Publ. Math. Debrecen
52 / 3-4 (1998), 247–274

What to do until (and when) the functional equationist arrives

By JÁNOS ACZÉL (Waterloo)

Dedicated to Professors Zoltán Daróczy and Imre Kátai on their 60th birthday

Abstract. Using recent results as examples, the paper presents remarks on how to (and how not to) apply functional equations in the social and behavioral sciences. The examples concern a.o. visual space, memoryless processes, choice (selection) probabilities, dimensional analysis, consistent aggregation, utility, gambles, Cauchy equations and smoothing.

1. What kind of problems lead to functional equations? It often happens that theories in empirical sciences are formalized by equations involving unknown functions. For instance, the theorist may be reluctant to make specific assumptions regarding the form of the functions involved in a mathematical model. The equations themselves often reduce the possibilities, however, and occasionally the restrictions are so severe as to restrict

Mathematics Subject Classification: 39B22, 90A10, 90A11, 90A15, 92G05, 92J30, 92J45, 26A42.
Key words and phrases: functional equations, Cauchy equations, domain of equation, codomain, injectivity, differentiable, locally integrable, smoothing, psychophysics, visual space, memoryless, choice probability, dimensional analysis, consistent aggregation, production functions, compatibility, representativity, gambles, additivity, separability.
This paper originated in a course I taught in an NSF sponsored workshop for junior behavioral scientists in July 1997 at the University of California, Irvine. Thanks are due to the National Science Foundation (U.S.A.) and to the Natural Sciences and Engineering Research Council (Canada) for support and to R. Duncan Luce (UC Irvine) for advice. I found questions and remarks by participants of the workshop thought provoking and useful.

all forms to but a few. A celebrated case involves the connection between Weber's law and the so-called Fechner problem in psychophysics. The restrictions there are such as to yield a logarithm as the only possible form for a function, which Fechner interpreted as measuring the magnitude of the sensation evoked by a stimulus (J.-C. Falmagne, 1985).

What are functional equations and what is their role? Children learn most (and annoy their parents most) with two questions: "why?" and "what makes it tick?" (usually while taking the device apart). In science, whether natural, behavioral or social, these are the questions asked by researchers of mathematics, in particular of axiomatics and of functional equations. Usually there is a formula, equation, function, used in practice among others because it has properties favorable for the purpose. The "why" or "what makes it tick" question here is whether these functional properties (functional equations and/or inequalities) determine the function ("formula"): Is it the only one having these properties? If there are others, could they also be used for the same purpose? If not, the reason is probably that they lack some important property or properties which we overlooked. Maybe if we add these we have the function characterized. On the other hand, if there is more than one function (usually a whole family or families of functions) possessing all the qualities which can be reasonably expected for purposes of the application, so much the better: we have gained choice and flexibility; but again it is useful to know all possibilities. Functional equations can also be used to determine the exact conditions under which a situation favourable for purposes of an application can prevail. A recent example of such ongoing research aims at finding the conditions under which the inputs and also the outputs of several producers in an industry can be consistently aggregated into a fictive "representative" producer's inputs and output. Another strives to characterize the so-called Luce choice model in the framework of selection probabilities. Other, even more recent ones deal with distinct ways of measuring gains and utilities (or gambles with several useful properties). At present, we work on functional equations motivated by the problem in psychophysics of describing how we see the physical space.

In what follows we concentrate on finding the functional equations for the applied problems and on the conditions and domains. In some cases we furnish the (short) technical details of the solution process; in every case we give references.

Figure 1: Vieth–Müller circles and hyperbolæ of Hillebrand. The left eye is centered at L, the right eye at R.

Example 1. ***Visual space – physical space.*** There are several theories on this subject.

In particular, according to R.K. Luneburg (1947), we see as *equidistant* – not the points on concentric circles but – *the points on "Vieth–Müller (V-M) circles"* going through the points L and R, where the left and right eye are centered, respectively (centre of each circle on the vertical axis), see Figure 1.

We also see *under constant direction* – not points of rays through a point but – *points on "Hildebrand (H) hyperbolas"* going through L and R (centre on the vertical axis); the intersection with each V-M circle appears under constant angle from the intersection of the V-M circle with the negative vertical axis, see Figure 1.

If the straight lines from L or R to a point bend from the vertical left by β and α, respectively, (if they bend to the right they are considered negative) then α and β are the *bipolar coordinates* of the point (α, β). For

Figure 2: The bipolar coordinates α, β denote the monocular directions with respect to the right and the left eye. The bipolar parallax $\psi = \alpha - \beta$ associated to a stimulus is the angle subtended by the visual axis when the eyes converge on the stimulus. The bipolar latitude $\varphi = (\alpha + \beta)/2$ describes the lateral deviation of a stimulus from the x-axis.

each V-M circle $\psi = \alpha - \beta$ is constant, for each H hyperbola ($\varphi = (\alpha+\beta)/2$ or) $\theta = \alpha + \beta$ is constant. See α, β, ψ and φ in Figure 2.

So it is natural to introduce the following *order*:

$$(L_\rho) \qquad (\alpha, \beta) \precsim_\rho (\alpha', \beta') \Leftrightarrow \alpha - \beta \geq \alpha' - \beta'$$

$$(L_\theta) \qquad (\alpha, \beta) \precsim_\theta (\alpha', \beta') \Leftrightarrow \alpha + \beta \geq \alpha' + \beta'$$

As a consequence, $\succsim_\rho$ and $\succsim_\theta$ are defined too. Furthermore

$\sim_\rho$ means $\succsim_\rho$ *and* $\precsim_\rho$ (i.e.: on the same V-M circle) and

$\sim_\theta$ means $\succsim_\theta$ *and* $\precsim_\theta$ (i.e.: on the same H hyperbola)

Figure 3: Locus of equidistant points according to the Luneburg theory and in two observed cases (Foley, J.M., 1966)

Experiments show *qualitative agreement but quantitative deviations* (consistently: the observer notes points equidistant to her always some-

what to the right or somewhat above (etc.) the V-M circle), see Figure 3.

Therefore J. Heller (1997) proposed that, in place of $\alpha-\beta$, $\alpha+\beta$, we take $f(\alpha)-g(\beta)$, $f(\alpha)+g(\beta)$, respectively, with *strictly increasing* f, g *mapping* $I=\,]-\pi/2,\pi/2[$ (we denote open intervals by $]\cdot,\cdot[$) *onto real intervals* (so f,g are *continuous*). This leads to the following *definitions*:

$$(D_\rho)\qquad (\alpha,\beta)\precsim_\rho(\alpha',\beta')\Leftrightarrow f(\alpha)-g(\beta)\geq f(\alpha')-g(\beta'),$$

$$(D_\theta)\qquad (\alpha,\beta)\precsim_\theta(\alpha',\beta')\Leftrightarrow f(\alpha)+g(\beta)\geq f(\alpha')+g(\beta').$$

Notice that $\precsim_\rho$ goes over into $\precsim_\theta$ if g is replaced by $-g$. Therefore we usually state just one of the two parallel definitions or (in)equations. We want to preserve some of the following invariance properties of (L_ρ) and (L_θ) for the more general (D_ρ) and (D_θ).

Definition. $\precsim$ on S is (cf. Figure 2):

ψ-shift invariant if

$$(\alpha,\beta)\precsim(\alpha',\beta')\Leftrightarrow(\alpha+\tau,\beta-\tau)\precsim(\alpha'+\tau,\beta'-\tau);$$

φ-shift invariant if

$$(\alpha,\beta)\precsim(\alpha',\beta')\Leftrightarrow(\alpha+\tau,\beta+\tau)\precsim(\alpha'+\tau,\beta'+\tau);$$

α-shift invariant if $(\alpha,\beta)\precsim(\alpha',\beta')\Leftrightarrow(\alpha+\tau,\beta)\precsim(\alpha'+\tau,\beta')$;

β-shift invariant if $(\alpha,\beta)\precsim(\alpha',\beta')\Leftrightarrow(\alpha,\beta+\tau)\precsim(\alpha',\beta'+\tau)$

All pairs are in $S=\{(\alpha,\beta)\in I^2\mid\alpha>\beta\}$ where

$$I=\,]-\pi/2,\pi/2[.$$

For instance, ψ-shift invariance for (D_ρ) means

$$f(\alpha)-g(\beta)\geq f(\alpha')-g(\beta')\Leftrightarrow$$
$$f(\alpha+\tau)-g(\beta-\tau)\geq f(\alpha'+\tau)-g(\beta'-\tau).$$

Thus we have

$$\text{(i)}\qquad f(\alpha+\tau)-g(\beta-\tau)=H[f(\alpha)-g(\beta),\tau]$$

which, finally, is a *functional equation*.

We get similarly the equations (ii)–(iv) and (i′)–(iv′) (see Table 1). Thus the question, when the relations remain invariant under these shifts,

Invariance applied to	*Equation*	*No.*
ψ-shift $\precsim_\rho$	$H[f(\alpha)-g(\beta),\tau]=f(\alpha+\tau)-g(\beta-\tau)$	(i)
$\precsim_\theta$	$H[f(\alpha)+g(\beta),\tau]=f(\alpha+\tau)+g(\beta-\tau)$	(i')
φ-shift $\precsim_\rho$	$H[f(\alpha)-g(\beta),\tau]=f(\alpha+\tau)-g(\beta+\tau)$	(ii)
$\precsim_\theta$	$H[f(\alpha)+g(\beta),\tau]=f(\alpha+\tau)+g(\beta+\tau)$	(ii')
α-shift $\precsim_\rho$	$H[f(\alpha)-g(\beta),\tau]=f(\alpha+\tau)-g(\beta)$	(iii)
$\precsim_\theta$	$H[f(\alpha)+g(\beta),\tau]=f(\alpha+\tau)+g(\beta)$	(iii')
β-shift $\precsim_\rho$	$H[f(\alpha)-g(\beta),\tau]=f(\alpha)-g(\beta+\tau)$	(iv)
$\precsim_\theta$	$H[f(\alpha)+g(\beta),\tau]=f(\alpha)+g(\beta+\tau)$	(iv')

Table 1
Functional equations induced by the shift invariance properties

leads to functional equations. Notice that in each equation H is strictly increasing in its first variable. This helps in the solution of these equations to which we will return later.

2. The domains of the equations (not the same as the domains of the unknown functions) ***are important***, for instance in Example 1 it is important that $-\pi/2<\beta<\alpha<\pi/2$ and $-\pi/2<\beta-\tau<\alpha+\tau<\pi/2$. Another example (J. Aczél, 1987) is the following

Example 2. Solutions of the basic ***Cauchy equation***

$$g(x+y)=g(x)+g(y) \tag{1}$$

for $x, y \in \mathbb{R}$ (the set of all real numbers) bounded on a (no matter how small) proper interval $J \subset \mathbb{R}$ are of the form

$$g(x) = cx \tag{2}$$

for all $x \in \mathbb{R}$. The same is true if (1) is supposed only for, say, $x, y, x+y \in [0,1]$ (of course, then $J \subset [0,1]$ and (2) holds for $x \in [0,1]$). But we have to be careful: Suppose (1) holds for $x, y \in [2,3]$. What does

$$g(x+y) = g(x) + g(y) \quad \text{for } x \in [2,3], y \in [2,3]$$

(domain of the equation: $[2,3]^2$) mean for the domain of g? Certainly $[2,3]$ should be in the domain. But if $x \in [2,3], y \in [2,3]$ then $x+y \in [4,6]$ and $g(x+y)$ stands on the left hand side of the equation. So the domain of g should be (at least)

$$[2,3] \cup [4,6]. \tag{3}$$

On this domain

$$g(x) = \begin{cases} x+5 & \text{if } x \in [2,3] \\ x+10 & \text{if } x \in [4,6] \end{cases}$$

is a bounded solution (check) which is *not* of the form (2) (cf. Z. Daróczy and L. Losonczi, 1967).

Maybe this was caused by the separate domains for x, y and for $x+y$ (see (3))? No: another example is the following. Define g on $T = \{0,2,3,4,5\}$ by

$$\begin{aligned} g(x) = x \quad & \text{if } x = 0,\ x = 2,\ x = 4, \\ = 2.5 \quad & \text{if } x = 3, \\ = 4.5 \quad & \text{if } x = 5. \end{aligned}$$

It is easy to check that $g(x+y) = g(x) + g(y)$, for $x, y, x+y \in T$, but g is not of the form (2). Incidentally, the finiteness of T and the vacuous continuity of g are not the issue here. Essentially the same counterexample is obtained if we define an extension g^* of g on $\{0\} \cup [2,5]$ by joining by segment the successive points $(i, g(i))(2 \le i \le 5)$ of the graph of g. (Thus, $(0,0)$ belongs to the graph of g^* but not the open segment $](0,0),(2,2)[$) (J.-C. Falmagne, 1981).

Many equations coming from applications can be reduced to the Cauchy equation (1).

Example 3. ***A memoryless phenomenon.*** An observer is watching a display. His or her task is to detect the realization of an event. The phenomenon is memoryless in the following sense: if the event does not occur between times 0 and t, the probability that it does not occur between times t and $t+s$ only depends on s. Formally, let T be a random variable representing the time elapsed until the occurrence of the event. The lack of memory is then represented by the equation (J.-C. Falmagne, 1985)

$$P\{T > t+s \mid T > t\} = f(s),$$

where P represents the probability measure. Thus $0 \le P \le 1$, so

$$0 \le f(s) \le 1. \tag{4}$$

The non-occurrence of the event in two distinct time intervals is independent. Therefore

$$\begin{aligned} &P\{\mathbf{T} > t+r+s \mid \mathbf{T} > t\} \\ &= P\{\mathbf{T} > t+r \mid \mathbf{T} > t\} P\{\mathbf{T} > t+r+s \mid t+r\}, \end{aligned}$$

that is,

$$f(r+s) = f(r)f(s). \tag{5}$$

As lengths of time, the variables r, s may be supposed *positive* or *nonnegative*. If (5) (and (4)) is supposed to hold for all positive r, s, then there are just *two* kinds of solutions:

$$f(s) \equiv 0 \quad \text{and } f(s) = e^{-ks} \quad (s \ge 0)$$

for some nonnegative constant k. But if the domain is $s, t \ge 0$ then there are *three* kinds of solutions:

$$f(s) = \begin{cases} 1 & \text{if } s = 0 \\ 0 & \text{if } s > 0 \end{cases}, \quad f(s) \equiv 0 \text{ and } f(s) = e^{-ks} \quad (s \ge 0)$$

(k again a nonnegative constant). We again see the importance of the *domain*.

3. Reduction to known equations. If, instead of (4), we suppose $0 < f(s) \le 1$ then we can take, in Example 3, logarithms on both sides of (5) and we get with $g(s) = \log f(s)$

$$g(r+s) = g(r) + g(s),$$

the Cauchy equation (1). The *domains* are again $r, s > 0$ or $r, s \ge 0$ but now *the value of the function* f cannot be 0. Then

$$g(s) = cs, \quad f(s) = e^{cs} = e^{-ks} = P\{T > s \mid T > 0\}$$

($k \ge 0$) and so

$$P\{T \le s\} = 1 - e^{-ks}.$$

Sometimes we get, however, more from the original equation without reduction.

Example 1a. In Example 1 (the *physical space – visual space problem*), applying (i) twice, we get the *translation equation*

$$H[H(z,\sigma),\tau] = H(z,\sigma+\tau). \tag{T}$$

The domains are given by

$$z \in Z = \left\{ f(\alpha) - g(\beta) \;\Big|\; -\frac{\pi}{2} < \beta < \alpha < \frac{\pi}{2} \right\},$$

$$\sigma, \sigma+\tau \in I(z)$$

$$= \bigcup \left\{ \left]\frac{\beta-\alpha}{2}, \min\left(\frac{\pi}{2}-\alpha, \beta+\frac{\pi}{2}\right)\right[\;\Big|\; f(\alpha) - g(\beta) = z \right\},$$

$$\tau \in I[H(z,\sigma)].$$

On such domains there were no result at hand for *(T)* but Heller (1997) added the following *condition* (cf. Figure 4).

(C) For all $(\alpha,\beta) \in S$ *there exists* γ *with*

$$(\alpha,\beta) \sim_\rho (\gamma,-\gamma) \in S$$

i.e. with $f(\alpha) - g(\beta) = f(\gamma) - g(-\gamma),\ \gamma \in \left]0, \frac{\pi}{2}\right[$.

Figure 4: Condition (C) postulates that (in bipolar coordinates) to each point (α, β) *there exists a* $\gamma > 0$ *such that* $(\alpha, \beta) \sim_\rho (\gamma, -\gamma)$.

This leads to $H(z, \tau) = h[h^{-1}(z) + \tau]$ (h strictly monotonic, continuous), which transforms (i) into

$$(I) \qquad h^{-1}[f(\alpha + \tau) - g(\beta - \tau)] = h^{-1}[f(\alpha) - g(\beta)] + \tau \qquad ((\alpha, \beta), (\alpha + \tau, \beta - \tau) \in S).$$

Applying known results we get as the general strictly increasing solutions f, g, mapping $]-\frac{\pi}{2}, \frac{\pi}{2}[$ onto intervals,

$$(I_1) \qquad f(\alpha) = a\alpha + b, \quad g(\beta) = c\beta + d \quad (a > 0,\ c > 0)$$

and

$$(I_2) \qquad f(\alpha) = ae^{k\alpha} + b, g(\beta) = -ce^{-k\beta} + d \qquad (a > 0,\ c > 0,\ k > 0 \quad \text{or} \quad a < 0,\ c < 0,\ k < 0).$$

But the *condition (C) is satisfied only if* $a = c$ *in* (I_1) *resp.* $a = c > 0$ *in* (I_2) (the former being the Luneburg case).

However, (i) and the other equations can be solved directly (actually, by reduction to a result not involving the translation equation) and we get the same solutions (I_1), (I_2) *without assumption (C)* (J. Aczél, Z. Boros, J. Heller and C.T Ng, 1998). So: ***relate*** also ***your intermediate steps***.

4. What if our equation is ***not sufficient*** to get the result we hope for? As mentioned before, either there are more functions solving our problem than we thought (for example (I_2) above, not only (I_1)) or ***further conditions***, appropriate for the problem, ***have to be added***. Some of them may again be functional equations. This is the case in the following example.

Example 4. ***Selection probabilities.*** (J. Aczél, Gy. Maksa, A.A.J. Marley and Z. Moszner, 1997).

The probability $P(e : E)$ of selecting options (elements) from a set E (selection or choice probability) is aggregated from m component (or individual) probabilities $P_i(e : E)$ $(i = 1, \dots, m)$ obtained in different contexts or according to different benchmarks or from different individuals: For $E = \{e_1, \dots, e_n\}$ there exist n functions $(H_1, \dots, H_n) : \Gamma_n^m \to \Gamma_n$ such that

$$P(e_j : E) = H_j[P_1(e_1 : E), \dots, P_1(e_n : E), \dots, P_m(e_1 : E), \dots, P_m(e_n : E)]. \tag{6}$$

Note that, since probabilities here are positive and add up to 1,

$$\Gamma_n = \big\{(p_1, \dots, p_n) \;\big|\; p_j > 0;\ j = 1, \dots, n;\ \sum_j p_j = 1\big\}.$$

It is supposed that the probabilities depend upon the options $e_j \in E$ through ratio scale (invariant under linear transforms) values $v_1(e_j), \dots, v_m(e_j)$, $v(e_j)$ $(j = 1, \dots, n)$:

$$P_i(e_j : E) = F_j[v_i(e_1), \dots, v_i(e_n)], \tag{7}$$

$$P(e_j : E) = F_j[v(e_1), \dots, v(e_n)] \tag{8}$$

(same F_j for $P_1, \dots P_m, P$; $i = 1, \dots, m$; $j = 1, \dots, n$; $(F_1, \dots, F_n)$: $\mathbb{R}_{++}^n \to \Gamma_n$). The "overall" scale value is also obtained by accumulation:

$$v(e_j) = G[v_1(e_j), \dots, v_m(e_j)] \tag{9}$$

with the same $G : \mathbb{R}_{++}^m \to \mathbb{R}_{++}(=]0, \infty[)$ for all j.

We suppose that for all $x_{ij} \in \mathbb{R}_{++}$ there exist e_j with $x_{ij} = v_i(e_j)$ (n-nonatomicity). From (8), (9), (6), (7) we get a system of functional equations and inequalities:

$$\begin{aligned} &F_j[G(x_{11}, \dots, x_{m1}), \dots, G(x_{1n}, \dots, x_{mn})] \\ &= H_j[F_1(x_{11}, \dots, x_{1n}), \dots, F_n(x_{11}, \dots, x_{1n}), \dots, \\ &F_1(x_{m1}, \dots, x_{mn}), \dots, F_n(x_{m1}, \dots, x_{mn})] \end{aligned} \tag{10}$$

(in vector form we get the generalized bisymmetry equation, see Section 6,

$$\begin{aligned} &\mathbf{F}[G(x_{11}, \dots, x_{m1}), \dots, G(x_{1n}, \dots, x_{mn})] \\ &= \mathbf{H}[\mathbf{F}(x_{11}, \dots, x_{1n}), \dots, \mathbf{F}(x_{m1}, \dots, x_{mn})]; \end{aligned}$$

$\mathbf{F} = (F_1, \dots, F_n)$, $\mathbf{H} = (H_1, \dots, H_n)$). We list also the consequences of $\mathbf{F} : \mathbb{R}_{++}^n \to \Gamma_n$, $\mathbf{H} : \Gamma_n^m \to \Gamma_n$.

$$\sum_j F_j = 1, \tag{11}$$

$$0 < F_j < 1, \tag{12}$$

$$\sum_j H_j = 1, \tag{13}$$

$$0 < H_j < 1. \tag{14}$$

The scales are ratio scales (i.e. (7), (8), (9) are invariant – really: covariant – under linear transformations). So there exist M, N such that

$$G(\alpha_1 y_1, \dots, \alpha_m y_m) = M(\alpha_1, \dots, \alpha_m) G(y_1, \dots, y_m) \tag{15}$$

$$F_j(\alpha z_1, \dots, \alpha z_n) = N(\alpha) F_j(z_1, \dots, z_n) \quad (j = 1, \dots, n) \tag{16}$$

(since there is just one scale in each F_j but m scales in G). From (11) and (16), $N(\alpha) \equiv 1$ and so

$$F_j(\alpha z_1, \dots, \alpha z_n) = F_j(z_1, \dots, z_n) \quad (j = 1, \dots, n) \tag{17}$$

Z. Moszner determined the general solution of (10)–(16) (or –(17)) by an abstract *construction*; the result seems too general for applications.

A.A.J. Marley wanted to *characterize Luce's choice model* (R.D. Luce, 1959, 1977), where

$$P(e_j : E) = \frac{v(e_j)}{\sum_k v(e_k)}, \quad P_i(e_j : E) = \frac{v_i(e_j)}{\sum_k v_i(e_k)}, \quad v(e_j) = \prod_i v_i(e_j)^{a_i}$$

$$\left(\sum_i a_i = 1;\ a_i > 0,\ i = 1, \dots, m;\ j, k = 1, \dots, n\right),$$

and

$$P(e_j : E) = \prod_i P_i(e_j : E)^{a_i} \ / \ \sum_k \prod_i P_i(e_k : E)^{a_i},$$

that is,

$$\text{(18)} \qquad \begin{cases} F_j(z_1, \dots, z_n) = \dfrac{z_j}{\sum_k z_k}, \quad G(y_1, \dots, y_m) = \prod_i y_i^{a_i} \\ \left(\sum_i a_i = 1;\ a_i > 0\right), \\ H_j(x_{11}, \dots, x_{1n}, \dots, x_{m1}, \dots, x_{mn}) = \dfrac{\prod_i x_{ij}^{a_i}}{\sum_k \prod_i x_{ik}^{a_i}}, \end{cases}$$

or its generalization

$$\text{(19)} \qquad \begin{gathered} F_j(z_1, \dots, z_n) = \frac{z_j^\gamma}{\sum_k z_k^\gamma}, \quad G(y_1, \dots, y_n) = \prod_i y_i^{a_i}, \\ H_j(x_{11}, \dots, x_{mn}) = \frac{\prod_i x_{ij}^{a_i}}{\sum_k \prod_i x_{ik}^{a_i}} \end{gathered}$$

For this (6)–(9) are *not enough by far*. So we required also that new 'scales', depending on probabilities $P_1, \dots, P_m$,

$$\text{(20)} \qquad \bar{v}(e_j : E) = \Phi[P_1(e_j : E), \dots, P_m(e_j : E)]$$

satisfy (8), that is,

$$\text{(21)} \qquad P(e_j : E) = F_j(\bar{v}(e_i : E), \dots, \bar{v}(e_n : E))$$

(same F_j as in (7), (8)). Thus

$$\begin{aligned} P(e_j : E) &= F_j(\Phi[P_1(e_1 : E), \dots, P_m(e_1 : E)], \dots, \\ &\quad \Phi[P_1(e_n : E), \dots, P_m(e_n : E)]), \\ &\quad F_j[G(x_{11}, \dots, x_{m1}), \dots, G(x_{1n}, \dots, x_{mn})] \\ &= F_j(\Phi[F_1(x_{11}, \dots, x_{1n}), \dots, F_1(x_{m1}, \dots, x_{mn})], \dots \\ &\quad \dots, \Phi[F_n(x_{11}, \dots x_{1n}), \dots, F_n(x_{m1}, \dots, x_{mn})]). \end{aligned} \tag{22}$$

An even stronger supposition is $\Phi = G$ (the Φ in (20) is the same as G in (9)). We have also regularity conditions: If G is bounded locally (= in a neighbourhood) then, from (15), $G(y_{11}, \dots, y_m) = b \prod_i y_i^{a_i}$. If G is strictly monotonic in each variable and

$$G(y, \dots, y) = y \tag{23}$$

then, as in (18),

$$G(y_1, \dots, y_m) = \prod_i y_i^{a_i} \quad \left(\sum_i a_i = 1,\ a_i > 0 \right). \tag{24}$$

From (17)

$$F_j(z_1, \dots, z_n) = f_j \left(\frac{z_1}{\sum_k z_k}, \dots, \frac{z_n}{\sum_k z_k} \right). \tag{25}$$

Here $\mathbf{f} = (f_1, \dots, f_n) : \Gamma_n \to \Gamma_n$. Suppose this is injective: $\mathbf{f}(\mathbf{t}) = \mathbf{f}(\mathbf{u}) \Rightarrow \mathbf{t} = \mathbf{u}$ $(\mathbf{t}, \mathbf{u} \in \Gamma_n)$; then (22) is equivalent to

$$\begin{aligned} &\frac{G(x_{1j}, \dots, x_{mj})}{\sum_k G(x_{1k}, \dots, x_{mk})} \\ &= \frac{\Phi[F_j(x_{11}, \dots, x_{1n}), \dots, F_j(x_{m1}, \dots, x_{mn})]}{\sum_k \Phi[F_k(x_{11}, \dots, x_{1n}), \dots, F_k(x_{m1}, \dots, x_{mn})]}. \end{aligned}$$

Put $x_{i\ell} = x_\ell$ $(i = 1, \dots, m;\ \ell = 1, \dots, n)$, $\varphi(x) = \Phi(x, \dots, x)$ then, by (23),

$$\frac{x_j}{\sum_k x_k} = \frac{\varphi[F_j(x_1, \dots, x_n)]}{\sum_k \varphi[F_k(x_1, \dots, x_n)]}.$$

There are still many solutions.

Under the additional condition $\Phi = G$ we get $\varphi(x) = x$ and, by (11) ($\sum_k F_k = 1$):

$$F_j(x_1, \dots, x_n) = \frac{x_j}{\sum_k x_k} \quad (j = 1, \dots, n).$$

Thus, with

$$G(y_1, \dots, y_m) = \prod_{i=1}^{m} y_i^{a_i} \quad \left(\sum_i a_i = 1, a_i > 0 \right), \tag{24}$$

we have characterized the Luce choice model (18) (the form of H_j follows by substitution into (10))

So: ***think of all relevant conditions*** (or of as many as you can). If an unwanted solution (for instance (I_2) in Example 1a, space perception) cannot be eliminated, maybe a condition (ψ-shift invariance) is ***not sufficiently relevant***.

5. Reduction of the number of variables and of unknown functions. Some reductions are quite easy. Consider the following (see e.g. Aczél, 1966).

Example 5. The simplest problem of ***classical (naive) dimensional analysis***, independent ratio scales for the independent variables ($x_k \mapsto r_k x_k$; $k = 1, 2, \dots, n$) resulting in ratio scales for the dependent variable ($u \mapsto Ru$), gives the functional equation

$$u(r_1 x_1, r_2 x_2, \dots, r_n x_n) = R(r_1, r_2, \dots, r_n) u(x_1, x_2, \dots, x_n) \tag{26}$$

(the domain for $r_1, r_2, \dots, r_n, x_1, x_2, \dots, x_n$ is, say, the set $\mathbb{R}_{++}$ of positive reals). The number of unknown functions is easily reduced.

Substitute $x_1 = \cdots = x_n = 1$. With $u(1, \dots, 1) = a$ we get

$$u(r_1, r_2, \dots, r_n) = aR(r_1, r_2, \dots, r_n). \tag{27}$$

By the nature of the problem, u is positive, so $a = u(1, 1, \dots, 1) > 0$. Thus (26) with (27) yields

$$R(r_1 x_1, r_2 x_2, \dots, r_n x_n) = R(r_1, r_2, \dots, r_n) R(x_1, x_2, \dots, x_n) \tag{28}$$

which contains only one unknown function (though the "wrong" one, R, but then (27) will give u in which we are interested). The number of variables can also be reduced simply: Put into (28) $x_1 = r_2 = r_3 = \cdots = r_n = 1$ and get

$$R(r_1, x_2, \dots, x_n) = R(r_1, 1, \dots, 1)R(1, x_2, \dots, x_n)$$

and, by repeating the process,

$$\begin{aligned} &R(r_1, r_2, \dots, r_n) \\ &= R(r_1, 1, \dots, 1)R(1, r_2, 1, \dots, 1) \dots R(1, \dots, 1, r_n) \\ &= R_1(r_1)R_2(r_2) \dots R_n(r_n). \end{aligned}$$

From (28) we see also that

$$R_k(r_k x_k) = R_k(r_k)R_k(x_k) \quad (k = 1, \dots, n).$$

If we know, for example, that u is increasing (not necessarily strictly) in each variable then we get (for instance by reduction to (5))

$$R_k(r_k) = r_k^{c_k}, \ R(r_1, r_2, \dots, r_n) = r_1^{c_1} r_2^{c_2} \cdots r_n^{c_n},$$

and

$$u(x_1, \dots, x_n) = a x_1^{c_1} x_2^{c_2} \cdots x_n^{c_n},$$

with constant $c_k \geq 0$ $(k = 1, \dots, n)$ as the general increasing solution of (26). So ***some reductions can be safely done***.

6. Interpretation after solution. Art of the possible. Can we do better?

Example 6. ***Consistent aggregation*** (for instance of inputs of material, capital, labor and output of products in economics, or two-way average of averages per subjects and per number of repetitions of responses, say reaction times, to stimuli in psychology) means that aggregates of microeconomical (maximal) outputs depend only upon aggregates of microeconomical inputs through a macroeconomical relation (production function). This, see Table 2, is equivalent to the functional equation of rectangular $(m \times n)$ generalized bisymmetry

$$(B_{mn}) \qquad \begin{aligned} &G(F_1(x_{11}, \dots, x_{1n}), \dots, F_m(x_{m1}, \dots, x_{mn})) \\ &= F(G_1(x_{11}, \dots, x_{m1}), \dots, G_n(x_{1n}, \dots, x_{mn})). \end{aligned}$$

Producers	Inputs (goods and serices) 1	$\cdots$	k	$\cdots$	n	(Maximal) outputs (production functions)
1	x_{11}	$\cdots$	x_{1k}	$\cdots$	x_{1n}	$y_1 = F_1(x_{11}, \dots, x_{1n})$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
j	x_{j1}	$\cdots$	x_{jk}	$\cdots$	x_{jn}	$y_j = F_j(x_{j1}, \dots, x_{jn})$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
m	x_{m1}	$\cdots$	x_{mk}	$\cdots$	x_{mn}	$y_m = F_m(x_{m1}, \dots, x_{mn})$
agg-regates	$z_1 = G_1(x_{11}, \dots, x_{m1})$	$\cdots$	$z_k = G_k(x_{1k}, \dots, x_{mk})$	$\cdots$	$z_n = G_n(x_{1n}, \dots, x_{mn})$	$z = G(y_1, \dots, y_m) = G(F_1(x_{11}, \dots, x_{1n}), \dots, F_m(x_{m1}, \dots, x_{mn})) \overset{?}{=} y = F(z_1, \dots, z_n) = F(G_1(x_{11}, \dots, x_{m1}), \dots, G_n(x_{1n}, \dots, x_{mn}))$

Table 2
Consistent aggregation of inputs and outputs

Nevertheless, the two topics evolved from 1946 on up to now rather independently, except for a "close encounter" in W.M. Gorman, 1968. In that work the continuous and monotonic solutions of $(B_{m,n})$ have been determined through set theoretical – combinatorial arguments by reduction to the functional equation

$$H(K(x,y),z) = L(x,M(y,z))$$

of generalized associativity. As another, maybe more appropriate starting point one can find the equally well known functional equation of (2×2) generalized bisymmetry

$$(B_{2,2}) \qquad G(F_1(x_{11},x_{12}),F_2(x_{21},x_{22})) = F(G_1(x_{11},x_{21}),G_2(x_{12},x_{22})),$$

which is clearly the $m = n = 2$ case of $(B_{m,n})$. It seems natural to solve $(B_{m,n})$ by using results on $(B_{2,2})$ and applying induction with respect to both m and n.

On the other hand, there is no compelling reason why inputs, outputs, etc. have to be measured by money or other real valued measures. The following result (J. Aczél and Gy. Maksa, 1996) describes the facts for quite *general sets*, so one can take for instance the set consisting of each and every input of j-th kind for the k-th producer as X_{jk} through which

x_{jk} in $(B_{m,n})$ goes. We then specialize the result to continuous functions on real intervals and give an interpretation.

The conditions under which we solve $(B_{m,n})$ are of the following types. *Injectivity*: The equation $F(s, z_2^0, \dots, z_n^0) = y$ has *at most* one solution s; similar requirements for the second, third, ... variable and for G. *Surjectivity*: The equation $F_j(t_j, u_2^0, \dots, u_n^0) = y_j$ has *at least* one solution t_j $(j = 1, \dots, m)$; similarly for the second, third, ... variable and for G_k $(k = 1, \dots, n)$. (Injectivity *and* surjectivity together is *bijectivity*). *Under such conditions the solutions of* $(B_{m,n})$ *are of the form*

$$(P) \qquad y = F(z_1, \dots, z_n) = \varphi^{-1}(\alpha_1(z_1) + \cdots + \alpha_n(z_n)),$$

$$(A) \qquad z = G(y_1, \dots, y_m) = \varphi^{-1}(\gamma_1(y_1) + \cdots + \gamma_m(y_m)),$$

$$(P_j) \qquad y_j = F_j(x_{j1}, \dots, x_{jn}) = \gamma_j^{-1}(\beta_{j1}(x_{j1}) + \cdots + \beta_{jn}(x_{jn})) \quad (j = 1, \dots, m),$$

$$(A_k) \qquad z_k = G_k(x_{1k}, \dots, x_{mk}) = \alpha_k^{-1}(\beta_{1k}(x_{1k}) + \cdots + \beta_{mk}(x_{mk})) \quad (k = 1, \dots, n).$$

Here the β_{jk} are surjections, the α_k, γ_j and φ are bijections, and $+$ is an abelian group operation (commutative, associative and there exist unit and inverse elements). If the underlying sets are *real intervals* and F, G, G_k are *continuous* then $+$ is the usual addition of real numbers and φ, a_k, γ_j, β_{jk} are also continuous $(j = 1, \dots, m;\ k = 1, \dots, n)$.

A possible interpretation of (A) and (A_k) is the following. The inputs x_{jk} and the outputs y_j are "rightly" measured by $\beta_{jk}(x_{jk})$ and by $\gamma_j(y_j)$, respectively $(j = 1, \dots, m;\ k = 1, \dots, n)$. Then aggregation is by addition resulting in

$$\alpha_k(z_k) = \beta_{1k}(x_{1k}) + \cdots + \beta_{mk}(x_{mk}) \quad (k = 1, \dots, n)$$

and

$$\varphi(z) = \gamma_1(y_1) + \cdots + \gamma_m(y_n).$$

We give as examples the CD (Cobb–Douglas) production functions, defined by

$$(\mathrm{CD}) \qquad F(z_1, \dots, z_n) = a z_1^{c_1} z_2^{c_2} \cdots z_n^{c_n} \quad (a, c_1, \dots, c_n > 0 \text{ const.})$$

on $]0,\infty[^n = \mathbb{R}^n_{++}$, and the CES (Constant Elasticity of Substitution) production functions, defined by

$$\text{(CES)} \qquad F(z_1,\dots,z_n) = (c_1 z_1^b + \cdots + c_n z_n^b)^{1/b}$$
$$(c_1,\dots,c_n > 0,\ b \neq 0 \text{ const.}),$$

which can be *extended* to

$$\overline{F}(z_1,\dots,z_n) = \varphi_b^{-1}(c_1\varphi_b(z_1) + \cdots + c_n\varphi_b(z_n))$$
$$\text{($\overline{\text{CES}}$)} \qquad \left(\varphi_b(z) = \begin{cases} |z|^b \operatorname{sign} z & (z \neq 0) \\ 0 & (z = 0) \end{cases}\right)$$

on $\mathbb{R}^n$. Surprisingly, these are *incompatible* with aggregation by addition (except for $b = 1$) but *compatible* with aggregation by multiplication or by

$$G_k(x_{1k},\dots,x_{mk}) = (x_{1k}^b + \cdots + x_{mk}^b)^{1/b},$$

respectively (meaning inputs should "rightly" be measured by $\log x_{jk}$ or x_{jk}^b, outputs by $\log y_j$ or y_j^b, respectively!).

Representativity: Can aggregates be considered as a (fictive) representative "producer"'s inputs and outputs connected by a "macroeconomical" production function F "of the same form" as the "microeconomical" ones $F_1,\dots,F_m$? Yes and no: $F(z_1,\dots,z_n) = \varphi^{-1}(\alpha_1(z_1) + \cdots + \alpha_n(z_n))$ is "of similar structure" as $F_j(u_1,\dots,u_n) = \gamma_j^{-1}(\beta_{j1}(u_1) + \cdots + \beta_{jn}(u_n))$ $(j = 1,\dots,n)$ but φ, α_k can be chosen independently of $\gamma_j, \beta_{j1},\dots,\beta_{jn}$. E.g. the aggregate of CES (CD) producers need not have CES (CD) production functions.

The surjectivity and injectivity conditions in the above results imply *overall bijectivity* in the following sense. For *any* fixed $z_2^0,\dots,z_n^0$, the equation $F(s, z_2^0,\dots,z_n^0) = y$ has *exactly one* solution s in the domain Z_1 for the first variable in F, for each y in the codomain (set of values) T of F, *all variables considered* and the same holds for the second, third, ... variable for the same T. This is satisfied for (CD) but not for (CES) on $\mathbb{R}^n_{++}$. It is satisfied for the extended ($\overline{\text{CES}}$) functions on $\mathbb{R}^n$. On real intervals, however, we had the additional condition of *continuity*. The ($\overline{\text{CES}}$) functions are continuous on $\mathbb{R}^n$ if $b > 0$ but *not* if $b < 0$. Even for $b < 0$ they are continuous on $\mathbb{R}^n_{++}$ but, as just said, not bijective in the above sense. Not long ago weaker conditions have been found (J. Aczél,

Gy. Maksa, and M.A. Taylor, 1997), which are satisfied also by the ($\overline{\text{CES}}$) functions but still guarantee the above result (notice that ($\overline{\text{CES}}$) *is* of the form (P)). They require the functions to be continuous on subintervals (like the subinterval $\mathbb{R}_{++}$ of $\mathbb{R}$) and that the sets of function values on the subintervals assigned to different variables have points in common (they are not supposed to be the same anymore).

Furthermore, we were able to weaken the surjectivity conditions for general sets (in particular for real intervals) also in that sense that we require only that they hold for certain fixed values of the variables and for certain fixed function values.

However, even these weakened surjectivity conditions exclude such simple and important aggregation functions as the arithmetic mean $G_1(x_1, \dots, x_m) = (x_1 + \cdots + x_m)/m$ on $\mathbb{R}^m_{++}$. Gy. Maksa (1998) recently achieved a result where no surjectivity condition has been assumed at all. It thus holds also for arithmetic means and, more generally, quasi-linear means

$$G_1(x_1, \dots, x_m) = g^{-1}(q_1 g(x_1) + \cdots + q_m g(x_m))$$
$$\left(q_j > 0;\ j = 1, \dots, m;\ \sum_{j=1}^{m} q_j = 1\right),$$

which are rather important in some applications.

Moral: if you obtain results under conditions which exclude some applications, ***try harder***. Not that one always succeeds. Sometimes it can be shown that the suppositions cannot be further reduced, at other times one just is not able to do so (yet).

Example 7. R.D. Luce was led in the quest for conditions under which ***measuring the utility of gains*** (losses) via risky or riskless choices, gains (losses) alone or trade-offs between gains and losses to several functional equations among which

$$\begin{aligned} &f^{-1}[f(x) + f(y) - f(x)f(y)]z \\ &= f^{-1}[f(xz) + f(yP(x,z)) - f(xz)f(yP(x,z))] \end{aligned} \tag{29}$$

($x, y \in [0,1[$, $z \in [0,1]$, $f : [0,1[\to [0,1[$ (strictly increasing, onto), $P : [0,1[\times[0,1] \to [0,1]$) proved to be the most difficult nut to crack.

Actually we (J. Aczél R.D. Luce, and Gy. Maksa, 1996) could solve (29) only under the supposition that f is differentiable with nonzero derivative on $]0,1[$. This was so much the more annoying since we already determined P without any differentiability assumption: It satisfies the functional equation

$$P(x,zw)=P(x,z)P(xz,w) \quad (x\in[0,1[,\ z,w\in[0,1]) \tag{30}$$

and the solution is

$$P(x,z)=\frac{g(x)}{g(xz)} \quad (x\in\,]0,1[,\ z\in\,]0,1]),\ P(x,0)=0,\ P(0,z)=z. \tag{31}$$

This was not quite easy either (the main difficulty was caused by the fact that $x=1$ is *not in the domain* of (30)). What we needed differentiability for is the equation

$$h\left(y\frac{g(x)}{g(xz)}\right)=\frac{h[H(x,y)z]}{h(xz)},$$

($h(x)=1-f(x)$, $H(x,y)=h^{-1}[h(x)h(y)]$; $x,y\in[0,1[$, $z\in[0,1]$), resulting from (29) and (31). Under this condition the general solution of (29) is

$$f(x)=1-(1-x^b)^a \quad \text{and thus} \quad P(x,z)=z\,\frac{(1-x^b)^{1/b}}{(1-x^bz^b)^{1/b}} \tag{32}$$

(a, b arbitrary positive constants). We conjecture that the solution is the same if instead of differentiability only continuity of f is supposed (in addition to strict monotonicity) but could not prove it, no matter how hard we tried. This shows that ***weakening the conditions may not be easy or even possible***.

7. Wrangling differentiability out of weaker conditions. Actually there exist several methods for *deriving differentiability* from weaker conditions. (Unfortunately, none we know could be applied to the previous example.) We apply one to the Cauchy equation (1):

Example 2a. The basic ***Cauchy equation***

$$g(x+y)=g(x)+g(y) \quad (x,y\in\mathbb{R}) \tag{1}$$

is very easy to solve if we know that g is *differentiable on a (no matter how small) proper interval* $[a, b]$ $(b > a)$. First, g will then be differentiable also on $[0, b-a] = [0, d]$. Indeed, (1) implies $g(x) = g(x+a) - g(a)$ so, if $x \in [0, b-a]$, then $x + a$ is in $[a, b]$, where g is differentiable, thus g is differentiable also on $[0, b-a] = [0, d]$. From this we get that g is *everywhere differentiable*, by repeated application of

$$g(x+d) = g(x) + g(d) \quad \text{and } g(t-d) = g(t) - g(d).$$

If, however, g is everywhere differentiable then we simply differentiate (1) with respect to x:

$$g'(x+y) = g'(x), \quad \text{that is, } g'(x) = \text{constant} = c,\ g(x) = cx + C.$$

Substituting this into (1) we get $C = 0$, that is,

$$g(x) = cx \tag{2}$$

as the general differentiable solution of (1).

For (1) and many other functional equations, differentiability can be obtained from much weaker conditions (see, for instance, M. Kac, 1937 and J. Aczél, 1966). One such condition is local integrability on a (small) interval $[a, b]$ $(b > a)$. By an argument similar to that given above, integrability everywhere follows. We do not go here into the definition of Lebesgue integrability which is more general than Riemann integrability. The only property of locally integrable functions which we will need is that their *antiderivatives*

$$G_A(x) = \int_A^x g(t)dt$$

are continuous. If g is continuous then $G_A(x)$ is *differentiable* in x for all A.

So, for locally integrable g, this time we *integrate* (1) with respect to y from a to b $(> a)$.

$$\int_a^b g(x+y)dy = g(x)(b-a) + \int_a^b g(y)dy.$$

The last term is a constant, say B. Introducing the new variable $t = x + y$ we get

$$g(x) = \frac{1}{b-a}\left(\int_{x+a}^{x+b} g(t)dt + B\right)$$

$$= \frac{1}{b-a}\left(G_A(x+b) - G_A(x+a) + B\right).$$

Since g is *integrable*, the right hand side is *continuous*, so also the left hand side, which is $g(x)$. But for continuous g, the right hand side is *differentiable*, so also $g(x)$ on the left. Now we can apply the above argument to *get (2) as the general integrable solution of (1).*

This method has been applied recently with surprising efficiency to the following problem.

Example 8. R.D. Luce (1998) reduced the search for all ***utility measures*** over binary gambles that are both ***additive and separable*** to the task of solving the functional equation (done in J. Aczél, R. Ger, and A. Járai, 1998)

$$f(v) = f(vw) + f[vQ(w)] \quad (v \in [0,k[, w \in [0,1]). \tag{33}$$

It is natural for the problem to suppose f to be *strictly increasing*, $f(0) = 0$, $f(1) = 1$. Somewhat less natural is the condition that Q *is strictly decreasing*, $Q(0) = 1$, $Q(1) = 0$ and even less so (though tolerable) that both f and Q map their domain *onto intervals* (f onto a $[0,K[$, Q onto $[0,1]$) so they are *continuous*. It is certainly not natural to suppose f and Q to be *differentiable* (compare example 7) but it sure makes the solution easy:

Differentiate (33) with respect to v or w and get

$$f'(v) = wf'(vw) + Q(w)f'[vQ(w)],$$

$$0 = vf'(vw') + vQ'(w)f'[vQ(w)],$$

respectively. We multiply the first equation by $Q'(w)$, the second by $Q(w)/v$ and subtract in order to obtain, with

$$H(w) = \frac{Q'(w)}{wQ'(w) - Q(w)}$$

(it is easy to show that the denominator cannot be 0 under the conditions of our problem), the so called Pexider equation (generalization of an analogue of (1) and (5))

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

On this domain and under weak conditions easily satisfied in this case, the general solution is

$$f'(v) = Av^B, \quad \text{so } f(v) = av^\beta + c.$$

(We are not interested in H or in the constant or logarithmic solutions f or in those with $\beta < 0$; these are not nonnegative or not strictly increasing; thus we exclude $a = 0$ and $\beta \leq 0$; otherwise β is *arbitrary*). Now, if we wish, $f(0) = 0$ and $f(1) = 1$ give $c = 0$ and $a = 1$. Using also (33), we have as general solution ($\beta > 0$)

$$f(v) = v^\beta, \quad Q(w) = (1 - w^\beta)^{1/\beta} \quad (v \in [0,k[,\ w \in [0,1]) \tag{34}$$

What if we do not have differentiability? By an argument similar to but more sophisticated than that in Example 2a, weak conditions like continuity (with f nonconstant on $]0,k[$) guarantee that the function F defined by

$$F(u) = \frac{1}{u}\int_0^u f(t)dt \ (u \neq 0), \quad F(0) = 0, \tag{35}$$

is *continuously differentiable* (the integration in (35) does the *"smoothing"*) with $F' > 0$. From (33) we get by integration

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

(we substituted $s = vw$, $t = vQ(w)$) for $u > 0$, $w > 0$ but by continuity also at $u = 0$, and/or $w = 0$. Comparison to (33) furnishes the surprise that F satisfies the same equation as f but F is *differentiable*, so the above

differentiating method gives us $F(u) = au^\beta + c$. Moreover, since the limit of F at 0 is 0, we have $c = 0$ and $\beta > 0$. Now from (35)

$$f(v) = \frac{d}{dv}[vF(v)] = \frac{d}{dv}[av^{\beta+1}] = a(\beta+1)v^\beta = \alpha v^\beta.$$

If also $f(1) = 1$ is supposed, then *we get (34) again.*

The miracles do not stop even here: also *monotonicity and continuity follow from the nonnegativity of the function values*, which is thus the *only* supposition we need for the solution of the functional equation (33), a very obvious and weak assumption indeed. We do not go into the somewhat intricate details, only point out that, since $vw \le v$ for $w \in [0,1]$ and $f \ge 0$, equation (33) instantly yields $f(v) \ge f(vw)$, that is, f is *increasing* (though not yet strictly).

The message here is the flip side of that in Example 7: ***Weakening the conditions may be possible, even if difficult: if you don't succeed first, keep trying***. *"Smoothing" methods are particularly effective* (but not effective enough for Example 7).

Moreover, *while there are some broad ideas applicable to several classes of functional equations, even these may need essential modifications for individual equations*. Compare, for instance, the methods applied in Examples 2a and 8.

8. There seems to be a widespread disdain for ***exact conditions*** and ***proofs***, mainly among appliers of mathematics to fields other than measurement theory, mathematical psychology and mathematical economics. However, if a result is used under the wrong conditions, which do not hold (or we do not know whether they hold) in that situation, trouble lurks, up to bridges and mine shafts collapsing. In more theoretical fields of endeavor it just leads to bad science. For instance, in example 7, the fast way to "solve"

$$P(x, zw) = P(x, z)P(xz, w) \quad (x \in [0,1[;\ z, w \in [0,1]), \tag{30}$$

and getting the desired (31), $P(v, w) = g(v)/g(vw)$, would be to substitute $x = 1, g(v) = 1/P(1, v)$ – except that (30) is not (supposed to be) valid for $x = 1$. So what? "the goal justifies the means". Well, the final result is

$$P(x, v) = v\frac{(1 - x^b)^{1/b}}{(1 - x^b v^b)^{1/b}} \tag{32}$$

($b > 0$) and substituting here $x = 1$, $g(v) = 1/P(1, v)$ would give the absurd $g(v) = \infty$ for all v, hardly a confidence generating process (the importance of *domains* also shows again).

This leads to the role of proofs, which some in applications really find annoying, tedious and/or superfluous. Well, no matter how much experimental evidence we have, ***we know that a statement is true only when we have proved it***. Moreover, ***by analyzing proofs we can check which conditions were really needed*** (if a condition is not used in the proof, dispose of it) thus possibly weakening the suppositions and making the results more widely applicable.

But *this* can be done ***after the functional equationist arrived*** ...

References

[1] J. Aczél, Lectures on Functional Equations and Their Applications, *Academic Press*, *New York*, 1966.

[2] J. Aczél, A Short Course on Functional Equations Based Upon Recent Applications to the Social and Behavioral Sciences, *Reidel, Kluwer*, *Dordrecht, Boston*, 1987.

[3] J. Aczél, Z. Boros J. Heller and C. T. Ng, Functional equations in binocular space perception, *Journal of Mathematical Psychology* (1998).

[4] J. Aczél, R. Ger and A. Járai, Solution of the functional equation arising from utility that is both separable and additive, Proceedings of the American Mathematical Society, 1998.

[5] J. Aczél, R. D. Luce and Gy. Maksa, Solutions to three functional equations arising from different ways of measuring utility, *Journal of Mathematical Analysis and Applications* **204** (1996), 451–471.

[6] J. Aczél and Gy. Maksa, Solution of the rectangular $m \times n$ generalized bisymmetry equation and of the problem of consistent aggregation, *Journal of Mathematical Analysis and Applications* **203** (1996), 104–126.

[7] J. Aczél, Gy. Maksa, A. A. J. Marley and Z. Moszner, Consistent aggregation of scale families of selection probabilities, *Mathematical Social Sciences* **33** (1997), 227–250.

[8] J. Aczél, Gy. Maksa and M. A. Taylor, Equations of generalized bisymmetry and consistent aggregation: Weakly surjective solutions which may be discontinuous at places, *Journal of Mathematical Analysis and Applications* **214** (1997), 22–35.

[9] Z. Daróczy and L. Losonczi, Über die Erweiterung der auf einer Punktmenge additiven Funktionen, *Publicationes Mathematicae Debrecen* **14** (1967), 239–245.

[10] J.-C. Falmagne, On a recurrent misuse of a classical functional equation result, *Journal of Mathematical Psychology* **23** (1981), 190–193.

[11] J.-C. Falmagne, Elements of Psychophysical Theory, *Clarendon, Oxford University Press*, *Oxford, New York*, 1985.

[12] J. M. Foley, Locus of perceived equidistance as a function of viewing distance, *Journal of the Optical Society of America* **56** (1966), 822–827.

[13] W. M. Gorman, The structure of utility functions, *Review of Economic Studies* **35** (1968), 367–390.

[14] J. Heller, On the psychophysics of binocular space perception, *Journal of Mathematical Psychology* **41** (1997), 29–43.

[15] M. Kac, Une remarque sur les équations fonctionnelles, *Commentarii Mathematici Helvetici* **9** (1937), 170–171.

[16] R. D. Luce, Individual Choice Behavior, *Wiley*, *New York*, 1959.

[17] R. D. Luce, The choice axiom after twenty years, *Journal of Mathematical Psychology* **15** (1977), 215–233.

[18] R. D. Luce, Coalescing event commutativity and theories of utility, *Journal of Risk and Uncertainty* (1998).

[19] R. K. Luneburg, Mathematical Analysis of Binocular Vision, *Princeton University Press*, *Princeton*, 1947.

[20] Gy. Maksa, Solution of generalized bisymmetry type equations without surjectivity assumptions, *Aequationes Mathematicae* (1998).

JÁNOS ACZÉL
FACULTY OF MATHEMATICS
UNIVERSITY OF WATERLOO
WATERLOO, ONT. N2L 3G1
CANADA

(Received September 5, 1997)