Math. Japonica **52**, No. 3(2000), 469–512

FUNCTIONAL EQUATIONS IN THE BEHAVIORAL SCIENCES[1]

JÁNOS ACZÉL*, JEAN-CLAUDE FALMAGNE** AND R.DUNCAN LUCE**

Received April 4, 2000

ABSTRACT. Functional equations are useful in the experimental sciences because they offer a tool for narrowing the possible models for a phenomenon. A model can be formulated by one or more not very restrictive equations, which when paired with an empirical or logical constraint of a general character, lead—via functional equation techniques—to precise quantitative relationships. The article reviews various applications of functional equations in some areas of the behavioral sciences such as sensory psychology (psychophysics), utility theory under uncertainty, and aggregation of inputs and outputs in an economic or social context. We also provide enough basic material on functional equations to make this review self contained.

1. INTRODUCTION

1.1 A Behavioral Example

Functional equations arise in the sciences because they allow researchers to formulate mathematical models in general terms, via some not very restrictive equations that only stipulate basic properties of functions appearing in these equations, without postulating the exact forms of such functions. However, the data or the experimental situation itself may exhibit regularities or symmetries that can be captured by some other equation(s) involving the same functions, thereby constraining their forms and specifying the model.

We begin with an example dating back to the 19th century. In a famous study, the physicist Plateau[2] gave a pair of painted disks—one white, one black—to each of eight artists and asked them to return to their studios and paint a grey disk appearing midway between the two. According to Plateau, the eight resulting grey disks were virtually identical. A possible formalization of such data is as follows (cf. Falmagne, 1985). Label each disk by its luminance in conventional units (lux). Denote by $M(x,y)$ the luminance of a disk appearing midway between the disks x and y, with M in the same units as x and by. Plateau's data can then be formalized by the homogeneity equation

$$M(\lambda x, \lambda y) = \lambda M(x,y) \qquad (\lambda, x, y > 0), \tag{1.1}$$

where the value of λ reflects the differing conditions of illumination. (Realistically, the domain of M should be restricted to a suitable positive region near the origin. Here and also later in this paper, we sometimes simplify the presentation and assume that the relevant functions are defined on idealized domains such as $]0,\infty[$. Usually, such idealizations have no

2000 *Mathematics Subject Classification.* Primary: 39B22, 91A30, 91B16, 91E30. Secondary: 26A24, 26A48, 26A51, 39B12, 39B72, 91C05.

Key words and phrases. Functional equations. Cauchy and Pexider equations. Invariance. Weber's law and generalizations. Fechner-Thurstone model. Gain control model. Joint receipt, segregation Separability. Utility and value functions.

[1]We thank Chris Doble for his comments on the first draft of this paper.

[2]The same Plateau (1801-1883) whose name was given to the classical problem of determining the minimal surface with a given twisted curve as its boundary.

substantial impact on the results.) A conceivable mechanism for the operation performed by the artists in Plateau's experiment is that the grey disk results from some kind of mental averaging of the values of the two disks in the pair. This averaging, however, is not necessarily carried out in the lux scale. We can suppose, for instance, that there is some sensory scale u mapping the lux scale into the reals, such that

$$u[M(x,y)] = \frac{u(x)+u(y)}{2}. \tag{1.2}$$

A function M satisfying this equation for some continuous strictly monotonic function u is called a *quasiarithmetic mean*. As mentioned, we assume that the scale u is defined on $\mathbb{R}_+ :=]0,\infty[$. Combining (1.1) and (1.2) leads to the functional equation

$$u^{-1}\left[\frac{u(\lambda x)+u(\lambda y)}{2}\right] = \lambda u^{-1}\left[\frac{u(x)+u(y)}{2}\right] \qquad (\lambda,x,y>0), \tag{1.3}$$

which has only two families of solutions for the function u. We will solve a more general equation, (2.20), in Section 2.2.1. Here we sketch the solution process for (1.3) to illustrate, by way of introduction, some functional equations methods in a simple case. We write $I := u(\mathbb{R}_+)$, an open interval, and define the function $j : I \times \mathbb{R}_+ \to I$, with $j_\lambda(s) = j(s,\lambda)$, by the equation

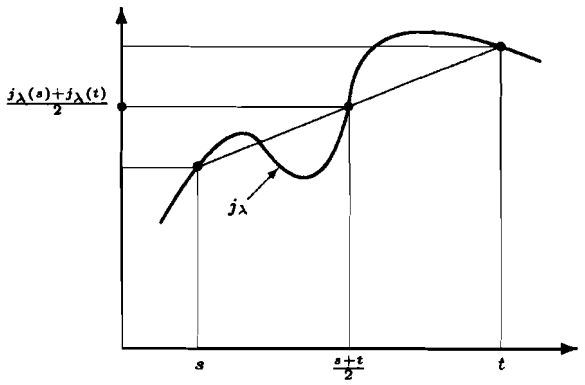
$$j_\lambda(s) = u[\lambda u^{-1}(s)]. \tag{1.4}$$

Note that j_λ is continuous and nonconstant in s. Applying u on both sides of (1.3) and rewriting the result in terms of j_λ, with $s = u(x)$ and $t = u(y)$, we get

$$j_\lambda\left(\frac{s+t}{2}\right) = \frac{j_\lambda(s)+j_\lambda(t)}{2} \qquad (s,t\in I,\lambda>0). \tag{1.5}$$

For any fixed λ, (1.5) is a *Jensen equation* (the equality case of Jensen's inequality). This equation is a particular case of Pexider's equation that will be solved in Section 2.1.4. The property expressed by (1.5) is immediately clear, namely: the midpoint of the two points $(s, j_\lambda(s))$ and $(t, j_\lambda(t))$ also lies on the graph of j_λ (cf. Fig. 1).

FIGURE 1. Jensen's equation.



Iterating this observation for s, $\frac{s+t}{2}$, for $\frac{s+t}{2}$, t, etc. we understand that the graphs of continuous functions satisfying (1.5) for some fixed λ are straight lines segments. Thus, the

general continuous solution of (1.5) is

$$j_\lambda(t) = m(\lambda)t + n(\lambda) \qquad (m(\lambda) \neq 0).$$

With (1.4) we obtain

$$u(\lambda x) = m(\lambda)u(x) + n(\lambda)$$

[cf. (2.23)]. Subtracting from this equation its particular case $x = 1$ and defining

$$l(x) = u(x) - u(1), \tag{1.6}$$

we get

$$l(\lambda x) = m(\lambda)l(x) + l(\lambda) \qquad (\lambda, x > 0). \tag{1.7}$$

There are two cases. If $m(\lambda) \equiv 1$ then we have [cf. (2.5)] *Cauchy's logarithmic equation*

$$l(\lambda x) = l(\lambda) + l(x) \qquad (\lambda, x > 0) \tag{1.8}$$

which has the general nonconstant continuous solution

$$l(x) = \gamma \ln x \qquad (x > 0), \tag{1.9}$$

with $\gamma \neq 0$, as we will see in Section 2.1.3. On the other hand, if there exists a λ_0 with $m(\lambda_0) \neq 1$, then, from

$$m(\lambda)l(x) + l(\lambda) = l(\lambda x) = m(x)l(\lambda) + l(x),$$

we get with $\lambda = \lambda_0$, $\alpha = l(\lambda_0)/[m(\lambda_0) - 1]$

$$l(x) = \alpha[m(x) - 1] \tag{1.10}$$

($\alpha \neq 0$ because l is nonconstant). Putting this into (1.7) gives

$$m(\lambda x) = m(\lambda)m(x) \qquad (\lambda, x > 0). \tag{1.11}$$

This is *Cauchy's power equation.* As we will see, also in Section 2.1.3, its nonconstant continuous solutions are of the form

$$m(x) = x^\beta \qquad (x > 0) \tag{1.12}$$

($\beta \neq 0$). In view of (1.6), (1.9), (1.10), (1.12), and (1.2), we have shown that: *The general strictly monotonic solutions of the pair of functional equations* (1.1), (1.2) *are given by*

$$u(x) = \gamma \ln x + \delta, \qquad M(x, y) = (xy)^{1/2}, \tag{1.13}$$

and

$$u(x) = \alpha x^\beta + \delta, \qquad M(x, y) = \left(\frac{x^\beta + y^\beta}{2}\right)^{1/\beta}, \tag{1.14}$$

with arbitrary constants $\alpha \neq 0$, $\beta \neq 0$, $\gamma \neq 0$, and δ. For strictly increasing solutions, we have $\gamma > 0$ and $\alpha\beta > 0$.

This example illustrates how a functional equation (or a system of such equations), derived in a particular situation, can be solved by deducing from it some functional equation(s)

whose solution is known. In the Plateau case, (1.1) and (1.2) jointly led first to (1.5), and then successively to (1.7) and to Cauchy's logarithmic equation (1.8) and Cauchy's power equation (1.11) which are two of the four fundamental Cauchy equations which will be reviewed in Section 2, together with a collection of basic functional equation results.

1.2 Overview

The purpose of this review is to describe, in the style of the example in 1.1, some of the uses of functional equations in the behavioral and social sciences. No pretense is made that the coverage is in any way exhaustive. Rather, we demonstrate a number of techniques using additional examples taken from areas in which the authors have done appreciable work. Applications of functional equations in three empirical fields are discussed: Section 3 gives examples in sensory psychology (in particular, psychophysics) resembling the Plateau example, but more complex and covering different situations. Section 4 is devoted to individual decision making under uncertainty (utility theory); and Section 5 covers consistent aggregation of inputs and outputs in a social or economic context. Except for some proofs, which are either omitted or abbreviated, our discussion is self contained, drawing on the mathematical material in Section 2. We start there with a summary of some classic types of functional equations, solved long ago (for details see Aczél, 1987), that have proved useful in the analysis of behavioral models from a functional equation viewpoint. In all cases that we consider, the functional equations are numerical ones, generated by a mathematical model or a numerical representation of the phenomenon under consideration, in which the modelling is incomplete and at least some of the functions are unspecified. A substantial literature on representational measurement theory describes classes of qualitative systems that give rise to numerical representations (Krantz, Luce, Suppes, and Tversky, 1971; Luce, Krantz, Suppes, and Tversky, 1990; and Suppes, Krantz, Luce, and Tversly, 1989). From this starting point, functional equations enter the picture in at least three distinct ways.

One occurs when some invariance property holds. An example of such an invariance is some type of homogeneity of one of the functions, as illustrated by the function M in (1.1), and by several cases in Section 3. Other examples of invariance, from 4.9, involve utility and weighting functions from utility theory.

The second arises when two independent measurement schemes give rise to distinct measures of the same underlying attribute. An example in physics is mass. It can be measured by concatenating objects on the pans of a pan balance (in a vacuum) which is used to determine the order of greater mass. The resulting mass measure is additive over concatenation. It can also be measured fundamentally by varying the container size and the choice of homogeneous substances within them on a pan balance. This measure of mass exhibits a product structure with corresponding measures of volume and density. Clearly, because both mass measures represent the same ordering, they are related by a strictly increasing function. Any additional empirical law linking the two measurement structures imposes a constraint on that function in the sense that it must satisfy a functional equation. In the physical case, the link is a qualitative distribution law. This approach, with quite different linking laws, is illustrated repeatedly in Section 4 on utility theory.

The third concerns what may be thought of as consistency principles. One such principle is the commutative property used in modelling aggregation. Suppose we have variables x_{ij} that can be aggregated over either i or j and once done, then over the remaining subscript. An economic example concerns several types of inputs, such as raw materials, energy, and labor, to some production. Often one wishes to speak of an aggregated result over each type of input and over the outputs. The question is under what circumstances does the order of aggregation not matter. This is discussed in Section 5 on consistent social aggregation.

We keep notations consistent within sections, but not necessarily accross them. Achieving total consistency would have been at the cost of contravening well established notational conventions in the three empirical fields considered.

2. Several Basic Functional Equations

2.1 The Cauchy Family

2.1.1. The Four Equations

The example discussed in Section 1.1 yielded two of four closely related functional equations that constitute the *Cauchy family*. Later applications give rise to the other two. So we treat all of them as a unit.

The first, known as *Cauchy's fundamental equation* or, for short, *Cauchy's equation*, is

$$f(s+t) = f(s) + f(t) \qquad (s > 0, t > 0). \tag{2.1}$$

It could be considered for all real s, t but here it is restricted to positive s, t. This restriction is easily removed (other possible restrictions may not be so easily dropped, if at all). Define a "new" f (we could have introduced a new symbol for it) to be the old f on $\mathbb{R}_+ =]0, \infty[$, $f(0) = 0$, and

$$f(v) = -f(-v) \qquad (v < 0), \tag{2.2}$$

(which is defined because $-v > 0$, and the old f is defined at $-v$). It is easy to check that the new f satisfies

$$f(s+t) = f(s) + f(t) \qquad (s \in \mathbb{R}, t \in \mathbb{R}), \tag{2.3}$$

where $\mathbb{R}$ stands for the set of real numbers. Functions satisfying either version are called *additive*.

A second, called *Cauchy's exponential equation*, is

$$g(s+t) = g(s)g(t) \qquad (s > 0, t > 0). \tag{2.4}$$

This functional equation is important because it formalizes the intuitive idea of a system having no memory of its past ("lack of memory property"). Indeed, with s and t measuring disjoint, contiguous time intervals, it implies that the value of $g(s+t)/g(s)$ is independent of s. Provided g is positive, it is reduced to the fundamental one by $f = \ln g$. Positivity will be shown to follow from (2.4).

The third is *Cauchy's logarithmic equation*

$$l(st) = l(s) + l(t) \qquad (s > 0, t > 0), \tag{2.5}$$

and the fourth *Cauchy's power equation*

$$m(st) = m(s)m(t) \qquad (s > 0, t > 0). \tag{2.6}$$

Functions $m : \mathbb{R}_+ \to \mathbb{R}$ satisfying (2.6) are called *multiplicative*.

By substitution $s = e^v, t = e^w$ equation (2.5) becomes (2.3) and (2.6) becomes the equation

$$g(v+w) = g(v)g(w) \qquad (v \in \mathbb{R}, w \in \mathbb{R}). \tag{2.7}$$

The difference between (2.4) and (2.7) is that the latter is supposed to hold for all real arguments whereas the former is only for positive ones.

It is clear that it suffices, first, to show that g of (2.4) or, similarly, of (2.7) is positive and, then, to solve the fundamental equation (2.3) for the function f.

2.1.2. The Four Smooth Solutions

These equations have three types of solutions of which only one interests us:

1. Trivial, constant ones that are excluded simply by assuming the unknown function to be nonconstant, which in the case of g in (2.4) or (2.7) excludes

$$g(t) \equiv 1 \quad \text{and} \quad g(t) \equiv 0.$$

2. Highly irregular solutions exist and are excluded in some fashion, which we will do by assuming that on some closed subinterval the solution is bounded. In particular, we assume that g is bounded from above on a subinterval $[a,b]$ of $\mathbb{R}_+$ (of positive length, no matter how small), say by Θ.

3. Smooth ones that are our main interest. We may list these (in all cases $c \neq 0$ is a constant):

(2.8)	$f(t) = ct$	$(t \in \mathbb{R}_+$ or $t \in \mathbb{R})$,	[cf. (2.1) or (2.3)]
(2.9)	$g(t) = e^{ct}$	$(t \in \mathbb{R}_+$ or $t \in \mathbb{R})$,	[cf. (2.4) or (2.7)]
(2.10)	$l(t) = c \ln t$	$(t \in \mathbb{R}_+)$,	[cf. (2.5)]
(2.11)	$m(t) = t^c$	$(t \in \mathbb{R}_+)$,	[cf. (2.6)].

It is trivial to see that each solves its corresponding functional equation. The only issue is to show that they are the unique nonconstant bounded ones.

2.1.3 The Proofs

We start with Equation (2.6) which arose in Section 1.1 as (1.11). We write it here in the form

$$m(xy) = m(x)m(y) \qquad (x \in \mathbb{R}_+, y \in \mathbb{R}_+).$$

The substitution $x = e^v$, $y = e^w$ and $g(v) := m(e^v)$ converts this into

$$g(v+w) = g(v)g(w) \qquad (v \in \mathbb{R}, w \in \mathbb{R}),$$

that is, into equation (2.7). Restricting (2.7) to $\mathbb{R}_+$, we get (2.4):

$$g(s+t) = g(s)g(t) \qquad (s > 0, t > 0).$$

First, we show that the non-constancy condition and (2.4) imply that g is positive on $\mathbb{R}_+$. By (2.4) with $s = t = \frac{r}{2}$,

$$g(r) = g(\tfrac{r}{2})^2 \qquad (r \in \mathbb{R}_+).$$

So g is nonnegative. If there existed an $s_0 \in \mathbb{R}_+$ with $g(s_0) = 0$, then $g \equiv 0$. Indeed,

$$g(r) = g(s_0)g(r - s_0) = 0 \qquad (r > s_0).$$

This does not yet exclude the possibility of $g(r_0) > 0$ for some $r_0 < s_0$. But that also leads to a contradiction: Repeated application of (2.4) gives

$$g(nv) = g(v)^n \qquad (v \in \mathbb{R}_+, n = 1, 2, \ldots) \tag{2.12}$$

and for a sufficiently large positive integer n, we have $nr_0 > s_0$, whence $0 < g(r_0)^n = g(nr_0) = 0$, a contradiction. [The proof is even simpler for (2.7).]

Thus, g is positive and we can take logarithms on both sides of (2.4) to reduce the problem to solving (2.1). Because g was supposed to be bounded from above on an interval $[a, b]$ by Θ, so $f = \ln g$ is bounded from above on $[a, b]$ by $\theta = \ln \Theta$ (the positivity of g has no other consequence than that f can be so defined). Moreover, by (2.2) the "new" f is bounded from below on $[-b, -a]$ by $-\theta$. But then f is bounded from both sides on every interval $[a', b']$ of length $b' - a' = b - a$. Indeed, by (2.3),

$$f(t) = f(t + a - a') - f(a - a') = f(t - b - a') + f(b + a').$$

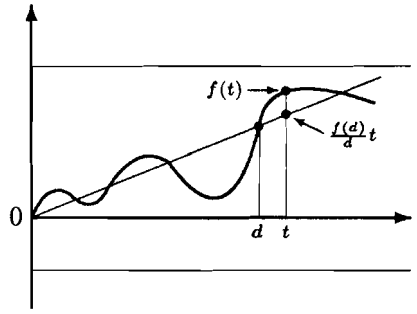
If $t \in [a', b']$, then $t + a - a' \in [a, b]$, $t - b - a' \in]-b, -a]$ and so

$$-\theta + f(b + a') \leq f(t) \leq \theta - f(a - a').$$

Clearly, $f_0(t) = ct$ gives solutions of (2.3), bounded on every finite interval. We prove that this is the general such solution. The function $h = f - f_0$, formed from $f_0(u) = cu$ and from an arbitrary solution f of (2.3), also satisfies (2.3). We choose c in $h(t) = f(t) - ct$ so that $h(d) = 0$ with $d = b - a$, i.e. (cf. Fig. 2),

$$h(t) = f(t) - ct = f(t) - \frac{f(d)}{d}t. \tag{2.13}$$

FIGURE 2. Reduction of the Cauchy Equation; $h(t) = f(t) - \frac{f(d)}{d}t$.



Since this h satisfies $h(s + t) = h(s) + h(t)$ and $h(d) = 0$, we have

$$h(t + d) = h(t) + h(d) = h(t), \tag{2.14}$$

which means that h is periodic with period d. But with f bounded on $[0, d]$ (Fig. 2), so is $h = f - f_0$ (remember that $d = b - a$) and thus by the periodicity, (2.14), it is bounded on the whole line $\mathbb{R}$. If this h were not identically 0, then there would exist a t_0 with $h(t_0) \neq 0$, say $h(t_0) < 0$. However, $h(s + t) = h(s) + h(t)$ implies $h(nt_0) = nh(t_0)$ for all positive integers n. Since $h(t_0) < 0$, for large enough n the number $nh(t_0) = h(nt_0)$ could be as small (as large a negative number) as we want, thus contradicting the boundedness of h on $\mathbb{R}$. Therefore, $h(t) \equiv 0$, and we see from (2.13)

$$f(t) = ct \qquad (t \in \mathbb{R}),$$

as asserted, and f is nonconstant iff $c \neq 0$. Thus, we have proved the following.

Theorem 1. *The general nonconstant solutions of* (2.3) *or* (2.1), *bounded from above on an interval, are given by* (2.8).

REMARKS:

1. The proof makes it clear that "bounded from below" may replace "bounded from above." Bounded from one side on an interval is also called "*locally bounded* on one side."
2. Obviously, the only constant solution of (2.3) or (2.1) is $f(t) \equiv 0$.

Corollary 1. *The result is unchanged if boundedness is replaced by either f is continuous at a point or is monotonic on an interval. In the latter case, f is strictly increasing or strictly decreasing if and only if $c > 0$ or $c < 0$, respectively, but otherwise c is arbitrary.*

Returning to (2.4), because $f = \ln g$ satisfies (2.1), we have:

Corollary 2. *The general nonconstant solutions of both* (2.4) *and* (2.7), *bounded on an interval from below by a positive constant or from above by any constant or monotonic on an interval or continuous at a point are given by* (2.9).

REMARKS:

1. We need g bounded from below by a positive number if we want the $f = \ln g$ to be bounded from below at all.
2. For the solutions f of (2.3) and g of (2.4) or (2.7), the assumptions that they are monotonic and nonconstant implies that they are strictly monotonic.
3. The solution g will be strictly increasing or strictly decreasing according to whether $c > 0$ or $c < 0$, respectively.

As was noted earlier, the substitution $s = e^v, t = e^w$ converts (2.5) to (2.3) and (2.6) to (2.4). Thus, we have the following:

Corollary 3. *The general nonconstant solutions of* (2.5), *bounded from one side on an interval are given by* (2.10), *and the general solutions of* (2.6), *locally bounded from below by a positive constant or locally bounded from above, are given by* (2.11). *The solutions l and m are strictly increasing or strictly decreasing according to whether $c > 0$ or $c < 0$, respectively.*

REMARK:

Some new things happen when either (2.4) or (2.5) is considered for nonnegative instead of positive variables. In addition to (2.9), there is one (and only one) further locally bounded nonconstant solution of (2.4) for $s \geq 0$, $r \geq 0$, namely,

$$g(t) = \begin{cases} 1 & \text{if } t = 0, \\ 0 & \text{if } t > 0. \end{cases}$$

(Similarly, as before, a function is "locally bounded" if there exists a proper interval on which it is bounded.) Indeed, $s = 0$ in $g(s + t) = g(s)g(t)$ $(s \geq 0, t \geq 0)$ shows that either $g(0) = 1$ or $g(t) \equiv 0$. The former permits both $g(t) = 0$ for $t > 0$ and $g(t) = e^{ct}$ for all $t \geq 0$. On the other hand, $l(xy) = l(x) + l(y)$ for all $x \geq 0, y \geq 0$ permits only the trivial solution $l(x) \equiv 0$, as can be seen by choosing $y = 0$.

2.1.4 The Pexider Equation

We consider now the equation

$$f(x + y) = g(x) + h(y), \tag{2.15}$$

where (x, y) lie in an open region (i.e., connected open set) of $\mathbb{R}^2$, and one of f, g, h, say g, is strictly monotonic. This equation, called the *Pexider equation*, is similar to Cauchy's fundamental equation (2.3), except that it contains three unknown functions rather than one. Yet, as we shall see, they are all determined (a typical feature for functional equations but rather unusual for other equations). Pexider equation has already arisen in Section 1 and will arise again in several applications. Solving the Pexider equation is easily reduced to solving (2.3).

If a Pexider equation holds on an open region in $\mathbb{R}^2$, then its validity can be extended to all $(v, w) \in \mathbb{R}^2$ (see, e.g., Aczél, 1987, p. 80):

$$f(v + w) = g(v) + h(w) \qquad (v \in \mathbb{R}, w \in \mathbb{R}). \tag{2.16}$$

Substituting $w = 0$ and, separately, $v = 0$, we get, respectively,

$$g(v) = f(v) - b,\ h(w) = f(w) - a, \tag{2.17}$$

where $a = g(0)$, $b = h(0)$. Putting these back into (2.16), we see that $F(t) = f(t) - a - b$ satisfies Cauchy's equation $F(v + w) = F(v) + F(w)$. Because F is strictly monotonic, $F(t) = ct$ $(c \neq 0)$. Thus [see (2.17)] we have the following:

Theorem 2. *The general solution of the Pexider equation* (2.15) *on an open region R of $\mathbb{R}^2$ with at least one of f, g, h strictly monotonic, is given by*

$$f(t) = ct + a + b, \quad g(v) = cv + a, \quad h(w) = cw + b \tag{2.18}$$

for $v \in \{v \mid (v, w) \in R\}$, $w \in \{w \mid (v, w) \in R\}$, $t \in \{v + w \mid (v, w) \in R\}$. Here $c \neq 0$, a, b are arbirary constants.

As with the Cauchy equation, there are three additional variants involving multiplication as well as addition. For example, we will encounter $f(vw) = g(v)h(w)$ $(v \in \mathbb{R}_+, w \in \mathbb{R}_+)$ in Sections 2.2.3 and 4.4.3. The locally bounded, nonconstant solutions are obviously

$$f(t) = abt^c,\ g(v) = av^c,\ h(w) = bw^c \qquad (abc \neq 0). \tag{2.19}$$

2.2 The Invariance Equation

2.2.1 The Equation

We solve the following functional equation, an example of which arose in Section 1.1 [see (1.3)], and later we show applications to proposed invariance assumptions in the behavioral sciences:

$$\varphi[f(\lambda x) + r(\lambda y)] = \lambda\varphi[f(x) + r(y)] \qquad ((x, y) \in T, \lambda \in \mathbb{R}_+), \tag{2.20}$$

where f, r and φ are real valued functions and T is an open cone with $(x, y) \in \mathbb{R}^2_+$, i.e., a connected open set of pairs of positive reals, closed under multiplication by any $\lambda > 0$ in the sense that if $(x, y) \in T$, then $(\lambda x, \lambda y) \in T$. Clearly $T = \{(x, y) \in \mathbb{R}^2_+ \mid Ax < y < Bx\}$ where $B > A > 0$ are constants. The functions f, r, φ are, by supposition, strictly monotonic and map their domains, which are intervals, onto intervals. Thus, they are continuous. The domain of f and r is $]0, \infty[$. Let their range (set of function values) be I and J, respectively.

By the above supposition, these are open intervals (open because of the strict monotonicity). Thus, φ is defined on $I + J = \{v + w \mid v \in I, w \in J\}$.

2.2.2 Reduction to Cauchy Equations

For all $v \in I, w \in J$, there exist, by the above, $x \in \mathbb{R}_+, y \in \mathbb{R}_+$ such that $v = f(x)$, $w = r(y)$. Because (2.20) is supposed to hold for all (x, y) in the open cone T, which is an open region, and because f, r are continuous and strictly monotonic, it follows that

$$T' = \{(v, w) \mid v = f(x), w = r(y), (x, y) \in T\}$$

is an open region. Considering λ to be a parameter, we define

$$f_\lambda(v) = f[\lambda f^{-1}(v)], \quad r_\lambda(w) = r[\lambda r^{-1}(w)], \quad \varphi_\lambda(t) = \varphi^{-1}[\lambda\varphi(t)], \tag{2.21}$$

yielding

$$\varphi_\lambda(v + w) = f_\lambda(v) + r_\lambda(w) \qquad ((v, w) \in T'),$$

a Pexider equation (2.15). Thus, the solution is

$$\varphi_\lambda(t) = m_\lambda t + a_\lambda + b_\lambda, \quad f_\lambda(v) = m_\lambda v + a_\lambda, \quad r_\lambda(w) = m_\lambda w + b_\lambda \tag{2.22}$$

Considering λ as a variable again, we get, in view of (2.21),

$$f(\lambda x) = m(\lambda) f(x) + a(\lambda) \qquad (x, \lambda \in \mathbb{R}_+), \tag{2.23}$$

and similar equations for r and φ. As in Section 1.1 we get two distinct solutions

$$f(x) = \alpha_1 \ln x + \beta_1 \qquad (x \in \mathbb{R}_+, \alpha_1 \neq 0), \tag{2.24}$$

and

$$f(x) = \alpha_1 x^\gamma + \beta_1 \qquad (x \in \mathbb{R}_+, \alpha_1 \neq 0, \gamma \neq 0). \tag{2.25}$$

Because they satisfy (2.23) for $m(\lambda) = 1$ or $m(\lambda) = \lambda^\gamma$, respectively and for all $\alpha_1 \neq 0, \gamma \neq 0$ and β_1, the general solutions of (2.23) for strictly monotonic f are given by (2.24) and (2.25). Substitution yields $a(\lambda) = \alpha_1 \ln \lambda$ or $a(\lambda) = \beta_1(1 - \lambda^\gamma)$, respectively.

Similarly, the equations for r and φ [cf. (2.18), (2,20) and (2.21)]

$$r(\lambda y) = c(\lambda) r(y) + b(\lambda), \ \varphi^{-1}(\lambda s) = c(\lambda)\varphi^{-1}(s) + a(\lambda) + b(\lambda)$$

have as general solutions for strictly monotonic r and φ^{-1}

$$r(y) = \alpha_2 \ln y + \beta_2 \qquad (\alpha_2 \neq 0), \tag{2.26}$$
$$\varphi^{-1}(s) = (\alpha_1 + \alpha_2) \ln s + \beta_3 \qquad (\alpha_1 + \alpha_2 \neq 0),$$

and

$$r(y) = \alpha_2 y^\gamma + \beta_2 \qquad (\alpha_2 \neq 0, \gamma \neq 0), \tag{2.27}$$
$$\varphi^{-1}(s) = \alpha_3 s^\gamma + \beta_1 + \beta_2 \quad (\gamma \neq 0, \alpha_3 \neq 0),$$

respectively. The φ^{-1} expressions yield

$$\varphi(t) = \exp\left(\frac{t-\beta_3}{\alpha_1+\alpha_2}\right) \qquad (\alpha_1+\alpha_2 \neq 0), \tag{2.28}$$

$$\varphi(t) = \left(\frac{t-\beta_1-\beta_2}{\alpha_3}\right)^{1/\gamma} \qquad (\alpha_3 \neq 0, \gamma \neq 0). \tag{2.29}$$

2.2.3 The Result

The following summarizes what we have just shown:

Theorem 3. *The general strictly monotonic and continuous solutions f, r, φ of the functional equation* (2.20), *where T is an open cone, are given by* (2.24), (2.26), *and* (2.28) *and by* (2.25), (2.27), *and* (2.29), *with the constants restricted as indicated but otherwise arbitrary.*

Corollary. *The general strictly monotonic and continuous solutions f of the functional equation*

$$f^{-1}[f(\lambda x)+f(\lambda y)] = \lambda f^{-1}[f(x)+f(y)] \qquad ((x,y) \in T, \lambda \in \mathbb{R}_+), \tag{2.30}$$

where T is an open cone, are given by

$$f(x) = \alpha x^\gamma \qquad (x \in \mathbb{R}_+),$$

where α and γ are arbitrary nonzero constants.

Equation (2.30) is, of course, the particular case of (2.20) where $r = f$ and $\varphi = f^{-1}$. Substitution shows that (2.25) satisfies (2.30) only with $\beta_1 = 0$ and (2.24) does not satisfy (2.30) with any $\alpha_1 \neq 0$ and β_1 at all.

The equation

$$F^{-1}[F(v^\lambda)+F(w^\lambda)] = \left(F^{-1}[F(v)+F(w)]\right)^\lambda \qquad (v,w,\lambda \in \mathbb{R}_+) \tag{2.31}$$

is transformed by $v = e^x, w = e^y, f(x) = F(e^x)$ into (2.30). Here the open cone T is replaced by $\mathbb{R}^2$. The same argument as above and as in Section 1.1 yields

$$f(\lambda x) = m(\lambda) f(x) \qquad (x \in \mathbb{R}, \lambda \in \mathbb{R}_+). \tag{2.32}$$

(Note that it is not restricted to $x \in \mathbb{R}_+$.) As in (2.19), for $\lambda \in \mathbb{R}_+$, $x \in \mathbb{R}_+$, the general continuous, strictly monotonic solution is given by $m(\lambda) = \lambda^\gamma$, $f(\lambda) = \alpha\lambda^\gamma$ ($\alpha\gamma \neq 0$). On the other hand, $x = -1, \delta = f(-1)$ in (2.32) give $f(-\lambda) = \delta m(\lambda) = \delta\lambda^\gamma$ ($\delta \neq 0$). Because setting $x = 0$ in (2.32) yields $f(0) = 0$ for strictly monotonic m, we get

$$f(x) = \begin{cases} \alpha x^\gamma & \text{for } x \geq 0, \\ \delta|x|^\gamma & \text{for } x < 0. \end{cases}$$

The function f is strictly monotonic only for positive γ and negative $\alpha\delta$. For $F(p) = f(\ln p)$, we get

$$F(p) = \begin{cases} \alpha(\ln p)^\gamma & (p \geq 1), \\ \delta|\ln p|^\gamma & (p < 1), \end{cases}$$

with $\gamma > 0$ and $\alpha\delta < 0$, as the general strictly monotonic and continuous solution of (2.31).

Two behavioral examples of such invariance arguments are given in Sections 4.9.1 and 4.9.2.

3. Psychophysics

Many applications of functional equations in psychophysics are in the style illustrated by the Plateau's situation discussed in Section 1.1. On the experimental side, one starts with some homogeneity equation that seems to capture an important regularity in the data. In the case of Plateau, the function M in (1.1) is homogeneous of degree 1. Next, one introduces some reasonable model, involving one or more unknown functions, formalizing a possible mechanism for explaining the behavior of the individuals in the study. In Plateau's example, such a model is obtained by assuming that the function M is a quasi-arithmetic mean [see (1.2)], with the unknown function u. Combining the homogeneity equation with the model results in a considerable specification of the possible forms of the unknown functions. In our example, u must be either a logarithmic or a power function [cf. (1.13) and (1.14)].

The cases reviewed in this section all involve a binary discrimination situation in which an individual must decide which of the two stimuli presented has more of some sensory attribute. For instance, the individual compares the loudness of pure tones differing only by their intensities, which is measured in standard ratio scale units (such as sound pressure level).

In a classic 19th century case, involving E. E. Weber and G. T. Fechner, the individual is presented with a pair of stimuli (x, y), and the researcher's task is to estimate experimentally the probability $P(x, y)$ that a stimulus of intensity x is judged as louder (or brighter, longer, etc.), than a stimulus of intensity y. In mathematical terminology, the so-called *Weber Law* states that the function $P : \mathbb{R}_+^2 \to]0, 1[$ is homogeneous of degree 0:

$$P(\lambda x, \lambda y) = P(x, y) \qquad (x, y, \lambda > 0). \tag{3.1}$$

As mentioned in Section 1.1, for the purpose of this review, we indulge in some idealization and suppose that the variables take their values on sets such as $\mathbb{R}$ or $\mathbb{R}_+$, even though humans cannot realistically be presented with very intense stimuli, (e.g., the brightness of the sun). In the case dealt with here, such assumptions have no substantive impact on the results, and can be generalized considerably (see Falmagne, 1985). There are, however, equations where the exact form of the results depends upon the domains of the functions (cf. Falmagne, 1981).

Fechner's contribution consisted in proposing that the individual's judgment resulted from a computation of the differences between the two sensations produced by x and y, this difference being evaluated in terms of some unknown sensation scale. Formally, Fechner's idea is expressed by the equation

$$P(x, y) = F\big(u(x) - u(y)\big), \tag{3.2}$$

where the functions u and F are real valued, strictly increasing and continuous, but otherwise not a priori specified. In the context of Weber's Law, (3.1), the possible forms of u and F in (3.2) are severely constrained. Putting (3.2) into (3.1), we get the functional equation

$$F\big(u(\lambda x) - u(\lambda y)\big) = F\big(u(x) - u(y)\big) \qquad (x > 0, y > 0, \lambda > 0),$$

which, since F is strictly increasing, is equivalent to

$$u(\lambda x) - u(\lambda y) = u(x) - u(y) \qquad (x > 0, y > 0, \lambda > 0).$$

With $\lambda = 1/y$, $z = x/y$, $l(z) = u(z) - u(1)$, the above equation gives $l(yz) = l(y) + l(z)$, $(y > 0,\ z > 0)$, that is, (2.5), whose strictly increasing solutions are given by (2.10) as

$l(x) = c \ln x$ $(c > 0)$. With $B = u(1)$, we see that all the strictly increasing continuous solutions u of the pair of functional equations (3.1), (3.2) are given by

$$u(x) = c \ln x + B \qquad (x \in \mathbb{R}_+), \tag{3.3}$$

while P and F are related by the equation

$$F(t) = P(e^{t/c}, 1) \qquad (t \in \mathbb{R}) \tag{3.4}$$

$(c \in \mathbb{R}_+, B \in \mathbb{R})$.

The other examples in this section are variations on this theme, but require more work because we deal with cases in which each of the two arguments of the function P is a real vector. Because each component of the two vectors may contribute separately to the sensation, various models describing such contributions may be considered, in the context of several possible forms of homogeneity. As in the above example, we take loudness discrimination as our experimental situation. The next three sections summarize results by Falmagne and Iverson (1979; see also Falmagne, Iverson & Marcovici, 1976, and Falmagne, 1985).

3.1 The Conjoint Weber Law

Suppose that an individual wearing earphones is presented with a 1000 Hz tone delivered simultaneously, with different intensities, to the two auditory channels. The impression created by such a stimulus is that of a single sensation, the loudness of which depends upon the combination of two inputs. (The location of the sensation inside the individual's head also depends upon the combination of the two intensities, but this aspect of the phenomenon is not relevant here.) We write (a, x) for such a two-dimensional stimulus, where a and x stand for positive real numbers denoting the sound pressures in the left and right auditory channels, respectively. Extending our earlier notation, we write $P(a, x; b, y)$ for the probability that the individual judges the two-dimensional stimulus (a, x) to be louder than the 2-dimensional stimulus (b, y). Various concepts concerning the function P are gathered in the definitions below.

3.1.1 Definitions

We suppose that $P : \mathbb{R}_+^4 \to]0, 1[$ is continuous, with $P(a, x; b, y)$ strictly increasing in a and strictly decreasing in b, and strictly contra-monotonic in x and y [that is, $P(a, x; b, y)$ is either strictly increasing in x and strictly decreasing in y, or vice versa]. We will call such a function a *discrimination probability*. Note that the hypothesis that P is strictly co-monotonic in the second and fourth variables is weaker than what is suggested by the binaural loudness summation example, but makes sense because the formalism is then also applicable to other empirical cases. For instance, a pair (a, x) could represent a stimulus of intensity a presented over a 'noisy' background of intensity x, cf. Section 3.3.

A discrimination probability P satisfies the *Conjoint Weber Law* if it is homogeneous of degree zero, that is,

$$P(\lambda a, \lambda x; \lambda b, \lambda y) = P(a, x; b, y) \qquad (\lambda, a, x, b, y > 0). \tag{3.5}$$

It is easily shown that (3.5) holds if and only if there is a function $Q : \mathbb{R}_+^3 \to]0, 1[$, such that

$$P(a, x; b, y) = Q(a/x, x/y, b/y), \tag{3.6}$$

with Q strictly increasing in the first variable and strictly decreasing in the third variable.

It is tempting to conceptualize the combination of the loudnesses of the two tones as an addition which is somehow performed by the auditory system in terms of some sensory

scales. In the literature, this phenomenon is in fact referred to as "binaural loudness summation." Falmagne and Iverson (1979) assumed that the function P satisfies the equation

$$P(a,x;b,y) = H[f(a)+r(x), f(b)+r(y)] \tag{3.7}$$

where f, r and H are continuous functions, with f strictly increasing, r strictly monotonic, and H strictly increasing in the first variable and strictly decreasing in the second variable. When (3.7) holds for some functions f, r and H, we say that P satisfies *component additivity.*

Falmagne and Iverson (1979) also considered the generalization of (3.7) represented by the equation of *simple scalability*

$$P(a,x;b,y) = H[g(a,x), g(b,y)], \tag{3.8}$$

with H as above and g continuous, strictly increasing in the first variable and strictly monotonic in the second variable. When a function P satisfies (3.8) for some functions g and H satisfying the stated conditions, we say that P is *simply scalable*, and we call (g,H) a *simple scale* representation of P. The Conjoint Weber Law puts stringent constraints on the functions entering the component additivity and even the simple scalability equations.

3.1.2 Results

We begin with a preparatory result which is of some intrinsic interest.

Lemma 1. *Suppose that a discrimination probability* $P : \mathbb{R}_+^4 \to]0,1[$ *has a simple scale representation* (g,H)*; cf.* (3.8)*. Then, the following two conditions are equivalent.*

(i): *The discrimination probability* P *satisfies the Conjoint Weber Law* (3.5).

(ii): *Either the function* g *in* (3.8) *is homogeneous of degree* 0*, or there exists a constant* $\beta \neq 0$*, a continuous function* $h : \mathbb{R}_+^2 \to \mathbb{R}_+$ *homogeneous of degree* β*, strictly increasing in the first variable, strictly monotonic in the second variable, and a continuous, strictly increasing function* G*, such that*

$$P(a,x;b,y) = G\left(\frac{h(a,x)}{h(b,y)}\right) \qquad (a,x,b,y>0). \tag{3.9}$$

Sketch of proof. Clearly, each of the two cases of (ii) implies (i). Assume that

$$P(a,x;b,y) = H\big(g(a,x), g(b,y)\big).$$

for some simple scale representation (g,H). Notice that $g(a,a)$ is defined for any $a \in \mathbb{R}_+$. For the moment, assume that $g(a,a)$ is either constant or strictly monotonic in a. If $g(a,a)$ is constant, we obtain successively

$$\begin{aligned} P(a,a;b,y) &= H\big(g(a,a), g(b,y)\big) \\ &= H\big(g(\lambda a,\lambda a), g(\lambda b,\lambda y)\big) \qquad \text{(Conjoint Weber Law)} \\ &= H\big(g(a,a), g(\lambda b,\lambda y)\big), \end{aligned}$$

and by the strict monotonicity of H in the second argument, we derive $g(\lambda b,\lambda y) = g(b,y)$ for all $\lambda, b, y > 0$. In other words, g is homogeneous of degree 0, and the first case of (ii) follows.

Suppose, on the other hand, that $g(a,a)$ is strictly monotonic. By an argument of continuity, which is omitted here, one shows that there must exist a solution $q(a,x) > 0$ to the equation

$$g\big(q(a,x),q(a,x)\big) = g(a,x) \qquad (a>0,\ x>0), \tag{3.10}$$

where the function q is continuous and nonconstant on $\mathbb{R}_+^2$. With $\alpha = q(a,x)$ and $\alpha' = q(b,y)$, this yields

$$P(a,x;b,y) = H\big(g(a,x),g(b,y)\big) = H\big(g(\alpha,\alpha),g(\alpha',\alpha')\big) = P(\alpha,\alpha;\alpha',\alpha').$$

Applying the Conjoint Weber Law, we obtain with $\lambda = 1/\alpha'$,

$$\begin{aligned} P(a,x;b,y) &= P(\alpha,\alpha;\alpha',\alpha') \\ &= P(\lambda\alpha,\lambda\alpha;\lambda\alpha',\lambda\alpha') \\ &= P\Big(\frac{\alpha}{\alpha'},\frac{\alpha}{\alpha'};1,1\Big) \\ &= M(\alpha/\alpha'), \end{aligned}$$

where $M(s) = P(s,s;1,1)$ is strictly monotonic, since $g(s,s)$ is strictly monotonic. We have thus, applying once more the Conjoint Weber Law

$$P(a,x;b,y) = M\left(\frac{q(a,x)}{q(b,y)}\right) = M\left(\frac{q(\lambda a,\lambda x)}{q(\lambda b,\lambda y)}\right),$$

which implies

$$\frac{q(a,x)}{q(b,y)} = \frac{q(\lambda a,\lambda x)}{q(\lambda b,\lambda y)}.$$

With $m(\lambda) = q(\lambda,\lambda)/q(1,1)$, we get

$$q(\lambda a,\lambda x) = m(\lambda)q(a,x)$$

with the function m nonconstant, continuous on $\mathbb{R}_+$, and satisfying $m(\lambda\lambda') = m(\lambda)\,m(\lambda')$, which is Cauchy's power equation [cf. Section 2.1.1, (2.6)]. According to (2.11), $m(\lambda) = \lambda^\beta$ for some $\beta \neq 0$. Thus, q is homogeneous of degree $\beta \neq 0$. If $g(t,t)$ is strictly increasing, then M is also strictly increasing, and the second case of (ii) obtains with $h = q$, $G = M$ and $\beta \neq 0$. In the other case, we define $h = 1/q$ and $G(s) = M(1/s)$.

To complete the proof, it remains to show that if $g(t,t)$ is nonconstant, it must be strictly monotonic. This fact also results from the continuity of g and the Conjoint Weber Law, but we do not include the argument here. □

REMARK. The result in Lemma 1 also holds under weaker conditions concerning the domain of the function P; for instance: $P(a,x;b,y)$ is defined for all $(a,x),(b,y) \in D$, where D is a nonempty, positive, convex open cone (see Falmagne and Iverson, 1976).

Lemma 1 is instrumental for establishing the next two results.

Theorem 4. *Suppose that $P : \mathbb{R}_+^4 \to]0,1[$ is a simply scalable discrimination probability satisfying the Conjoint Weber Law* (3.5). *Then, one of the two following possibilities holds:*

(a): *There are some real valued, continuous, strictly monotonic functions $T > 0$ and M satisfying*

$$P(a,x;b,y) = M\left(\frac{T(a/x)x}{T(b/y)y}\right) \qquad (a,x,b,y>0). \tag{3.11}$$

(b): *There is a continuous function Q_0, strictly increasing in the first argument, strictly decreasing in the second argument, such that*

$$P(a,x;b,y) = Q_0(a/x, b/y). \tag{3.12}$$

Proof. Clearly, Case (b), which describes homogeneity of degree $\beta = 0$, is compatible with the hypotheses of the theorem. Let (h, H) be the simple scale representation of P obtained in Lemma 1(ii). Thus, g or h is homogeneous of degree β. Assume that Case (b) does not hold. It is easily verified that we must have $\beta \neq 0$. Thus,

$$P(a,x;b,y) = G\left(\frac{h(a,x)}{h(b,y)}\right) \qquad (a,x,b,y > 0)$$

for some continuous, strictly increasing function G, with $h(a,x) = x^\beta h(a/x, 1)$. Defining $T(s) = h(s,1)^{1/\beta}$, $M(s) = G(s^\beta)$, gives (3.11). Thus, Case (a) is implied by the hypotheses of the theorem when (b) does not hold. □

Theorem 5. *Suppose that $P : \mathbb{R}_+^4 \to]0,1[$ is a discrimination probability satisfying the Conjoint Weber Law* (3.5)*, together with the component additivity equation* (3.7)

$$P(a,x;b,y) = H\big(f(a) + r(x), f(b) + r(y)\big) \qquad (a,x,b,y > 0),$$

where f, r and H are continuous functions, with f strictly increasing, r strictly monotonic, and H strictly increasing in the first vari able and strictly decreasing in the second variable. Then, one of the three equations

$$P(a,x;b,y) = G\left(\frac{a^\beta + \delta x^\beta}{b^\beta + \delta y^\beta}\right), \tag{3.13}$$

$$P(a,x;b,y) = G\left(\frac{a^\beta x^\gamma}{b^\beta y^\gamma}\right), \tag{3.14}$$

$$P(a,x;b,y) = Q_0\left(\frac{a}{x}, \frac{b}{y}\right), \tag{3.15}$$

must hold for $a,x,b,y > 0$. In (3.13) *and* (3.14)*, G is continuous and strictly increasing, and $\beta > 0$, $\delta \neq 0$ and $\gamma \neq 0$ are constants. In* (3.15)*, Q_0 is continuous, strictly increasing in its first argument, and strictly decreasing in its second argument.*

Thus, the function P is necessarily decreasing in its second argument, and increasing in its fourth argument.

Proof. The hypotheses of the theorem imply those of Theorem 4. Accordingly, either (3.15) holds, which with $f = -r = \ln$ and $H(s,t) = Q_0(e^s, e^t)$ is clearly of the additive form (3.7), or else Case (a) of Theorem 4 holds, that is:

$$P(a,x;b,y) = M\left(\frac{T(a/x)x}{T(b/y)y}\right) = H\big(f(a) + r(x), f(b) + r(y)\big) \tag{3.16}$$

with T and M continuous and strictly monotonic. Holding $b = b_0$, $y = y_0$ constant and letting

$$K(t) = T\left(\frac{b_0}{y_0}\right) y_0 M^{-1}\big(H[t, f(b_0) + r(y_0)]\big),$$

we get from the last equation in (3.16)

$$K[f(a)+r(x)] = T\left(\frac{a}{x}\right)x.$$

Thus, for all $\lambda > 0$,

$$K[f(\lambda a)+r(\lambda x)] = \lambda K[f(a)+r(x)],$$

which is the invariance equation (2.20). Accordingly [cf. (2.24), (2.26) and (2.25), (2.27)], we have the following two solutions:

$$f(a) = A_1 a^\beta + B_1, \qquad r(x) = A_2 x^\beta + B_2, \tag{3.17}$$

where A_1,B_1, A_2, B_2 and β are constants, $A_1\beta > 0$, $A_2 \neq 0$;

$$f(a) = A_1 \ln a + B_1, \qquad r(x) = A_2 \ln x + B_2, \tag{3.18}$$

where A_1,B_1, A_2, B_2 are constants, $A_1 > 0$, $A_2 \neq 0$. If (3.17) holds, then

$$\begin{aligned} P(a,x;b,y) &= H(A_1a^\beta + B_1 + A_2x^\beta + B_2, A_1b^\beta + B_1 + A_2y^\beta + B_2) \\ &= H_1(a^\beta + \delta x^\beta, b^\beta + \delta y^\beta), \end{aligned}$$

with $\delta = A_2/A_1$ and $H_1(s,t) = H(A_1s + B_1 + B_2, A_1t + B_1 + B_2)$. So, the two functions $h : (a,x) \mapsto a^\beta + \delta x^\beta$ and H form a simple scale representation of the discrimination probability P. Notice that h is homogenous of degree $\beta \neq 0$, and that P satisfies the Conjoint Weber Law. The conditions of Case (ii) of Lemma 1 are thus satisfied, and there must exist a strictly increasing function G satisfying (3.13). A similar argument leads to (3.14) in the case of the solution (3.18). □

3.2 The Conjoint Weber Inequality

Empirically, the Conjoint Weber Law may be satisfied only approximately. For instance, it may fail for small values of the stimuli. In such a case, the following *Conjoint Weber Inequality* may give a more adequate representation of the data

$$P(a,x;b,y) \leq P(\lambda a, \lambda x; \lambda b, \lambda y) \qquad (\lambda \geq 1, a \geq b > 0, x \geq y > 0). \tag{3.19}$$

It is then natural to ask whether the type of expressions obtained for the discrimination probability P in Theorems 1, 4 and 5 could hold asymptotically, that is, for large λ in (3.19). An example of what can be obtained is outlined here, without proof.

Suppose that P is a discrimination probability; thus, P is a continuous function mapping $\mathbb{R}^4_+$ into $]0,1[$. We also suppose that $P(a,x;b,y)$ is strictly increasing in a, x, strictly decreasing in b, y and satisfies (3.9) of Lemma 1, that is

$$P(a,x;b,y) = G\left(\frac{h(a,x)}{h(b,y)}\right) \qquad (a,x,b,y > 0)$$

for a continuous, strictly increasing function G, and a function $h > 0$ strictly increasing in both variables, and satisfying continuity and other regularity conditions; in particular, for large a, we have $h(a,a) \leq \gamma a^\mu$ for some constants $\gamma, \mu > 0$. Suppose moreover that the Conjoint Weber Inequality holds. Then

$$\lim_{\lambda\to\infty} P(\lambda a, \lambda x; \lambda b, \lambda y) = M\left(\frac{T(a/x)x}{T(b/y)y}\right) \qquad (a,x,b,y > 0).$$

The proof is similar to that of Theorem 4 but uses some new ideas [cf. Falmagne, 1977; for details see Falmagne and Iverson, 1979]. This type of asymptotic result is sometimes referred to as a *stability* result [cf. the survey by Forti (1995)].

3.3 The Strong Conjoint Weber Law

In some situations, a stronger version of homogeneity can hold for a discrimination probability P. As before, suppose that two pairs of stimuli (a,x) and (b,y) are presented to an individual, who has to decide which pair seems louder. In this case, however, the numbers x and y represent the intensity of a 'noisy' background. The relative loudness of (a,x) with respect to (b,y) may not change if, on one hand, a and b are multiplied by the same positive constant, and on the other hand, both x and y are multiplied by a possibly different positive constant. This idea leads to the following homogeneity condition:

$$P(a,x;b,y) = P(\lambda a,\tau x;\lambda b,\tau y) \qquad (a,x,b,y,\lambda,\tau>0). \tag{3.20}$$

We refer to (3.20) as the *Strong Conjoint Weber Law*[1]. Obviously, this condition implies the Conjoint Weber Law. So, the results of the previous sections are applicable, with possibly stronger consequences. The following theorem is one of the results.

Theorem 6. *Suppose that P is a simply scalable discrimination probability, which is strictly increasing in the second variable and strictly decreasing in the fourth variable. Then, the following two conditions are equivalent.*

(i): *The Strong Conjoint Weber Law* (3.20) *holds;*

(ii): *Equation* (3.11) *holds, with $T(s)=s^\beta$ for some constant β, $0<\beta<1$. Accordingly,*

$$P(a,x;b,y) = M\left(\frac{a^\beta x^{1-\beta}}{b^\beta y^{1-\beta}}\right), \qquad (a,x,b,y>0), \tag{3.21}$$

where M is as in (3.11).

Proof. Clearly, (ii) implies (i). The conditions of the theorem together with Condition (i) imply the hypotheses of Theorem 4. However, Case (b) of Theorem 4 cannot hold because here $P(a,x;b,y)$ is strictly increasing in x by hypothesis. Thus, Case (a) of Theorem 4 holds, and we must have (3.11)

$$P(a,x;b,y) = M\left(\frac{T(a/x)x}{T(b/y)y}\right)$$

for some real valued, continuous function $T>0$. Note that we can assume, without loss of generality, that $T(1)=1$. Applying the Strong Conjoint Weber Law, we get (after simplification)

$$\frac{T(a/x)}{T(b/y)} = \frac{T(\lambda a/\tau x)}{T(\lambda b/\tau y)} \qquad (a,x,b,y,\lambda,\tau>0).$$

Setting $b/y=1$, $s=a/x$, $t=\lambda/\tau$, we obtain

$$T(s)T(t) = T(st) \qquad (s,t>0),$$

a Cauchy power equation [cf. (2.6)]. Accordingly (cf. Corollary 3 to Theorem 1), we have $T(s)=s^\beta$ for some constant β. Thus, (3.21) must hold, with $0<\beta<1$ because $P(a,x;b,y)$ is strictly increasing in both a and x. □

[1] In Falmagne and Iverson (1979), a related condition was also investigated, represented by the equation $P(a,x;b,y)=P(\lambda a,\lambda x;\tau b,\tau y)$.

3.4 The Near-Miss-to-Weber's Law

We turn now to some recent functional equation results and generalize Weber's Law considerably in terms of a monomial equation [see (3.28)], to which for historical reasons we refer as the *Near-Miss-to-Weber's Law.* (In the experimental literature of psychophysics, a one-dimensional version of the monomial equation (3.28) is referred to by that name; cf. Falmagne, 1985).

In Section 3.4.1, which summarizes results from Aczél and Falmagne (1999), we consider a situation in which an individual is presented with a pair $(\mathbf{x}_n, \mathbf{y}_m)$ of stimuli where $\mathbf{x}_n$ and $\mathbf{y}_m$ denote real positive vectors, i.e., $\mathbf{x}_n \in \mathbb{R}_+^n$, $\mathbf{y}_m \in \mathbb{R}_+^m$, with $n \geq 2$. For example, $\mathbf{x}_n$ is a stimulus described by a number of components, $\mathbf{y}_m$ denotes a 'noisy' background, which may also be described by components, and $P(\mathbf{x}_n, \mathbf{y}_m)$ is the probability that the individual detects stimulus $\mathbf{x}_n$ in the backgound $\mathbf{y}_m$. We suppose that the probabilities P satisfy a *Fechner-Thurstone* difference representation

$$P(\mathbf{x}_n, \mathbf{y}_m) = F\big(u(\mathbf{x}_n) - g(\mathbf{y}_m)\big) \qquad (\mathbf{x}_n \in \mathbb{R}_+^n,\ \mathbf{y}_m \in \mathbb{R}_+^m), \tag{3.22}$$

where u and g denote real valued functions. The more general *Gain Control* representation

$$P(x, y) = F\left(\frac{u(x) - g(y)}{h(x)}\right) \qquad (x \in \mathbb{R}_+,\ y \in \mathbb{R}_+) \tag{3.23}$$

is examined in Section 3.4.2, which describes results from Falmagne and Lundberg (2000). Thus, the functional equations (3.22) and (3.23) play here a role similar to that of (3.7) and (3.8) in Section 3.1, while a monomial equation [see (3.28) and (3.37)] replaces the Conjoint Weber Law.

3.4.1 Fechner-Thurstone Model

We begin by simplifying the notation. For reasons that soon will be apparent, we single out the last component of $\mathbf{x}_n$ in (3.22) and write

$$\mathbf{x}_n = (\mathbf{x}, x), \qquad \text{with } \mathbf{x} \in \mathbb{R}_+^{n-1}, x = x_n \in \mathbb{R}_+$$
$$\mathbf{y}_m = \mathbf{y} \in \mathbb{R}_+^m.$$

To avoid multiplying the parentheses, $P(\mathbf{x}, x; \mathbf{y})$ denotes probability of detecting stimulus $(\mathbf{x}, x)$ in the background $\mathbf{y}$. The function P is assumed to be continuous, strictly increasing in its first n variables, and strictly decreasing in its last m variables. Close attention is paid to the domains of variations of all the variables involved. For $1 \leq i \leq n$ and $1 \leq j \leq m$, let $]a_i, a_i'[$ and $]b_j, b_j'[$ be $n + m$ real open intervals, with $0 < a_i < 1 < a_i'$ and $0 < b_j < 1 < b_j'$. Singling out the interval $]a_n, a_n'[$, we define the Cartesian products

$$A_{n-1} =]a_1, a_1'[\times \ldots \times]a_{n-1}, a_{n-1}'[\qquad (n > 1), \tag{3.24}$$

$$B_m =]b_1, b_1'[\times \ldots \times]b_m, b_m'[\qquad (m \geq 1). \tag{3.25}$$

We also suppose that the probabilities P satisfy a *Fechner-Thurstone* difference representation

$$P(\mathbf{x}, x; \mathbf{y}) = F\big(u(\mathbf{x}, x) - g(\mathbf{y})\big) \qquad (\mathbf{x} \in A_{n-1},\ x \in]a_n, a_n'[,\ \mathbf{y} \in B_m), \tag{3.26}$$

which is a special case of (3.23), with u, g and F continuous, real valued, and strictly increasing in all relevant variables. Researchers are typically much more interested in the forms of the functions u and g than in that of F. For that reason, they routinely study the

phenomenon represented in (3.26) by estimating empirically x so that $P(\mathbf{x}, x; \mathbf{y}) = \rho$, for some values of ρ, and for many values of the variables involved in $\mathbf{x}$ and $\mathbf{x}$. In other terms, they study the function $\xi : (\mathbf{x}, \mathbf{y}, \rho) \mapsto \xi(\mathbf{x}, \mathbf{y}; \rho)$ satisfying

$$\xi(\mathbf{x}, \mathbf{y}; \rho) = x \iff P(\mathbf{x}, x; \mathbf{y}) = \rho; \tag{3.27}$$

thus $P(\mathbf{x}, \xi(\mathbf{x}, \mathbf{y}; \rho); \mathbf{y}) = \rho$. Notice that for any fixed $\mathbf{x}$ in A_{n-1} and $\mathbf{y}$ in B_m, the function $x \mapsto P(\mathbf{x}, x; \mathbf{y})$ is strictly increasing and continuous on $]a_n, a'_n[$. Accordingly, its range

$$S_{\mathbf{x},\mathbf{y}} = P(\mathbf{x},]a_n, a'_n[; \mathbf{y})$$

must be an open interval, and so $\xi(\mathbf{x}, \mathbf{y}; \rho)$ is defined for all points $\mathbf{x} \in A_{n-1}$, $\mathbf{y} \in B_m$ and $\rho \in S_{\mathbf{x},\mathbf{y}}$. A simple model for the function ξ is offered by the product

$$\xi(\mathbf{x}, \mathbf{y}; \rho) = \prod_{i=1}^{n-1} x_i^{-\eta_i(\rho)} \prod_{j=1}^{m} y_j^{\zeta_j(\rho)} C(\rho) \tag{3.28}$$

$$(\mathbf{x} = (x_1, \ldots, x_{n-1}) \in A_{n-1},\ \mathbf{y} = (y_1, \ldots, y_m) \in B_m,\ \rho \in S_{\mathbf{x},\mathbf{y}}).$$

This monomial representation has the form of the laws of classical physics and is a natural one to consider here if each of the components is measured on a ratio scale, as is the case in the empirical example mentioned at the beginning of Section 3. Aczél and Falmagne (1999) investigated the compatibility of the representations (3.26) and (3.28), which are linked by the equivalence (3.27). They showed that, under reasonable background assumptions concerning the domains for the variables x_i and y_j in (3.28), the equations (3.26) and (3.28) force all functions η_i in (3.28) to be constant. Moreover, either all functions ζ_j must also be constant and $C = \exp \circ F^{-1}$, where F^{-1} is the inverse of the function F in (3.26); or, if at least one of the ζ_j's is nonconstant, then all of them must have the form $\zeta_j(\rho) = \theta_j \exp[\delta F^{-1}(\rho)]$, for some constants $\theta_j > 0$ $(1 \leq j \leq m)$ and $\delta \neq 0$. None of these results hinges on the assumption that the function P is measuring a probability, i.e. is bounded above by 1 and below by 0. (It certainly need not be a 'discrimination probability' in the sense used in Section 3.1.1.) This can be achieved just by choosing the otherwise arbitrary continuous and strictly increasing function F so that its value lie between 0 and 1. Aczél and Falmagne's main result is reproduced below. Notice that F turns out to be arbitrary as in the solution of (3.1)-(3.2).

Theorem 7. *Suppose that (P, ξ) is a pair of functions related by the equivalence* (3.27). *The following three conditions are then equivalent.*

(i) *The function P satisfies* (3.26) *for some functions u, g and F strictly increasing and continuous in all arguments, but otherwise arbitrary. Moreover, ξ satisfies* (3.28) *for some positive functions C, η_i $(1 \leq i \leq n-1)$, and ζ_j $(1 \leq j \leq m)$, all defined on $J = P(A_{n-1} \times]a_n, a'_n[\times B_m)$, with at least one of the ζ_j nonconstant.*

(ii) *The function P satisfies* (3.26) *with F strictly increasing and continuous, and with u, g specified by*

$$u(\mathbf{x}, x) = \frac{1}{\delta} \ln \ln \left(\frac{\tilde{A}}{x \prod_{i=1}^{n-1} x_i^{\alpha_i}} \right)^{\gamma} \tag{3.29}$$

$$g(\mathbf{y}) = \frac{1}{\delta} \ln \ln \left(\frac{\tilde{B}}{\prod_{j=1}^{m} y_j^{\theta_j}} \right)^{\gamma} \tag{3.30}$$

for some constants $\alpha_i > 0$ $(1 \le i \le n-1)$, $\theta_j > 0$ $(1 \le j \le m)$, γ, δ, $\tilde{A}$ *and* $\tilde{B}$, *the latter four satisfying either* Case [a] *or* Case [b] *below:*

$$\text{(Case [a])} \qquad \delta < 0 < \gamma, \qquad \tilde{A} \ge a'_n \prod_{i=1}^{n-1} a_i'^{\alpha_i} > 1, \qquad \tilde{B} \ge \prod_{j=1}^{m} b_j'^{\theta_j},$$

$$\text{(Case [b])} \qquad \delta > 0 > \gamma, \qquad 0 < \tilde{A} \le a_n \prod_{i=1}^{n-1} a_i^{\alpha_i}, \qquad 0 < \tilde{B} \le \prod_{j=1}^{m} b_j^{\theta_j}.$$

Accordingly, the function P *takes the form*

$$(3.31)\quad P(\mathbf{x}, x; \mathbf{y}) = F\left(\frac{1}{\delta}\ln\ln\left(\frac{\tilde{A}}{x\prod_{i=1}^{n-1} x_i^{\alpha_i}}\right)^{\gamma} - \frac{1}{\delta}\ln\ln\left(\frac{\tilde{B}}{\prod_{j=1}^{m} y_j^{\theta_j}}\right)^{\gamma}\right)$$

$$= \begin{cases} F\left[\frac{1}{\delta}\ln\ln\left(\frac{\tilde{A}}{x\prod_{i=1}^{n-1} x_i^{\alpha_i}}\right) - \frac{1}{\delta}\ln\ln\left(\frac{\tilde{B}}{\prod_{j=1}^{m} y_j^{\theta_j}}\right)\right] & \text{if } \gamma > 0, \\ F\left[\frac{1}{\delta}\ln\ln\left(\frac{x\prod_{i=1}^{n-1} x_i^{\alpha_i}}{\tilde{A}}\right) - \frac{1}{\delta}\ln\ln\left(\frac{\prod_{j=1}^{m} y_j^{\theta_j}}{\tilde{B}}\right)\right] & \text{if } \gamma < 0. \end{cases}$$

(iii) *The function* ξ *satisfies* (3.28) *for some positive functions* C, η_i *and* ζ_j, *all defined on* $J = P(A_{n-1} \times]a_n, a'_n[\times B_m)$, *with constant* $\eta_i = \alpha_i$ $(1 \le i \le n-1)$, *and nonconstant* ζ_j $(1 \le j \le m)$. *Moreover, there exist constants* δ, $\theta_j > 0$ $(1 \le j \le m)$, $\tilde{A}$, $\tilde{B}$ *satisfying either* Case [a] *or* Case [b] *above, so that for all* $\rho \in J$

$$(3.32)\qquad \zeta_j(\rho) = \theta_j \exp[\delta G(\rho)], \quad (1 \le j \le m),$$

$$(3.33)\qquad C(\rho) = \tilde{B}^{-\exp[\delta G(\rho)]}\tilde{A},$$

where G *is a strictly increasing and continuous (but otherwise arbitrary) function on* J. *Consequently,* (3.28) *takes the form*

$$(3.34)\qquad \xi(\mathbf{x}, \mathbf{y}; \rho) = \tilde{A}\prod_{i=1}^{n-1} x_i^{-\alpha_i}\left(\frac{1}{\tilde{B}}\prod_{j=1}^{m} y_j^{\theta_j}\right)^{\exp[\delta G(\rho)]}.$$

The proof of this result contained in Aczél and Falmagne (1999) is quite long and we only summarize it here.

They begin by establishing the implication "(i) ⇒ (iii)", and prove that, when one of the ζ_j's in (3.28) is nonconstant, then they all must be nonconstant and of the form specified by (3.32), with $G = F^{-1}$. They then prove (3.33). Finally, they show that all η_i must be constant if one of the ζ_j's is nonconstant. Equation (3.34) obtains. The representations (3.31) and (3.34) follow easily from each other, with $G = F^{-1}$ and γ arbitrarily positive or negative in Case [a] or [b], respectively. We have thus "(i) ⇒ (iii) ⟺ (ii)". It remains to establish "(ii) and (iii) ⇒ (i)", which is readily obtained by observing that (3.31) has the form (3.26), with u and g defined by (3.29) and (3.30), and that (3.34) has the form (3.28), with the η_i constant, and the ζ_j and C defined by (3.32) and (3.33), respectively.

3.4.2 Gain Control Model

The motivation here is similar to that of Aczél and Falmagne's work reviewed in the last section. The two major differences are: (1) only the case $n = m = 1$ of the monomial equation (3.28) is considered; (2) the Fechner-Thurstone model (3.26) is replaced by the Gain Control model embodied in the equation

$$P(x,y) = F\left(\frac{u(x)-g(y)}{h(x)}\right) \qquad (x \in]a,a'[,\ y \in]b,b'[), \tag{3.35}$$

where $h > 0$, $]a,a'[$ and $]b,b'[$ are real, open, positive intervals, and the function P is continuous, strictly increasing in the first variable, and strictly decreasing in the second variable.

To be specific, for any $x \in]a,a'[$, let $I_x = P(x,]b,b'[) \subseteq]0,1[$ stand for the range of the function $y \mapsto P(x,y)$, and define the function ξ by the equivalence

$$\xi(x;\rho) = y \iff P(x,y) = \rho \qquad (x \in]a,a'[,\ \rho \in I_x). \tag{3.36}$$

(This is the equivalence (3.27) for $m = n = 1$.) A simple model for the function ξ is offered by the power law

$$\xi(x;\rho) = x^{\zeta(\rho)}\psi(\rho)\,. \tag{3.37}$$

with ζ and ψ real valued, positive, continuous functions.

In the style of Theorem 7, Falmagne and Lundberg (2000) derive the consequences of (3.35) and (3.37) holding jointly, on the form of the functions u, g, h, ζ and ψ. They prove the following.

Theorem 8. *Suppose that the functions P and ξ are linked by the equivalence* (3.36). *Suppose moreover that the following two conditions are satisfied:*

(i): *the function P satisfies* (3.35) *for some continuous and strictly increasing functions u, g, h, and F, with $h > 0$;*

(ii): *the function ξ satisfies* (3.37) *for some continuous, positive functions ζ and ψ, both of them defined on $R(P)$ and strictly decreasing.*

Then, there exist five positive constants $\alpha, \beta, \gamma, \delta$ and ν, and two constants θ and ς such that

$$u(x) = \alpha\nu(\ln x + \delta)^{\beta} + \varsigma, \tag{3.38}$$

$$g(y) = \nu(\gamma \ln y - \theta)^{\beta} + \varsigma, \tag{3.39}$$

$$h(x) = \nu(\ln x + \delta)^{\beta}, \tag{3.40}$$

with

$$\ln a + \delta \geq 0, \quad \gamma \ln b - \theta \geq 0. \tag{3.41}$$

Accordingly, (3.35) *takes the particular form*

$$P(x,y) = F\left(\alpha - \left(\frac{\gamma \ln y - \theta}{\ln x + \delta}\right)^{\beta}\right). \tag{3.42}$$

Moreover, the functions ζ and ψ in (3.37) *can be written as*

$$\zeta(\rho) = (1/\gamma)\left[\alpha - F^{-1}(\rho)\right]^{1/\beta}, \tag{3.43}$$

$$\psi(\rho) = \exp\left((1/\gamma)\left(\delta(\alpha - F^{-1}(\rho))^{1/\beta} + \theta\right)\right). \tag{3.44}$$

The proof is based on the solution of the functional equation

$$\varphi(t\,v(s) + w(s)) = s\,k(t) + f(t), \tag{3.45}$$

for real variables s and t, which we state here as a lemma. It is assumed that the domain of (3.45) is a region (open, connected) $D \subseteq \mathbb{R}^2$.

Lemma 2. *Suppose that* (3.45) *is satisfied by the real valued continuous functions* φ, w, v, k, *and* f *on a region* $D \subseteq \mathbb{R}^2$. *Suppose further that* k *and* v *are nonconstant and have no zeroes. Then, there are constants* α, δ, ς, *and* κ *such that*

$$w(s) = \alpha v(s) + \varsigma, \tag{3.46}$$

$$f(t) = \delta k(t) + \kappa, \tag{3.47}$$

and constants β, ν, *and* γ *where* $\beta\gamma\nu \neq 0$ *such that*

$$v(s) = \nu|s + \delta|^{\beta}, \tag{3.48}$$

$$k(t) = \frac{1}{\gamma}|t + \alpha|^{1/\beta}. \tag{3.49}$$

For a proof, see Falmagne and Lundberg (1999).

Sketch of a proof of Theorem 8. Applying the inverse F^{-1} of F on both sides of (3.35), we get, with $P(x, y) = \rho$ and after rearranging,

$$g(y) = u(x) - F^{-1}(\rho)h(x).$$

Solving for $y = \xi(x; \rho)$ and using also (3.37)

$$\xi(x; \rho) = g^{-1}[u(x) - F^{-1}(\rho)h(x)] = x^{\zeta(\rho)}\psi(\rho).$$

Taking logarithms in the last equation (both sides are positive) yields

$$(\ln \circ g^{-1})[u(x) - F^{-1}(\rho)h(x)] = \zeta(\rho)\ln x + (\ln \circ \psi)(\rho). \tag{3.50}$$

With

$$\varphi = \ln \circ g^{-1}, \quad s = \ln x, \quad w(s) = u(e^s), \quad t = -F^{-1}(\rho) \tag{3.51}$$

$$v(s) = h(e^s), \quad k(t) = (\zeta \circ F)(-t), \quad f(t) = (\ln \circ \psi \circ F)(-t), \tag{3.52}$$

(3.50) leads to (3.45). Moreover, it can be verified that all conditions of Lemma 2 are satisfied here. Thus, the functions w, f, v and k of (3.45) have the form given by (3.46), (3.47), (3.48) and (3.49), for the relevant constants.

Since $v(s) = h(e^s)$ by (3.52), v is strictly increasing; thus $s + \delta$ in (3.48) cannot change sign. It can in fact be shown that $s + \delta < 0$ for all values of s leads to a contradiction. Thus, we can drop the abslolute value symbols in (3.48) and write $v(s) = \nu(s + \delta)^{\beta}$. A similar argument of sign preserving permits to rewrite (3.49) as $k(t) = \frac{1}{\gamma}(t + \alpha)^{1/\beta}$. Using the expressions obtained for w, f, v and k together with (3.51) and (3.52) leads to the forms of the functions u, g and h given in Theorem 8, and thus to (3.42), (3.43) and (3.44) □

4. Utility Theory

4.1 General Background—Binary Gambles

A key notion in classical utility theory in the presence of uncertainty is that of a 'binary gamble.' It is formalized in terms of two primitive concepts: (1) a set $\mathcal{C}$ of valued goods (and bads) generically called *consequences*; (2) an algebra $\mathcal{E}$ of events on a set $\Gamma = \cup\mathcal{E}$. (No probability measure is postulated in the following developments.) A *binary (first order) gamble* is a triple

$$(x, C; y) \qquad (x, y \in \mathcal{C},\ C \in \mathcal{E}).$$

The empirical interpretation is that an individual who chooses gamble $(x, C; y)$ over other gambles receives consequence x if the event C is realized in an experiment, and consequence y otherwise. It is understood that some more or less precise information regarding the chances of the event C occuring is available to the individual. In many cases, no actual experimental trial takes place, and gambles are compared by the individual using 'thought experiments.' Whereas, in many situations, the gambles compared by an individual may arise from different sample spaces, we only consider here the case of gambles defined with respect to a fixed sample space Γ and algebra of events $\mathcal{E}$. (For a full presentation of the general theory, involving n-ary gambles and different samples, see Luce, 2000a.) We also deal here with *compound binary gambles* such as

$$((x, C; y), D; z) \qquad (x, y, z \in C;\ C, D \in \mathcal{E}), \tag{4.1}$$

denoting a situation where the individual holding this gamble gets gamble $(x, C; y)$ if the event D is realized in an experiment. In this case, a new experiment is performed and the individual receives x if C occurs, and y otherwise. We may also consider cases in which, in (4.1), z itself may be a gamble. All theoretically feasible sets of compound gambles of any level can be defined recursively, with

1. $\mathcal{B}_0 = C$;
2. for $n \geq 0$, $\mathcal{B}_{n+1} = \mathcal{B}_n \cup \{(x, C; y) | x, y \in \mathcal{B}_n,\ C \in \mathcal{E}\}$.

The set of all compound gambles of any level is then defined by $\mathcal{B} = \cup_{n=0}^{\infty} \mathcal{B}_n$. In practice, however, it is not realistic to suppose that individuals can conceptualize compound gambles of levels higher than 2. Accordingly, in the rest of this section, we only consider the case of gambles in $\mathcal{B}_2$.

A preference order $\succsim$ over $\mathcal{B}_2$ is assumed to exist, which is transitive and connected; thus, $\succsim$ is a *weak order*. Denote the converse order by $\precsim$, indifference by $\sim = \succsim \cap \precsim$, and strict preference by $\succ = \succsim \setminus \sim$. Within C, there is assumed to be a unique element e, which may be thought of as "no change from the status quo," neither a gain nor a loss. Any element $x \in \mathcal{B}_2$ is called a *gain* (or good) if $x \succsim e$ and a *loss* (or bad) if $x \prec e$.

Much of utility theory has studied behavioral conditions (axioms) that describe interlocks between the gambling structure and the preference order. In many cases, one is able to show that the preference order $\succsim$ can be represented numerically. In all such situations, the representations are order preserving in the sense that there is a *utility* function $U : \mathcal{B}_2 \to \mathbb{R}$ satisfying the following two formulas

$$x \succsim y \iff U(x) \geq U(y) \qquad (x, y \in \mathcal{B}_2), \tag{4.2}$$

$$U(e) = 0. \tag{4.3}$$

Conditions for the existence of such an order-preserving representation are well known, namely, $\succsim$ must indeed be a weak order and $\langle \mathcal{B}_2, \succsim \rangle$ must include a countable order-dense subset (e.g., Krantz, Luce, Suppes, and Tversky, 1971, Ch. 2). What makes utility theory interesting is the investigation of possible decompositions of $U[(x, C; y)]$ in terms of components $U(x)$, $U(y)$, and a weighting function $W : C \mapsto W(C)$ mapping the algebra of events $\mathcal{E}$ into $[0, 1]$.

4.2 Basic Behavioral Assumptions

The following assumptions, which all seem a priori rational and for which there is some empirical support, underlie the work. They hold for any $x, y, z \in \mathcal{B}_1$ and $C \in \mathcal{E}$, with $\overline{C} = \Gamma \setminus C$:

Certainty: $(x, \Gamma; y) \sim x$.

Idempotence: $(x, C; x) \sim x$.

Complementarity: $(x, C; y) \sim (y, \overline{C}; x)$.

Consequence Monotonicity on First Component: If $\emptyset \subset C \subset \Gamma$, then

$$x \succsim y \Longleftrightarrow (x, C; z) \succsim (y, C; z).$$

Order Independence of Events: If $x, y \succ e$, then

$$(x, C; e) \succsim (x, D; e) \Longleftrightarrow (y, C; e) \succsim (y, D; e).$$

Status-Quo Event Commutativity: If $x \in \mathcal{B}_0$ and $x \succsim e$, then

$$((x, C; e), D; e) \sim ((x, D; e), C; e), \tag{4.4}$$

where the successive gambles are from independent realizations of the chance experiment or phenomenon. Note that, in (4.4), x is the consequence resulting from the occurence of both C and D in independently run experiments, regardless of the order in which they occur; in all other cases, the result is e.

In the rest of this section, except when otherwise indicated, the variables x, y, z etc. appearing in the notation of gambles or alone denote elements in $\mathcal{B}_1$.

4.3 Bilinear Representations

One goal of the theory has been to understand the circumstances that give rise to the class of bilinear representations

$$U[(x, C; y)] = U(x)W(C) + U(y)\varphi[W(C)], \tag{4.5}$$

where φ is some strictly decreasing function from $[0, 1]$ to $[0, 1]$. In doing so, we assume that the range of U is a half-open interval $[0, k[$, where $k = \infty$ is a possibility although for simplicity here we assume $k < \infty$, and the range of W is the closed interval $[0, 1]$. Conditions are known that are sufficient to ensure these ranges for U and W. For simplicity in the sequel, we use the abbreviation $U(x, C; y)$ to mean $U[(x, C; y)]$.

Two classes of bilinear models have received a great deal of attention: those giving rise to the *rank-independent utility* (RIU) representation in which there are no constraints on the (x, y) pairs and those that give rise to the rank-dependent utility (RDU) ones in which we require $x \succsim y$ and assume that the other case is covered by the behavioral indifference of complementarity stated above. We will take up RIU and RDU in that order.

4.4 Axiomatization of Binary RIU Representations

4.4.1 Two Trade-offs

Luce (1998) explored the fact that (4.5) consists of two important kinds of trade-offs. The one is the additive trade-off between the pairs (x, C) and $(y, \overline{C})$. It has been axiomatized in the literature under the name *additive conjoint measurement* (e.g., Krantz, Luce, Suppes, and Tversky, 1971, Ch.6) and yields an additive form

$$U(x, C; y) = \Psi_1(x, C) + \Psi_2(y, C), \tag{4.6}$$

where, for $i = 1, 2$, $\Psi_i : \mathcal{C} \times \mathcal{E} \to [0, k[$ is surjective (onto) and $\Psi_i(e, C) = 0$. By certainty and (4.6),

$$U(x) = U(x, \Gamma; y) = \Psi_1(x, \Gamma) + \Psi_2(y, \Gamma).$$

Thus, $\Psi_2(y, \Gamma)$ is a constant, which must be 0 because $\Psi_2(e, \Gamma) = 0$.

The other trade-off in (4.5) arises when one sets $y = e$, yielding a multiplicatively separable trade-off between x and C, namely, $U_1(x)W_1(C)$. The key behavioral property underlying the latter trade-off, called the Thomsen condition, was shown by Luce (1996) to

derive from the above assumption of status-quo event commutativity, (4.4). Moreover, W_1 preserves the order of event inclusion, and so $W_1(\emptyset) = 0$ and $W_1(\Gamma) = 1$.

Because both U_1W_1 and Ψ_1 preserve the same order there is a strictly increasing and surjective $f : [0, k[\to [0, k[$ such that for all $x \in \mathcal{C}$, $C \in \mathcal{E}$,

$$\Psi_1(x, C) = f[U_1(x)W_1(C)], \quad f(0) = 0. \tag{4.7}$$

Note that setting $C = \Gamma$ and using $\Psi_2(y, \Gamma) = 0$ yields from (4.6)

$$\begin{aligned} f[U_1(x)] &= f[U_1(x)W_1(\Gamma)] \\ &= \Psi_1(x, \Gamma) \\ &= \Psi_1(x, \Gamma) + \Psi_2(y, \Gamma) \\ &= U(x, \Gamma; y) \\ &= U(x). \end{aligned}$$

Under suitable structural conditions, one can also justify the existence of certainty equivalents $CE(x, C; y) \sim (x, C; y)$, where $CE : \mathcal{B}_2 \to \mathcal{B}_0$, and so we can write

$$U(x, C; y) = f\,(U_1[CE(x, C; y)]) = f[U_1(x, C; y)].$$

Thus, from (4.6)

$$f[U_1(x, C; y)] = f[U_1(x)W_1(C)] + \Psi_2(y, C), \tag{4.8}$$

where $f(0) = 0$, and f is strictly increasing. We explore three approaches—one in this section and the others in 4.5 and 4.6—that are based on this equation.

4.4.2 The RIU Functional Equation

Setting $x = e$ in (4.8) which we can do because there is no constraint on the pair x, y beyond being gains, and using the fact that $U_1(e) = 0$ and complementarity,

$$\begin{aligned} \Psi_2(y, C) &= f[U_1(e)W_1(C)] + \Psi_2(y, C) \\ &= f[U_1(e, C; y)] \\ &= f[U_1(y, \overline{C}; e)] \\ &= f[U_1(y)W_1(\overline{C})] + \Psi_2(e, \overline{C}) \\ &= f[U_1(y)W_1(\overline{C})]. \end{aligned}$$

Substituting this into (4.8)

$$f[U_1(x, C; y)] = f[U_1(x)W_1(C)] + f[U_1(y)W_1(\overline{C})]. \tag{4.9}$$

Setting $x = y$, using idempotence, and introducing the notations $v = U_1(y), w = W_1(C), q(w) = W_1(\overline{C}) = q[W_1(C)]$, we arrive at the functional equation

$$f(v) = f(vw) + f[vq(w)], \quad (v \in [0, k[, w \in [0, 1]), \tag{4.10}$$

where

$$f(0) = 0. \tag{4.11}$$

Since (4.10) is supposed to hold for all $v \in [0, k[$, necessarily $vq(w) \in [0, k[$, and so $q(w) \in [0, 1]$ for all $w \in [0, 1]$.

The derivation of (4.10) forces f to be a strictly increasing mapping of $[0, k[$ onto $[0, k[$ and q to be a strictly decreasing mapping of $[0, 1]$ onto $[0, 1]$. So they are continuous. In the published proof (Aczél, Ger, and Járai, 1999), however, a much weaker assumption sufficed to solve (4.10). This assumption is that f is nonnegative on its domain $[0, k[$. We omit the proof that, except for two trivial solutions of (4.10) [constant on $]0, k[$, cf. (4.30) and

(4.31)], f is continuous and strictly increasing, but we do show the equally surprising result that the problem can then be reduced to the case where f is continuously differentiable.[2] We show that, *under either of these assumptions, the nontrivial solutions of* (4.10) *are*

$$f(v) = \alpha v^{\beta},\ q(w) = (1 - w^{\beta})^{1/\beta}, \quad (v \in]0, k[, w \in]0, 1[, \alpha > 0, \beta > 0). \tag{4.12}$$

If we define $U = U_1^{\beta}$ and $W = W_1^{\beta}$, then from this solution and (4.9) we obtain the rank-independent representation, (4.5), with $W(\overline{C}) = q[W(C)] = 1 - W(C)$. Note that this reduces to binary subjective expected utility if and only if W is finitely additive over disjoint unions.

4.4.3 Solution of RIU Functional Equation

The proof is divided into two major parts.

Reduction to the Differentiable Case: We introduce the integral mean of f :

$$\Theta(u) = \frac{1}{u}\int_0^u f(v)dv, \text{ if } u > 0 \text{ and } \Theta(0) = 0. \tag{4.13}$$

Because f is continuous and positive on $]0, k[$, its integral mean exists, and it is itself strictly monotonic, positive, and continuously differentiable on $[0, k[$ if we define $\Theta'(0) = f(0) = 0$. Surprisingly, integrating (4.10) with respect to v, dividing by u, and setting $s = vw$, $t = vq(w)$, yields the same equation (4.10), this time with Θ in place of f:

$$\begin{aligned}\Theta(u) &= \frac{1}{u}\int_0^u f(v)dv \\ &= \frac{1}{u}\int_0^u f(vw)dv + \frac{1}{u}\int_0^u f[vq(w)]dv \\ &= \frac{1}{uw}\int_0^{uw} f(s)ds + \frac{1}{uq(w)}\int_0^{uq(w)} f(t)dt \\ &= \Theta(uw) + \Theta[uq(w)],\end{aligned}$$

(for $u > 0$ but, with $\Theta(0) = 0$, also for $u = 0$). Because $\Theta'(u) = f(u) \neq 0$, Θ^{-1} is differentiable and so

$$q(w) = \frac{1}{u}\Theta^{-1}[\Theta(u) - \Theta(uw)] \tag{4.14}$$

is differentiable on $]0, 1[$.

Observe that if we can show the differentiable solutions are those given by (4.12), then

$$\Theta(u) = \alpha u^{\beta},\ f(v) = v\Theta'(v) = \alpha\beta v^{\beta} = Av^{\beta}, \quad (v \in [0, k[, A > 0, \beta > 0),$$

and (4.14) yields the same expression as in (4.12).

Solution in the Differentiable Case: So we now assume that f is strictly increasing and continuously differentiable on $[0, k[$ and q is differentiable on $[0, 1]$. Differentiating (4.10) with respect to v and w, we then get

$$f'(v) = wf'(vw) + q(w)f'[vq(w)], \tag{4.15}$$

$$0 = vf'(vw) + vq'(w)f'[vq(w)], \tag{4.16}$$

respectively. Eliminating $f'[vq(w)]$ yields

$$q'(w)f'(v) = [wq'(w) - q(w)]f'(vw). \tag{4.17}$$

[2]The actual history of the proof went the opposite way: It was first solved on the assumption of differentiability, then continuity and strict increasing was shown to suffice, and finally that was established from the equation and non-negativity.

The expression in square brackets is nowhere 0 on $]0,1[$. So, (4.17) can be divided by $wq'(w) - q(w)$ yielding the Pexider equation

$$f'(vw) = f'(v)\psi(w), \quad (v \in]0,k[, w \in]0,1[),$$

where $\psi(w) = q'(w)/[wq'(w) - q(w)]$. From (2.19) in Section 2.1.4 (or Aczél, 1987, pp. 73-74), we know its general continuous solution f' is given by

$$f'(v) = av^b \quad (v \in]0,k[),$$

where $a \neq 0$ because $f'(v) \equiv 0$ has been excluded. Also, $b \neq -1$ because for $b = -1$ the resulting $f(v) = a \ln v + c$ is either negative or decreasing in the (right) neighborhood of 0, which is impossible. So integration yields

$$f(v) = \alpha v^\beta + \gamma \quad (v \in]0,k[).$$

Because f is strictly increasing, continuous at 0, and $f(0) = 0$, we may conclude $\alpha > 0, \beta > 0, \gamma = 0$. Substituting this into (4.10), we obtain (4.12) as the general solution of (4.10) under the assumptions that f is strictly increasing and continuously differentiable on $[0,k[$ and q is differentiable on $]0,1[$. Notice that q turned out to be continuously differentiable on $]0,1[$ (but not necessarily at 0 or 1) and that $q(0) = 1$ and $q(1) = 0$ followed.

As was noted earlier, the weaker assumption that f is nonnegative and nonconstant on $]0,k[$ is sufficient for the conclusion.

4.5 Rank-Dependent Utility: Version 1

4.5.1 The Functional Equation

In the rank-dependent case, we work with the constraint $x \succsim y$ and obtain the form for $x \prec y$ from the assumption of complementarity. The argument is valid up to (4.8), i.e.,

$$f[U_1(x, C; y)] = f[U_1(x)W_1(C)] + \Psi_2(y, C),$$

under the constraint $x \succsim y$. What does not work in the rank-dependent case is setting $x = e$ [this corrects Luce (1998)]. Luce and Marley (2000) explored three approaches to the ranked case.

The first simply postulates without any real axiomatization that the (y, C) terms also have a separable representation, say U_2W_2. So, (4.8) can be written

$$f[U_1(x, C; y)] = f[U_1(x)W_1(C)] + h[U_2(y)W_2(C)]. \tag{4.18}$$

In (4.18) let $x = y$, $v = U_1(y)$, $w = W_1(C)$, and define the function g by $U_2(x) = g[U_1(x)] = g(v)$ and the function q by $q(w) = W_2(C)$. Note that because of order preserving properties and the assumption of order independence of events, g is strictly increasing and q is strictly decreasing with $q(0) = 1$ and $q(1) = 0$. Setting $C = \emptyset$ in (4.18) yields

$$f(v) = h[g(v)], \tag{4.19}$$

and so (4.18) with $x = y$ implies

$$f(v) = f(vw) + f\left(g^{-1}[g(v)q(w)]\right) \quad (v \in [0,k[,\ w, q(w) \in [0,1]), \tag{4.20}$$

where we know that g^{-1} is defined because $g : [0,k[\to [0,k[$ is strictly monotonic and surjective. As a consequence, it is also continuous.

Although we know and will use the fact that f and q are strictly monotonic and continuous, the published proof supposes only that f is nonnegative and nonconstant on $]0,k[$, and derives strict monotonicity and continuity from that and the equation. Equation (4.20) reduces, of course, to (4.10) if g = identity. We need, however, new ideas to solve (4.20). Because f and g both increase strictly and $f(0) = g(0) = 0$, it follows that $q(0) = 1, q(1) = 0$

hold. Ignoring these values for the time being, we sketch the novel road map leading to all nontrivial solutions of (4.20) (for more details, see Aczél, Maksa, Ng, and Páles, 2000).

4.5.2 Solution of Version 1 Functional Equation (4.20)

Once again, the proof is developed in two stages.

Linearizing the Equation and Convexity: First, we "linearize" the functional equation (4.20) with the aid of new functions defined as follows:

$$Q(s) = -\ln q(e^{-s}) \qquad (s \in]0,\infty[=: \mathbb{R}_+), \tag{4.21}$$

and

$$F(t) = f(e^{-t}), \qquad G(t) = -\ln g(e^{-t}), \qquad H(t) = f[g^{-1}(e^{-t})] \qquad (t \in]-\ln k,\infty[\,). \tag{4.22}$$

Because the nontrivial solutions f of (4.20) are strictly increasing and g also increases strictly, F and H both strictly decrease and G strictly increases. With these functions and with $s = -\ln w$, $t = -\ln v$, (4.20) is "linearized" as

$$F(t) - F(t+s) = H[G(t) + Q(s)] \qquad (s \in]0,\infty[, t \in\ ]-\ln k,\infty[\,). \tag{4.23}$$

From this equation, $t \mapsto F(t) - F(t+s)$ is strictly decreasing because on the right hand side H decreases and G increases, both strictly. But then

$$F(t) - F(t+s) > F(t+s) - F[(t+s)+s] \qquad (s > 0),$$

or with $z = t + 2s$,

$$F(\frac{t+z}{2}) < \frac{F(t)+F(z)}{2} \qquad (z > t),$$

that is, F is strictly Jensen (midpoint) convex. Because F is also monotonic, it is strictly convex (see e.g., Roberts and Varberg, 1973, p. 219).

This idea (of Zsolt Páles) and result proved to be very useful because we know a lot about convex functions. For instance, they are continuous; they have everywhere left-side and right-side derivatives except maybe at the boundary of their interval of convexity; moreover, they are differentiable except for at most countably many points (see Roberts and Varberg, 1973, pp. 4-7). Of course, because G and H are monotonic, they are almost everywhere differentiable (see, e.g., Riesz and Szökefalvi-Nagy, 1990, pp. 5-9). We write (4.23) as

$$H^{-1}[F(t) - F(t+s)] = G(t) + Q(s) \qquad (s \in]0,\infty[, t \in\ ]-\ln k,\infty[\,), \tag{4.24}$$

which we can do because H (and thus H^{-1}) is strictly monotonic decreasing and surjective. Thus, H^{-1} is almost everywhere differentiable on its domain

$$\{F(t) - F(t+s)\,|s \in \mathbb{R}_+,\ t \in\ ]-\ln k,\infty[\,\}$$

which, because F is continuous and $t \mapsto F(t) - F(t+s)$ is strictly decreasing, is a nondegenerate interval. One can manipulate the two variables s and t and the differentiability of H almost everywhere and that of F up to at most countably many places, so that the left-hand side of (4.24) turns out to be everywhere differentiable in s. Thus, Q on the right and finally H^{-1} prove to be differentiable on the entire domain. We do not yet know that F and G are everywhere differentiable, but the right-side derivatives F'_+ and G'_+ exist everywhere (as do the left-side ones): the former because of the convexity of F and the latter because of (4.24) and the chain rule. We differentiate (4.24) from the right with respect to s and t, respectively, and then eliminate $(H^{-1})'[F(t) - F(t+s)]$ in order to get

$$Q'(s)[F'_+(t+s) - F'_+(t)] = G'_+(t)F'_+(t+s) \quad (s \in]0,\infty[, t \in\ ]-\ln k,\infty[\,). \tag{4.25}$$

Because F is strictly decreasing and strictly convex, we know that F'_+ is negative everywhere and (see e.g., Kuczma, 1985, p. 156, or Roberts and Varberg, 1973, p. 5) $F'_+(t+s) - F'_+(t)$ is sign preserving (everywhere positive or everywhere negative). Looking at (4.25) more thoroughly, we see that also G'_+ and Q' are sign preserving. With the notation

$$L = \frac{1}{F'_+},\ M = \frac{G'_+}{F'_+},\ N = -\frac{1}{Q'}, \tag{4.26}$$

we obtain from (4.25)

$$L(s+t) = L(t) + M(t)N(s) \qquad (s \in]0,\infty[,\ t \in]-\ln k, \infty[). \tag{4.27}$$

This is related to the equation (2.23) and is a particular case of several known functional equations (see Aczél and Chung, 1982, and Járai, 1984). By our previous results, M and N preserve their signs and so (4.27) implies that L is strictly monotonic. Accordingly, the general solutions for $s \in]0,\infty[,\ t \in]-\ln k, \infty[$ are given by the two sets of functions

$$L(t) = Ae^{ct} + B,\ M(t) = Ce^{ct},\ N(s) = \frac{A}{C}(e^{cs} - 1), \tag{4.28}$$

$$L(t) = At + B,\ M(t) = Ct,\ N(s) = \frac{A}{C}s. \tag{4.29}$$

The Solutions: All we have to do now is "play the tape in reverse" through (4.26), (4.22), and (4.21) and, after passing some hurdles, we arrive at the general solution of the functional equation (4.20) under the given conditions.[3] We recall also $f(0) = g(0) = q(1) = 0$ and $q(0) = 1$. The final result is the following theorem.

Theorem 9. *If f is nonnegative and g is continuous and strictly monotonic, then all solutions of the functional equation* (4.20) *are given first by the trivial (degenerate) solutions:*

$$f(v) = 0 \quad (v \in [0,k[), \quad q : [0,1] \to [0,1], \quad g : [0,k[\to [0,k[, \tag{4.30}$$

(with g strictly increasing and surjective, but otherwise both g and q arbitrary);

$$f(v) = \begin{cases} 0 & \text{if } v = 0, \\ c & \text{if } v \in]0,1[; \end{cases} \quad q(w) = \begin{cases} 0 & \text{if } w \in]0,1[, \\ d, & \text{if } w = 0, \alpha \in]0,1[, \end{cases} \quad g : [0,k[\to [0,k[, \tag{4.31}$$

(with g strictly increasing and surjective but otherwise arbitrary);

second, by

$$f(v) = \alpha v^\beta,\ g(v) = k^{1-\beta\gamma} v^{\beta\gamma},\ q(w) = (1 - w^\beta)^\gamma \qquad (v \in [0,k[, w \in [0,1]), \tag{4.32}$$

and third, by

$$f(v) = \alpha \ln(1 + \mu v^\beta),\ g(v) = k(\mu + k^{-\beta})^\gamma,\ q(w) = (1 - w^\beta)^\gamma \qquad (v \in [0,k[, w \in [0,1]), \tag{4.33}$$

where $\alpha > 0, \beta > 0, \gamma > 0$, and $\mu > -k^\beta$ are constants.

We have thus completely solved (4.20) under weak conditions. Note that the solution (4.12) of (4.10) is the particular case $\gamma = \frac{1}{\beta}$ of (4.32), whereas (4.10) has no solution corresponding to (4.33).

[3]The "hurdles" arise by those conditions and by the unusual domain of validity of (4.27), which is why the theorem of Aczél and Chung (1982) was needed because it holds for arbitrary intervals. At the point where we get F'_+ and G'_+ from (4.26), (4.28), and (4.29), we see that these right-side derivatives are continuous; thus F and G are everywhere differentiable (see Kuczma, 1985, p. 156), and they as well as Q can be determined by integration.

Utility representations: With (4.32), the representation in utility terms becomes

$$U(x,C;y) = U(x)W(C) + U(y)[1 - W(C)] \qquad (x \succsim y),$$

which, except for the condition $x \succsim y$, is the same as the RIU representation, i.e., (4.5) with $g(z) = 1 - z$. This is only half of the story. To get what happens when $x \prec y$, we invoke the assumption of complementarity, i.e., $(x,C;y) \sim (y,\overline{C};x)$, yielding the full *binary rank-dependent* representation

$$U(x,C;y) = \begin{cases} U(x)W(C) + U(y)[1 - W(C)] & (x \succsim y) \\ U(x)[1 - W(\overline{C})] + U(y)W(\overline{C}) & (x \precsim y). \end{cases} \tag{4.34}$$

However, the solution (4.33) yields, after some algebraic manipulation, a new representation of the utility of a binary gamble for $x \succsim y$ and $\mu > -\frac{1}{k^2}$,

$$U(x,C;y) = \frac{U(x)W(C) + U(y)[1 - W(C)] + \mu U(x)U(y)W(C)}{1 + \mu U(y)W(C)}, \tag{4.35}$$

which is called *ratio rank-dependent utility*, or more briefly *RRDU*. Interestingly, the particular case $\mu = 0$ is just (4.34), whereas (4.32) is not a particular case of (4.33) (although it is, in a sense, a limiting case). Luce and Marley (2000) asked what behavioral properties force $\mu = 0$. Three were found of which the most natural is mentioned in Section 4.8.1.

Despite the nonsymmetric form of RRDU, (4.35), it exhibits the important symmetry property called *event commutativity*

$$((x,C;y),D;y) \sim ((x,D;y),C;y), \tag{4.36}$$

which plays a role below. Note that each side says that x is the consequence if in two independent realizations of the chance phenomenon C and D both occur and otherwise y is the consequence. The difference is only in the order of occurrence, and event commutativity assumes that does not matter to the decision maker. Also note that the earlier assumed status-quo event commutativity, used to insure separable representations, is the special case of (4.36) with $y = e$.

4.6 Rank-Dependent Utility: Version 2

Luce and Marley (2000) took another approach that again began with (4.8), but then followed a more principled route. Using the idempotence axiom and setting $x = y$ in (4.8), we get

$$f[U_1(y)] - f[U_1(y)W_1(C)] = \Psi_2(y,C).$$

Substituting this back into (4.8) yields: for $x \succsim y \succsim e$,

$$f[U_1(x,C;y)] = f[U_1(x)W_1(C)] + f[U_1(y)] - f[U_1(y)W_1(C)]. \tag{4.37}$$

This constraint on f seems inadequate for our purpose; we must assume more.

A plausible, experimentally supported, added condition is event commutativity, (4.36), which was just mentioned as a property of RRDU. To see what that implies, let

$$u = U_1(x),\ v = U_1(y),\ w = W_1(C),\ t = W_1(D),\ k = \limsup U_1.$$

Then, for all u, v, w, t such that $k > u \geq v \geq 0$ and $w, t \in [0,1]$, (4.18) yields

$$\begin{aligned} &f[U_1((x,C;y),D;y)] \\ &\quad = f[U_1(x,C;y)W_1(D)] + f[U_1(y)] - f[U_1(y)W_1(D)] \\ &\quad = f[f^{-1}(f[U_1(x,C;y)])W_1(D)] + f[U_1(y)] - f[U_1(y)W_1(D)] \\ &\quad = f\left(f^{-1}[f[U_1(x)W_1(C)] + f[U_1(y)] - f[U_1(y)W_1(C)]]W_1(D)\right) \\ &\qquad + f[U_1(y)] - f[U_1(y)W_1(D)] \\ &\quad = f\left(f^{-1}[f(uw) + f(v) - f(vw)]t\right) + f(v) - f(vt). \end{aligned}$$

Thus, event commutativity yields

$$f\left(f^{-1}[f(uw) + f(v) - f(vw)]t\right) - f(vt) = f\left(f^{-1}[f(ut) + f(v) - f(vt)]w\right) - f(vw), \tag{4.38}$$

where one has to show that $f(uw) + f(v) - f(vw)$ is in the range of f.

Aczél and Maksa (2000) solved this equation:

Theorem 10. *If the domain and the range of f are intervals, f is strictly monotonic and twice differentiable, then for some constants $\alpha, \beta, \gamma, \eta$, and μ*

$$f(v) = \eta \ln(\mu v^\beta + \gamma), \tag{4.39}$$

$$f(v) = \eta v^\beta + \gamma, \tag{4.40}$$

$$f(v) = \eta \ln(\mu v^\beta + \gamma). \tag{4.41}$$

If f is also supposed to be strictly increasing and $f(0) = 0$, then there are just the following two solutions:

$$f(v) = \eta \ln(\mu v^\beta + 1) \qquad (\mu > -\frac{1}{k^\beta},\ \mu\gamma > 0,\ \beta > 0), \tag{4.42}$$

$$f(v) = \eta v^\beta \qquad (\eta > 0,\ \beta > 0). \tag{4.43}$$

From (4.42) one derives again the ratio rank-dependent form (4.35) with the added restriction $\mu \neq 0$. And (4.43) yields RDU, i.e., the case with $\mu = 0$. For the cases where $x \precsim y$, we simply use complementarity and the representation with the roles of x and y interchanged.

We hope that future research will remove the differentiability assumption, as was the case for RIU and the first version of RDU (Section 4.5).

4.7 Rank-Dependent Utility: Version 3

Marley and Luce (2000) gave an axiomatization based on two major properties, event commutativity and the following assumption called *gains partition:* There exists a permutation M on $\mathcal{E}$ that inverts the order induced by $\succsim$ on $\mathcal{E}$ so that for $x, x', y, y' \in C$, with $x \succsim y$, $x' \succsim y'$, and $C, C' \in \mathcal{E}$, if

$$(x,C;e) \sim (x',C';e) \quad \text{and} \quad (y,M(C);e) \sim (y',M(C');e),$$

then

$$(x,C;y) \sim (x',C';y').$$

At first blush, gains partition may seem trivial; however, it need not hold in a representation where the pieces $U(x)$, $W(C)$, $U(y)$, and $W[M(C)]$ play separate roles, not just as $U(x)W(C)$ and $U(y)W[M(C)]$. Representations of these more complex types have arisen, for example, in Luce's (1997, 2000a) treatment of binary mixed gains and losses, which we do not go into here.

A rather intricate argument, which we do not repeat here, gives rise to the functional equation

$$\frac{z}{p}g^{-1}[zg(p))] = f^{-1}[f(z)q(p)] \qquad (z,p \in]0,1[\,). \tag{4.44}$$

Aczél, Maksa, and Páles (2000) proved the following

Theorem 11. *The general solutions* $f, q :]0,1[\to]0,\infty[$, *both strictly monotonic and* $g :]0,1[\to]0,\infty[$ *strictly decreasing and surjective, of the functional equation* (4.44) *are given by*

$$g(p) = \frac{(1-p^\kappa)^{\frac{1}{\kappa}}}{p}, \quad q(p) = p^{\kappa c}, \quad f(z) = A\left(\frac{1-z^\kappa}{z^\kappa}\right)^c \qquad (p,z \in]0,1[\,),$$

where $A > 0, \kappa > 0, c < 0$ *are otherwise arbitrary constants.*

Equation (4.44) leads to an equation that is similar to (4.25) in Section 4.5.2, but here there is no solution corresponding to (4.32).

Marley and Luce (2000) used Theorem 11 to show that the RDU representation holds with $W[M(C)] = 1 - W(C)$.

4.8 Utility of Joint Receipt

4.8.1 Joint Receipt

Luce (1991) and Luce and Fishburn (1991) were the first to study theoretically the concept of a binary operation $\oplus$ over gambles, called *joint receipt*. Earlier empirical work had invoked the concept, which in one study was called a duplex lottery. It is the very natural idea of having or receiving two (or more, as it turns out) things at once. So, if x and y are two gambles (including pure consequences as a special case), then $x \oplus y$ means having or receiving both of them. Several natural assumptions are made: For all $x, y, z \in \mathcal{B}_1$,

JR Commutativity: $x \oplus y \sim y \oplus x$.

JR Monotonicity: $x \succsim y \iff x \oplus z \succsim y \oplus z$.

JR Identity: $x \oplus e \sim x$.

A major scientific question is: How does joint receipt interlock with the gambling structure? The answer proposed by these authors is the highly rational, and empirically supported, property called (binary) *segregation:*

$$(x,C;e) \oplus y \sim (x \oplus y, C; y) \qquad (x \succsim e,\ y \succsim e,\ C \in \mathcal{E}). \tag{4.45}$$

Luce and Marley (2000) showed that segregation added to rational rank-dependent utility, (4.35), forces $\mu = 0$, i.e., binary RDU.

4.8.2 RDU and Segregation

Given the concept of joint receipt $\oplus$, binary rank-dependent utility, and segregation, Luce and Fishburn (1991, 1995) noted that by applying the utility function U of binary RDU to this expression and using the separability assumption, we obtain

$$U\left(U^{-1}[U(x)W(C)] \oplus U^{-1}U(y)\right) = U(x \oplus y)W(C) + U(y)[1 - W(C)].$$

With

$$u = U(x), \quad v = U(y), \quad w = W(C), \quad \text{and} \quad G(u,v) = U\left[U^{-1}(u) \oplus U^{-1}(v)\right], \tag{4.46}$$

the last equation becomes

$$G(uw,v) = G(u,v)w + v(1-w) \qquad (0 < v \le u < k, 0 < w < 1), \tag{4.47}$$

where, by the monotonicity of $\oplus$, G is strictly increasing in each argument, which fact, however, is not used in what follows. We prove that the solution is

$$G(u,v) = \alpha(v)u + v, \tag{4.48}$$

where $\alpha :]0,k[\to]0,k[$ is an arbitrary function. If we define

$$g_v(u) = G(u,v) - v, \tag{4.49}$$

then (4.47) can be written as

$$g_v(uw) = wg_v(u).$$

By fixing $u = c$ close to k and setting $\alpha(v) = g_v(c)/c$, we get

$$g_v(t) = \frac{t}{c} g_v(c) = \alpha(v)t \qquad (0 < t < c).$$

This can be extended to all $t \in]0,k[$. From (4.49) we then get (4.48). Conversely, (4.47) is satisfied by (4.48) for arbitrary α.

If we also require, as follows from JR commutativity and (4.47),

$$G(u,vw) = G(u,v)w + u(1-w), \tag{4.50}$$

then we get similarly $G(u,v) = \beta(u)v + u$. Comparison with (4.48) gives

$$\alpha(v)u + v = \beta(u)v + u,$$

or rewriting

$$\frac{\alpha(v) - 1}{v} = \frac{\beta(u) - 1}{u} = -\delta \quad \text{(constant)}.$$

Thus,

$$G(u,v) = u + v - \delta uv \tag{4.51}$$

is the general solution of (4.47) and (4.50), where δ is an arbitrary constant.

Rewriting (4.51) in terms of the function U and using (4.46), we find

$$U(x \oplus y) = U(x) + U(y) - \delta U(x)U(y), \tag{4.52}$$

which is called a *p−additive* representation because it is the only polynomial form with $U(e) = 0$ that can be transformed into an additive representation. Thus, the assumptions of RDU, JR idempotence, JR commutativity, and segregation imply:

JR Associativity:: $x \oplus (y \oplus z) \sim (x \oplus y) \oplus z \qquad (x,y,z \in \mathcal{B}_1)$.

4.8.3 p-Additive and Separable Utility

Of course, we also know from the assumption of RDU that gambles of the form $(x, C; e)$, where e is the identity of $\oplus$, have the (multiplicative) separable, order-preserving representation $U(x)W(C)$. From separability and p-additivity with a common utility function and segregation, Luce and Fishburn (1991) showed that RDU follows. So that is a fourth way to axiomatize the RDU representation. This has the virtue of making very transparent the source of rank dependence, namely, segregation.

But it also has a substantial weakness. It is easy to axiomatize separability as noted above. It is also easy to axiomatize the existence of an additive representation (Krantz, Luce, Suppes and Tversky, 1971, Ch. 3), but it is far from clear that both can be done with the same utility function U. Luce (1996) explored the question. A necessary and sufficient condition was found, namely, the structure is said to be *joint-receipt decomposable* if for each $x \in \mathcal{B}_1$ and $C \in \mathcal{E}$, there exists an event $D = D(x,C) \in \mathcal{E}$ such that for all $y \in \mathcal{B}_1$,

$$(x \oplus y, C; e) \sim (x, C; e) \oplus (y, D; e). \tag{4.53}$$

It is not very difficult to show that this property follows when both U is p-additive and UW forms a separable representation. In the process, one shows that D satisfies

$$W(D) = W(C)\frac{1-\delta U(x)}{1-\delta U(x)W(C)}. \tag{4.54}$$

The deeper question concerns the other direction. Suppose we have a p-additive representation U_1 and a separable one U_2W_2, each of which we know how to axiomatize. Does (4.53) mean that there exists a U that is both p-additive and, with some W, also UW is a separable representation? The answer is that for some $\beta > 0$, $U = (U_2)^\beta$ is p-additive and (U, W), where $W = (W_2)^\beta$, is separable. To show this, one proceeds as follows.

Let f be defined by $U_1 = f(U_2)$. Since they both preserve the same order, f is strictly increasing. Then applying these assumptions to joint receipt decomposability and defining $p = U(x)$, $q = U(y)$, $w = W(C)$, $P(p,w) = W\big(D[U^{-1}(p), W^{-1}(w)]\big) = W[D(x,C)]$, one is led to the functional equation

$$H(p,q)w = H[pw, qP(p,w)] \qquad (p,q \in [0,1[, w \in [0,1]), \tag{4.55}$$

where P maps $[0,1[\times[0,1]$ into $[0,1]$ and

$$H(p,q) = F^{-1}[F(p) + F(q) - F(p)F(q)], \tag{4.56}$$

with $F = \delta f$ so that F maps $[0,1[$ onto $[0,1[$ and is strictly increasing.

First, we simplify (4.56) to

$$H(p,q) = \Phi^{-1}[\Phi(p)\Phi(q)] \qquad (p,q \in [0,1[), \tag{4.57}$$

by introducing

$$\Phi(p) = 1 - F(p). \tag{4.58}$$

Clearly, $\Phi : [0,1[\to]0,1]$ is a surjection and is strictly decreasing and so continuous. By (4.57), H is strictly increasing and continuous in each variable, and thus is also continuous as a function of two variables.

Functional equation for P : Repeated application of (4.55) gives

$$H[pws, qP(p,ws)] = H(p,q)ws = H[pw, qP(p,w)]s = H[pws, qP(p,w)P(pw,s)].$$

Because H is strictly increasing in the second variable, P satisfies

$$P(p,ws) = P(p,w)P(pw,s) \qquad (p \in [0,1[, w,s \in [0,1]). \tag{4.59}$$

So our solution plan is: first solve the functional equation (4.59), next plug that solution into (4.55), and then try to use that functional equation to find the general solution H and so, by (4.56), find F.

Note that it would be easy to solve (4.59) if we could substitute $p = 1$, but we cannot (and, as we will see in (4.73) below, $\lim_{p \nearrow 1} P(p,w) = 0$, for which (4.59) reduces to $0 = 0$.)

Solution of equation (4.59): First, one can show that

$$P(0,w) = 0 \tag{4.60}$$

and for $p \in]0,1[$

$$P(p,w) = 0 \text{ iff } w = 0. \tag{4.61}$$

After excluding $p = 0$, $w = 0$, and $s = 0$ from (4.59), we do the next best thing to substituting $p = 1$, namely, we substitute $p = c < 1$ and, to counterbalance, we choose $w = r/c$. Then, with the notation

$$\lambda_c(r) = \frac{1}{P(c, r/c)} \qquad (r \in]0, c[\,),$$

we get

$$P(r,s) = \frac{\lambda_c(r)}{\lambda_c(rs)} \qquad (r \in]0,c[, s \in]0,1]).$$

One easily verifies that λ_c is determined up to a multiplicative constant. So, for $c < c' < 1$, we can choose $\lambda_{c'}$ (defined on $]0,c']$) so that $\lambda_{c'}(c) = \lambda_c(c)$ and, thus, $\lambda_{c'}(s) = \lambda_c(s)$ for all $s \in]0,c[$. Thus, these λ_c's are restrictions to $]0,c]$ of a function $\lambda :]0,1[\to]0,\infty[$ such that

$$P(r,s) = \frac{\lambda(r)}{\lambda(rs)} \qquad (r \in]0,1[, s \in]0,1]). \tag{4.62}$$

This, (4.60), and (4.61) satisfy (4.59), and so we have solved that functional equation. One can show that $\lambda :]0,1[\to]0,\infty[$ is strictly decreasing and continuous, so P is strictly increasing in its second variable and continuous.

First approach to (4.55): Up to here we summarized work in Aczél, Luce, and Maksa (1996). That paper accomplished the second step of the solution plan, determining H, G, and F from (4.60), (4.61), (4.62), (4.55), (4.57) and (4.58), only under the assumption of differentiability of both F and F^{-1} (or equivalently Φ and Φ^{-1}). Several attempts were made to eliminate this condition, which certainly is not inherent to the original utility problem. Partial results were achieved, e.g., Ng (1998) replaced it by assuming differentiability of P in one variable. It is easy to see the difficulties when one substitutes, as planned, (4.62) into (4.55) with (4.57):

$$\Phi^{-1}[\Phi(p)\Phi(q)]w = \Phi^{-1}\left[\Phi(pw)\Phi\left(q\frac{\lambda(p)}{\lambda(pw)}\right)\right] \qquad (p,q \in]0,1[, w \in]0,1]), \tag{4.63}$$

which is a rather intimidating equation.

Second approach to (4.55): One idea that pushed the solution forward was replacing (4.63) by a simpler limiting case obtained by taking the limit as $q \nearrow 1$ in (4.63) which, it turns out, does not add any new solutions (Aczél, Maksa, and Páles, 1999). Notice that $\Phi : [0,1[\to]0,1]$ is decreasing, surjective, and continuous; therefore, in view of (4.55) and (4.57), the following limit equations hold

$$\lim_{q \nearrow 1} \Phi(q) = 0, \quad \lim_{q \nearrow 1} H(p,q) = 1, \quad \Phi(w) = \Phi(pw)\Phi[P(p,w)]. \tag{4.64}$$

[Observe that we did not substitute 1, we took limits in two of the functions as $q \nearrow 1$, but definitely not of $P(p,w)$ as $p \nearrow 1$.] Putting (4.62) into (4.64), we get an equation much simpler than (4.63):

$$\frac{\Phi(w)}{\Phi(pw)} = \Phi\left(\frac{\lambda(p)}{\lambda(pw)}\right) \qquad (p \in]0,1[, w \in [0,1]). \tag{4.65}$$

Now we are ready to execute the somewhat modified second step of our solution plan of determining Φ, and thus H and f, from (4.65).

Determining Φ *from* (4.65): As with (4.20) earlier, we "linearize" the functional equation (4.65) to

$$K(v) - K(u+v) = K[\ell(u) - \ell(u+v)], \tag{4.66}$$

by introducing $u = -\ln p, v = -\ln w$, and defining

$$K(u) = -\ln \Phi(e^{-u}), \ \ell(u) = -\ln \lambda(e^{-u}). \tag{4.67}$$

Because Φ and λ are continuous and strictly decreasing, so too are K, ℓ and, as well, the function $v \mapsto K[\ell(v) - \ell(u+v)]$. Thus,

$$\begin{aligned} K(v) - K(u+v) &= K[\ell(u) - \ell(u+v)] \\ &\geq K\left(\ell(u) - \ell[u + (u+v)]\right) \\ &= K(u+v) - K(2u+v). \end{aligned}$$

So K and, similarly, ℓ are strictly convex, the one-sided derivatives K'_+, K'_-, ℓ'_+, ℓ'_- exist and are strictly increasing everywhere. The "convexity method" of Section 4.5.2 works again, and even more simply. One differentiates (4.66) from the right or left with respect to u or v, respectively, and eventually gets for $L = 1/K'_+$ and for some function Λ the equation [cf. (4.27)]

$$L(u+v) = L(v) + L(u)\Lambda(v),$$

with the strictly decreasing solutions

$$K'_+(u) = \frac{1}{L(u)} = \frac{1}{du} \text{ and, accordingly, } \ell'_+(u) = \frac{\gamma}{u}, \tag{4.68}$$

and

$$K'_+(u) = \frac{1}{L(u)} = \frac{c}{1 - e^{bu}} \text{ and, accordingly, } \ell'_+(u) = \frac{\delta}{e^{\beta u} - 1}. \tag{4.69}$$

But these are continuous and if the right derivatives of the convex functions K and ℓ are continuous, then they are differentiable everywhere.

Integration gives K and ℓ, and the equations (4.67) yield Φ and λ. Substitution into (4.65) and taking into account the fact that the strictly decreasing Φ maps $[0,1[$ onto $]0,1]$ does three things: It eliminates (4.68); it restricts the constants in (4.69) to $b = -\beta > 0$; and it eliminates the constant of integration in Φ. Finally, one gets $g(p) = d(p^{-b} - 1)^{1/b}$ $(b > 0)$ and for $p \in]0,1[, w \in]0,1[$,

$$G(p) = (1 - p^b)^a, \; P(p,w) = w\left(\frac{1 - p^b}{1 - p^b w^b}\right) \qquad (a > 0, b > 0). \tag{4.70}$$

Continuity extends the validity of (4.70) to $p = 0$, $w = 0$, and $w = 1$. Equations (4.57) and (4.58) finally leads to

$$H(p,q) = \left(p^b + q^b - p^b q^b\right)^{1/b}, \; F(p) = 1 - \left(1 - p^b\right)^a \qquad (a > 0, b > 0). \tag{4.71}$$

Equations (4.55) and (4.56) are satisfied by (4.70) and (4.71). This concludes the sketch of the proof of the following.

Theorem 12. *Suppose that the strictly increasing function F maps $[0,1[$ onto $[0,1[$ and P maps $[0,1[\times[0,1]$ into $[0,1]$. Then* (4.55) *and* (4.56) *are satisfied by*

$$F(p) = 1 - (1 - p^b)^a \qquad (p \in [0,1[\,), \tag{4.72}$$

$$P(p,w) = w\left(\frac{1 - p^b}{1 - p^b w^b}\right)^{1/b} \qquad (p \in [0,1[, w \in [0,1]), \tag{4.73}$$

for arbitrary positive constants a and b, and there are no other solutions.

For more detail, see Aczél, Luce, and Maksa (1996) and Aczél, Maksa, and Páles (1999). Using Theorem 12, Luce (1996) showed that if there is a p-additive representation U_1, a separable representation $U_2 W_2$, and the structure is joint-receipt decomposable, then there exists a $\beta > 0$ such that $U = U_1^\beta$ is p-additive and for $W = W_2^\beta$, that UW is a separable representation.

4.9 Invariance Assumptions

In applying utility theory to data, having completely open ended monotonic utility and weighting functions leaves the models insufficiently specified. So one is led to ask whether the forms of these functions can be seriously constrained. We know of such constraints only when the consequences are money and the chance events can be characterized in terms of probabilities. Further work needs to be done for, at least, weights over events, not just probabilities.

4.9.1 Utility of money

Consider money consequences, x and y, $x > 0, y > 0$. Although it may seem plausible that $x \oplus y \sim x + y$, some empirical evidence suggests otherwise. A rather weaker assumption is that money and joint receipt $\oplus$ are related by the a multiplicative invariance condition of the form [cf. (1.1)]

$$\lambda x \oplus \lambda y \sim \lambda(x \oplus y) \qquad (x > 0, y > 0, \lambda > 0). \tag{4.74}$$

If, as arises under the conditions at the end of Section 4.8.2, $\oplus$ has an additive representation, which we denote by V, then Luce (2000a) pointed out that

$$V^{-1}[V(\lambda x) + V(\lambda y)] = \lambda V^{-1}[V(x) + V(y)],$$

which is (2.30) with $V = f$ and so for some $\beta > 0$

$$V(x) = \alpha x^{\beta} \quad \text{and} \quad x \oplus y = \left(x^{\beta} + y^{\beta}\right)^{1/\beta}.$$

4.9.2 Form of Weighting Functions

For known probabilities in gambles of the form $(x, p; 0)$, $p \in [0, 1]$, consider a separable representation UW, where both U and W are strictly increasing in their arguments. Prelec (1998) imposed an invariance condition which he showed to be equivalent to the following form for the weighting function

$$W(p) = \exp\left[-\gamma(-\ln p)^{\eta}\right] \qquad (p \in [0, 1], \gamma, \eta > 0). \tag{4.75}$$

Luce (2000b) showed the same result using the following simpler invariance condition called *reduction invariance*: For all $x \in \mathcal{C}_+$, $p, q, r = \rho(p, q) \in [0, 1]$, and integers $N = 2, 3,$

$$((x, p; 0), q; 0) \sim (x, r; 0) \Longrightarrow ((x, p^N; 0), q^N; 0) \sim (x, r^N; 0). \tag{4.76}$$

Note that $((x, p; 0), q; 0)$ means that x arises with probability pq, and $(x, r; 0)$ means that x arises with probability r; therefore, a person understanding probability theory will have $r = pq$, in which case reduction invariance is automatically satisfied and in (4.75) we have $\eta = 1$ and $W(p) = p^{\gamma}$ $(p \in [0, 1])$.

The proof involves, first, showing that reduction invariance holds with N replaced by an arbitrary real number λ, which follows by induction and a limiting process. Using that and separability we are led to the functional equation

$$W^{-1}[W(p^{\lambda})W(q^{\lambda})] = r^{\lambda} = \left(W^{-1}[W(p)W(q)]\right)^{\lambda} \qquad (p, q \in [0, 1]). \tag{4.77}$$

By simple logarithmic transformations this reduces to (2.31) and the solution follows with a little algebra.

5. Consistent Aggregation

We present this topic by an example involving n different kinds of inputs contributing to the outputs of m producers. (Other examples could pertain to investments, employment, consumption, etc.) The jth producer's output (or maximal output by some measure) depends upon the inputs $x_{j1},\dots,x_{jn}$ to that producer through possibly producer-specific *(microeconomic) production functions* F_j $(j=1,\dots,m)$. We ask whether there exist $n+1$ *aggregator functions* G and G_k $(k=1,\dots,n)$ so that the aggregated output depends only upon the n aggregated inputs through a *macroeconomic function* F, that is,

$$G\left(F_1(x_{11},\dots,x_{1n}),\dots,F_m(x_{m1},\dots,x_{mn})\right) = F\left(G_1(x_{11},\dots,x_{m1}),\dots,G_n(x_{1n},\dots,x_{mn})\right). \tag{5.1}$$

This is the *generalized bisymmetry*[4] functional equation in $m \times n$ variables. Table 1 below may help understanding the situation. For convenience, we use in this table the vector notation

$$\begin{aligned} \mathbf{x}_{j\bullet} &= (x_{j1},\dots,x_{jk},\dots,x_{jn}) && (j=1,\dots,m),\\ \mathbf{x}_{\bullet k} &= (x_{1k},\dots,x_{jk},\dots,x_{mk}) && (k=1,\dots,n),\\ \mathbf{y} &= (y_1,\dots,y_j,\dots,y_m), && \mathbf{z} = (z_1,\dots,z_k,\dots,z_n). \end{aligned}$$

TABLE 1. Consistent Aggregation of Inputs and Outputs.

Producers	Inputs (commodities and services) 1	...	k	...	n	(Maximal) Outputs (production functions)
1	x_{11}	...	x_{1k}	...	x_{1n}	$y_1 = F_1(\mathbf{x}_{1\bullet})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
j	x_{j1}	...	x_{jk}	...	x_{jn}	$y_j = F_j(\mathbf{x}_{j\bullet})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	x_{m1}	...	x_{mk}	...	x_{mn}	$y_m = F_m(\mathbf{x}_{m\bullet})$
Aggregators	$z_1 = G_1(\mathbf{x}_{\bullet 1})$	...	$z_k = G_1(\mathbf{x}_{\bullet k})$	...	$z_n = G_1(\mathbf{x}_{\bullet n})$	$G(\mathbf{y}) \stackrel{?}{=} \mathbf{F}(\mathbf{z})$

Two questions arise naturally: what are the domains and ranges of the functions $F_1,\dots,F_m$, $G_1,\dots,G_n$, F and G; and which of these function should be regarded as given and which as unknown? We will consider both real and quite general domains and ranges. As to the second question, it is often argued that, if outputs can be measured by their monetary value, then they can be "aggregated by addition," that is,

$$G(y_1,\dots,y_m) = y_1 + \dots + y_m, \tag{5.2}$$

and if the kinds of outputs are "completely separated", even

$$G_k(x_1,\dots,x_m) = x_1 + \dots + x_m \qquad (k=1,\dots,n) \tag{5.3}$$

can be assumed. In this case, (5.1) leads to a Pexider equation [cf. (2.15), Section 2.1.4], and can easily be solved: under reasonable regularity conditions, $F_1,\dots,F_m$ and F will be

[4]Bisymmetry is the $m = n = 2$, $F_1 = F_2 = F = G = G_1 = G_2$ case which is important in measurement theory (cf. Krantz, Luce, Suppes, and Tversky, 1971).

affine (linear) functions. However, the production functions most often used in practice, such as the CD (Cobb-Douglas) function

$$F(z_1,\dots,z_n) = az_1^{c_1}z_2^{c_2}\dots z_n^{c_n} \tag{5.4}$$

with $c_1,\dots,c_n,a>0$, and the CES (Constant Elasticity of Substitution) function

$$F(z_1,\dots,z_n) = a(c_1z_1^b + c_2z_2^b + \dots + c_nz_n^b)^{1/b} \tag{5.5}$$

with $b\neq 0$, are not linear (except for $b=1$) or affine.

This problem can be solved if we do not require that the G's be additions. We show here that (5.4) and (5.5) are incompatible with (5.2) and (5.3), in the general framework of (5.1).

On the other hand, there is no compelling reason for assuming that inputs, outputs, etc. have to be measured by money or other real valued indices, or even that these variables can be so measured. Accordingly, it makes sense to solve (5.1), as we do in the sequel, on quite general sets without imposing any order or topological properties (cf. von Stengel, 1991; Aczél and Maksa, 1996; Taylor, 1999). In particular, the sets could be discrete, for instance consisting of all possible collections of inputs or outputs. This may give a more realistic approximation of the empirical situation. We shall see that a cornerstone of the solution is an abelian structure. In addition to surjective (onto) functions, which we used before, we use the standard notions of injective (1-to-1), bijective (1-to-1 and onto) functions f from X into (onto) Y. We denote by $f(X)$ the image of X under f (i.e. the range of f). [Note that if f is an injection, then it is a bijection of X onto $f(X)$.] When, in a function $F: Z_1\times\dots\times Z_n\to S$ in n variables $z_1,\dots,z_n$, we fix all the variables except z_l, we get a *partial function* $F^l: Z_l\to S$ (really, a family of such partial functions: one for each possible fixing of the values of $z_1,\dots,z_{l-1},z_{l+1},\dots,z_n$). For arbitrary sets, we have the following (Aczél and Maksa, 1996).

Theorem 13. *Let $m\geq 2$, $n\geq 2$ be integers. For $1\leq j\leq m$ and $1\leq k\leq n$, let X_{jk}, Y_j, Z_k, and S be nonempty sets, with the functions*

$$F_j: X_{j1}\times\dots\times X_{jn}\to Y_j, \qquad G_k: X_{1k}\times\dots\times X_{mk}\to Z_k,$$
$$F: Z_1\times\dots\times Z_n\to S, \qquad G: Y_1\times\dots Y_m\to S.$$

Then, the two following conditions are equivalent.

(i) *The partial functions $F_j^k: X_{jk}\to Y_j$ and $G_k^j: X_{jk}\to Z_k$ are surjections, $F^k: Z_k\to S$ and $G^j: Y_j\to S$ are injections and* (5.1) *is satisfied for all $x_{jk}\in X_{jk}$, $1\leq j\leq m$, $1\leq k\leq n$.*

(ii) *There exists an abelian group $(T,\circ)$ ($T\subseteq S$), surjections $f_{jk}: X_{jk}\to T$ and bijections $g_j: Y_j\to T$ and $h_k: Z_k\to T$ such that*

$$F(z_1,\dots,z_n) = h_1(z_1)\circ\dots\circ h_n(z_n) \qquad (z_k\in Z_k; 1\leq k\leq n), \tag{5.6}$$

$$G(y_1,\dots,y_m) = g_1(y_1)\circ\dots\circ g_m(y_m) \qquad (y_j\in Y_j; 1\leq j\leq m), \tag{5.7}$$

$$F_j(x_{j1},\dots,x_{jn}) = g_j^{-1}\big(f_{j1}(x_{j1})\circ\dots\circ f_{jn}(x_{jn})\big) \tag{5.8}$$

$$G_k(x_{1k},\dots,x_{mk}) = h_k^{-1}\big(f_{1k}(x_{1k})\circ\dots\circ f_{mk}(x_{mk})\big) \tag{5.9}$$
$$(x_{jk}\in X_{jk}; 1\leq j\leq m, 1\leq k\leq n).$$

Others results on consistent aggregation have been reported before and after this one. For earlier results, see Aczél's (1997) survey. As to more recent ones, Taylor (1999) has weakened the surjectivity conditions of Theorem 13 from supposing the surjectivity of the partial functions for *all* choices of the remaining variables to just *one* choice for each partial

function. Accordingly, $(T, \circ)$ becomes a cancellative ($a \circ x = a \circ y \Rightarrow x = y$) abelian semigroup (closed under the commutative and associative operation $\circ$) with unit element, rather than an abelian group. This result is more appropriate for applications to real intervals. (For example, the conditions of Theorem 13 exclude the set of all nonnegative reals equipped with addition, but this result does not.) At the same time, Maksa (1999) proved Theorem 14 below for arbitrary real intervals without assuming any surjectivity. We do not give here the details of the rather intricate proofs of either of these three results beyond mentioning that they rely on a double induction and on an adjustment of the solution on different domains. In the theorem below, we use CM for the class of all real valued functions defined on a subset of $\mathbb{R}^l$ for some positive integer l, which are continuous, and also monotonic in all variables. We also write $f(W_1 \times \ldots \times W_l)$ for the image of the Cartesian product $W_1 \times \ldots \times W_l$ by the function f. [This is sometimes denoted by $f(W_1, \ldots, W_l)$.]

Theorem 14. *Let $m \geq 2$, $n \geq 2$ be integers, and let X_{jk} $(1 \leq j \leq m,\ 1 \leq k \leq n)$ be real intervals with*

$$F_j : X_{j1} \times \ldots \times X_{jn} \longrightarrow \mathbb{R}, \qquad G_k : X_{1k} \times \ldots \times X_{mk} \longrightarrow \mathbb{R},$$
$$F_j(X_{j1} \times \ldots \times X_{jn}) =: J_j \qquad G_k(X_{1k} \times \ldots \times X_{mk}) =: I_k,$$
$$F : I_1 \times \ldots \times I_n \longrightarrow \mathbb{R}, \qquad G : J_1 \times \ldots \times J_m \longrightarrow \mathbb{R},$$

and $F_j, G_k, F, G \in CM$. Suppose also that (5.1) *holds for all x_{jk}, $1 \leq j \leq m$, $1 \leq k \leq n$. Then there exist a real interval I and functions*

$$\psi : I \to \mathbb{R}, \quad \alpha_k : I_k \to \mathbb{R}, \quad \gamma_j : J_j \to \mathbb{R}, \quad \beta_{jk} : X_{jk} \to \mathbb{R},$$

with $\psi, \alpha_k, \beta_{jk} \in CM$ for $1 \leq j \leq m$, $1 \leq k \leq n$ such that

$$F(x_1, \ldots, x_n) = \psi^{-1}\left(\sum_{k=1}^{n} \alpha_k(z_k)\right) \qquad (z_1, \ldots, z_n) \in I_1 \times \ldots \times I_n, \tag{5.10}$$

$$G(y_1, \ldots, y_m) = \psi^{-1}\left(\sum_{j=1}^{m} \gamma_j(y_j)\right) \qquad (y_1, \ldots, y_m) \in J_1 \times \ldots \times J_m, \tag{5.11}$$

$$F_j(x_{j1}, \ldots, x_{jn}) = \gamma_j^{-1}\left(\sum_{k=1}^{n} \beta_{jk}(x_{jk})\right) \tag{5.12}$$

$$G_k(x_{1k}, \ldots, x_{mk}) = \alpha_k^{-1}\left(\sum_{j=1}^{m} \beta_{jk}(x_{jk})\right) \qquad x_{jk} \in X_{jk}, 1 \leq j \leq m, 1 \leq k \leq n. \tag{5.13}$$

In other words: *For consistent aggregation, under the conditions of Theorems 13 and 14, the production functions are given either by* (5.6) *and* (5.8), *or by* (5.10) *and* (5.12), *and the aggregation functions either by* (5.7) *and* (5.9), *or by* (5.11) *and* (5.13), *on general sets or on real intervals, respectively.*

We can now answer the question raised earlier concerning the incompatibility of CD functions (5.4) and CES functions (5.5) with an aggregation by addition as in (5.3): *If even one aggregation function is a sum, say*

$$G_1(x_1, \ldots, x_m) = x_1 + \ldots + x_m, \tag{5.14}$$

then no production function (5.12) can be CD. Indeed, putting (5.14) into (5.13) gives

$$\alpha_1(x_1 + \ldots + x_m) = \beta_{11}(x_1) + \ldots + \beta_{m1}(x_m). \tag{5.15}$$

This is a Pexider equation. A result proved in Aczél (1987, p. 80) applies if (5.15) is satisfied for x_j on real intervals J_j $(1 \leq j \leq m)$ and yields that all continuous solutions are given by [cf. (2.6) in Section 2.1.4]

$$\beta_{j1}(x) = rx + s_j \qquad (1 \leq j \leq m),$$

$$1(x) = rx + \sum_{j=1}^{m} s_j$$

where $r > 0$, $s_1, \ldots, s_m$ are constants; so from (5.12), with $w_k = x_{jk}$ $(1 \leq k \leq n)$,

$$F_j(w_1, \ldots, w_n) = \gamma_j^{-1}\left(rw_1 + s_j + \sum_{k=2}^{n} \beta_{jk}(w_k)\right). \tag{5.16}$$

Another application of a Pexider equation shows that (5.16) cannot equal a CD function $aw_1^{c_1} \ldots w_n^{c_n}$ (cf. (5.4)). The situation for CES functions is similar and they too are incompatible with aggregation by addition. We leave the details to the reader. This pair of results does not say that inputs (or outputs) cannot actually be additively aggregated. Rather, it says that such an additive aggregation is not consistent with common and realistic production functions. (As mentioned earlier, it is consistent only with affine or linear production functions.)

The following question is often asked: if all of the F_j functions are in some sense of the same form, does that dictate that the macroeconomic function F must also be of the same form? This is called the *representativeness* problem. The answer of course depends upon exactly what is meant by 'of the same form.' In a trivial sense, the answer is 'Yes' because the F and F_j's in the above (5.10) and (5.12) solutions have the same structure. A somewhat more sophisticated answer is that it need not be so. For example, if all the F_j are CD functions, then in (5.12)

$$\gamma_j(y) = \ln\left(\frac{y}{a_j}\right), \quad \beta_{jk}(x) = c_k \ln x;$$

or if they are all CES functions, then in (5.12)

$$\gamma_j(y) = y^b, \quad \beta_{jk}(x) = c_k x^b.$$

Neither conclusion restricts ψ and α_k in (5.10) in any way. Thus, one may choose F to be a CD function, a CES function, or some other function of the form (5.10).

The general conclusion from solving the aggregation problem is that consistent or even representative aggregation is feasible only for appropriately chosen functions. In general, neither is possible if the aggregating functions are pre-selected the "wrong way." The implications for the possibility of macroeconomic models are considerable.

Results of this kind have applications elsewhere. For instance, they put stringent constraints on models having legitimate application to data aggregated over trials of an experiment, or over individuals tested. For some results concerning aggregation of probabilistic models of choice, which, among others, characterize Luce's (1959) choice model as a special case, see Aczél, Maksa, Marley and Moszner (1997).

6. Conclusions

As demonstrated in this article, the solution of many functional equation problems consists in reducing a functional equation to another one, which belongs to a running list of all those already solved. Over time, the list grows, extending the reach of the techniques. This process is supplemented by general results concerning broad classes of equations, such as uniqueness

theorems and theorems strengthening the regularity of functions (e.g., from integrability to differentiability of arbitrary order). Uniqueness considerations are implicit in Sections 1.1, 2.2 and 3, and a method of elevating regularity from continuity to differentiablity was explicitly used in Section 4.4.3.

The examples of functional equations techniques described in this paper were taken from three quite different areas of behavioral sciences: sensory psychology, micro and macroeconomics, and utility theory. This diverse choice was deliberate, and intended to suggest that such methods have potentially wide ranging applicability, not only in the behavioral sciences (obviously) but in all the sciences. In spirit, these methods resemble those used in dimensional analysis (cf., e.g., Krantz, Luce, Suppes, and Tversky, 1971), but they are considerably more general in scope: the expressions in the equations need not be monomials, and the goal is to nail down the forms of some functions, and not simply to uncover the values of some exponents. In many cases, they can be used to convince oneself and others that the functions specifying a model are the only feasible ones within a given framework.

REFERENCES

[1] Aczél, J. (1987). *A Short Course on Functional Equations Based on Applications to the Social and Behavioral Sciences*. Dordrecht-Boston-Lancaster-Tokyo: Reidel-Kluwer.

[2] Aczél, J. (1997). Bisymmetry and consistent aggregation: Historical review and recent results. In A. A. J. Marley (Ed.) *Choice, Decision, and Measurement: Essays in Honor of R. Duncan Luce*. Mahwah, NJ: Lawrence Erlbaum Associates, 225-233.

[3] Aczél, J., & Chung, J. K. (1982). Integrable solutions of functional equations of a general type. *Studia Sci. Math. Hungar.*, **17**, 51-67.

[4] Aczél, J., & Falmagne, J.-Cl. (1999). Consistency of monomial and difference representations of functions arising from empirical phenomena. *J. Math. Anal. Appl.*, **234**, 632-659.

[5] Aczél, J., Ger, R., & Járai, A. (1999). Solution of a functional equation arising from utility that is both separable and additive. *Proc. Amer. Math. Soc.*, **127**, 2923-2929.

[6] Aczél, J., Luce, R. D., & Maksa, G. (1996). Solutions to three functional equations arising from different ways of measuring utility. *J. Math. Anal. Appl.*, **204**, 451-471.

[7] Aczél, J., Maksa, G. (1996). Solution of the rectangular $m \times n$ generalized bisymmetry equation and the problem of consistent aggregation. *J. Math. Anal. Appl.*, **203**, 104-126.

[8] Aczél, J., Maksa, G. (2000). A functional equation generated by event commutativity in separable and additive utility theory. *Aequationes Math.* In press.

[9] Aczél, J., Maksa, G., Marley, A.A.J., & Moszner, Z. (1997). Consistent aggregation of scale families of selection probabilities. *Math. Soc. Sci.*, **333**, 227-230.

[10] Aczél, J., Maksa, G., Ng, C. T., & Páles, Z. (2000). A functional equation arising from ranked additive and separable utility. *Proc. Amer. Math. Soc.* In press.

[11] Aczél, J., Maksa, G., & Páles, Z. (1999). Solution of a functional equation arising from different ways of measuring utility. *J. Math. Anal. Appl.*, **233**, 740-748.

[12] Aczél, J., Maksa, G., & Páles, Z. (2000). Solution of a functional equation arising in an axiomatization of the utility of binary gambles. *Proc. Amer. Math. Soc.* In press.

[13] Falmagne, J.-Cl. (1977). Note: Weber's inequality and Fechner's Problem. *J. Math. Psychol.*, **16**, 267-271.

[14] Falmagne, J.-Cl. (1981). On a recurrent misuse of a classic functional equation result. *J. Math. Psychol.*, **23**, 190-193.

[15] Falmagne, J.-Cl. (1985). *Elements of Psychophysical Theory*, New York: Oxford University Press.

[16] Falmagne, J.-Cl. & Iverson, G. (1979). Conjoint Weber's laws and additivity. *J. Math. Psychol.*, **20**, 164-183.

[17] Falmagne, J.-Cl., Iverson, G. & Marcovici, S. (1979). Binaural "loudness" summation: Probabilistic theory and data. *Psychol. Review*, **86**, 25-43.

[18] Falmagne, J.-Cl., & Lundberg, A. (1999). Compatibility of gain control and power law representations—A János Aczél connection. *Aequationes Math.*, **58**, 1-10.

[19] Forti, G.L. (1995). Hyers-Ulam stability of functional equations in several variables. *Aequationes Math.*, **50**, 143-190.

[20] Járai, A. (1984). A remark to a paper of J. Aczél and J. K. Chung. *Studia Sci. Math. Hungar.*, **19**, 273-274.

[21] Krantz, D. M., Luce, R. D, Suppes, P., & Tversky, A. (1971). *Foundations of Measurement, Vol. 1.* New York: Academic Press.
[22] Kuczma, M. (1985). *An Introduction to the Theory of Functional Equations and Inequalities.* Państw. Wyd. Nauk.-Uniw. Śląski, Warszawa-Kraków-Katowice.
[23] Luce, R. D. (1959). *Individual Choice Behavior.* New York: Wiley.
[24] Luce, R. D. (1991). Rank- and sign-dependent linear utility models for binary gambles. *J. Econ. Theory*, **53**, 75-100.
[25] Luce, R. D. (1996). When four distinct ways to measure utility are the same. *J. Math. Psychol.*, **40**, 297-317.
[26] Luce, R. D. (1997). Associative joint receipts. *Math. Soc. Sci.*, **34**, 51-74. Erratum, **35**, 81.
[27] Luce, R. D. (1998). Coalescing, event commutativity, and theories of utility. *J. Risk Uncert.*, **16**, 87-114. Correction in 1999, **19**, 99.
[28] Luce, R. D. (2000a). *Utility of Gains and Losses: Measurement-Theoretical and Experimental Approaches.* Mahwah, NJ: Lawrence Erlbaum Associates.
[29] Luce, R. D. (2000b). Reduction invariance and Prelec's weighting functions. *J. Math. Psychol.* In press.
[30] Luce, R. D., & Fishburn, P. C. (1991). Rank- and sign-dependent linear utility models for finite first-order gambles. *J. Risk Uncert.*, **4**, 29-59.
[31] Luce, R. D., & Fishburn, P. C. (1995). A note on deriving rank-dependent utility using additive joint receipts. *J. Risk Uncert.*, **11**, 5-16.
[32] Luce, R. D., Krantz, D. H., Suppes, P., & Tversky, A. (1990). *Foundations of Measurement, Vol. III.* San Diego: Academic Press.
[33] Luce, R. D., & Marley, A. A. J. (2000). Separable and additive representations of binary gambles of gains. *Math. Soc. Sci.* In press.
[34] Maksa, G. (1999). Solution of generalized bisymmetry type equations without surjectivity assumptions. *Aequationes Math.*, **57**, 50-74.
[35] Marley, A. A. J., & Luce, R. D. (2000). A simple axiomatization of binary rank-dependent expected utility of gains (losses). Submitted.
[36] Ng, C. T. (1998). An application of a uniqueness theorem to a functional equation arising from measuring utility. *J. Math. Anal. Appl.*, **228**, 66-72.
[37] Prelec, D. (1998). The probability weighting function. *Econometrica*, **66** , 497-527.
[38] Riesz, F., & Szökefalvi-Nagy, B. (1990). *Functional Analysis.* New York: Dover.
[39] Roberts, A. W., & Varberg, D. E. (1973). *Convex Functions.* New York, London: Academic Press.
[40] Suppes, P., Krantz, D. H., Luce, R. D., & Tversky, A. (1989). *Foundations of Measurement, Vol. II.* San Diego: Academic Press.
[41] Taylor, M. A. (1999). The generalized equation of bisymmetry: Solutions based on cancellative abelian monoid. *Aequationes Math.*, **57**, 288-302.
[42] von Stengel, B. (1991). *Eine Dekompositionstheorie für mehrstellige Funktionen.* Frankfurt/M: Hain.

* Department of Pure Mathematics, University of Waterloo, Waterloo, ON N2L 3G1, Canada
E-mail address: jdaczel@math.uwaterloo.ca or janos@aris.ss.uci.edu

** Institute for Mathematical Behavioral Sciences, SSP, University of California, Irvine, CA 92697-5100, USA
E-mail addresses: rdluce@uci.edu and jcf@uci.edu , respectively.