $\mu$: $I$ is merely another representation of $\mu$ (for details, see Grabisch, 1997).

**Definition 3** A capacity $\mu$ is said to be $k$-additive if its interaction transform satisfies $I(A) = 0$ for any $A$
such that $|A| > k$, and there exists at least one subset $A$ of $N$ of exactly $k$ elements such that $I(A) \neq 0$.

Thus, $k$-additive capacities can be represented by a limited set of coefficients, at most $\sum_{i=1}^{k} \binom{n}{i}$ coefficients. A 1-additive capacity is an additive capacity.

In practice, 2-additive capacities are a good compromise between flexibility and complexity. In this case, the Choquet integral can be rewritten as follows:

$$
C_{\mu}(a) = \sum_{i,j>0} (a_i \land a_j) I_{ij} + \sum_{i,j<0} (a_i \lor a_j) I_{ij}
+ \sum_{i=1}^{n} a_i (\phi_i - \frac{1}{2} \sum_{j \neq i} |I_{ij}|), \forall a \in [0,1]^n, \quad (5)
$$

with the property that $\phi_i - \frac{1}{2} \sum_{j \neq i} |I_{ij}| \geq 0$ for all $i$. This formula shows clearly the disjunctive and conjunctive effects of negative and positive interaction between criteria. Note also that the Shapley index lies in the linear part, and that the above formula is in fact a convex sum.

3 Application to workload assessment

The above multicriteria decision model can be readily applied to workload assessment, and to any subjective evaluation problem, and can be used instead of the NASA-TLX: it suffices to consider as set $N$ of criteria the set of 6 ratings used in the NASA-TLX. Then provided $\mu$ is known, one can calculate the overall score (subjective workload $SW$) of any vector $f$ of scores on criteria, namely

$$
SW(a) = C_{\mu}(a) \tag{6}
$$

where $a = (a_1, \ldots, a_6)$ is a vector of ratings.

The problem which remains to be solved is the identification of $\mu$, that is, $2^n - 2$ coefficients, or less if a $k$-additive capacity is used. Contrarily to the NASA-TLX, these coefficients are not elicited from the subjects (this would be too difficult, and most probably meaningless), but determined as an optimal solution of an error minimization criterion. Having it working requires to know the overall evaluation for each datum, but it is usually asked in the NASA-TLX. The error criterion chosen is the quadratic error between desired output (the subjective workload $SW$ given by the subjects) and the output of the model. Several methods exist for the optimization of this criterion, we use here quadratic programming (QUAD) and a heuristic suboptimal method (HLMS) (see details in Grabisch and Roubens, 2000).

Once the model $\mu$ identified, it can be mainly applied at two aims:

- analysis of the reasons of the workload in the given situation: Computing the Shapley index gives the relative contribution of each of the 6 ratings into the workload, and computing the interaction index for all pairs of ratings permits to detect which pair of criteria are substitutive or complementary. This is of invaluable aid for understanding how the 6 criteria intervene in the overall workload.

- Formula (6) provides another mean to compute $SW$. Its discriminative power in distinguishing different situations can be tested and compared to other methods of computing $SW$, including the classical NASA-TLX.

4 The experiment

We applied the above methodology to the assessment of workload in two well known computer games, Tetris and the minesweeper.

Basically the test consisted in placing participants in situations presenting objective differences in work demands, in order to test which of the various aggregation operators are the most sensitive to those differences. Differences in demands were manipulated by introducing two tasks on one hand, and two versions of each task (easy and difficult) on other hand. The primary tasks were two computer games, Tetris and minesweeper. The former is dynamic (the situation evolves rapidly even if the player does not act) whereas the later is not (the situation changes only in response to the player actions). Since each of the two primary tasks had two difficulty levels, each participant played in four conditions.

In addition, participants had to fulfill a secondary task in parallel to the primary tasks: they had to press the “Esc” key as soon as possible when the background of the screen started blinking. Response times to the secondary tasks were recorded. As it is classical in cognitive psychology, degradation of response times to the secondary task can be used to estimate the overall mental workload. At the end of each experimental condition, participants filled a subjective workload questionnaire.

Participants. 49 students (15 men, 34 women) from the University of Toulouse-II volunteered. All had already played Tetris and minesweeper before the experiment.

Experimental tasks. The first task was the minesweeper, a popular game distributed with Microsoft Windows\textsuperscript{TM}. It initially appears as an array of grey cells. A number of invisibles mines are randomly placed throughout the array. The player can uncover a cell by left-clicking on it. If the cell contains a mine, the game is lost. If it does not, the number of mines in the adjacent cells is shown within the clicked cell. The purpose of the game is to uncover every cell that contains no mine, without exploding, as fast as
possible. The level of difficulty depends on the size of the array and on the number of mines that are placed in it.

The second game was "Tetris". It appears under the form of a column from the top of which geometrical shapes vertically fall. The speed of the fall increases with difficulty levels. When a shape reaches the floor or a previously fallen shape, it stops. In the meanwhile, the player can use the keyboard to move the shape horizontally, rotate it, or make it fall instantaneously. When a horizontal line is full, it disappears, increasing the score. The purpose of the game is to reach a score as high as possible. The game ends when the pile of fallen shapes reaches the top of the column.

The grey border of the board blinked red every 30s, a very salient stimulation. The participant had to stop the blink as soon as possible by pushing the "Esc" key, on the upper left corner of the keyboard. At the beginning of each condition, participants started by the familiarization task alone. In order to avoid irrelevant lags participants were told to permanently leave their finger on the "Esc" key. The primary task only required the right hand, and the secondary task the left hand.

**Procedure.** After reading instructions, the participant proceeded with four blocks of tasks corresponding to the experimental conditions defined by the crossing of two within factors: Game (Tetris / minesweeper) x Difficulty (easy / difficult). The order of the conditions was randomized. For a given game, the difficult condition always appeared after the easy condition. Each block started with a familiarization to the secondary task, followed by an eight-minute experimental phase (game plus secondary task). Then participant read instructions about workload measurement and answered a translation in French of the NASA-TLX questionnaire (Human Systems Information Analysis Center, 2004). Finally participants answered an overall workload question (see rationale in the next section).

**Measurements.** In accordance with the NASA-TLX procedure, participants provided 6 ratings and 15 pairwise comparisons of importance. These ratings are: mental demand ($r_1$), temporal demand ($r_2$), physical demand ($r_3$), effort ($r_4$), performance ($r_5$) and level of frustration ($r_6$). The ratings were produced using a sliding cursor whose position was converted into a 0 to 100 integer. Two measurements served here as adjustment criteria for the capacity identification algorithm. The first criterion $c_1$ was an overall subjective rating, i.e., a direct subjective measure of the overall workload taken on scales of the same kind used for the NASA-TLX. The second criterion $c_2$ was the mean of the four worst reaction times to the secondary task.

**Identification of the model.** Two models $\mu_1, \mu_2$ were built using the identification procedure as explained in Sec. 2, these models corresponding to criteria $c_1$ and $c_2$ respectively. We both applied QUAD and HLMS on the various data set, and used also 2-additive capacities or general capacities. Using the two models, we were able to compute two workload scores:

$$SW_1(a) = C_{\mu_1}(a), \quad SW_2(a) = C_{\mu_2}(a).$$

Besides, for each model $\mu_1, \mu_2$, an analysis was done, computing the Shapley indices and the interaction indices, to get a behavioural interpretation of the models. The Shapley indices were also used as weights in the weighted average models; they provide in a sense a “best” linear approximation of the Choquet integral model.

**Statistical Analyses.** The sensitivity of the aggregation models was assessed using repeated-measure ANOVAs with two within-factors, Task (Tetris vs. Minesweeper), and Difficulty (Easy vs. Difficult). Because there was no expectation about interactions between Task and Difficulty, they could not teach us anything about the sensitivity of the models and they are not reported. With regard to the topic of this paper, it was crucial to check the effect size associated with the differences of means between experimental conditions. Thus, Cohen's $d$ values (Cohen, 1988) are reported. According to the widely accepted conventions proposed by Cohen, a standardized effect is deemed negligible when it is below .20, "small" in the range [.20, .49], "medium" in the range [.50, .79], and "large" above. Because we know that differences actually exist between experimental conditions, effect sizes reflect the capacity of aggregation operators to detect those differences.

5 Experimental results

5.1 Analysis of the model

We give the result of analysis of the model for criterion $c_1$ on the whole data set, which was the most significative. The following table give error modeling (sum of squared errors) for the respective identification methods.

<table>
<thead>
<tr>
<th></th>
<th>QUAD</th>
<th>QUAD 2-additive</th>
<th>HLMS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.707</td>
<td>2.972</td>
<td>3.096</td>
</tr>
</tbody>
</table>

Performances are close to each other, although QUAD is clearly the best, since optimal and using a general capacity. However, given the number of data and of criteria, we think that the HLMS model should be the best in terms of interpretation (no overfitting). Hence we restrict to the analysis of $\mu_1$ given by this method (results on Shapley and interaction indices are anyway similar in all three methods). The following matrix gives the Shapley indices $\phi_1, \ldots, \phi_6$ (on the
diagonal, in bold face), and the interaction indices $I_{ij}$, $i, j = 1, \ldots, 6, i \neq j$.

<table>
<thead>
<tr>
<th></th>
<th>$r_1$</th>
<th>$r_2$</th>
<th>$r_3$</th>
<th>$r_4$</th>
<th>$r_5$</th>
<th>$r_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_1$</td>
<td>2.29</td>
<td>-0.13</td>
<td>-0.00</td>
<td>0.062</td>
<td>-0.059</td>
<td>-0.027</td>
</tr>
<tr>
<td>$r_2$</td>
<td>0.67</td>
<td>-0.01</td>
<td>-0.04</td>
<td>-0.027</td>
<td>0.054</td>
<td>0.029</td>
</tr>
<tr>
<td>$r_3$</td>
<td>0.59</td>
<td>-0.025</td>
<td>0.054</td>
<td>0.027</td>
<td>0.029</td>
<td></td>
</tr>
<tr>
<td>$r_4$</td>
<td>1.1</td>
<td>0.072</td>
<td>0.027</td>
<td>0.027</td>
<td>0.029</td>
<td></td>
</tr>
<tr>
<td>$r_5$</td>
<td>0.38</td>
<td>-0.040</td>
<td>0.027</td>
<td>0.027</td>
<td>0.029</td>
<td></td>
</tr>
<tr>
<td>$r_6$</td>
<td>0.97</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The Shapley indices show that "mental demand" is by far the most important criterion, while "performance" and "physical demand" are the least significant. An analysis of the interaction indices shows that most substitutive criteria are mental and temporal demands, also temporal demand and frustration, while the most complementary criteria are effort and performance, and mental demand and effort. Also, a favor effect for "temporal demand" was detected, meaning that an overall workload is felt as soon as there exists a temporal demand.

5.2 Statistical analysis

Table 1 on the next page reports results about the sensitivity to tasks differences (i.e., minesweeper vs. Tetris). Tetris induced more workload than the minesweeper. Descriptively, all methods found this effect. However, some methods were statistically more sensitive than others. The overall subjective measure was barely significant, with a small effect size. The NASA-TLX performed only slightly better. Large effects were obtained by the Choquet integral and the weighted mean computed from the capacities based on the secondary task response times.

Sensitivity to difficulty Differences. Unsurprisingly, easy conditions systematically induced less workload than difficult conditions of the same tasks. Descriptively, all methods found this effect although some methods were more sensitive than others. The overall rating was the best at detecting difficulty differences, with medium effect sizes for both Tetris and minesweeper tasks. None of the aggregation models, including the NASA-TLX, could produce better than little effects (yet significant in many cases). The weighted mean could generally detect effects for the two games.

6 Discussion

The analysis of the model $\mu_1$ shows clear conclusions for the relative importance of ratings (namely, mental demand is the most significant, and physical demand and performance are not relevant), and for their interaction. Most notably, it was shown that a workload was felt by the subject once a temporal demand was present.

Concerning comparisons of situations, in conditions where comparison is easy (e.g., comparing two conditions of a same task, no aggregation model including the NASA-TLX could outperform the participants’ direct subjective estimate. But when comparison is not easy for a human (e.g., comparing the workload in two different tasks), the direct subjective measure does not seem to perform efficiently. This is where aggregation operator become relevant. In that condition, the Choquet integral performed fairly well since the effect size were clearly better than the effect size derived from the global rating or the NASA-TLX. The fact that the best effect sizes were those obtained with the RTs as adjustment criterion confirms another strength of the approach we propose : neither the NASA-TLX nor the direct subjective measure can integrate such objective and external data in the integration, even though it is relevant to do so. Interestingly, it can be remarked that the weighted means computed with Shapley values as weights outperformed the Choquet integral. This suggest that in this case interactions between coalitions were not relevant, or even constitute noise. As suggested by a referee, it will be interesting in the future to investigate tasks where interactions between workload sources loom larger.

References


Table 1: Statistical capacity of operators to detect task differences.

<table>
<thead>
<tr>
<th>Model</th>
<th>Minesweeper mean (SE)</th>
<th>Tetris mean (SE)</th>
<th>Effect size (d)</th>
<th>F (1,48)</th>
<th>Sig (p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global rating</td>
<td>55.0 (2.9)</td>
<td>62.0 (2.6)</td>
<td>0.30</td>
<td>4.47</td>
<td>.040</td>
</tr>
<tr>
<td>NASA-TLX</td>
<td>55.4 (1.8)</td>
<td>62.7 (2.3)</td>
<td>0.39</td>
<td>7.42</td>
<td>.009</td>
</tr>
<tr>
<td>Choquet on global measure</td>
<td>53.8 (2.3)</td>
<td>62.0 (2.5)</td>
<td>0.47</td>
<td>10.63</td>
<td>.002</td>
</tr>
<tr>
<td>Choquet on response time</td>
<td>36.2 (1.8)</td>
<td>50.4 (2.0)</td>
<td>1.02</td>
<td>50.73</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Weighted Mean on global measure</td>
<td>49.7 (1.9)</td>
<td>58.4 (2.3)</td>
<td>0.56</td>
<td>15.19</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Weighted Mean on response time</td>
<td>40.7 (1.8)</td>
<td>55.3 (2.1)</td>
<td>1.01</td>
<td>50.07</td>
<td>&lt; .001</td>
</tr>
</tbody>
</table>

Values were aggregated on a 0-100 scale, where 0 represents no workload felt and 100 the worst workload possible.

SE stands for Standard Error.

Effect sizes represent the standardized mean difference between workload estimates obtained for Tetris and Minesweeper tasks. Because the two tasks were actually different, higher effect sizes represent a better discrimination. Conventionally, an effect is deemed “small” when \( .20 \leq d \leq .50 \), “medium” when \( .50 \leq d \leq .80 \), and “big” when \( d \geq .80 \).

\( F(1,48) \) is the usual value of Fisher’s \( F \) in ANOVAs, computed as a signal to noise ratio. \((1, 48)\) means that there was 1 degree of freedom in the comparison (2 groups minus 1) and 48 degrees of freedom for the noise. This information is provided for experimentation-oriented readers.

Sig(\( p \)) means that significance is evaluated by the probability \( p \) that this effect size has occurred by the genuine effect of randomness. Conventionally, a result is deemed significant when \( p < .05 \).


