The Spanish Journal of Psychology
2001, Vol. 4, No. 2, 111-122

1138-7416

The Decade 1989-1998 in Spanish Psychology: An Analysis of Research in Statistics, Methodology, and Psychometric Theory

Miguel A. García-Pérez
Universidad Complutense de Madrid

This paper presents an analysis of research published in the decade 1989–1998 by Spanish faculty members in the areas of statistical methods, research methodology, and psychometric theory. Database search and direct correspondence with faculty members in Departments of Methodology across Spain rendered a list of 193 papers published in these broad areas by 82 faculty members. These and other faculty members had actually published 931 papers over the decade of analysis, but 738 of them addressed topics not appropriate for description in this report. Classification and analysis of these 193 papers revealed topics that have attracted the most interest (psychophysics, item response theory, analysis of variance, sequential analysis, and meta-analysis) as well as other topics that have received less attention (scaling, factor analysis, time series, and structural models). A significant number of papers also dealt with various methodological issues (software, algorithms, instrumentation, and techniques). A substantial part of this report is devoted to describing the issues addressed across these 193 papers–most of which are written in the Spanish language and published in Spanish journals–and some representative references are given.

En este artículo se presenta un análisis de los trabajos de investigación publicados durante la década 1989–1998 por profesores numerarios españoles en las áreas de métodos estadísticos, metodología de investigación y psicometría. La búsqueda en bases de datos y la correspondencia directa con profesores del área de Metodología de las Ciencias del Comportamiento dio como resultado una lista de 193 artículos publicados por 82 profesores. Éstos y otros profesores del área han publicado en realidad 931 artículos durante la década objeto de análisis, pero 738 de ellos abordaban materias que no encajan con lo analizado en el presente trabajo. La clasificación y análisis de estos 193 artículos reveló una serie de temas que se han abordado profusamente (psicofísica, teoría de respuesta a los ítems, análisis de varianza, análisis secuencial y meta-análisis) así como otros que han recibido una menor atención (escalamiento, análisis factorial, series temporales y modelos estructurales). Un número importante de artículos ha abordado problemas metodológicos (software, algoritmos, instrumentación y técnicas experimentales). La mayor parte del presente artículo está dedicada a describir los asuntos abordados en estos 193 artículos–la mayoría de los cuales están escritos en español y publicados en revistas españolas–y se citan algunos artículos representativos.

I thank all the colleagues who have answered the request to return a complete list of their publications. Special thanks to Pere J. Ferrando, Sofía Fontes, Ana Garriga-Trillo, José Muñiz, Vicente Ponsoda, Vicenç Quera, Julio Sánchez-Meca and Guillermo Vallejo for their cooperation during the writing of the description of their work. I also thank David Clark-Carter, Edgar Erdfelder, Juan Fernández, Ulrich von Hecker, Jürgen Heller and Sandy MacRae for their comments on an earlier draft of this paper.

Address correspondence to: Miguel A. García-Pérez, Departamento de Metodología, Facultad de Psicología, Universidad Complutense, Campus de Somosaguas, 28223 Madrid (Spain). E-mail: miguel@psi.ucm.es

This report covers the areas of behavioral research methods, statistical methods, mathematical psychology, and psychometric theory. Issues addressed in the papers included in this analysis fall into the categories listed as appropriate for presentation of materials at meetings covering these topics (see Table 1 for a list), which find outlets in journals such as *Applied Measurement in Education*, *Applied Psychological Measurement*, *Behavior Research Methods, Instruments, & Computers*, the *British Journal of Mathematical and Statistical Psychology*, *Educational and Psychological Measurement*, the *Journal of Educational Measurement*, the *Journal of Mathematical Psychology*, *Multivariate Behavioral Research*, *Psychological Methods*, and *Psychometrika* among others. Thus, this analysis covers material similar in content to what is often published in those journals, although only a small amount of this material was indeed published in them (see below).

Noticeably missing in Table 1 are the core topics of mathematical psychology: preference behavior, probabilistic choice behavior, fundamental measurement, reaction times, and models thereof. One reason for this absence is that a certain number of papers testing or proposing mathematical models have been distributed for analysis elsewhere in this issue (see below), where they belong more properly because of the processes being modeled. Thus, research involving mathematical models of psychological processes (attention, perception, learning, memory, etc.) is dealt with by Igoa (this issue), attesting to the slow but continuous transfer of the quantitative and theoretical approach of mathematical psychology to all fields of cognitive and experimental psychology (Batchelder & Riefer, 1999; Luce, 1997; Ratcliff, 1998). We have nonetheless kept for analysis here a subset of theoretical and empirical work in psychophysics and psychometric theory. A second reason for the absences in Table 1 is a little more distressing: a mere lack of research in those areas of mathematical psychology.

In line with the nature of this special issue, only papers published by Spain-based scholars were included for analysis, a decision that should not be misconstrued in this smoldering world of growing nationalisms. Indeed, topics of scientific interest do not know about political or administrative borders (except for very regional research; consider issues involving social minorities or cultural idiosyncrasies), and it thus seems unreasonable to break scientific research into separate categories according to the nationality of the contributing authors. Yet, this report is not a topical review. In addition, national policies on scientific research act as a catalyst (or deterrent) that may explain differences in scientific productivity across countries; thus, national analyses serve as indirect indicators for a cross-national comparison of the effects of these policies. On another count, publication of research in non-standard languages (i.e., non-English) implies inaccessibility of the results to the worldwide community, either by lack of familiarity with the language itself or by limited availability of the journals where the research is published.

Table 1
Research Areas Covered in this Report

Sensory and Cognitive Psychophysics
Psychometric Theory and Applications
– Item Response Theory Models
– Performance of Parameter Estimation Methods
– Adaptive and Self-Adapted Testing
– Differential Item Functioning
Quantitative and Statistical Methods
– Analysis of Variance
– Sequential Analysis
– Meta-Analysis
– Time series
– Scaling
– Factor Analysis
– Structural Models
Software and Algorithms
Instrumentation and Techniques

Our analysis thus aims at describing the research that has been carried out by Spanish scholars in the decade 1989–1998, research of limited accessibility because it has been published mostly in Spanish journals and in the Spanish language (see below), despite the fact that the topics that were addressed are indeed of general interest.

A Brief Note on the Method

Our method is described in detail by Fernández (this issue). Briefly, database search (PsycLIT, ERIC, MEDLINE) was used to look for papers published between 1989 and 1998 (inclusive) by each of the 154 faculty members in Departments of Methodology in all 22 eligible Spanish Universities.[1] An initial, naïve search based on each faculty member's first last

[1] This includes 20 universities offering a degree in Psychology (out of 22 universities offering it; two of them do not have faculty members in the Department of Methodology) and two universities not offering degrees in Psychology but having faculty members in Departments of Methodology.

name[2] revealed mis-spellings and inaccuracies in the compilation of these databases that are reminiscent of well-known citation errors (Brown, 1999; Kotiaho, 1999; Kotiaho, Tomkins, & Simmons, 1999; Price, 1998). This anomaly prompted further exhaustive searches based on each faculty member's second last name and combination of last names.

Once all possible variants of search appeared exhausted and a seemingly satisfactory list of references was collected, each individual faculty member was approached directly by mail to ask for their confirmation of each of the items on their list, and possibly to supply other references that the database search did not yield. None of these letters returned undeliverable. Sixty-nine (44.8%) of the faculty members (from 20 of the 22 universities) replied, and a final by-author list of 1258 papers resulted. Yet, many entries were listed multiple times (once under each qualifying co-author[3]).

Each entry in this list was then checked for compliance with the requirement that the paper describes research in areas within the broad topics covered in this analysis. As a result, 869 entries (69%) were deemed inappropriate,[4] but they were distributed for analysis by the authors of other reports in this issue.[5] Also, 88 of the remaining entries (62 unique) did not meet eligibility requirements described by Fernández (this issue), and they were discarded too. Altogether, this screening process rendered a final by-author list of 301 entries (including multiple occurrences of co-authored papers) for analysis and classification.

In our subsequent analysis of these papers, no attempt was made to judge the relevance of the issues under study, or the significance of the contributions. The journals in which those papers were published have a review process that should guarantee conformity to minimal quality standards, and we simply relied on the evaluation carried out by the journals to accept papers for publication.[6] Then, inclusion of papers for analysis in the present report should not be misunderstood as implying that the papers succeeded in passing any form of quality assessment by the present author. Likewise, no paper was excluded for a presumed failure to pass it.

Descriptive Analysis of Published Research

Removal of multiple occurrences across the 301 entries in the final by-author list yielded an overall figure of 193 unique papers and 82 different faculty members. To arrive at these figures, each individual paper was counted just once (whether or not it was co-authored) and each qualifying (co-) author was also counted once (whether or not other faculty members or non-faculty co-authored the same papers). This final number of publishing faculty members is relatively small compared to 154 members in Departments of Methodology across Spain, but this does not mean that faculty members were unproductive over the period of this analysis: many of them just carry out their research in other areas of psychology (see footnote 4), and that research is analyzed elsewhere in this issue. Also, the reduced final number of 193 individual papers (out of 301 when these were listed under each of the qualifying faculty members) reveals a substantial amount of cooperation resulting in co-authored papers.

For a first glance at this research, Figure 1a shows the number of faculty members (regardless of other co-authors) who have published various numbers of papers. Note that more than half of the faculty members (46 of 82) have published only one or two papers in their nominal field, indicating a somewhat scarce group interest in developing the field. (It should be kept in mind that most of these faculty members have research interests in areas of psychology that are analyzed elsewhere in this issue.) Figure 1b shows a histogram of the number of authors per paper (regardless of whether co-authors were non-faculty or faculty members in these departments, in other departments, or in foreign institutions). The number of papers with two or three authors represents about 73% of the total number of papers published in this period, indicating a healthy degree of cooperation that also involves foreign colleagues: 10% of the co-authored papers (16 of 159) included at least one co-author affiliated with a non-Spanish institution.

---

[2] Unlike other nationals, Spaniards bear two last names: the first being our father's first last name and the second being our mother's first last name (legislation was recently passed that allows spouses to fight over the order in which their children will bear these two last names). As authors of journal articles, some of us connect our two last names with a dash so the Gestalt looks like a single last name to those expecting everyone to have just one; others omit their second last name altogether; yet others write them both without connection, and their papers usually end up indexed in databases under their second last name.

[3] The term "qualifying co-author" refers to the 154 faculty members in Departments of Methodology. Co-authors not affiliated with these departments did not have an individual section in this by-author list.

[4] The break-up of these 869 entries is as follows: 364 (267 unique) corresponded to research in cognitive psychology; 211 (161 unique) to research in social and organizational psychology; 172 (142 unique) to personality and clinical psychology; 83 (70 unique) to educational and developmental psychology; and 39 (36 unique) to physiological and biological psychology.

[5] As a result of this exchange, an additional section in our by-author list was opened for a faculty member in another department who nevertheless published research in the area of our analysis without cooperation with faculty members in Departments of Methodology.

[6] Seven book chapters and a book are also included in the set of papers that we will next analyze. It may be argued that these items do not go through the same process of quality assessment as journal articles, but their inclusion here will hardly bias our results.

Figure 1c finally shows the number of papers published in each of the years covered in this analysis. Besides a relatively small number of papers in the earlier years, the overall rate of publication is fairly stable at slightly above 20 papers per year since 1992.

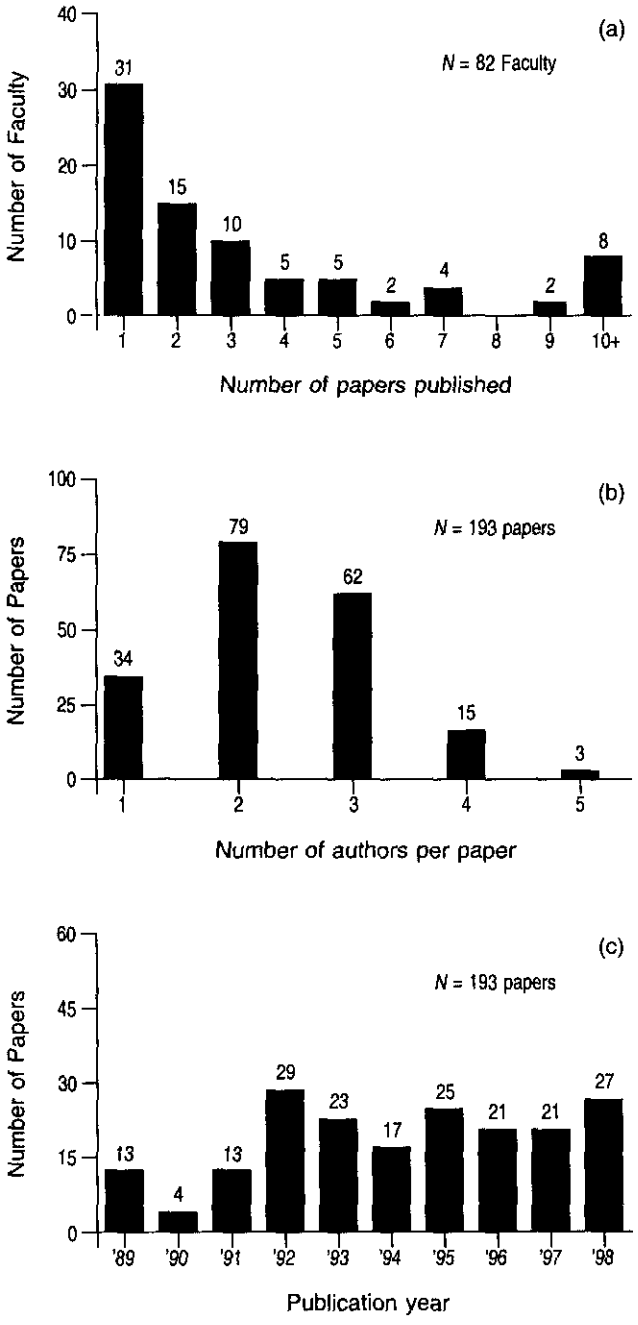

Figure 1. (a) Number of qualifying faculty members who have published different numbers of papers in the areas covered in this report (see Table 1), whether the papers were single- or multiple-authored. (b) Distribution of the number of authors per paper, whether co-authors in multiple-authored papers were other faculty members, non-faculty, or foreign colleagues. (c) Number of papers published in each of the years covered in this analysis.

Table 2
Outlet and Language of 193 Papers

Outlet (language)	No. of papers (%)
International Journals (English) [a]	49 (25.4%)
Spanish Journals (English)	3 (1.5%)
Spanish Journals (Spanish)	141 (73.1%)

[a] The term "international journal" refers to scientific journals (a) published by major publishers, (b) carried by major distribution agents, (c) having a supra-national editorial board, (d) with worldwide contributors, and (e) having a worldwide readership which, in turn, implies that most (if not all) published papers are written in English.

Table 2 summarizes the outlet and accessibility of this research by indicating the numbers of papers published in international versus Spanish journals and, within the latter, written in English versus Spanish. It stands right out that the widespread practice is for Spanish authors to write their manuscripts in Spanish and submit them for publication by Spanish journals. The obvious consequence of this practice is that most of this research remains hidden from view of the worldwide community.

Table 3 shows the distribution of papers by journal and area, differentiating Spanish and international titles. Only journals in which five or more papers were published across the entire list of areas have a separate entry in Table 3; all other journals have been aggregated in the "other" categories (which, again, exist separately under the Spanish and the international journal listings).

The bulk of the research published in Spanish journals (97 of 144 papers; 67%) has come out in just two journals of broad scope (*Psicológica* and *Psicothema*), whereas research published in international journals appears more evenly distributed across more specialized journals. Note also that the distribution of research in each of the areas across international journals is uneven, something that merely reflects the defining area of coverage of each journal.

Excluding topics within the two areas listed at the bottom of Table 3 (software, algorithms, instrumentation, and techniques), the approach taken to address specific topics in each of the three remaining areas can be theoretical/analytical (focusing on theoretical issues or analytical developments), empirical (with recourse to empirical data from actual subjects), or by simulation (generating artificial data to address theoretical or practical issues that may not lend themselves to analytical methods or empirical research). Table 4 shows the number of papers in each of these three areas that relied primarily on each of the three approaches. Quite clearly, research in sensory and cognitive psychophysics has exclusively been empirical, whereas research in quantitative and statistical methods has

Table 3
Distribution of Papers by Journal and Area

	Spanish Journals							International Journals					
Area	Psicológica	Psicothema	RPGA [a]	Anales de Psicología	Revista de Psicología	Investigaciones Psicológicas [b]	Other [c]	Ed. & Psych. Measurement	BRMIC [d]	Quality & Quantity	BJMSP [e]	Other [f]	Total
Sensory and Cognitive Psychophysics	1	3	6				1					2	13
Psychometric Theory and Applications	7	11	2		2	2	5	4			3	9	45
Quantitative and Statistical Methods	25	16	1	7	1	1	3	1	1	3	2	5	66
Software and Algorithms	11	13			3	2	4	3	3			2	41
Instrumentation and Techniques	6	4	1	1			5		4	3		4	28
Total	50	47	10	8	6	5	18	8	8	6	5	22	193

[a] *Revista de Psicología General y Aplicada.*

[b] Discontinued in 1994.

[c] These 18 papers came out in 9 different journals (*Análisis y Modificación de Conducta, Anuario de Psicología, Cuadernos de Bioestadística y sus Aplicaciones Informáticas, Estudios de Psicología, Revista de Historia de la Psicología, Revista de Informática y Automática, Revista de Investigación Educativa, Revista Latinoamericana de Psicología,* and *Revista Mexicana de Análisis de la Conducta*).

[d] *Behavior Research Methods, Instruments, & Computers.*

[e] *British Journal of Mathematical and Statistical Psychology.*

[f] Includes 7 book chapters, 1 book, and 14 papers in 11 different journals (*Applied Psychological Measurement, Applied Statistics, European Journal of Psychological Assessment, Journal of Educational Measurement, Journal of Experimental Education, Journal of Neuroscience Methods, Multivariate Behavioral Research, Psychological Bulletin, Psychological Methods, Psychological Reports,* and *Vision Research*).

mainly used the simulation approach; on the other hand, research in psychometric theory and applications has mostly used empirical and simulation approaches, although empirical studies seem to prevail. All areas combined, theoretical or analytical developments have been relatively rare, whereas the number of empirical and simulation studies is balanced and represents about 85% of the total number of papers across the three areas.

Specific Areas of Research

This section describes the major topics addressed across the set of papers within each of the broad areas listed in Table 3. Since we cannot possibly mention each and every paper (or contributing author), only a few papers will explicitly be referred to, those which more broadly describe salient contributions. This selection of papers was primarily made in accordance with the general criteria described by Fernández (this issue), namely, capitalizing on the work of individual authors who (independently or in cooperation with others) have published five or more papers on a specific topic over the period covered in this analysis. Although the analysis itself covered only the decade 1989–1998, some references are given here to papers published later if these are more comprehensive, afford a better perspective, or provide more pointers to related literature.

Table 4
Number of Papers (and Percent Within Each Area) Using Each Approach

	Approach			
Area	Theoretical/Analytical	Empirical	Simulation	Total
Sensory and Cognitive Psychophysics		13 (100)		13
Psychometric Theory and Applications	6 (13.3)	22 (48.9)	17 (37.8)	45
Quantitative and Statistical Methods	13 (19.7)	18 (27.3)	35 (53.0)	66
Total	19 (15.3)	53 (42.7)	52 (41.9)	124

However, the above criteria would leave out a significant amount of research. Indeed, there are areas in which no single individual has published five or more papers, and yet the overall number of papers published on those topics is well in excess of five. It seems unreasonable to overlook such research areas on grounds of the absence of an identifiable leader and, therefore, we will also describe areas of miscellaneous authorship under appropriately marked headings. In this case, papers are cited which, in the opinion of the present author, may be more useful for the interested reader to get a flavor of the status of this research area in Spain. Also, everything else equal, we have chosen to cite papers in Spanish journals because papers in international journals will be easier to track down by the interested reader.

Finally, topics addressed in fewer than five papers (across the entire set of 193) will simply be enumerated for completeness, but no references will be given.

Sensory and Cognitive Psychophysics

Sensory psychophysics is believed to mark the origin of mathematical psychology, and its empirical approach is certainly the origin of modern experimental psychology (Fechner, 1987; Scheerer, 1987). Despite its century-and-a-half history, sensory psychophysics continues to be an active research area that gathers scholars at Fechner Day celebrations yearly. One empirical way to approach the issues involved in sensory scaling is to find out the functional form of the psychophysical law describing the relationship between the physical magnitude of some stimulus and its subjective magnitude as reported by human observers. The traditional dispute over this issue pertains to the universality of the psychophysical law: whether a single functional relationship holds for all subjects and stimuli regardless of the empirical method that is used to obtain the subjective estimates and also regardless of contextual effects. Fontes, Garriga, and Barbero (1993) used a magnitude estimation task to gather data on the subjective distance between two vertical lines in order to compare the fit of linear, power (Stevens), and logarithmic (Fechner) laws. Subsequently, Fontes, Barbero, and Fontes (1994) studied whether the range of magnitudes in the stimulus set affects the fit of these various laws.

Sensitivity measures can be obtained with a variety of empirical methods, including cross-modality matching, magnitude estimation, or discrimination tasks (e.g., the triangular method) among others. Garriga-Trillo (1992) used regression analysis to determine whether magnitude estimates or cross-modality matches are more related to actual physical measures of the stimuli. Each of the empirical methods used to gather psychophysical data further involves a different balance of pure sensory processes and cognitive components, and the interplay of these two factors might account for some differences found across studies, for example, as to the confidence expressed by experimental subjects on the quality of their own judgements. Garriga Trillo, Villarino, González Labra, and Arnau (1994) proposed an indirect index that would allow the calibration of psychophysical judgements obtained in magnitude estimation tasks, and they also studied its behavior for the empirical assessment of confidence from magnitude estimation data.

Psychometric Theory and Applications

A small subset of the research in this area has dealt with the empirical comparison of test properties under classical test theory versus item response theory (IRT), with the assessment of unidimensionality under IRT, or with the IRT parameterization of conventional aptitude and personality tests. A significantly larger amount of work has been devoted to four specific IRT areas that are described next.

IRT Models

Item response functions (IRFs) are the essential building block of IRT. An IRF specifies the probability that an examinee will give the correct answer to a multiple-choice item, as a function of examinee and item parameters. Current IRT applications are almost exclusively based on logistic IRFs. García-Pérez and Frary (1991) elaborated on a finite state theory of performance in multiple-choice tests that gives rise to a distinctly new set of IRFs, all of which turn up having the mathematical form of a polynomial. Finite state polynomic IRFs arise naturally when test taking is considered within the context of multinomial process tree models (see Batchelder & Riefer, 1999), and it directly incorporates into the mathematical form of the IRF such characteristics as the examinees' guessing strategies, the number of options per item, the relative identifiability of correct answers versus distractors, the format of administration of the test, and other item characteristics such as use of "none of the above" as an option. Besides IRFs, finite state theory provides expressions for the probability of each of the response outcomes (not only correct/incorrect) that may arise under any format of administration of a multiple-choice item (e.g., answer-until-correct). Thus, under finite state theory, an IRF coexists with a number of other functions each expressing the probability of one of the remaining response outcomes as a function of examinee and item parameters and characteristics. Over the years, work with this model has consisted of testing it against several sets of empirical data, developing and studying the properties of goodness-of-fit and parameter estimation methods, and comparing the theoretical psychometric properties of diverse item formats (see García-Pérez, 1999).

The use of conventional (i.e., logistic) IRFs initially implies the assumption that item responses are dichotomous (correct/incorrect), but further theoretical developments have allowed the use of IRT methods when item responses are still discrete but polytomous, or when they are continuous (but possibly discretized upon recording). The latter occurs in personality and attitude inventories consisting of Likert-type items, whose response options define an ordered set of categories revealing the strength of an underlying continuous trait. One of the approaches to deal with these items is by recourse to the linear factor analysis model. Ferrando (1996) proposed an extension to the factor-analytic, continuous item response model that allows for item calibration and multi-group analyses towards the assessment of parameter invariance. Ferrando (1999) further compared the characteristics of linear (factor-analytic) and non-linear (IRT) continuous models applied to actual responses to Likert-type items, using criteria such as goodness of fit, item and subject parameter estimates and criterion-related validity.

Performance of Parameter Estimation Methods

Practical application of IRT requires estimation of the parameters that best describe each individual item in a test, given a convenient IRF that is often chosen beforehand. A large number of parameter estimation methods have been developed over the past three decades (mostly for use with logistic IRFs; see Baker, 1987), and computer software implementing these methods is commercially available. Besides the structural assumption of a mathematical form for the IRF, all of these parameter estimation methods make strong assumptions about the structure of the data, most notably the unidimensionality of the examinee parameter space (i.e., the assumption that performance on each item in the test depends on a single examinee trait), the dimensions of the item parameter space (in the case of logistic models, whether the target IRFs should include only one or up to four distinct parameters), and the absence of item responses that logistic IRFs cannot accommodate (e.g., omissions). A major concern in the application of IRT parameter estimation methods is how violation of the characteristics assumed during parameter estimation affect the performance of the algorithms and, therefore, the extent to which estimation methods are insensitive to these violations. Muñiz, Rogers, and Swaminathan (1989) studied the capability of the Rasch model to accurately estimate item difficulties and examinee abilities (the two only parameters that are estimated under the Rasch model) when the data were generated using the three-parameter logistic model. In a similar vein, Cuesta and Muñiz (1995) studied the effects of trait multidimensionality on ability and item parameter estimates obtained through methods which assume that item responses depend on a single trait.

Adaptive and Self-Adapted Testing

The invariance of item and examinee parameters under IRT is the basis for computerized adaptive testing (CAT), whereby each examinee's ability is measured with a tailored (and possibly unique) set of items that are chosen on-line along the testing process in order to obtain ability estimates with the highest possible precision at the lowest possible cost. Adaptive testing requires that a calibrated item pool exists to choose items from, and it also demands the use of computers to carry out the heavy on-line computation that the item selection process requires. Without control of item exposure, CAT may wind up administering some items in the pool much more often than others across the set of examinees. Exposure control methods aim at preventing this evil without compromising the precision of CAT ability estimates. Revuelta and Ponsoda (1998) proposed two new exposure control methods and compared their performance with that of previously existing methods.

Self-adapted testing (SAT) is a variant of CAT in which the difficulty of the item to be administered next is chosen by the examinees themselves, instead of being determined by a suitable item selection algorithm. This practice may not be psychometrically optimal, but it may solve some motivational and anxiety issues that CAT seems to generate. Ponsoda, Olea, Rodríguez, and Revuelta (1999) carried out an empirical study comparing CAT and SAT as to their psychometric properties (the characteristics of ability estimates obtained with either method) and their psychological effects (anxiety caused by either method).

Differential Item Functioning (Miscellaneous Authorship)

Although item parameters are presumed invariant under IRT, there is empirical evidence that individuals with the same ability but different group membership (e.g., culture, gender, etc.) do not have the same probability of responding correctly to specific items, as if the parameters describing the IRF of those items varied across groups. These items were initially referred to as "biased," but current terminology dubs them DIF (for *Differential Item Functioning*; see the discussion in Angoff, 1993, pp. 3-5). Because of the social and legal issues that DIF raises, research on statistical methods for its detection has sprouted considerably in the last decades. In this tradition, Gómez and Navas (1996) devised a stepwise method for the detection of DIF, and Hidalgo Montesinos and López Pina (1997) compared the performance of several methods for the detection of DIF.

Quantitative and Statistical Methods

A small number of papers in this area addressed a variety of issues including analyses of the statistical power in research published in several journals, methods for the analysis of reaction times, simulation studies of empirical

Type I and Type II error rates under violations of the assumptions of miscellaneous statistical tests, or simulation studies on the sampling distribution of test statistics for which analytical results are lacking. Some issues in seven other areas have received more thorough attention, as described next.

Analysis of Variance

Analysis of variance (ANOVA) is perhaps the single statistical method most frequently used in all areas of experimental psychology. Like all parametric methods, ANOVA was set up under some restrictive distributional assumptions that empirical data do not always meet. This raises concerns about the appropriateness and robustness of ANOVA, that is, the extent to which its application can yield dependable conclusions when the data violate these assumptions. ANOVA is only a general term referring to a quite diverse set of methods each of which assumes that the data satisfy a specific set of constraints, from the relatively simple case of fixed-effects, balanced designs with a single between-group factor, to the more sophisticated unbalanced, incomplete, and/or multivariate designs with random effects and including a number of between-group and within-group (repeated-measures) factors. Therefore, any investigation into the appropriateness of ANOVA when the data violate its defining assumptions must necessarily be limited in scope to some specific version of this general procedure.

In repeated-measures designs, where each experimental unit provides multiple responses along the levels of the trial factor, ANOVA requires that the data meet the assumption of sphericity: all levels of the trial factor have the same variance and all pairs of levels have the same correlation. Yet, alternative ANOVA models have been devised that allow other structures on the covariance matrix of the repeated measures. Sphericity is a strong assumption in designs in which the treatments along the trial factor are likely to introduce serial dependence into what otherwise should be random error with identical distribution. Fernández and Vallejo (1997) carried out a simulation study to compare the results of multivariate ANOVA on data with these characteristics with the alternative strategy of using univariate ANOVA with the error structure modeled through a first order auto-regressive process.

Rejection of any of the null hypotheses tested through ANOVA prompts the use of multiple comparison procedures that test for pairwise differences between means, and an analysis of their performance under violation of their defining assumptions also seems appropriate. Vallejo and Menéndez (1998) carried out a simulation study in which the empirical Type I and Type II error rates of six multiple comparison procedures were examined for correlated data in between-group one-way ANOVA designs, as a function of sample size and the pattern of departure from the null hypothesis.

Sequential Analysis

Repeated measures may involve categorical variables that are not appropriate for ANOVA. Sequential analysis aims at uncovering temporal patterns in these sequences of categorical data. If the experimental hypotheses involve processes or if researchers are interested in the interaction between participants, observing them systematically and representing their behavior as it unfolds over time seems the logical approach. Bakeman and Quera (1995a) proposed the Sequential Data Interchange Standard (SDIS), a standard for classifying such sequential data, and a syntax for representing them in computer files for analysis. They also developed the General Sequential Querier (GSEQ), general-purpose software for analyzing sequential data in SDIS format. SDIS can represent a variety of sequential data, from simple, non-concurrent event sequences to complex, concurrent, timed event sequences. GSEQ can perform event lag-sequential analyses, and concurrent analysis of time windows anchored to specific behaviors, depending on whether hypotheses about sequential or synchronicity patterns must be tested. More sophisticated analyses are possible when log-linear models are applied to multidimensional lag-sequential tables, and when winnowing techniques are used for detecting main significant residuals in those tables (see Bakeman & Quera, 1995b).

Meta-Analysis

Meta-analysis is now a well-established method for the quantitative integration of results obtained across independent empirical studies on the same topic. Quite often, meta-analysis is used as a tool to approximate a larger sample size than used in any of the independent studies that are thus integrated, and its broad goal is obtaining a more accurate estimate of effect size. A fundamental preliminary step in the overall procedure is the statistical assessment of homogeneity of results across studies, something that justifies their integration on the (plausible) assumption that variations across studies merely reflect sampling error. If homogeneity appears untenable, the usual strategy is to test the hypothesis that some moderator variables (which must be identified) explain the heterogeneity of effect sizes found across studies. Several statistical approaches can be used to test this hypothesis, and Sánchez-Meca and Marín-Martínez (1998) carried out a simulation study to compare three of them as to their bias, efficiency, and Type I and Type II error rates as a function of such factors as the number of independent studies involved, the sample size in each of them, and the distribution of effect size.

One other problem in meta-analysis is how to summarize the results of an empirical study that includes several dependent variables all of which are believed to be indicators of a single construct. One approach consists of averaging effect sizes separately calculated on each variable, although

the averaging can be done in various ways. Marín-Martínez and Sánchez-Meca (1999) evaluated theoretical and empirical differences among various statistical approaches to carry out this averaging.

Time Series (Miscellaneous Authorship)

Clinical and behavioral designs often involve data recorded over long periods of time, usually with an interleaved interruption corresponding to the application of some treatment whose effectiveness must be assessed. These data are then subjected to statistical methods of time series analysis. The use of spectral methods requires the data to satisfy the assumption of stationarity, something that is untenable in many cases. Time-domain methods are free of this requirement, but they make specific assumptions about the serial dependence and trend in a series, and these assumptions may not hold for the data at hand. Moreover, application of these multi-stage methods requires comparatively larger amounts of data. Vallejo Seco (1994) studied the consequences of omitting the identification stage in the application of multi-stage methods, and Arnau and Bono (1998) compared two alternative approaches to deal with short time series.

Scaling (Miscellaneous Authorship)

Determining an appropriate metric and empirical procedure for the ranking of stimuli along psychological dimensions is an old problem whose solution has many practical ramifications for the measurement of psychological variables. Unidimensional scaling is the simplest approach, whereby the responses of subjects confronted with a suitable task are used to place each stimulus in the experimental set at a specific location along a single, continuous underlying dimension. Sospedra, Molina, and Meliá (1994) proposed a new method for unidimensional scaling that they also compared with four long-existing alternatives, and Cañadas Osinski and Sánchez Bruno (1998) determined empirically the interval-scale values of linguistic quantifiers of frequency used in Likert-type items in Spanish.

Factor Analysis (Miscellaneous Authorship)

In its heyday, factor analysis (FA) was believed to be the method that would disclose the structure of human intelligence and, by extension, of all psychological aptitudes. Nowadays, FA is regarded as a general-purpose statistical tool with a confirmatory or exploratory rather than a theory-building role. As such, FA is subject to the same scrutiny and developments as other statistical methods, the more so when FA actually describes a general approach that can be implemented in many diverse ways which, in turn, may provide different factorial solutions for the same data. Most of the papers in this category compared FA methods as to the results that they produce (see Ferrando & Lorenzo, 1993; González-Romá, Hernández, & Ferreres, 1997; Oliver, Sancerni, Tomás, & Lis, 1995).

Structural Models (Miscellaneous Authorship)

Together with FA, structural equations (or covariance structure) models offer a theory-testing methodology for elucidating the structure underlying an array of data. These models formalize hypotheses about patterns of relations between a set of empirically measured variables and a set of unobservable, latent variables. Their practical use requires the obvious statistical elaboration and recourse to parameter-estimation methods and goodness-of-fit statistics, both of which make more or less strong distributional assumptions. This is, then, another area in which an assessment of the workings of these methods seems mandatory. Papers in this area described research in that direction (see Hernández & Ramírez, 1996; Hernández Cabrera, San Luis Costas, & Guardia Olmos, 1995; Oliver, Tomás, & Meliá, 1993).

Software and Algorithms

Strictly speaking, the heading of this section does not qualify as a research area. Yet, research in almost any area is increasingly dependent on software and algorithms that are not general-purpose items and, hence, are not developed commercially. Most of the papers in this category represent contributions from authors with a well-defined research interest, something that must have prompted them to make available the computational tools they developed for their work.

By its nature, this is an area of miscellaneous authorship. Even authors who have published five or more papers in this broad category have addressed quite different issues across the board. For this reason, a detailed description (with references) of all this work is not appropriate here. Instead, Table 5 gives a summary description of the specific areas where this software and algorithms are relevant. Broadly speaking, papers in this category present pieces of software for second-stage analyses not included in general-purpose statistical packages or in other application software that is commercially available (e.g., for IRT). The software described in these papers has characteristics similar to that in papers published in journals such as *Applied Psychological Measurement*, *Applied Statistics*, *Behavior Research Methods, Instruments, & Computers*, or *Educational and Psychological Measurement*, and 8 of these 41 papers (19.5%) were indeed published in those journals (see Table 3).

Table 5
Break-up of 41 Papers Describing Software and Algorithms

Topic	No. of papers
Item Response Theory	10
Factor Analysis	7
Unidimensional Scaling	3
Miscellaneous Statistical Software	9
Experimental Routines	6
Reference Analysis	3
Image Processing	3

Table 6
Break-up of 28 Papers Describing Instrumentation and Techniques

Topic	No. of papers
Statistical / Numerical Methods	14
Experimental Methods	9
Other	5

Instrumentation and Techniques

This area is similar to the previous one (including its miscellaneous authorship), but here the emphasis is on hardware and general methodology (excluding data analysis methods). The diversity of topics is larger here, and Table 6 gives the break-up of the 28 papers in this category into two major areas of application: statistical/numerical methods (including evaluation of software, comparative analyses of alternative computational procedures, studies on the performance of algorithms under limiting circumstances, comparative analyses of random number generators, etc.) and experimental methods (including instrumentation and experimental protocols or designs).

Discussion

This report has presented an analysis of research published by Spanish faculty members within the areas of behavior research methods, mathematical psychology, statistical methods, and psychometric theory. Considering that this analysis covers a full decade (1989–1998, inclusive) and that there are 154 faculty members in Departments of Methodology across Spain, the overall figure of 193 papers published by 82 faculty members seems a very small group contribution, although the analysis also revealed a much greater contribution from these 82 and the remaining faculty members to other areas of psychology described elsewhere in this issue.

The research presented in these 193 papers addressed topics that are of current interest worldwide, but about 73% of the papers (see Table 2) have been published in Spanish journals and in the Spanish language. This practice clearly puts this research in scientific isolation: the worldwide community is highly likely to remain unaware of research published in a non-standard language and in a barely accessible outlet. The validity of this claim is corroborated by a cited-reference search on the web version of the *Social Sciences Citation Index* (SSCI), which was performed on July 8, 2000, using the database reportedly updated on July 6, 2000. *Psicológica*, the Spanish journal with the largest publication record in the area and period of our analysis (see Table 3) is not listed in SSCI, and out of the 47 papers published in *Psicothema*, 20 (43%) came out as having been cited in 24 papers altogether. An analysis of these 24 citations revealed that all of them were self-references in other papers published by the same authors in *Psicothema* itself (21 cases), in other Spanish journals (1 case) or in international journals (2 cases).

Also, although most of the Spanish journals listed in Table 3 are indexed in international electronic databases as well as in *Psychological Abstracts*, hard copies hardly reach institutions in non-Spanish-speaking countries, and full-text electronic versions of some of these journals are only starting to be made available.

Luckily, the way out of the current state of isolation is very easy. At least in the areas covered in this analysis, a challenge for Spanish psychologists is to report to their worldwide peers by writing their manuscripts in English and submitting (at least some of) them to international journals for wider accessibility.

Postscript

Because this special issue was conceived long before it has been published, a follow-up of the situation in 1999 and 2000 seems appropriate. A similar database search was carried out which revealed 20 papers published in 1999 and 31 in 2000 on topics that are within the scope of this report, besides a larger number of papers describing research on topics covered elsewhere in this issue. These figures continue to support the claim of a publication rate slightly above 20 papers per year since 1992 (Figure 1c). The topics addressed in these 51 papers cover all of the areas described in this report, with a significant proportion on issues related to psychometric theory (21 of 51 papers; 41.2%). Of the 20 papers published in 1999, 11 came out in international journals, whereas 9 of the 31 papers published in 2000 came out in international journals. These figures imply that nearly 40% of the production during 1999 and 2000 reached an international audience, a percentage that is meaningfully larger than that during the decade 1989–1998 (25%; Table 2).

References

Angoff, W.H. (1993). Perspectives on differential item functioning methodology. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillsdale, NJ: Erlbaum.

Arnau, J., & Bono, R. (1998). Short time series analysis: *C* statistic vs Edgington model. *Quality and Quantity, 32*, 63-75.

Bakeman, R., & Quera, V. (1995a). *Analyzing interaction: Sequential analysis with SDIS and GSEQ*. New York: Cambridge University Press.

Bakeman, R., & Quera, V. (1995b). Log-linear approaches to lag-sequential analysis when consecutive codes may and cannot repeat. *Psychological Bulletin, 118*, 272-284.

Baker, F.B. (1987). Methodology review: Item parameter estimation under the one-, two-, and three-parameter logistic models. *Applied Psychological Measurement, 11*, 111–141.

Batchelder, W.H., & Riefer, D.M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review, 6*, 57-86.

Brown, N.L. (1999). On the trail of the prolific Dr Path. *Nature, 398*, 555.

Cañadas Osinski, I., & Sánchez Bruno, A. (1998). Categorías de respuesta en escalas tipo Likert. *Psicothema, 10*, 623-631.

Cuesta, M., & Muñiz, J. (1995). Efectos de la multidimensionalidad en la estimación de parámetros desde modelos unidimensionales de teoría de respuesta a los ítems. *Psicológica, 16*, 65-86.

Fechner, G.T. (1987). Outline of a new principle of mathematical psychology (1851). *Psychological Research, 49*, 203-207. (Translated and edited by Eckart Scheerer.)

Fernández, J. (this issue). Research trends in Spanish psychology (1989–1998).

Fernández, P., & Vallejo, G. (1997). Diseño de medidas repetidas con dependencia serial en el error. *Psicothema, 9*, 619-635.

Ferrando, P.J. (1996). Calibration of invariant item parameters in a continuous item response model using the extended LISREL measurement submodel. *Multivariate Behavioral Research, 31*, 419-439.

Ferrando, P.J. (1999). Likert scaling using continuous, censored and graded response models: Effects on criterion-related validity. *Applied Psychological Measurement, 23*, 161-175.

Ferrando, P.J., & Lorenzo, U. (1993). Relación entre las soluciones factoriales MINRES y P.A.F.: algunas consideraciones. *Revista de Psicología Universitas Tarraconensis, 15*, 7-14.

Fontes, S., Barbero, I., & Fontes, A.I. (1994). Efecto del rango del estímulo en la función de Stevens. *Revista de Psicología General y Aplicada, 47*, 253-257.

Fontes, S., Garriga, A.J., & Barbero, I. (1993). Funciones psicofísicas de la estimación de distancias entre dos rectas. *Revista de Psicología General y Aplicada, 46*, 23-32.

García-Pérez, M.A. (1999). Fitting logistic IRT models: Small wonder. *The Spanish Journal of Psychology, 2*, 74-94. [Available online at www.ucm.es/sjp]

García-Pérez, M.A., & Frary, R.B. (1991). Finite state polynomic item characteristic curves. *British Journal of Mathematical and Statistical Psychology, 44*, 45-73.

Garriga Trillo, A.J., Villarino, A., González Labra, M.J., & Arnau, M.A. (1994). La calibración de juicios psicofísicos: estimación de magnitudes. *Psicothema, 6*, 525-532.

Garriga-Trillo, A. (1992). How much of the physical continuum is explained by magnitude estimates and cross-modality matches? In G. Borg & G. Neely (Eds.), *Fechner day 92* (pp. 81-85). Stockholm: Stockholm University.

Gómez, J., & Navas, M.J. (1996). Detección del funcionamiento diferencial de los ítems mediante regresión logística: purificación paso a paso de la habilidad. *Psicológica, 17*, 397-411.

González-Romá, V., Hernández, A., & Ferreres, A. (1997). Análisis factorial confirmatorio de matrices multirrasgo-multimétodo: análisis y comparación de tres parametrizaciones. *Psicológica, 18*, 105-118.

Hernández, J.A., & Ramírez, G. (1996). Procedimiento bootstrap modificado para la evaluación de los índices de ajuste en el entorno de los modelos de estructura de covarianza. *Psicológica, 17*, 41-54.

Hernández Cabrera, J.A., San Luis Costas, C., & Guardia Olmos, J. (1995). Acerca de la robustez de los estimadores multinormales y elípticos bajo ciertas condiciones de asimetría, tamaño muestral y complejidad de los modelos de estructuras de covarianza. *Anales de Psicología, 11*, 203-217.

Hidalgo Montesinos, M.D., & López Pina, J.A. (1997). Comparación entre las medidas de área, el estadístico de Lord y el análisis de regresión logística en la evaluación del funcionamiento diferencial de los ítems. *Psicothema, 9*, 417-431.

Igoa, J.M. (this issue). The decade 1989–1998 in Spanish psychology: An analysis of research on basic psychological processes, history of psychology, and other related topics.

Kotiaho, J.S. (1999). Papers vanish in mis-citation black hole. *Nature, 398*, 19.

Kotiaho, J.S., Tomkins, J.L., & Simmons, L.W. (1999). Unfamiliar citations breed mistakes. *Nature, 400*, 307.

Luce, R.D. (1997). Several unresolved conceptual problems of mathematical psychology. *Journal of Mathematical Psychology, 41*, 79-87.

Marín-Martínez, F., & Sánchez-Meca, J. (1999). Averaging dependent effect sizes in meta-analysis: A cautionary note about procedures. *The Spanish Journal of Psychology, 2*, 32-38. [Available online at www.ucm.es/sjp].

Muñiz, J., Rogers, J., & Swaminathan, H. (1989). Robustez de las estimaciones del modelo de Rasch en presencia de aciertos al azar y discriminación variable de los ítems. *Anuario de Psicología, 43*, 81-97.

Oliver, A., Tomás, J.M., & Meliá, J.L. (1993). Análisis de los efectos del error de medida sobre las estimaciones en modelos de ecuaciones estructurales. *Psicológica, 14*, 293-306.

Oliver, A., Sancerni, M.D., Tomás, J.M., & Lis, R. (1995). Métodos de estimación y tamaños muestrales en análisis factorial confirmatorio: implicaciones en la validez factorial del GHQ. *Psicológica, 16*, 101-113.

Ponsoda, V., Olea, J., Rodríguez, M.S., & Revuelta, J. (1999). The effects of test difficulty manipulation in computerized adaptive testing and self-adapted testing. *Applied Measurement in Education, 12*, 167-184.

Price, N.C. (1998). What's in a name (or a number or a date)? *Nature, 395*, 538.

Ratcliff, R. (1998). The role of mathematical psychology in experimental psychology. *Australian Journal of Psychology, 50*, 129-130.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 311-327.

Sánchez-Meca, J., & Marín-Martínez, F. (1998). Testing continuous moderators in meta-analysis: A comparison of procedures. *British Journal of Mathematical and Statistical Psychology, 51*, 311-326.

Scheerer, E. (1987). The unknown Fechner. *Psychological Research, 49*, 197-202.

Sospedra, M.J., Molina, J.G., & Meliá, J.L. (1994). Estudio comparativo de cinco métodos de escalamiento unidimensional: ajuste y divergencia de los valores escalares. *Psicológica, 15*, 427-437.

Vallejo, G., & Menéndez, I. (1998). Efectos de la dependencia entre las observaciones en diversos procedimientos de comparación múltiple. *Psicológica, 19*, 53-71.

Vallejo Seco, G. (1994). Evaluación de los efectos de la intervención en diseños de series temporales en presencia de tendencias. *Psicothema, 6*, 503-524.