CITATION CLASSICS FROM SOCIAL INDICATORS RESEARCH

Social Indicators Research Series

Volume 26

This new series aims to provide a public forum for single treatises and collections of papers on social indicators research that are too long to be published in our journal *Social Indicators Research*. Like the journal, the book series deals with statistical assessments of the quality of life from a broad perspective. It welcomes the research on a wide variety of substantive areas, including health, crime, housing, education, family life, leisure activities, transportation, mobility, economics, work, religion and environmental issues. These areas of research will focus on the impact of key issues such as health on the overall quality of life and vice versa. An international review board, consisting of Ruut Veenhoven, Joachim Vogel, Ed Diener, Torbjorn Moum, Mirjam A.G. Sprangers and Wolfgang Glatzer, will ensure the high quality of the series as a whole.

The titles published in this series are listed at the end of this volume.

CITATION CLASSICS FROM SOCIAL INDICATORS RESEARCH

The Most Cited Articles Edited and Introduced by Alex C. Michalos

Edited by

ALEX C. MICHALOS

University of Northern British Columbia, Canada

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 1-4020-3722-8 (PB)
ISBN 978-1-4020-3722-1 (PB)
ISBN 1-4020-3742-2 (e-book)
ISBN 978-1-4020-3742-9 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

www.springeronline.com

Printed on acid-free paper

Printed in the Netherlands.

Table of Contents

PREFACE

The idea of publishing some sort of a volume celebrating the first thirty years of publishing *Social Indicators Research* came to me some time early in the twenty-ninth year. When I shared the idea with the publisher's representative, Welmoed Spahr, she was very enthusiastic about it. We both thought we should try to produce a volume in 2004, preferably in time for the November meeting of the International Society for Quality of Life Studies. That implied that the publisher would need the complete manuscript by May 2004. So, we had a clear time frame for getting the job done.

As I reflected on the variety of volumes we might produce, I thought it would be particularly helpful for contemporary and future social indicators researchers to assemble a set of papers that would represent the best that had come out of our journal. The question was: How should we define 'best'? A few fairly elaborate decision procedures based on different criteria were constructed, and each seemed to require more time and other resources than we could spare. I rapidly settled on the idea of using article citation counts as a measure of article quality. I knew it was problematic, but so was everything else I could think of, and citation counts had four distinct advantages. First, such counts were relatively easy to obtain and objectively observable by other researchers. Second, there were about 40 years' of constructive and critical research on the strengths and weaknesses of the use of citation counts as indicators of the quality of research publications. Third, it was possible to provide an overview of that research which would be sufficient to allow readers to make their own judgments about my decision to use this approach. Finally, it would be wonderful if anyone unhappy with this approach would undertake an alternative analysis that would provide some confirmation or disconfirmation of the results reported here.

All the papers reprinted here appear as they were originally published, except for minor corrections of typos. In 1978, at the ninth World Congress of Sociology in Uppsala, I presented a short paper describing the origins of the journal. Because most of the people interested in the papers in this collection will have some interest in knowing how it all started and that paper has never been published, I have included it immediately following my introductory essay. Since it was written over 25 years ago, both its author and the publisher (D. Reidel) have changed quite a bit. I must say that it gives me considerable pleasure to be able to report that I have enjoyed our long relationship immensely, and I look forward to many more years of collaboration. When I read that John Maynard Keynes took over the *Economic Journal* from F.Y. Edgeworth in 1912 and edited it for the next 33 years, it occurred to me that 30 years is not too long a time to edit a journal.

Besides, for me, next to writing up my own research, editing other people's research is the most enjoyable academic activity I do.

I would like to thank Welmoed Spahr and Kluwer Academic Publishers for helping me produce this volume. Also, I would like to thank Bob Cummins and Ruut Veenhoven for providing me with electronic versions of their papers, which helped reduce our production costs.

1. Citation Classics: The Idea and the Collection

ALEX C. MICHALOS

After about three years in graduate school studying the history of religions, my interests turned to logic and the philosophy of science. My doctoral thesis was on a dispute between Karl Popper and Rudolf Carnap over the construction of quantitative measures of the acceptability of scientific theories (Michalos, 1971). For most of the 40 years that I was a university teacher, I taught courses in the philosophy of science, always including issues related to the evaluation of the activities and products of social and natural scientists. Since there are no generally accepted definitions of such key concepts as science, scientific explanation, scientific laws, scientific theories and scientific acceptability, but there are particular groups of researchers that tend to accept certain concepts while rejecting others (Michalos 1980), I take a fairly pragmatic approach to the construction of my own scientific vocabulary. For present purposes, the only part of that vocabulary that requires explanation concerns the term 'citation classic'.

I first encountered the term some years ago reading something by Eugene Garfield, the founder of the Institute for Scientific Information (ISI), the *Science Citation Index*, *Social Sciences Citation Index* and several other important works. According to Garfield (1985), "By definition, a *Citation Classic* is a paper or book that has been highly cited in its field" (p.404). From 1977 to 1993 the ISI publication *Current Contents* regularly featured an article called Citation Classics Commentary, which was subsequently replaced by In-Cites. Because different fields are characterized by different institutional arrangements (e.g., numbers and kinds of communication media, practitioners, standardized practices and rules of procedure), different numerical thresholds are used to identify classics in each field. In an earlier paper, Garfield (1976) wrote that

> ". . .less than 25% of all papers will be cited ten times in all eternity! . . . any paper cited ten times in one year is *ipso facto* significant. Occasionally there is an anomaly. But a paper cited ten times in each of two successive years is well on its way to citation stardom. Whether the author is on the way to immortality depends on how well he or she does in other papers" (p.419).

An average paper in the *Science Citation Index* is cited about 1.7 times per year (Garfield, 1972). For the years 1955-1987, Garfield (1989) claimed that for the whole *Index* database,

> ". . .more than 56 percent of the source items are uncited – not even self-cited. (Many of these source items are abstracts, letters, and editorials, of limited interest; nevertheless, a huge number of papers go uncited.)" (p.7).

Alex C. Michalos (ed.), Citation Classics from Social Indicators Research, 1–56.

In 1973 Garfield distinguished three kinds of "uncitedness", namely, "the incitedness of the mediocre, the unintelligible, the irrelevant", then that "of the meritorious but undiscovered or forgotten", and finally that

> "of the *distinction* that comes to those whose work has become so well known (and presumably been previously so heavily cited) that one finds it at first tedious, then unnecessary, and finally actually gauche to cite such men at all" (p.413).

Hamilton (1990) reported that ISI data revealed that about 55% of the "papers published between 1981 and 1985 in journals indexed by the institute received no citations at all in the 5 years after they were published" (p.1331). In response, Pendlebury (1991) wrote that the precise figures were "47.4% uncited for the sciences, 74.7% for the social sciences, and 98.0% for the arts and humanities". However, more importantly, he explained that

> "These statistics represent every type of article that appears in journals indexed by the Institute for Scientific Information (ISI) in its *Science Citation Index*, *Social Sciences Citation Index*, and *Arts & Humanities Citation Index*. The journals' ISI indexes contain not only articles, reviews, and notes, but also meeting abstracts, editorials, obituaries, letters like this one, and other *marginalia*, which one might expect to be largely un-cited. In 1984, the year of the data quoted by Hamilton, about 27% of the items indexed in the *Science Citation Index* were such *marginalia*. The comparable figures for the social sciences and arts and humanities were 48% and 69%, respectively.
>
> If one analyzes the data more narrowly and examines the extent of uncited articles alone (this information was not yet available when Hamilton wrote his articles), the figures shrink, some more than others: 22.4% of 1984 science articles remained uncited by the end of 1988, as did 48.0% of social sciences articles and 93.1% of articles in arts and humanities journals. . .Only 14.7% of 1984 science articles by U.S. authors were left uncited by the end of 1988. . .articles published in the highest impact journals like Science are almost never left uncited" (pp.1410-1411).

Exhibit 1 illustrates a few thresholds for citation classics from different fields, ranging from papers with 50 or more citations covering the fields of geography and marine biology to 500 or more citations covering all fields listed in the *Science Citation Index*. Plomp (1990) used a threshold of 25 citations to identify "highly cited papers", which is a more modest label than 'citation classic' and perhaps not quite the same idea. His rationale for the figure was interesting.

> "Considering the (average) number of references in a paper as its 'input' and the number of citations achieved by that paper as its 'output', the ratio citations/references may be interpreted as the 'gain factor' of the paper; it sounds reasonable that a gain factor of 1 is a sort of watershed between papers recognized by the scientific community as important and papers not recognized as important. As the average number of references in scientific articles is about 20 (according to the 1986 SCI), I consider

> N = 25 a good compromise for a paper to be considered as a highly cited paper (from here on labeled as HCP). This category includes only 3% of all articles cited between 1961 and 1980 (Garfield, 1984). Extensive explorations by the author have confirmed that by using HCP rather than the total number of citations as an indicator of scientific impact, the injustice to the many excellent scientists working in small scientific fields, so abundantly clear in lists of most cited authors (e.g., Garfield, 1981a), is greatly eliminated" (p.187).

As others had before him (e.g., Garfield, 1972), Plomp noticed, for example, that because in 1982 the 40 core biochemistry journals produced 13,500 articles while the 25 core astrosciences journals produced 4,500 articles, there were many more opportunities for citations in the former field than in the latter. In fact, "the most cited biochemistry papers obtained ten times as many citations as the most cited astrosciences papers".

In the *Encyclopedia of Library and Information Science*, the entry for 'citer motivations' says

> "Citations are examples of unobtrusive or nonreactive social science measures. Unobtrusive measures are physical evidences of activity that exist independently of their source: the private act of authorship produces citations that are public objects available for scrutiny and analysis. As with many of these unobtrusive measures, it is difficult to ascertain in any given application what social or psychological construct the citation counts are measuring" (Brooks, 1988, p.48).

For social indicators researchers, this quotation has a remarkably familiar ring. In our terms the author was saying that citations are objective indicators providing unclear indications. They must be supplemented by subjective indicators revealing not only the motivation or aims of authors/citers, but also the meaning or interpretation of citations for those who read them. People like Garfield shared the view of sociologists of science like Merton (1977), who believed that the status of any scientist's research output

> ". . .resides only in the recognition accorded his work by peers in the social system of science through reference to his work. . .Since recognition of the worth of one's work by qualified peers is, in science, the basic form of reward (all other rewards deriving from it) and since it can only be widely accorded within the social system of science when the attributed work is widely known, this provides institutional incentive for the open publication, without direct financial reward, of scientific work" (pp.47-48, as quoted by Lawani and Bayer, 1983).

Lawani and Bayer (1983) go on to say that

> "The pressure for public diffusion of one's work through open publication is accompanied by the obligation within the institutional structure of science for the user of that freely published knowledge to make open reference to the sources to which he is indebted. 'Not to do so is to incur the sanctions visited upon those judged guilty of stealing another's intellectual property (i.e., plagiary)' [Merton 1997, p.48]. This reflects the origin of the practice of citing and is a basic justification for

> its application in studies of the sociology and history of science. . . Despite the ambiguities of citation practices, the difficulties of ascertaining why a paper is or is not cited, and the potential malpractices in citing, considerable evidence has been accumulated to suggest that citations do indeed provide an objective measure of what is variously termed "productivity," "significance," "quality," "utility," "influence," "effectiveness," or "impact" of scientists and their scholarly products" (pp.60-61).

The evidence takes many forms. For example, Narin (1976) reviewed 24 studies published between 1957 and 1975 that generally confirmed the hypothesis that citation counts are positively correlated with peer rankings of the quality of scientific articles, eminence of scientists, graduate schools, graduate departments, editor evaluations, Nobel prizes and other awards, authors' incomes, access to resources, initial appointments and mobility. Eight of 12 studies that provided correlation coefficients had values of at least 0.6, with 5 of those above 0.7. The lowest value was 0.2. Brooks (1985) reported that Virgo (1977)

> ". . .found citation analysis to be a consistent and accurate predictor of important scientific papers, better on the average than the individual scientist's judgment which 'is a reasonable conclusion if one considers that citations actually reflect a consensus of a large group of readers as compared to the evaluation of a single individual' [Virgo, 1977, p.423]".

Lawani and Bayer (1983) undertook a very thorough study comparing peer assessments of cancer research papers with the papers' citation rates and concluded that "Highly rated papers are more highly cited than average papers".

Notwithstanding such supporting evidence for the usefulness of citation counts in the evaluation of published research, Garfield has often published cautionary remarks about the use of such counts. For example, he wrote that

> "Counts of this sort are strictly quantitative and objective. But even admitting this limitation, an author's or a paper's frequency of citation has been found to correlate well with professional standing. It is certainly not the *only* measure, nor one that can be used, for any purpose, in isolation. We do not claim for it the absolute reliability that critics of citation analysis have wrongly imputed to us when they have attacked it. The fact does remain, however, that it provides a useful objective criterion previously unavailable" (Garfield, 1981, p.135, reprinted in Garfield 1983).

Among the problems with using citation counts in the ISI databases to evaluate publications, Garfield (1983) mentioned the following.

> ". . .there are undoubtedly highly useful journals that are not cited frequently [e.g., Scientific American]. Scientists read many such journals for the same reason people read newspapers and other non-scientific periodicals – to keep up with what is going on generally. . .Another consideration is that citation frequency is sometimes – indeed to some extent must be – a function of variables other than scientific merit. Some such variables may be an author's reputation, the controversiality of

subject matter, a journal's circulation and its cost, reprint dissemination, its coverage by current-awareness and indexing and abstracting services, society memberships, the availability and extent of libraries' journal collections, national research priorities" (p.137); see also Garfield (1977, 1988).

In a series of papers, MacRoberts and MacRoberts (1986, 1987, 1989, 1989a) challenged the use of citation counts on a variety of grounds. From their papers and others I constructed the following list of criticisms. (1) Many individuals and works that have had an influence on the development of published papers are not cited in those papers (they estimated that "about 15 percent of the influence on a paper is contained in its references"; (2) sometimes important influences are mentioned in the text of published papers but not in their bibliographies; (3) the motivation for self-citations is suspect (Self-citations are generally estimated to be about 10% to 30% of all citations; Bonzi and Snyder (1991) surveyed 51 self-citing authors and found "very few differences in motivation" between self- and other-citations.); (4) when review papers are cited, it is unclear exactly who is being "rewarded"; (5) general references within a text also have unclear referents, e.g., "Mendelian genetics"; (6) ISI indexes contain many errors, e.g., when authors' names are spelled in different ways, or written sometimes with and sometimes without middle initials, separate entries may made for the same document; when page and/or volume numbers are changed, separate entries may be made; when page numbers alone or page numbers and volume numbers are inverted, new entries may be made; (7) ISI indexes cover only about 10% of scientific literature; (8) "English language journals and western science are clearly over-represented, whereas small countries, non-western countries, and journals published in non-Roman scripts are under-represented."; (9) different disciplines are more or less well represented (e.g., 6% of biology journals are included, compared to 14% of clinical medicine journals); (10) politics seems to have influenced inclusions and exclusions (e.g., *Review of Radical Political Economy* is excluded, but *Public Interest* is included); (11) later proponents of views may be cited as the views' inventers; (12) proponents of views may be cited as opponents, and vice-versa; (13) people often cite material that "they have not seen or that they have seen but have not read"; (14) people make ceremonial and perfunctory citations (Moravcsik and Murugesan (1975) estimated about 40% of the references in a sample of high energy physics papers were perfunctory); (15) private communications are often influential but do not provide a published title to cite; (16) authors are notorious for repeating themselves in different papers, each of which may be cited as an additional contribution to research; (17) people sometimes select only one of several similar papers that influenced them; (18) they also redundantly cite several similar papers because they are similar, although they may have read and/or been influenced by one (Moravcsik and Murugesan (1975) estimated about one-third of the references in their sample were redundant); (19) they may cite a well-known or fail to cite a relatively unknown author because these practices are perceived to impress their readers; (20) about 70% of citations appear to be "multi-motivated"; (21) different disciplines have different typical rates of citation, which are typically neglected by citation analysts (e.g., engineering and mathematics papers average 5 to 6 references per publication, psychology and biology average 8 to 10, earth and space science, physics, chemistry and clinical medicine average 12 to 15 and biomedical research papers average 18 to 20); (22) citation rates vary with countries of origin; (23) rates vary with the "size of the pool of available citers"; (24) methods and review papers "receive disproportionately more citations than theoretical or empirical papers"; (25) citation counts cannot allocate credit for papers with several authors; (26) researchers from some countries tend to be aware of and cite papers coming from their countries (e.g., Americans tend to cite papers by Americans); (27) citation counts do not distinguish positive or negative reasons for the citations; (28) authors tend to search the literature for and cite those papers that agree with their views; (29) citation rates vary with the physical accessibility of cited material; (30) changes in editors and editorial boards have produced changes in citation patterns (Sievert and Haughawout, 1989); (31) recent papers are cited more than older ones because there are relatively more of the former (Oppenheim and Renn, 1978); (32) rapidly developing fields like molecular biology and biochemistry are "more dependent on fresh data" and therefore generate relatively more citations (Vinkler, 1987).

Summarizing their criticisms, MacRoberts and MacRoberts, 1989a) wrote

> "Apparently, the pioneers of citation analysis were so intent on finding an 'objective' measure of quality that, in their enthusiasm, they neglected to check their assumptions against events. . .What is the consequence of this discovery for the practical application of citation analysis? It alone should suffice to exclude evaluative citation analysis from the arena of science policy. Add to this our empirical findings, and those of others, and the claims made for citation analysis largely collapse. It would therefore be the better part of wisdom to place a moratorium on the use of citations in science policy until the problems are thoroughly aired, which means attending to critics rather than ignoring or dismissing them as uninformed and misguided 'non-believers'" (p.10).

While I believe that most of the criticisms listed above are accurate and should not be ignored or dismissed, I feel the same way about all the correlational studies indicating some validity in the use of citation counts for evaluative purposes. In fact, I believe that many of the problems and limitations of citation counts are similar to problems and limitations encountered with objective indicators in other fields. While research by sociologists and psychologists of science (scientometricians) has not progressed as rapidly as that of other social indicators researchers into explorations of the subjective side of citation behaviour and motivations, it has certainly progressed. Good investigations and literature reviews of citer motivations may be found in Smith (1981), Prabha (1983), Cronin (1984), Brooks (1985, 1988), Puder and Morgan (1987), Moravcsik (1988) Cozzens (1989), Bonzi and Snyder (1991) and Liu (1993). Today, anyone writing an essay like this one relying heavily on citation counts must be aware of, and make readers aware, of their limitations. I hope that this brief overview of some of the literature has at least accomplished that.

More importantly, I hope that other researchers will regard this review of the citation classics from *Social Indicators Research* as merely a first step toward an evaluation of our research literature. Additional steps involving survey research and detailed, systematic appraisals of publications must be taken, and the sooner the better.

Given the background information just reviewed, and especially that summarized in Exhibit 1, it is obvious that we have some flexibility in specifying a threshold figure for citation classics. I exchanged some ideas with people at Kluwer and at ISI regarding a list of the top 50 most frequently cited papers or, alternatively, papers with 50 or more citations. David Pendlebury provided us first with a list of all papers with 35 or more citations, and then with a complete list of every cited paper from the journal for the whole period from March 1974 to December 2003. I cleaned up the material a bit, entered the whole dataset into an SPSS file and produced the statistics in Exhibit 2. There were a total of 1392 titles published in the first 30 years. Since the journal seldom published book reviews, editorials or letters, most of those titles represented articles. Eight hundred and twenty articles (58.9%) were not cited at all, which is a bit higher than the 55%-57% general average for the whole *Index* database, lower than the 74.7% for all social sciences material and higher than the 48% for social science articles alone. The 572 (41.1%) cited articles generated 4979 citations, with a classic hyperbolic distribution curve in which relatively few articles attract many citations and relatively many articles attract few citations. The mean number of citations per published article was 3.6, with a mode and median value of zero, and a standard deviation of 11.8. There were 34 articles with 35 or more citations each, and those 34 (2.4%) articles attracted 2208/4979 = 44.4% of all citations. The top 68 (4.9%) articles attracted 2997/4979 = 60.2% of all citations. Given their extraordinary contribution to the journal's total citation count, and the fact that articles with 35 or more citations were nearly three standard deviations above the mean, those articles form a fairly distinguished lot. Therefore, I decided to use 35 as the threshold figure for designating citation classics from *Social Indicators Research*. Of course, other researchers may prefer to use other figures.

The Appendix to this introduction contains the complete list of cited articles with their citation figures. The articles are divided into the decades of their publication, 1974-79, 1980-89, 1990-99 and 2000-03, and each is listed alphabetically by first authors' names. After spending some hours cleaning up the list, I became convinced that I would never be able to catch all the errors in the time that was available to me. But it is a pretty good list for others to begin with. At the end of the Appendix there is a short table summarizing publication and citation figures.

Exhibit 3 lists our 34 citation classics with their citation figures. Because it takes some time for articles to be discovered, used and cited, there are no classics from the 2000-03 period. There are 10 articles (29.4%) each from 1974-79 and 1990-99, and 14 (41.1%) from 1980-89. By authors' countries of origin, there are 16 (47.1%) from the USA, 5 (14.7%) each from Canada and Australia, 3 (8.8%) from the Netherlands, 2 (5.9%) from the UK and 1 (2.9%) from Israel.

By first authors' names, there are 5 (14.7%) by Andrews, 4 (11.8%) by Diener, 3 (8.8%) each by Veenhoven and Headey, 2 (5.9%) each by Cummins, McKennell and Michalos, and one each by all other authors. Collectively, 37 authors produced the 34 papers, and there are 16 single-authored papers. Using Plomp's threshold for highly cited papers, only 8 more papers would be added to the 34.

Nineteen (55.9%) of the 34 articles appeared as the lead article in some issue. This seemed quite remarkable to me because I was responsible for sequencing the set of articles for every issue and I always put what I regarded as the best available article first. Since we know that only 2.4% of all articles ended up as classics, either (a) I was a pretty good judge of quality, (b) the lead articles tend to be read more than others (like newspaper articles), or (c) both.

Examining the content of the articles, I was shocked to discover that all but one of them (McCall, 1975) focused on some aspect of subjective indicators. In view of the historical facts that the field was originally dominated by researchers interested in objective indicators and that today practically all researchers agree that objective and subjective indicators are equally important, the near total dominance of subjective indicators research in our classics is both surprising and disturbing. I would guess that the main reason for the dominance of subjective indicators research is that there are relatively more psychologists and people interested in personal reports about a good quality of life than there are others. While "Others" would include a wide variety of people, e.g., sociologists, demographers, gerontologists, geographers, environmentalists, economists, political scientists and population health researchers, each group would have a relatively narrow range of interests compared to those interested in the psychological structure of perceived well-being. Whether or not there is anything to that explanation, it would be a pity if objective indicators research came to be relatively neglected in the future. Hopefully, this publication can serve as a wake up call for researchers to redirect attention to redressing the balance.

Although the articles in this collection are assembled chronologically, they are discussed below in an order that emphasizes their interrelations and reveals the gradual expansion of the frontier of research in subjective well-being. Only 17 articles are included because they seemed to be sufficient to reveal the progress made over the past 30 years, with a minimum of duplication.

The article by McCall (1975) contains a brief but historically instructive account of the state of the discussion in the early 1970s about the definition of the phrase 'quality of life'. There are several nice distinctions drawn and some arguments are given in favour of a definition of 'quality of life' based on objective indicators. First, he claimed that "to define QOL in terms of the general happiness, of 'the greatest happiness of the greatest number,' would be merely to repeat Mill's work". If one's analysis went no further than the definition, that might be true, but no contemporary proponent of Mill's formula ends his or her analysis at that point, e.g., see the paper in this collection by Veenhoven (1994). Second, he claimed that a society that looked good on the basis of objective indicators (e.g., full employment, low crime rates) would have a higher QOL than one that looked bad on the same basis even though people in the latter society were happier (had higher subjective well-being ratings) than those in the former. This is patently question-begging, and he offers nothing stronger in its defence than “intuitively, one would think that the QOL of” the society with the superior objective indicators is superior. It has, in his view, “the *necessary conditions* for happiness”, and it is these conditions that he regards as defining QOL. In my view, people’s perceived well-being measured by subjective indicators constitutes a complementary and equally necessary defining condition of QOL. That is, the quality of life of an individual or a community should be defined and measured by both objective indicators of their living circumstances (Veenhoven’s ‘livability indicators’) and subjective indicators of their perceived well-being (happiness, satisfaction or subjective well-being) (Michalos 2003, 2004). (It must be remembered, of course, that some evaluation by someone is necessary in order to determined exactly what objective indicators indicate evaluatively.) While McCall allowed that subjective indicators might be *signs* of but not *constitutive* of QOL, my view is that high levels of perceived well-being, measured by subjective indicators, are *partly constitutive* of QOL, the other part being people’s living circumstances broadly construed. I regard my view as socially constructed and subject to revision. I think it is generally representative of the views of the majority of social indicators researchers today, but I have not done any systematic research to support that belief.

With some oversimplification, we might use the two necessary conditions to characterize four states of affairs. If one’s living conditions and perceived well-being are good, then the overall quality of one’s life is good. This we may call Real Paradise. If one’s living conditions are bad and one’s perceived well-being is good, then the overall quality of one’s life is not good. This would be the proverbial Fool’s Paradise. If one’s living conditions are good and one’s perceived well-being is bad, then the overall quality of one’s life is still not good. It might be called a Fool’s Hell. Finally, if one’s living conditions and perceived well-being are bad, then the overall quality of one’s life would certainly be bad. That, I suppose, could be called Real Hell. Since living conditions and perceived well-being admit of diverse levels or degrees of goodness or desirability in some sense, these blunt distinctions could easily be made more sophisticated and subtle. Fortunately, there is no need to pursue such complications here.

In another section of McCall’s paper he claims that because human needs may be objectively determined and are in principle limited while human wants may only be subjectively determined and are in principle unlimited, a notion of quality of life defined in terms of need fulfillment is preferable to one defined in terms of want fulfillment. He relied heavily on Maslow’s (1954) theory of a hierarchy of needs, and the strength of his position is directly proportionate to the persuasiveness of his case for the objectivity of needs versus wants. In Michalos (1978) I devoted a considerable amount of space to providing fairly rigorous definitions of ‘needs’ and ‘wants’, and concluded that it would be impossible to make such a case persuasive. I thought that if anyone could have made such a case, Braybrooke (1987) would have done it, but a careful reading of that volume only reinforced my earlier view. A definition of 'quality of life' based on some set of objectively determined needs seems to me to be unachievable and, in the light of my views explained in the preceding two paragraphs, undesirable.

The first and last articles in the collection were published 22 years apart, but they addressed the same basic problem, with some different additional hypotheses and methodologies. The basic problem was to empirically determine the total number of domains required for a full assessment of the perceived quality of life of individuals and communities.

The paper by Andrews and Withey (1974) was followed by several articles and by their fine book, Andrews and Withey (1976). They began by creating a "list of items which, ideally, would include all the significant 'concerns' of people". They found 800 items by examining published lists from national and international bodies, 8 different national samples of Americans, representative data from 12 other countries, structured interviews of a dozen people with diverse backgrounds and lists of published values. Using "some *ad hoc* clustering to combine concerns", they found 123 items that seemed to cover about 100 concerns. Then they applied Guttman (1968) Smallest Space Analysis to a 62-item subset of the 123 items, which was possible because the 62 items were used in the US national survey of May 1972 (N = 1297). The "62 items were reduced to 30 semi-independent domains", and finally to 12, on the basis of three criteria, namely, (1) predictive power, (2) dispersion in multi-dimensional space and (3) policy relevance. The 12 domains were house/apartment, spare time activities, things done with family, your health, amount of fun, time to do things, job index, national government index, efficacy index, family index, consumer index and money index. To test the robustness of the dozen domains, Andrews and Withey undertook separate analyses on 10 subgroups, including, ". . .men, women, blacks, four different age groups, two groups extreme with respect to socio-economic status, and a group of married, white, employed men in their middle years with children living at home".

Summarizing their work, they wrote,

> "A series of analyses, some of them replicated in more than one survey, showed that a particular subset of 12 domains could explain 50% to 60% of the variance in sense of overall life quality, that neither other domains nor standard classification variables contributed anything additional to this explanatory power, and that this level of explanation could be achieved in each of 22 different subgroups of the American population (defined in demographic terms) as well as in the population considered as a whole. . .one wonders how close to the actual upper limit is the achieved explanatory power of 50% -- 60%. Given the unreliability of the measures the upper limit is certainly not 100%. Further work will attempt to assess the reliability of the measures employed" (p.23).

The dependent variable used to measure "overall life quality" was called "Life #3", whose values were equal to the mean score resulting from asking "How do you feel about your life as a whole?" at two points of time 8 to 12 minutes apart in each interview. The two responses to the same question (Life #1 and Life #2) usually correlated with each other in the 0.6 to 0.7 range. The response scale used was the now-famous 7-point Delighted-Terrible scale (DT scale). For reasons that will be made clear below (in the discussion of Andrews and McKennell, 1980), the DT scale contains two words specifically designed to elicit an affect-based response, i.e., "pleased" and "unhappy". The "explanatory power" of the 12 domains was measured by Multiple Classification Analysis, which is a kind of ordinary least-squares regression, using the various mean DT scores for each domain as predictors. In a footnote the authors explained that they tried weighting domain DT scores with "importance scores", but "no predictive gain could be achieved". They also "thought there might be substantial interactions in the data, but so far, none of marked effect has been found". The "standard classification variables" included respondents' income, sex, race, family life cycle, age and education.

Cummins (1996) began by scanning 1,500 articles providing data on life satisfaction, looking for "different terms that had been used to describe domains of life satisfaction". For an article's terms to be used, the article had to have at least three domains purporting to "represent a broad indication of life quality", and a detailed description of the scales used and average scores obtained for each domain. Most importantly, in contrast to Andrews and Withey, "Responses to criteria of happiness were excluded". All together, Cummins found 32 studies meeting his criteria, and those studies used 351 different domain names. The 68 samples described in those studies were "of four broad types: general population probability or quota samples, general population samples based on a variety of specific criteria, samples of people with chronic medical conditions, and samples comprising people with a chronic psychiatric impairment". The inclusion of samples of people with chronic physical and mental health problems was a notable step beyond Andrews and Withey.

His first aim was to determine how many domain names could be categorized under one of the seven domain headings of his Comprehensive Quality of Life Scale (ComQol). The latter's domains include material well-being, health, productivity, intimacy, safety, community and emotional well-being. He found that 83% of the 351 domain names could be reasonably classified into one or another of ComQol's seven domains. For example, ComQol's category of "intimacy" includes things like family life, family relations, friendships, marriage, living partner and spouse. He listed the 56 domain names that did not fit into the ComQol categories and indicated that "the question as to the appropriate number of domains must remain open". Nevertheless, he expressed reservations about adding the particular domain of government because "Not only would the inclusion of such a domain exert a strong negative bias on life satisfaction measurement, but also people generally report these aspects of life to be unimportant to them personally".

His second aim was to see if "a hierarchy of domain satisfaction could be established" and, if so, third, to see if it was invariant among groups with high or low levels of life satisfaction. Examining mean values for each domain across four levels of reported life satisfaction, he found that there was indeed a hierarchy with the two domains of intimacy and health "consistently above the study mean [of Z-scores], while the other five domains lay consistently below.

Most remarkable in this respect was intimacy which lay from 1.29 to 1.53 standard deviations above the mean for each of the four levels of overall life satisfaction". While the hierarchy was invariant with respect to membership in the top two and bottom five domains, among the latter there was considerable shuffling about. For example, productivity and community occupied the third and fourth places, respectively, for those in the group with the highest levels of life satisfaction, sixth and third places for those in the group with the second highest levels of life satisfaction, fourth and seventh for those in the next group, and sixth and fifth for those with the lowest levels of life satisfaction. Contrary to Cummins's claim, it does not seem to me that "the data in Table IV indicate a high level of inter-domain consistency" for five of the seven domains.

Considering the results for his samples of people with chronic health problems, Cummins found that those with chronic physical problems had life satisfaction scores much like those in "normal population samples", while those with chronic psychiatric problems "had significantly lower overall life quality scores than either of the other two groups. Moreover, . . .the domain of intimacy no longer retained its pre-eminence. These data are consistent with the loss of intimacy as a buffer against diminished life circumstances".

Since the procedures and criteria for success used by Andrews and Withey and Cummins were different, one would not have expected both approaches to yield identical results. It is perhaps worthwhile to mention that while Cummins treated all the terms used by other researchers in the context of domain identification as domain names, Andrews and Withey distinguished (unfortunately without rigorous definitions) domains or "role situations" from values. As illustrated in the first exhibit of their paper, these authors had the idea that people might, for example, look for beauty (a value in their terms) in several domains (on the job, at home, in one's friends). Then, using many values and many domains, one could generate a matrix of value-by-domain satisfaction levels, which in the aggregate would constitute one's overall satisfaction with life. Inspection of the first table in Cummins's paper reveals that some of his domain names were regarded as value names by Andrews and Withey. For example, 'safety' is a domain name for Cummins but a value name for Andrews and Withey. Since there is no generally accepted rule book for determining what one ought to regard as a value versus a domain, one need not take sides on which approach is most appropriate. The bottom line is that the two approaches yielded fairly similar results.

I suppose anyone reading the philosophic literature since *The Republic* of Plato would have expected to see fairly similar results. A good overview may be found in Tatarkiewicz (1976). Around the world and across time, people regard good health, family and friends, beautiful things, financial and other forms of security as important features of a good life. Andrews and Withey and Cummins showed us that the core of important life domains is not as broad as the great numbers of possible domain names or items might suggest. Cummins's observations that "The larger the proportion of scale items that relate to satisfaction with family and friends, the higher will be the overall life satisfaction score" and that "the inclusion of items relating to 'Government'. . .

will produce a lower life satisfaction score" in aggregated indexes are well worth remembering.

Although the paper by Levy and Guttman (1975) was not included in Cummins's study, the problem addressed in the paper was closely related. Methodologically, the paper is closer to Andrews and Withey. Levy and Guttman began with what I would call a semantic view of theories (Michalos 1980). According to this view, a scientific theory is a system of definitions that one hypothesizes will match some aspect of the real world, and then one makes empirical observations to test one's hypothesis. The elaborate network of definitions contained in Levy and Guttman's comprehensive mapping sentence includes 6 facets identifying (A) a type of assessment (cognitive versus affective), (B) type of state or treatment, (C) group, (D) environmental framework, (E) level of specificity, and (F) life area (domain) according to some level of satisfactoriness. I suppose their life areas of security and communication would be values for Andrews and Withey. In any case, multiplying the elements available for each facet times all the others, Levy and Guttman could generate 5824 items ("structuples"), from which they selected 24 to test in their 1973 spring survey in Israel. Applying Smallest Space Analysis to correlation coefficients for pairs of variables in their data-set, they were able to obtain geometrical representations of relationships among the variables. The representations are difficult for me to interpret and to compare with the representations constructed by Andrews and Withey (1974, 1976) using similar software. While it is clear that the mapping sentence provided a useful heuristic device for generating items that corresponded to features of relationships among attitudes of their respondents, it is unclear what good the geometrical representations are in themselves or for explaining the relationships. One can look at the figures and see that this or that attitude/item is literally in space nearer or farther from some others, but the correlation coefficients themselves gave us information about their conceptual closeness and additional statistical analyses without geometrical representations could give us information about likely causal connections among the variables. I suppose proponents of the geometrical representations would say that the figures make it easier to "see" relationships and craft explanatory hypotheses, but in fact very few hypotheses seem to have been generated from all the figures presented, for example, in Andrews and Withey (1976).

Andrews and Inglehart (1979) applied Smallest Space Analysis to the U.S. survey data from May 1972 and data from national surveys in nine European countries in order to see how comparable the structure of subjective well-being was in the 10 countries. At the time they wrote, "the cross-cultural comparability of the phenomenon of perceived well-being [was] largely unexplored". In their view, there were "important practical uses of such structures". In particular, Andrews and Inglehart thought

> "They provide guides to the adequacy of coverage and statistical efficiency of indicators of perceived well-being. To the extent that people in different societies organize their thinking about well-being in basically similar ways, it is feasible and potentially productive to

> undertake cross-cultural research with standardized instruments and to make well-grounded comparative statements based on the results" (p.74).

Fortunately, they discovered that

> ". . .(a) there seems to be a basic similarity in structures among all nine of these western societies; (b) within this basic similarity the European structures are distinct from the American structure; (c) even within Europe there is modest heterogeneity; and (d) if one averages out the differences among the individual European structures, the result is a structure that is closer to the American structure than are most of the individual European structures" (p.83).

Besides these general conclusions, Andrews and Inglehart reported that the American structure was "most similar to that of The Netherlands. . .and least similar to the structures in Great Britain, France and Denmark".

Schneider (1975) provided data addressing a fundamental problem for all social indicators researchers in the 1960s and 1970s. As he put it,

> "It is certainly possible that individuals or social groups may be exposed to objectively better conditions of health care, environment, employment, etc. than other individuals or groups but subjectively feel that the quality of their personal life experiences are no better. Despite the often found assumption that objective social indicators data actually reflect the quality of life experienced by people, we have no reason *a priori* to assume that such a correlation exists. The connections between the 'quality of life' as measured by objective social indicators and the 'quality of life' subjectively experienced by people is really open to question. . .(pp.496-497)."

Comparing data on "objective life conditions" in 13 U.S. cities from a 1972 study with "measures of subjective life quality" in those cities obtained from a 1968 social survey, Schneider boldly concluded that

> ". . .no relationship exists between the level of well being found in a city as measured by a wide range of commonly used objective social indicators and the quality of life subjectively experienced by individuals in that city. Cities that are most well off as measured by objective indicators are not necessarily the same cities in which people are subjectively most satisfied with their life situations. Conversely, cities that are worst off objectively are not necessarily the same cities where subjective satisfaction is highest. . .Objective social indicators cannot be taken as direct measures of the welfare or the quality of life actually experienced by individuals" (p.505).

It is difficult to confirm this very strong statement on the basis of the data provided in the article. We are told that of some 416 correlations calculated for two racial groups, "11% (9% for whites, 13% for blacks) [were] significant at the 0.10 level". We are not given the full list of subjective indicators or any justification for each of the objective indicators. So it is unclear how 416 correlations were generated. Presumably Schneider counted percent changes in indicator values as negative or positive if the indicators themselves were negative or positive, e.g., if robbery rates are negative indicators, then increases in such rates are negative and decreases are positive indicators. It may be asking too much to expect percent changes in a single year to show a significant correlation with some relevant subjective indicator for that year, e.g., to have a negative percent change in air quality in 1968-69 (annual reduction in air pollution) show a significant correlation with residents' satisfaction with air quality, especially since we do not know exactly when the attitudes were measured and when the pollution levels were measured. It is indeed unclear why satisfaction with housing would be negatively correlated with United Fund contributions for "whites" (-0.48) but positively correlated for "blacks" (0.51), but it is also unclear to me why one would expect any correlation at all between housing satisfaction and such contributions. I suppose it may have been assumed that such contributions would be positively related to incomes which would be positively related to objective housing quality which was finally to be positively related to housing satisfaction. If that is the chain of reasoning, one would have thought a simple income-to-housing satisfaction correlation would have been more direct and therefore preferable. I was unable to figure out which half of the significant correlations were "in the wrong direction".

Although the interpretation of Schneider's evidence is a bit difficult, there is plenty of other evidence supporting the view that the association between people's objectively measured life conditions and their subjective assessments of those conditions is relatively weak. Personal disposable income, as self-reported in surveys or as estimated from a nation's National Income and Product Accounts in the form of Gross Domestic Product (GDP) per capita, is probably the best single, overall objective measure of people's access to good life conditions. Generally speaking, access to both public and private goods and services are increased as a population's GDP per capita increases, although every segment of the population may not enjoy an equal share of the increase. The UNDP's annual *Human Development Report* reveals that around the world, as per capita incomes increase, so do people's opportunities for better housing and house furnishings, better hospitals and schools, health care and education, longer lives free of disability, food and food security, jobs and job security, social safety nets (e.g., old age pensions), roadways and other means of communication, greater social and economic mobility, greater control over government and over their own lives. Thus, using personal disposable income as a good surrogate objective indicator of life conditions, one might regard the association between this indicator and a subjective indicator like reported life satisfaction (or happiness or subjective well-being) as central for confirming or disconfirming Schneider's view.

The relative insignificance to life satisfaction of half a dozen "standard classification variables", including income, in the presence of a dozen domain satisfaction variables was already documented by Andrews and Withey (1974, 1976), and similar results were reported by Inglehart and Rabier (1986) for 8 west European countries. Duncan (1975) put the question bluntly when he asked: Does money by satisfaction? To answer this question, he examined associations between respondents' reported family incomes and satisfaction with their "standard of living. . .the kind of house, clothes, care, and so forth" from the Detroit Area Studies of 1955 and 1971. He found that although respondents' median family "income in constant dollars increased by a factor of 1.42" in the period from 1955 to 1971, their mean levels of satisfaction with their standard of living did not significantly change. Nevertheless, he also noticed that "individual satisfaction typically increases with increasing incomes *in cross-section data*". Thus, he concluded (roughly as Aristotle concluded 2400 years earlier) that "the relevant source of satisfaction with one's standard of living is having more income than someone else, not just having more income" (p.273). These conclusions confirmed those of Easterlin (1973).

Diener and Biswas-Diener (2002) published an excellent review of a substantial chunk of the literature around this question and provided the following convenient summary.

> "Four replicable findings have emerged regarding the relation between income and subjective well-being (SWB): 1. There are large correlations between the wealth of nations and mean reports of SWB in them, 2. There are mostly small correlations between income and SWB within nations, although these correlations appear to be larger in poor nations, and the risk of unhappiness is much higher for poor people, 3. Economic growth in the last decades in most economically developed societies has been accompanied by little rise in SWB, and increases in individual income lead to variable outcomes, and 4. People who prize material goals more than other values tend to be substantially less happy, unless they are rich. Thus, more money may enhance SWB when it means avoiding poverty and living in a developed nation, but income appears to increase SWB little over the long-term when more of it is gained by well-off individuals whose material desires rise with their incomes. Several major theories are compatible with most existing findings: A. The idea that income enhances SWB only insofar as it helps people meet their basic needs [Maslow 1954], and B. The idea that the relation between income and SWB depends on the amount of material desires that people's income allows them to fulfill. We argue that the first explanation is a special case of the second one. A third explanation is relatively unresearched, the idea that societal norms for production and consumption are essential to understanding the SWB-income interface. In addition, it appears high SWB might increase people's chances for high income" (p.119).

Additional information on the theories "compatible with most existing findings" and the "relatively unresearched" explanation will be considered below in Michalos (1985) and Veenhoven (1994).

Beginning with the assumptions that measures of perceived well-being are measures of attitudes, and that attitudes are composed of cognitive, affective and conative elements, Andrews and McKennell (1980) applied structural equation modeling (LISREL III, Joreskog and Sorbom, 1976) to British and American data-sets to assess the average distribution of these elements in 23 measures of life-as-a-whole. Their effort represented

> ". . .a conceptual and statistical refinement over a common past practice of considering that all global measures are simply reflections (perhaps to differing degrees) of a single underlying factor, feelings about life-as-a-whole. Measures that ask about happiness, fun, and enjoyableness (in addition to Bradburn's scales of positive and negative affect) fall in the group for which relatively high affect-cognition ratios were observed. On the other hand, items that employ the term 'satisfaction' and/or that involve comparisons with implicit or explicit criteria tend to have lower affect-cognition ratios. . .satisfaction measures tend to tap more negative than positive affect and. . .in the United States (but not in Great Britain) the reverse is true for happiness measures. . .There was a rather consistent tendency for measures employing 3-point response scales to show lower validities than measures with scales having more response categories" (pp.150-151).

A better idea of the significance of their analyses may be grasped by considering their results for two life satisfaction scales mentioned earlier, namely, the DT version including the words 'pleased' and 'unhappy', and the 7-point satisfaction scale running from 'completely satisfied' to 'completely dissatisfied'. Although the DT version (Life #1) and the 7-point satisfaction scale each had affect-to-cognition ratios of 0.8, the DT version had a more even balance of positive and negative affect than the 7-point satisfaction scale. The validity coefficients for each measure were similar (0.77 versus 0.75, for the DT version and the 7-point satisfaction scale, respectively).

In a later paper, Andrews (1984) used the same structural equation modeling techniques to estimate the construct validity, method effects and residual error for a comprehensive set of survey research measures. I think it is the best paper he published, and perhaps the best paper in the field, although it was not published in *Social Indicators Research*. He examined 13 aspects of survey design, including characteristics of the response scales (e.g., number of scale categories, off scale options, explicit midpoint), item characteristics (e.g., length of introduction and question, absolute vs. comparative), questionnaire design (e.g., item positions, single items vs. sets), topic characteristics (e.g., social desirability sensitivity, content specificity), and data collection procedures (e.g., telephone, face-to-face interview). His overall assessment was that

> ". . .the 'typical' survey measure examined in this research consisted of 66 percent valid variance, 3 percent method variance, and 28 percent residual variance. . .Although no claim is made that the set of survey items examined here is representative of all items used in current surveys, this set of items is broader and more heterogeneous than any other we know of whose quality has been estimated, and hence these estimates for the quality of our 'typical' item probably provide the best available information about these aspects of measurement quality for single-item survey measures tapping rather specific attitudes and behaviors. . .For the validity estimates, which have a standard deviation of .10, one can infer that about two-thirds of all validity estimates fell in the range .71 to .91. Hence, about two-thirds of the survey measures

> examined here contained between 50 percent and 83 percent valid variance; roughly one-sixth contained more than 83 percent valid variance, and about one-sixth had less than 50 percent valid variance" (pp.424-425).

The papers by Kammann, Farry and Herbison (1984) and Fordyce (1988) are similar insofar as they each provide comparative assessments of a wide array of measures of subjective well-being. Their main difference is that the former paper is focused on showing that there is a "single general well-being factor", while the latter is focused on showing the virtues of a single measure of happiness.

Kammann and his colleagues developed an affect-balance scale that was similar to Bradburn's (1969) in having separate items for positive and negative affect, and in using a "balance or net scoring formula to obtain the overall well-being score". While Bradburn's scale has 10 items, Kammann's has a 96-item version called Affectometer 1 and a 40-item version called Affectometer 2. With Bradburn's scale, respondents are asked *whether or not* they have had a particular feeling "during the past few weeks", while with the affectometers they are asked *how often* they have had a feeling in that period, with five options: not at all, occasionally, some of the time, often, all the time. Many researchers, using many different samples, found that Bradburn's Negative Affect (NA) and Positive Affect (PA) scales were relatively independent, while the affectometers' NA and PA scales had an average association of $r = -0.66$. Thus, in contrast to Andrews and Withey (1974, 1976) and Andrews and McKennell (1980), Kamman, Farry and Herbison (1984) claimed that

> "Psychologically. . .we must expand or enrich the sense of satisfaction [with life as a whole] into the experience of really enjoying life, as having a high preponderance of positive feelings and few if any negative ones. In fact, our data suggest that global life satisfaction and the balance of affect come very close to an identity of meaning and that both of these interpret the phrase 'sense of well-being'. . .The most important finding, however, is that this dimension of well-being overlaps strongly with the well-known dimension of neuroticism that regularly arises as a primary factor in the factor analysis of personality items. To this it may be added that scales of depression and trait anxiety also seem to be measuring the negative region of an overall well-being spectrum. . .The hypothesis that positive and negative affect are two independent processes is now shown to be less than a universal structure for well-being. As an empirical result this orthogonality may depend very closely on the particular ten items and test format used in Bradburn's method; on the other hand, the general incompatibility of PA and NA reported here could be a peculiarity of the affectometer technique. . .Researchers who wish to defend the model of independence of positive and negative affect should now demonstrate that independence can be demonstrated by methods other than Bradburn's original procedure" (pp.111-113).

Fordyce (1988) reviewed results of several years of testing his 2-item Happiness Measures which, he claimed, revealed that the simple index showed "good reliability, exceptional stability, and a record of convergent, construct, and discriminative validity unparalleled in the field". The index is constructed of an 11-point happy/unhappy scale measuring the *intensity* of affect, and a *frequency* scale measuring the percent of time "on the average" that respondents feel happy, unhappy or neutral. The formula used to obtain a single combination score from the two items is "*combination* = (*scale score* x 10 + *happy %*)/2".

In support of his own results, Fordyce noted that Diener's examination of twenty well-being indexes led the latter to conclude that

> ". . .the 11-point Fordyce scale showed the strongest correlations with daily affect and with life satisfaction of any measure we assessed. . ." and that the HM's ". . .positive and negative frequency estimates provide convergent, construct and criteria validities that are equal to or superior to those found for the Bradburn scale. . ." (p.364, quoted from Diener (1984)).

Based on those results, Fordyce was able to construct the familiar profile of a happy person as one who has low levels of

> ". . .fear, hostility, tension, anxiety, guilt, confusion, anger. . .a high degree of energy, vitality and activity; a high level of self-esteem. . .emotionally stable personality; a strong social-orientation. . .healthy, satisfying and warm love and social relationships; a life-style typified as involved, active, social and meaningfully productive. . .optimistic, worry-free, present-oriented, internally-locused, and well-directed. . .But beyond this, this description also closely approximates what the literature in psychology views as the major criteria of optimal mental health" (pp.365-370).

Having illustrated the fact that our field has a vast variety of measures of subjective well-being emphasizing one or more of its many dimensions or elements, it should not be surprising to discover that someone tried to find a 'gold standard' to bring some order to the apparent disorder. That is what Cummins (1995) tried to accomplish, not only for whole populations (of western countries) but also for ten subgroups, namely, females, those over 65 years of age, physically disabled, intellectually disabled, those with chronic medical problems, post-organ transplant recipients, low and high income people, U.S. "whites" and "blacks".

Looking for appropriate samples, he scanned over 1,000 articles and books according to a set of criteria that were generally reasonable. The only criterion that I found questionable was the one excluding all scales measuring happiness. He did not indicate why they were all excluded, but I would have included at least some of them on the grounds of their high correlations with most measures of life satisfaction. He included the life satisfaction DT scale of Andrews and Withey although, as mentioned earlier, that scale explicitly includes the terms 'pleased' and 'unhappy' to elicit affective responses. Results based on the Satisfaction with Life Scale of Diener, Emmons, Larson and Griffin (1985) were excluded because the scale routinely yielded "data which fall at least 10%SM [Scale Maximum] points below those of comparable scales". The 17 data-sets finally used were drawn from Australia, Canada, England, Norway, Sweden and the U.S.A., and they employed 14 different measures of life satisfaction. The common statistic he constructed is called "the Percentage of Scale Maximum" (%SM) and is generally defined as "(score – 1) x 100/(number of scale points – 1)".

Examining the reported results from the various studies, he found that "The arithmetic mean and standard deviation of the %SM values is 75.02 ± 2.74". So, he ". . .proposed that, as a working hypothesis, the life satisfaction gold standard be considered as 75.0 ± 2.5%SM". Using that standard as the norm, he found, for example, that (1) the mean scores for "females, people who have survived an organ transplantation for a number of years, and people with a mild or moderate level of intellectual disability living in the community" did not differ from the normative range, (2) the mean scores for "people over the age of 65 years, people with a physical disability, or with a low income, are on the lower margins of the normal range", (3) "people with a chronic medical condition. . .have a mean value well below the two standard deviation range" and (4) people with a high income have means scores "above the normal range".

Atkinson (1982) presented results from a longitudinal (panel) study involving "a representative sample of 2162 Canadians interviewed in 1977 and again in 1979". Apart from some results dealing with a subset (N = 285) of the national sample used in Campbell, Converse and Rodgers (1976), this was the first published report "on the stability of QOL measures over time". The particular measures employed included an 11-point life satisfaction item similar to those used in Great Britain and in the European Community, a self-anchoring ladder scale adapted from Cantril (1965), the Gallup 3-point happiness scale and domain satisfaction scales for five domains; job, finances, housing, health and marriage/romance. A General Quality of Life Index was formed by combining responses for the life satisfaction, happiness and Cantril scales, and a Domain Satisfaction Index was formed by combining domain satisfaction scores. In order to assess levels of change in people's lives from the first to the second survey, respondents were asked in the second wave if their current status was the same, better or worse, for life in general and for specific domains. They were also asked which of 16 significant life events they had experienced in that period, e.g., divorce/separation, serious injury, new job or house.

His most important conclusions were admirably summarized as follows.

> "Opinions have been expressed in some quarters that subjective measures such as satisfaction are poor social indicators because they were so conditioned by expectations and restricted awareness as to be insensitive to changing circumstances. It is also argued that expectations and aspirations adjust very quickly to new situations and that satisfaction and other measures revert to their original levels immediately. This position would have led to predictions that (a) very few individuals would indicate any change in their situation, and/or (b) virtually no adjustment in subjective indicators would occur when changes did occur.
>
> Our findings contradict both hypotheses in that significant numbers of respondents perceive changes in their lives and those changes were reflected, for better or worse, in their satisfaction levels. The fact that

> these changes took place over a two year period indicates that, while adaptation probably does occur, it is not instantaneous and will be detected by an indicator series which utilizes fairly frequent measurements" (pp.128-129).

Because the questionnaire included respondents' perceived changes (subjective indicators) and numbers of specific life events experienced (objective indicators), Atkinson was able to compare levels of association between each of these measures and satisfaction measures. He found that

> "The hypothesis that the No Change group would have more stable QOL scores than those reporting change is supported when the perceptual indicator of change is used as the independent variable but not when the event measures are involved. . .[As well] Relationships between self-reported change and satisfaction levels are always higher than between events and changes in satisfaction" (pp.122, 126).

This phenomenon is especially interesting to people working with discrepancy theories (e.g., Michalos 1985) because it has been found that perceived discrepancies between what one has and wants, for example, are better predictors of satisfaction than researcher-calculated discrepancies (i.e., difference scores between things actually possessed and wanted), e.g., see Oliver and Bearden (1985), Wright (1985), and Michalos (1991). Rice, McFarlin and Bennett (1989) compared the predictive power of perceived discrepancy scores with calculated discrepancy scores for job satisfaction and concluded

> "It appears that the components of a discrepancy are not capable of totally capturing the psychological comparison process represented by have-want discrepancies. Rather, it seems that have-want discrepancies have a power to predict and explain facet satisfaction that goes above and beyond the predictive power of the two components" (p.597).

Appealing to the different ways to measure changes in people's lives (perceived changes versus numbers of significant life events), Atkinson thought that the often observed low correlations between objective and subjective indicators might be the result of the fact that "the objective event is too gross a measure to lead to a prediction of its effects". Furthermore, he noted that the long- and short-term consequences of events may be very different. For example,

> "Divorce and separation appear to have neutral or negative short-term impacts followed by a positive long-term consequence. Most analyses of the impact of changing conditions have not considered the possibility of time dependent effects and have assumed that the impact of change is constant, or at least in the same direction, over time" (pp.129-130).

Most have also assumed that the impact of certain kinds of events is the same for all individuals, which is contrary to what we know about "individual differences in values and judgement standards".

Headey, Veenhoven and Wearing (1991) reported results from four waves of an Australian panel study conducted in the state of Victoria, involving samples ranging from 942 in 1981 to 649 in 1987. In the original linear model employed by Andrews and Withey (1974) and many others, global well-being measured by life satisfaction, happiness or, more broadly, subjective well-being is explained by (regressed on) measures of satisfaction in several domains, e.g., satisfaction with one's family life, job, housing and health. Diener (1984) called this the Bottom-Up (BU) model. However, because all of the key domain satisfaction variables are highly intercorrelated, some people have suspected that the original model might have the causal story completely backwards. They think that global well-being is like a personality trait or stable disposition that influences people's individual assessments of every domain of their lives. Diener called this the Top-Down (TD) model. The paper by Veenhoven (1994) in this collection provides a thorough review of all the evidence for and against the TD model. Besides these two possibilities, there is a Bi-Directional (BD) model that posits the causal arrows running in both directions (Mallard, Lance and Michalos, 1997), and a Spurious Causation (SC) model that posits some other variable(s) as the cause of the observed correlations among domain and global variables (Costa and McCrae, 1980).

Headey, Veenhoven and Wearing provided some evidence for the top-down (TD), two-way (BD) and spurious causation (SC) models. Their global dependent variable was a 9-point version of Life #3 from Andrews and Withey (1994), and their 9-point domain satisfaction variables were for the domains of marriage, work, standard of living, leisure, friendship and health. They had two personality variables, extraversion and neuroticism, from Eysenck and Eysenck (1964), and gross family income, main breadwinner's occupational status, respondent's educational level, age and sex. Like Andrews, Atkinson and others, the software employed was a version of LISREL (Joreskog and Sorbom, 1978).

For the domain of marriage, they found two-way causation.

> "Being happily married increases one's life satisfaction (*BU* = 0.18), but it is also true that happy people are more likely to maintain happy marriages. . .while miserable people tend to have miserable marriages (*TD* = 0.12). . .Costa and McCrae's (1980) hypothesis that stable personal characteristics account for apparent causal relationships between domain satisfactions and life satisfaction appears incorrect in regard to the marriage domain" (pp.90-91).

Regarding the domains of work and standard of living, they found only a TD model was confirmed. Part of the relationship between leisure and life satisfaction was from the latter to the former (TD), and part of it was spurious (SC). The relationship between friendship and life satisfaction was spurious (SC), resulting from the influence of extraversion on both variables, while the relationship between health and life satisfaction was also spurious (SC), driven by both extraversion and neuroticism.

These interesting results are difficult to square with the conclusions of Kammann, Farry and Herbison (1984). If those authors were right in their