ELSEVIER

Acta Psychologica 92 (1996) 169–208

acta psychologica

# Numerical scaling techniques for evaluating generative models of orthographies [1]

Kimberly A. Jameson *

*Dept. of Psychology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0109, USA*

Received 9 November 1993; revised 23 March 1995; accepted 10 April 1995

## Abstract

Research presented introduces a new approach for empirically investigating generative or inductive-like systems. Two versions of a generative model of the uppercase English alphabet are examined. The approach is both formal and interdisciplinary, applying techniques developed by psychophysics to a problem that is typically linguistic. Experimental results presented illustrate empirical methods and analytic tools used, and demonstrate how the techniques advance the psychological study of the English alphabet. Two different theoretical models of the alphabetic system are evaluated strictly on the basis of empirically observed two-alternative forced-choice data. Scaling methods described produce a numerical scale for generated alphabetic items which permits informative comparisons between scale values and across independently derived scales. The scaling theory used is a variant of Thurstone's Case V model. The methods can also be utilized to further model construction. Implications of the findings for the existing body of writing-system research, and the generalizability of the approach to other domains of investigation are also discussed.

*PsycINFO classification:* 2100; 2300; 2340

*Keywords:* Semiotics; Symbol processing; Categorization; Scaling; Thurstonian models; Generative models; Inductive categories

* E-mail: kjameson@ucsd.edu.

[1] Some of the earlier analyses of this study were reported in Jameson (1994) and Jameson (1989). The 67 pseudoletters and the sample alphabet employed in the research are available as postscript typefaces for the Apple Macintosh. To obtain these typefaces e-mail the author for information.

*SSDI* 0001-6918(95)00012-7

## 1. Introduction

This paper presents new results in the formal study of symbolic information systems and cognition. The research is interdisciplinary in nature, applying empirical methods and analytical techniques from psychophysics to a linguistic model of the uppercase English alphabet.

The paper has two goals: First, to present new methods of data collection and analysis that provide stronger tests of generative models than are possible using techniques existing in the literature, and to illustrate how variation in methodology influences the descriptive power and generalizability of empirical findings. Second, empirical tests of a cognitive model of the uppercase English alphabet are presented, demonstrating how the new methods are employed and serving as a valuable substantive test of the model.

The present research is based on the work of immediate predecessors, but the aim is to go beyond typical experimental research on the nature and functions of generative cognitive models.

The approach employed here is easily generalized to testing theoretical models that have generative or inductive capabilities. Specifically, these methods are useful for investigating psychological processing in probabilistic categorization judgments, and for evaluating judgments for which correct answers may be unknown. The methods presented are interdisciplinary and are appropriate for use in many psychological applications.

Here these new methods are applied to investigate a generative model of the English uppercase letters which is similar in many respects to models commonly used to describe aspects of natural languages. This research demonstrates that, for the alphabetic model examined, the approach yields results that are detailed, informative and that can be used for model improvement.

The paper first gives a brief background of previous research on alphabetic models, discussing data collection and analysis methods. Second, is an overview of the tested alphabetic model, emphasizing possible ways it can be empirically tested. Third, rationale for the present methods is discussed, including the use of empirical judgments of acceptability, improvements the approach presents over existing methods, and advantages of generalizing these methods to examine other generative models. Fourth, data collection and analysis methods are described and the results from three empirical studies, and their bearing on the alphabetic model, are explained. The emphasis is on contrasting and comparing the effects of (a) variants of paradigm design, (b) different methods for deriving a Rating Structure from the data, and (c) calibration approaches for combining independent rating structures. Finally, empirical results are summarized and the usefulness of the approach in investigating other psychological domains is discussed.

## 2. A brief background of relevant writing systems research

Serious psychological experimentation using simple figures (letters and other designs of like complexity) began in the 1930's (e.g. Fehrer, 1935). Early work on typical errors that people make when trying to remember or perceive visual symbols is well illustrated by the work of Bartlett (1932) and by Carmichael et al. (1932), and later by the work of Bruner et al. (1952). Modern research into the analysis of pictorial symbols began with the well-known computational archaeologist J.C. Gardin (1958), who apparently was the first to analyze two-dimensional figures into their distinctive features.

Beginning in the early 1960's, strong interest in analyzing the letters of the alphabet as an organized symbol system became prevalent at various research centers in the United States and elsewhere. Two distinct but sometimes related paths can be distinguished: (a) pattern recognition, which had as its major goal the machine identification of printed or written characters; and (b) visual experimental psychology, which had as one goal to discover how those same characters are identified by humans. The first is typified by the efforts by Eden and Halle (1961) to produce an analysis capable of enabling a machine to read handwriting, whereas the second is typified by the work of E.J. Gibson and her associates and students, their goal being the reduction of printed capital letters to distinctive features modelled on those then under development for the sounds of language. The premise of Gibson's group was that such an analysis could map putative human visual detection features similar (or identical) to those then being discovered in the cat (e.g. Hubel and Wiesel, 1962, 1965); their ultimate goal was the development of an explanation for misreadings and other performance errors (e.g. Gibson et al., 1963; Gibson, 1965). During the 1960's this area was actively researched. At other centers experimentalists were subjecting letters, both capital and lower-case, to various sorts of analytic procedures, including multi-dimensional scaling (Künnepas, 1966), distorted and permuted reading experiments (esp. Kolers, 1968, 1969), and factor analyzes for linear scalings of letter similarity (e.g. Dunn-Rankin, 1968). Some of these efforts continued into the mid-1980's, at which time at least one reached its presumed zenith in the work of Townsend and his associates (Townsend et al., 1984), where a complete confusion matrix for all twenty-six capital letters was presented.

The empirical methods employed by, and in part created for, this early research, although formally sound, were not aimed at addressing experimental tests of the generative aspects of alphabets. Instead they sought to examine the perceptual, or discrimination, aspects of alphabet processing. In contrast, the present research provides formal methods of data collection and analysis designed to examine the generative aspects of alphabetic systems and the cognitive induction carried out by users of those systems. To achieve this end, paradigms typically employed in psychophysical investigations are generalized and modified so that they can apply to the issues at hand.

## 3. An overview of Watt's alphabetic model

The alphabetic models examined here are the work of Watt (1975, 1980, 1981, 1988a).[2] In these papers Watt characterized the twenty-six uppercase letters of the English alphabet via *generative rules* using a small number of *primitive attributes* to

[2] As discussed below, more than one variant of Watt's Model is examined in the present investigation.

account for learners' errors and therefore for the changes of written letter forms that people make during the course of mastering and using the alphabet. The general model in effect weights certain attributes and attribute-combinations in such a way as to predict that they will tend to be replaced, unintentionally, by other attributes or combinations, thus explaining why 'N' is often reversed and why 'J' is almost always by people just learning the alphabet, whether they are modern schoolchildren or ancient Greeks. (Ancient Greek had a backwards 'L' that acted like modern 'J' in this respect.) By accounting for the 'errors' that have changed the alphabet over time, Watt's analysis provides an answer to the question, *Why is the alphabet the way it is (i.e., contains the letter-forms that it does), instead of some other way?* Apart from explaining these 'errors' the generative rules offer a broader psychological application as well, since they also map into the compositional rules people use in constructing the letters (chiefly, in writing or printing them).

Inspired by the results of Townsend et al. (1984), Eden and Halle (1961), and Gibson et al. (1963; Gibson, 1965), Watt also aimed to provide a deeper analysis of what 'greater similarity' between letters consists. His conclusion: similarity increases as the number of attributes in common increases. (But *which* attributes?) Unfortunately, while Watt could show that his model was supported by perceptual data, he did not have appropriate methods to test the generative, or inductive, aspects of his model. The methods available for testing inductive-like systems (especially those used in linguistics for testing the grammaticality of generated sentences) were ill-suited for examining *degrees of grammaticality* which implicitly play an important role in his theory and model.

Watt's model, referred to as a 'generative grammar', consists of syntactic rules which involve three distinct, yet interactive, levels of evaluation for the 26 uppercase English letters and other highly letter-like forms. These three levels of description include a level of fundamental aspects (which Watt has referred to as 'Abyssal' (Watt, 1988a)), a visual-pattern level (or 'Phanemic Level' (Watt, 1981)), and a motor program level (or 'Kinemic Level' (Watt, 1980)). The three levels of principles interact and influence the formal descriptions put forward by the others.

The deepest of these levels, the Abyssal, is a set of abstract rules that underlie the visual and motor rules. The Abyssal rules describe the most fundamental aspects of the set of letter-forms: aspects such as homogeneity of the alphabetic set, criteria governing excessive visual confusibility of any two items in the set, and rules that constrain the complexity of items. At the Abyssal level letters are described syntactically as combinations of line-segments and concatenators. Abysally, features are assigned to letter components, or letter-form 'phonemes' (Watt, 1975, p. 322). Thus, the Abyssal description of the component parts of 'D' consists of the vexillum '|' and the cusp ' )' (the latter being composed, at this level of the grammar, of the two line segments: '\' and '/').

The next two levels of description, visual patterns and motor programs, both employ syntactic distinctive feature analyses and both are characterized by rules expressing generalizations over the set of letters.

In the visual-level description each letter-form is analyzed as a member of a set of coherent visual patterns, and each is characterized with respect to such properties as visual symmetry, redundancy of features among letter-forms, and the complexity of the distinctive feature description. Thus, the visual description only specifies the formal syntactic descriptions of the visual pattern aspects of the writing system (see Watt, 1981).

The motor-level descriptions of Watt's model analyze and modify the descriptions given by the preceding levels with the goal of satisfying criteria concerning the 'competence' of a letter-form's production program as a reasonable procedure for a user of the system. The syntactic descriptions of these motor programs provide vector, or 'stroke', descriptions of the letter components (Watt, 1981). At this level the kinemic rules respecify the values for direction and orientation of the line segments and the concatenators issued by earlier levels of analysis. Thus, both visible and invisible strokes (line segments and concatenators) are examined, and curved and rectangular values for conjoined segments are determined and made explicit in the description.

As in the preceding levels, generalizations about the distinctive feature components of the motor programs come into play. Thus, for the vexillum '|' which is the first distinctive feature in 'D', the following generalization applies: Always begin the production program with the left-most line segment, and if it is vertical or near-vertical in orientation then begin with a down-stroke. The formal model applies many such 'generalizations' in both the visual and motor analyses.

These three levels of description make up the generative model which characterizes the set of 26 uppercase English letters and all other highly letter-like extensions of that set.[3] The derivative syntactic descriptions of individual letters are quite intricate in detail and are not reproduced here. Readers interested in examples of the syntactic descriptions should consult Watt (1975, 1980, 1981, 1988a).

As mentioned above, Watt's system is supported by evidence from common acquisition errors of young children learning to write the capital English letters, as well as in historical examples (Watt and Jacobs, 1975; Watt, 1975). However, only recently (Jameson, 1989; Jameson and Romney, 1990; Jameson, 1994) has empirical support of Watt's alphabetic model been reported.

### *3.1. The relevance of Watt's alphabetic model to psychology*

The scope and intentions of Watt's model are straightforward. Watt aimed to describe the properties of the set of items composed by the Uppercase English Alphabet. In doing so, Watt also described the properties of the larger macro-set of all possible Uppercase English letters. Watt's goal was to explain *how* the alphabet came to be in its present form, and to describe *why* it became what it is as opposed to some other way. This latter point involves important cognitive issues, as Watt points out, in that the driving dynamic behind the numerous changes that led up to the English alphabet's present form is the peculiarities inherent in human cognitive processing of alphabetic forms. Thus, for the purposes of explaining the English writing system, Watt's model had to address the psychological processing of that system.

[3] Hereafter the term 'generative formal model' (or 'alphabetic model') will be used to identify what Watt calls a 'generative grammar' and that grammar's 'analysis'.

An important feature of Watt's model is that it is foremost a model of the cognitive aspects of writing system processing. This distinguishes Watt's model from other alphabetic models which are largely perceptual, or discrimination, models (cf., Townsend and Ashby, 1982; Townsend, 1971a,b; Gibson et al., 1963; Gibson, 1965, 1969; Eden and Halle, 1961). Much in the way linguists have developed grammars for generative phonology (e.g., see Chomsky and Miller, 1963; Chomsky and Halle, 1968), Watt devised his generative model of the English uppercase letters. Just as English phonology is generative and therefore must allow for phonologically-wellformed 'pseudowords' currently not existing in the language (such as *flin*, *prip*, *spiff*), so too Watt's model generates extensions of the alphabet beyond the set of canonical letters. Thus, Watt's putative cognitive model is generative and it proposes that subjects access this cognitive grammar of the alphabet similar to the way a native English speaker constructs and evaluates new words in the English language.

Justifying a generative alphabetic model may seem difficult since instances of 'new letters' are not common to everyday experience, whereas new words are. However, the somewhat arbitrary fact that the present day English alphabet is a static system, does not erase the fact that at one time the alphabet was dynamically 'evolving', nor does it exclude the possibility that individuals learning the alphabet might acquire a cognitive representation involving properties similar to those acquired for other generative language skills (e.g., English phonology, morphology, and syntax). The similarities between the generative properties of language and the generative aspects of writing systems have been aptly stated by Watt (1979, 1988b) and is a topic beyond the scope of this paper. The important point Watt's model makes for psychology is that a discrimination-based 'perceptual' model of the restricted set of letters known as the English alphabet may not be adequate to explain the cognitive processing carried out by a user of that alphabet. What is needed is a more 'cognitive' treatment of alphabetic processing, and Watt's model of the English uppercase letters attempts to provide just such a model.

Watt's model is the only model of the alphabet which fully explains the psychological acquisition errors frequently observed in young children in the processing of learning the alphabet; the production errors commonly observed in those same children; the execution and confusion errors found in written historical accounts employing alphabetic precursors to the present-day writing system; and, as will be shown below, subjects' preferences for possible extensions of the set of uppercase English letters. And the notion central to Watt's model, that of capturing the generalities of the alphabetic set as opposed to only explaining the actual letters contained in that set, was not examined by Watt's precursors.

## 4. Empirical tests of Watt's alphabetic model

Generative properties of Watt's model permit empirical tests beyond the twenty-six letters through a rule-based manufacture of *new letters*. New letters generated through proper applications of the rules are described as *well-formed* new letters. Other forms, generated imperfectly through misapplication of the rules, are referred to as *ill-formed*

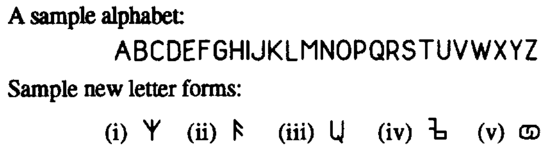


Fig. 1. An example of a 26-letter sample alphabet, as well as examples of new letter forms ranging in grammaticality from highly *grammatical*, (i), to highly *nongrammatical*, (v).

new letters. Fig. 1 shows a sample alphabet and some examples of new letters which we will hereafter refer to as 'pseudoletters'.

Watt's model is applied in pseudoletter classification to wellformedness categories (e.g., grammatical, semigrammatical, and nongrammatical), which depends strictly upon the degree of conformity or adherence to the model's rules and mechanisms. The model and analysis determines how easily a pseudoletter item is obtained from the underlying rules. If that item is redundant or confusible with an already existing item then it should be excluded on the basis of excessive homogeneity (Watt, 1979). If, on the other hand, a new item is deviates in form from the canonical set of alphabetic letters too much, then the item is excluded on the basis of excessive heterogeneity (Watt, 1979). In conjunction with this higher level analysis, each pseudoletter is analyzed for its distinctive feature properties and its coherence to the distinctive feature patterns and programs of the canonical uppercase letters. Together these two levels of analysis can be applied to *any* potential new letter for classifying that letter to a grammatical category.

The theory implicit in Watt (1988a) suggests that well-formed pseudoletters should be perceived by native speakers of English as more acceptable then ill-formed ones. The present studies assess this claim empirically by asking subjects to make judgments about the appropriateness of a candidate new-letter in a specified context of sample letters.

To assess the correctness of Watt's model an experimental paradigm of two-alternative forced-choice (2-AFC) design is employed to obtain judgments of acceptability for proposed 'new-letter candidates' as extensions of the set of existing 26 English letters. If Watt's model is correct, subjects should be much more likely to choose in a 2-AFC task well-formed pseudoletters over those ill-formed (see Appendix A).

Subjects were instructed that the 2-AFC task represented an attempt to extend the existing set of the 26 letters of the English alphabet by introducing newly created letter-forms. They were told to examine each of the two alternatives and to decide which alternative 'best belonged with' or could be considered 'a member of' an extended version of the uppercase English alphabet.

Judgments of acceptability were employed to assess choice behavior for two reasons: (a) they provided an easy format in which to assess subjects' awareness of the generative properties of Watt's model; and (b) because they permitted the comparative assessment of pseudoletters that varied greatly along a continuum of grammaticalness.

## 5. Advantages of the empirical methods presented

The empirical methods suggested here improve upon methods most often employed, in that they eliminate methodological flaws often found in investigations of 'acceptability' and 'grammaticality'. Problems typically inherent in studies of linguistic acceptability (discussed in Quirt and Svartnik, 1966) which are relevant to the present investigation are the following:

(1) What is the relationship between grammaticalness and acceptability? It is assumed that acceptability judgments can be employed to gauge degrees of wellformedness or grammaticalness. However, the suggested 2-AFC method does not require that degrees-of-grammaticality be explicitly accessible to the subject for evaluation, as is the case in direct-questioning techniques. In the present studies the subject need only form a judgment regarding which of two items is preferred. The present methods provide a means by which subjects may accurately access degrees of grammaticality even though they may not be able to describe mechanism(s) underlying such a judgment (as is often the case with inductive mechanisms). Additionally, the methods can be used to assess specific and subtle aspects of grammaticality, of which subjects may have no spontaneous conception, but which nevertheless play a relevant role in the preference for pseudoletters.

(2) The empirical methods presented here can be employed to insure that subjects' classification behaviors exhibit transitive properties. [4] Intransitive choice data may be due to confusion patterns often present in acceptability judgment data, and can thereby alert one to performance drop-off that frequently occurs with tasks involving much item similarity. In addition, transitive choices can also be used to monitor performance improvements due to habituation in the course of the test, as found by Miller and Isard (1963).

(3) Of course to fully assess the transitivity of a given subject's data, one should obtain responses for all possible $(N(N-1)/2)$ pairwise comparisons of $N$ stimuli. (This is often impractical empirically.) As discussed below, this problem is addressed through an overlapping-design paradigm and data analysis techniques that function very well on sparse and unsystematic data. The employed methods provide complete within-subject choice data for stimuli that otherwise would have required 2211 judgments per subject if collected as a single complete-design experiment.

(4) The three most often employed techniques from linguistics (e.g., the 'direct question technique', the 'translation task' and the 'operation test') have many drawbacks compared to the empirical methods suggested here. The more commonly used techniques (see Quirt and Svartnik, 1966) are subject to criticisms that are either avoided or resolved using the present methods. Among the more serious criticisms are: an over-reliance on individual idiosyncratic response data, insufficient objectivity with respect to data collection methods, and no underlying formal model for the treatment of collective data observations. These issues, and others, are addressed in the present research through principled data aggregation methods, objective and rigorous paradigm designs, and mathematically-modelled data analysis procedures.

[4] 'Transitive' is defined: given pseudoletters $a_i$, $a_j$, and $a_k$, and the conditional probability, denoted $P(a_x|a_x a_y)$ for any of three pseudoletters chosen in a pairwise comparison, the ordered relation:

If $P(a_i|a_i a_j) > 0.5$ and $P(a_j|a_j a_k) > 0.5$, then $P(a_i|a_i a_k) > 0.5$

reflects a transitive choice pattern.

(5) An additional strength of the suggested methods is that the results do not rely strictly upon within-subject choice data and thus permit the evaluation of data across subjects, eliminating reliance upon the idiosyncratic choices of a given individual. This is an improvement over standard empirical approaches for assessing linguistic acceptability in that it provides a principled procedure for handling aggregate data and understanding such data; comparatively, standard methods are oversimplistic and do not provide much external objectivity. This focus on aggregate group data is concordant with investigating alphabetic models for groups of individuals rather than investigating the responses of individuals – the latter being the focus of Watt's work and in general the focus of many other linguistic-like studies.

Data from 2-AFC paradigms are used to derive a numerical scale of Performance Ratings for the tested pseudoletters. To do this a performance rating algorithm is applied to the data, according to a mathematical theory, yielding a continuous-valued numerical scale with values for each pseudoletter tested. The resulting performance-rating rank-ordering, and the individual rating scale estimates, are then used to evaluate Watt's formal model.

To describe these Performance Ratings a generalization of Thurstonian scaling is presented (hereafter referred to as the *Rating System*). An important feature of these methods is that, in conjunction with 2-AFC paradigm, the Rating System method preserves the continuous-valued acceptability relations present in subjects' evaluations. This approach yields rich and highly-structured data, making it a major improvement over the standard paradigms used for assessing acceptability. Thus, the presented methods yield data that capture a continuous scale of wellformedness (free of constraints imposed by category classification tasks), which can be used to test either a formal model that posits a continuum of wellformedness, or a model which simply classifies pseudoletters into discrete grammatical categories. [5]

Using the suggested paired-comparison methodology to derive a numerical scaling of pseudoletters from subjects' choice data is considered, for theoretical and methodological reasons, a better method than an alternative scaling method frequently used in psychology, that is, Direct Scaling.

Direct Scaling was not used in this research for several reasons: While it is an easier way to collect data than paired-comparisons, it has a number of theoretical and methodological drawbacks not encountered in the paired-comparison paradigm.

The first drawback is that Direct Scaling lacks a theoretical foundation, and because of this, there are no good criteria for deciding when Direct Scaling is an inappropriate methodology. The second drawback is that the direct scalings of individuals can only be aggregated into a common scale by making assumptions that (a) individual direct scalings belong to a ratio or interval scale and that (b) the scale values of different individuals can be compared in a meaningful way. The assumption in (a) is usually a

[5] Although in theory Watt's model assumes a continuum of wellformedness, for the purpose of illustrating the proposed methodology, here Watt's model is only employed to analyze *categories* of wellformedness. Although finer continuous-level analyses are possible via these methods and are desirable, this simplifies the levels of analysis of Watt's model in cases where two or more pseudoletters are grammatical 'equals' and which complicate interval level comparisons with the rating scale.

completely *ad hoc* assumption based on no empirical evidence. (The rigorous checking of this hypothesis would require much more complicated experiments thus completely eliminating Direct Scaling's advantage of easy data collection.) Assumption (b) is a philosophically loaded assumption, generally discredited by researchers who have looked into the issue deeply. (See Aczél and Roberts, 1989, for a discussion of what is needed to aggregate subject's direct scales, and why the usual aggregation method via arithmetic means may be inappropriate; and Narens and Luce, 1983, for a discussion about comparing direct scale values of different individuals.) In addition, for complicated cognitive stimuli like the pseudoletters used in this study, there is enormous potential for nonconvergence of direct scale orderings with that of paired-comparisons. (See Bostic et al., 1990, for a discussion of this issue.)

## 6. A brief description of the Rating System and its scaling algorithm

In the present investigation a good scale of grammaticality is considered one which has the following properties: (a) it should accurately represent the empirically observed ordering of pseudoletters such that for any given pair of pseudoletters $i$ and $j$, with performance ratings $r_i$ and $r_j$ (where $r_i \neq r_j$) the performance rating scale should give accurate predictions for the empirical preference of $i$ over $j$. (b) A *good* scale would reflect the continuous nature of grammaticality through continuous-valued performance ratings. (c) To the extent that the preferences for pseudoletters are described by the Thurstone Case V model (Thurstone, 1927) the performance rating scale should also approximate empirical preferences when interval-scale information is used for predictions.

The Rating System Model (Batchelder and Bershad, 1979; Batchelder and Simpson, 1988) is a formal scaling model based on the paired-comparison methodology used by the international chess playing community to rate the performance abilities of players, as described by Elo (1978). Chess games, like many forms of two-player competition, have two opponents and the result is either a win, loss, or a draw. A system for measuring chess playing ability is called a 'chess rating system'. Batchelder and Bershad's goal was to create a system that overcame ''a number of methodological problems that have limited the applicability of paired-comparison scaling in psychology'' (Batchelder and Bershad, 1979, p. 40). Their Rating System represents a formal, yet simple, variant of Elo's chess rating system.

Batchelder and Bershad also show that Elo's system is, in essence, a ''system of approximations designed to render serviceable a modified version of Thurstone's Case V model'' (Batchelder and Bershad, 1979, p. 42), thus connecting the theory underlying Thurstone's model with Elo's algorithm.

To derive the performance-rating scale, the pseudoletters are treated as *players* in *games* of pairwise comparisons in which the judgments of human subjects determine which one *wins*, or is more alphabet-like. Using subjects' aggregate data, the Rating System algorithm produces Performance Ratings for each pseudoletter which can be compared with performance ratings of any other form incorporated into the same paradigm and scaled in the rating structure.

### 6.1. Advantages of the Rating System Model

There are quite a few advantages to using the Rating System algorithm over other types of numerical scaling schemes (cf. Batchelder and Bershad, 1979, pp. 41–42).

First, the Rating System algorithm can produce valid numerical performance-rating estimates from unsystematic and sparse data sets. Most estimation schemes typically used in psychology require multiple observations and complete paired-comparison data sets (here called *round-robin* data sets) – often an impractical demand in empirical settings. The Rating System algorithm eliminates the need for systematic sampling of stimuli and thereby greatly facilitates the application of paired-comparison methodology in many empirical domains.

Also the Rating System algorithm can efficiently incorporate a new stimulus item into the system of a set of already scaled objects using only a few new comparisons, thereby addressing the problem of introducing and scaling *newcomer* stimuli. Most psychological scale estimation methods currently used in psychology require deriving a stable estimate for the new object and then rescaling the entire system to adjust all the estimates. The algorithm used here provides procedures for the accurate estimation of new items as soon as they are introduced into the system.

Finally, the Rating System's easy-to-apply closed-form estimation methods permit explicit estimation of the performance-rating scale values. As Batchelder and Bershad point out "most parameter estimation schemes for paired-comparison systems in psychology involve complicated implicit equations for the estimated scale values, ... (and) the scale value for an object depends on the results of choices not involving that object" (1979, p. 41). Asymptotically the Rating System algorithm is a random-variable choice model similar to the discriminable dispersion models of Thurstone (1959). The difference is the underlying distribution assumed (Yellott, 1977), and the estimators for Thurstone's models are in general not unbiased (Batchelder and Bershad, 1979, p. 46).

### 6.2. The Uniform Model

The specific form of Rating System used here is what Batchelder and Bershad call the 'Uniform Model' (1979). It has the following properties: (1) it closely approximates the Thurstone Case V model, yet its estimates are unbiased (unlike the Thurstone model); (2) the formulae of the Uniform Model produce performance rating estimates that are consistent and asymptotically normal and include no more than .01 error per game; [6] (3) the estimators of the Uniform Model are simple to compute; and (4) one can directly obtain information about the sampling distribution of the joint estimator of the rating scale differences and a draw parameter (see Batchelder and Bershad, 1979, pp. 44–45).

As summed up by Batchelder and Bershad:

> "If one were interested in paired-comparison scaling where underlying choice probabilities were constrained away from 0 and 1 – even as loose a constraint as

[6] In the present application the '0.01 error per game' derives from the draw parameter being equal to zero.

> $p_{ij} \in (0.1, 0.9)$ – one could use the Uniform Model for both static and dynamical scaling. Even if the true model was quite different, such as the Thurstone model, fairly accurate scale values can be obtained from sparse and unsystematic data structures.'' (1979, p. 56)

Putting aside many technical details not essential to the present argument, a fixed theoretical function $F$ from the open interval $(-2, 2)$ into the open interval $(0, 1)$ is used, and a real-valued function $u$ on pseudoletters is empirically estimated so that for all forms $f$ and $g$ that have been compared empirically, the observed probability $P_{fg}$, that $f$ is chosen over $g$ is given by

$$P_{fg} \approx F[u(f) - u(g)], \tag{1}$$

where ' $\approx$ ' means approximately equal. In this sense $u$ measures the wellformedness of pseudoletters. The underlying rationale is that this probability is a function of the difference of values from the underlying wellformedness scale.

The advantage of estimating $u$ through this algorithm is that it does not require the complete item-by-item half matrix. (For the 67 pseudoletters a complete half would contain 2211 entries, whereas the designs of Experimental Series 1 and Series 2 presented below require only 833 pairwise comparisons to achieve a stable estimate of $u$.) Through the model expressed in Eq. (1) this procedure yields an interval-level wellformedness scale (of which $u$ is one of its representations) that produces a good approximation of $P_{fg}$. In cases where the representation in Eq. (1) can be obtained, $u$ is often called a scaling function and the system is called a monotone paired-comparison system.

## 7. The experimental studies

The experiments presented below employ the numerical scaling techniques discussed above in a test of Watt's cognitive model. Two separate Experimental Series are presented for deriving performance ratings for 67 tested pseudoletters. In addition, a third experiment is presented which tests the predictive capabilities of the empirically derived numerical scale.

### *7.1. Experimental series 1*

The procedure employed in Experimental Series 1 (hereafter *Series 1*) is a modified round-robin design which incorporates items into the system across a series of *overlapping* experiments, in which later experimental designs are contingent on the outcome of earlier experiments. The five experiments of Experimental Series 1 are now presented.

#### *7.1.1. Experiment 1.1 subjects and method*

This study collected acceptability data for complete pairwise matches between 14 pseudoletters, shown as items 1–14 in Table 1. These 91 paired comparisons were presented to all subjects in the same random order. Twenty college undergraduates participated in the experiment for partial course credit. Subjects recorded their acceptability judgments using pencil and paper examination forms. They were allowed one hour to complete the questionnaire, but in general finished the experiment within 30 to 45 minutes. An example of the Series 1 task is presented in Appendix A.

The experimental task consisted of 2-AFC questions involving pseudoletters. Subjects chose the better candidate as an extension of the sample alphabet provided. No criteria were provided on which to base their judgments, but subjects were instructed to guess if they could not easily choose between the two alternatives. Prior to the experiment, practice-trials and questions about the instructions were solicited.

The data of the 20 subjects yielded 1820 datapoints. The aggregate data of these 20 subjects were analyzed using the Unrated-Player formula from Case 4 in Batchelder and Bershad (1979) to determine the first performance-rating estimates, $r1_i$ (for $i = 1$ to 14), for the 14 pseudoletters in the experiment.[7] That formula is:

$$r1_i = \{\alpha(2S_i - N) + 2N\} + Q, \tag{2}$$

where,

$\alpha$ is the constant [sqrt $(2\pi)$],[8]

$S_i$ is the total observed frequency of 'wins' for pseudoletter $i$ plus $1/2$ for a hypothetical game against self,[9]

$N$ is the total observed frequency of games per pseudoletter in the aggregate round-robin tournament plus a single hypothetical game against self,

and, $Q$ is the mean rating of all pseudoletters involved in the computation, here equal to zero by the Standard Normal distribution assumption.

Based upon the initial performance ratings of the first 14 pseudoletters, denoted $r1_1 \ldots r1_{14}$, the next two experiments in Series 1 were designed.

### *7.1.2. Experiments 1.2 and 1.3 subjects and method*

Experiments 1.2 and 1.3 followed the same general design of Experiment 1.1 except that a heuristic was employed to select the pseudoletters for use in Series 1.2 and 1.3. That is, the $r1$'s of pseudoletters involved in Experiment 1.1 were rank-ordered; those pseudoletters in *odd* positions of the rank order were assigned to Experiment 1.2, and those in *even* positions were assigned to Experiment 1.3. This produced, in both series, a representative spread of scaled items. In addition, pseudoletters 15 to 21 (see Table 1) were assigned to Experiment 1.2 and pseudoletters 22 to 28 were assigned to Experi-

[7] Rather than 20 individual analyses, the aggregate data of the 20 subjects were analyzed to promote the testing of a *common cognitive model*. This is further discussed in Jameson (1994).

[8] The parameter $\alpha$ (equal to sqrt $(2\pi)$) is a scaling constant defined by the Taylor expansion of the cumulative distribution function of the Standard Normal Distribution, $\Phi(x)$. The value of $\alpha$ can depend on the range of $x$ considered, here $x$ is assumed $-1.75 \leq x \leq 1.75$ (see Batchelder and Bershad, 1979).

[9] For $S_i$, *wins* count $+1$, and *losses* count $+0$. We incorporate the scaling constant equal to 0.5 (for a hypothetical pairing against itself) in this initial estimate computation for consistency with the Case 4 scaling equation presented by Batchelder and Bershad (1979). In general, however, unless such a pairing is observed, or draws are permitted as response outcomes, then the hypothetical-game score can be eliminated, with no substantive impact, from the computation of the rating structure. One should note that the potential to use 2-AFC 'no preference' outcomes, or 'draws', is a valuable feature of this scaling model.

ment 1.3. Thus, Experiments 1.2 and 1.3 each incorporated seven *previously-rated* pseudoletters from Experiment 1.1 and seven *newcomer* pseudoletters from Table 1.

As in Experiment 1.1, Experiments 1.2 and 1.3 were designed to obtain acceptability data for tested pseudoletters. Ten undergraduate subjects were sampled for each

Table 1
Pseudoletter alphabetic model classifications and rating scale estimates

| Item Number | Pseudo-Letter | Model 1 | Model 2 | Series 1 Scale | Series 2 Scale | CRS1 Performance-Rating Estimate | CRS2 Performance-Rating Estimate |
|---|---|---|---|---|---|---|---|
| 1 | | G | G | -0.174 | -0.276 | -0.168 | -0.018 |
| 2 | | S | S | 0.103 | --- | 0.217 | 0.168 |
| 3 | | N | N | -0.502 | --- | -0.388 | -0.239 |
| 4 | | N | N | -0.256 | -0.191 | -0.167 | -0.073 |
| 5 | | G | G | 0.206 | -0.06 | 0.13 | 0.237 |
| 6 | | G | G | -0.184 | --- | -0.07 | -0.025 |
| 7 | | G | S | -0.02 | 0.493 | 0.294 | 0.085 |
| 8 | | G | G | 0.165 | 0.309 | 0.294 | 0.209 |
| 9 | | S | S | -1.117 | --- | -1.004 | -0.651 |
| 10 | | G | G | 0.903 | --- | 1.017 | 0.705 |
| 11 | | S | S | -0.01 | -0.399 | -0.147 | 0.092 |
| 12 | | G | G | 0.709 | --- | 0.822 | 0.574 |
| 13 | | N | N | -0.645 | -0.153 | -0.342 | -0.335 |
| 14 | | G | G | 0.452 | 0.247 | 0.406 | 0.402 |
| 15 | | G | N | 0.047 | -0.342 | -0.091 | 0.130 |
| 16 | | G | G | 0.232 | 0.432 | 0.389 | 0.254 |
| 17 | | G | G | 0.493 | --- | 0.607 | 0.429 |
| 18 | | G | N | -0.938 | -0.368 | -0.596 | -0.531 |
| 19 | | G | N | -0.338 | -0.491 | -0.358 | -0.129 |
| 20 | | G | N | -0.276 | 0.115 | -0.024 | -0.087 |
| 21 | | N | N | -0.892 | -0.559 | -0.668 | -0.500 |
| 22 | | G | G | 0.447 | 0.77 | 0.665 | 0.398 |
| 23 | | G | G | 0.37 | 0.687 | 0.585 | 0.347 |
| 24 | | S | S | -1.199 | -0.43 | -0.758 | -0.707 |
| 25 | | G | G | -0.292 | -0.399 | -0.288 | -0.098 |
| 26 | | G | N | -1.215 | -0.37 | -0.736 | -0.717 |
| 27 | | G | G | 0.432 | 1.109 | 0.827 | 0.388 |
| 28 | | G | G | 0.309 | --- | 0.422 | 0.306 |
| 29 | | G | S | -0.492 | -0.307 | -0.342 | -0.232 |
| 30 | | G | N | -0.892 | -0.707 | -0.742 | -0.500 |
| 31 | | G | G | 0.493 | --- | 0.607 | 0.429 |
| 32 | | G | G | 0.093 | 0.155 | 0.181 | 0.161 |

Table 1 (continued)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 33 | | S | G | -0.122 | -0.337 | -0.173 | 0.016 |
| 34 | | G | G | 0.309 | --- | 0.422 | 0.306 |
| 35 | | G | G | 0.647 | 0.893 | 0.827 | 0.533 |
| 36 | | S | S | -0.553 | --- | -0.439 | -0.273 |
| 37 | | S | S | 0.093 | --- | 0.207 | 0.161 |
| 38 | | N | N | -0.122 | -0.143 | -0.076 | 0.016 |
| 39 | | G | G | 0.739 | --- | 0.853 | 0.594 |
| 40 | | S | N | 0.062 | -0.122 | 0.027 | 0.140 |
| 41 | | G | G | 0.678 | 0.23 | 0.511 | 0.554 |
| 42 | | G | G | 0.709 | --- | 0.822 | 0.574 |
| 43 | | N | G | --- | 0.03 | 0.03 | 0.03 |
| 44 | | N | G | --- | 0.773 | 0.773 | 0.773 |
| 45 | | N | N | --- | 0.432 | 0.432 | 0.432 |
| 46 | | N | N | --- | 0.124 | 0.124 | 0.124 |
| 47 | | N | G | --- | 0.955 | 0.955 | 0.955 |
| 48 | | N | N | --- | -0.491 | -0.491 | -0.491 |
| 49 | | N | N | --- | -0.342 | -0.342 | -0.342 |
| 50 | | N | N | --- | -0.153 | -0.153 | -0.153 |
| 51 | | N | N | --- | -0.614 | -0.614 | -0.614 |
| 52 | | N | N | --- | -0.627 | -0.627 | -0.627 |
| 53 | | N | N | --- | 0.124 | 0.124 | 0.124 |
| 54 | | N | N | --- | 0.278 | 0.278 | 0.278 |
| 55 | | N | G | --- | 1.087 | 1.087 | 1.087 |
| 56 | | N | G | --- | 0.678 | 0.678 | 0.678 |
| 57 | | N | N | --- | -0.799 | -0.799 | -0.799 |
| 58 | | N | N | --- | -0.245 | -0.245 | -0.245 |
| 59 | | N | S | --- | 0.315 | 0.315 | 0.315 |
| 60 | | N | S | --- | 0.217 | 0.217 | 0.217 |
| 61 | | N | S | --- | -0.06 | -0.06 | -0.06 |
| 62 | | N | G | --- | -0.085 | -0.085 | -0.085 |
| 63 | | N | S | --- | 1.047 | 1.047 | 1.047 |
| 64 | | N | N | --- | -0.491 | -0.491 | -0.491 |
| 65 | | N | N | --- | -0.891 | -0.891 | -0.891 |
| 66 | | N | N | --- | -0.17 | -0.17 | -0.17 |
| 67 | | N | G | --- | 0.617 | 0.617 | 0.617 |

*Note:* In columns 3 and 4, *G* denotes grammatical, *S* denotes semigrammatical, and *N* denotes nongrammatical pseudoletters. *CRS1* and *CRS2* denote Calibrated Rating Scales 1 and 2 respectively.

Experiments 1.2 and 1.3. In every other respect these experiments were identical to the methods and design of Experiment 1.1.

Experiments 1.2 and 1.3 each produced 910 data points. A portion of these data were analyzed solely for the purpose of establishing rating estimates for the newcomer players (i.e., players 15–21 and 22–28). These newcomer rating estimates were computed independently for Experiments 1.2 and 1.3 and utilized only datapoints from newcomer items matched against previously-rated items from Experiment 1.1.[10] The estimation formula for evaluating newcomers is:

$$r1_i = \{(\alpha + 2N)(W_i - L_i)\} + Q, \tag{3}$$

where,

$\alpha$ is the constant defined in Eq. (2),

$N$ is the total number of observations per pseudoletter being utilized in the computation,

$W_i$ is the number of *wins* pseudoletter $i$ obtained, $L_i$ is the number of *losses*. And, $Q$ is the mean rating of all rated pseudoletters involved in the computation. The critical difference between this equation and Eq. (2) is that $Q$ in Eq. (3) reflects the established $r_j$'s for all $j$ paired against $i$. This follows recommendations in Batchelder and Bershad (1979, pp. 47–52). The rationale is that recursively re-estimating with updated $Q$ values will yield a closer approximation to 'true' scale values than will simply re-estimating with $Q$ set constant.

Using Eq. (3) pseudoletters 15–21 were given $r1$'s using a portion of Experiment 1.2 data and pseudoletters 22–28 were given $r1$'s using a portion of Experiment 1.3 data. This preliminary analysis was undertaken for the design of Experiments 1.4 and 1.5 which are now described.

### *7.1.3. Experiments 1.4 and 1.5 subjects and method*

Experiments 1.4 and 1.5 employed the same design as Experiment 1.1 with a few exceptions: First, as in Experiments 1.2 and 1.3, ten undergraduate subjects participated. Second, the pseudoletters selected for the experiments were determined by the rank-ordering of the newcomer pseudoletters from Experiments 1.2 and 1.3 – the items in the odd positions of the rank order were assigned to Experiment 1.4, those in the even positions were assigned to Experiment 1.5. In addition, pseudoletters 29–35 were assigned to Experiment 1.4 and pseudoletters 36–42 were assigned to Experiment 1.5, (see Table 1). Experiments 1.4 and 1.5 each consisted of complete pairwise matches between seven previously-rated pseudoletters (Experiment 1.2 and 1.3's 'newcomers') and seven newcomer pseudoletters.

Experiment 1.4 and 1.5 data analysis paralleled exactly that of Experiments 1.2 and 1.3. The combined data analyses for Experimental Series 1 (Experiments 1.1 to 1.5) produced initial scale values for each pseudoletter 1–42. However, these scale values are not intended for comparison against the alphabetic model because (a) the scale did not incorporate all available data (i.e., newcomers' games with each other and previously-rated players' against each other); and (b) the rating procedure, which uses a combined

[10] These are only 490 pairwise datapoints rather than the complete 910 datapoints available from each experiment.

rank-ordering of two independent scales across Experiments 1.2 and 1.3 for constructing Experiments 1.4 and 1.5, yields only rough approximations of the *true* scale values. [11]

Utilizing all data (including that mentioned in (a)), Series 1's initial scale estimates were refined through a method given in Appendix B. To summarize, Eq. (B1) is used to recursively to recompute all pseudoletter's estimates by randomly selecting from the 5460 datapoints and, for each comparison, recomputing the ratings for the two players relevant to the observations. Here recursive estimation was carried out until the variance between scale estimates reached, or approximated, zero. Theoretically the iterative method will almost always produce a better approximation of the *true* performance rating estimates than the initial estimates derived from Eq. (2) or (3) (Batchelder and Bershad, 1977).

### *7.1.4. Experimental Series 1 rating scale*

Frequently numerical scaling results in psychology are difficult to extend, or generalize, beyond the studies in which they are obtained. This is because usually those scales relate only to a specific set of experiments and are not easily updatable through additional data observations. A major goal of Experimental Series 1 was to (1) demonstrate that an initial rating structure can be easily established via these methods, and (2) provide the opportunity to subsequently demonstrate that independent data (from moderately different stimulus formats and items) can be used to extend the Series 1 rating structure, yielding informative empirical scale structures. The actual test of Watt's alphabetic model presented below depends upon the data of *both* Experimental Series 1 and 2, thus Series 1 results are reported in conjunction with Experimental Series 2 results below. [12]

## *7.2. Experimental Series 2*

The goal of Experimental Series 2 (hereafter *Series 2*) was to determine whether newcomer pseudoletters could be effectively introduced into the rating structure established by Experimental Series 1 with a minimum of empirical observations. The performance ratings from Series 1 served as a basis to scale the newcomers involved in Series 2. Paradigm modifications in Series 2 aimed to generally increase the flexibility of empirical designs and allow for easy extension of the Series 1 rating structure. Series 2's paradigm design is essentially the same as Series 1 with a few exceptions. For

[11] Implicit in the present research is the notion that there exists a *wellformedness continuum* along which the pseudoletters vary, and that this continuum can be quantified through the Rating System's approximations of the *true* scale values of this continuum.

[12] Jameson (1989) reports analyses of the 42 pseudoletter rating estimates from the Series 1 scale which suggest that the scale accords with a variant of Watt's alphabetic model (Model 2). As is discussed in detail again below, Scale 1 and the model are well correlated, showing the rating scale to be a good empirical measure of the continuum of wellformedness suggested by the alphabetic model.

example, the Series 2 experimental booklets employed enhanced graphic presentation of the pseudoletters.[13]

Series 2 produced performance ratings for additional pseudoletters beyond the 42 used in Series 1 (see Table 1). The new pseudoletters provide a more balanced sample from the pseudoletter categories (esp. 'semigrammatical' and 'nongrammatical'), compared with the mostly 'grammatical' set of Series 1, allowing a better test of the range of possible pseudoletters captured by Watt's model.

Finally, Series 2 was conducted to determine if the empirical rating-scale agreed with the alphabetic model in view of simplifying modifications in paradigm design and scaling procedures – changes which might also depress model agreement. The modifications are: (a) Series 2 is not an overlapping round-robin design, as was Series 1; and (b) the Series 2 rating-scale is an independently derived structure loosely based on the Series 1 rating structure. If the phenomenon remains tractable under these modifications, Series 2 represents an expedient way to introduce new stimuli in tests of a generative model and to compare the results with existing scales using a minimum of empirical manipulation. The expectation is that Series 2 will yield findings similar to those observed in Series 1 even under these modifications.

The Series 2 modifications were motivated by the need to practically and efficiently incorporate new stimuli into an existing rating structure. Intuitively, Series 1's overlapping design seems to involve an excessive number of pairwise matches. The Series 2 design reduces that number by taking advantage of the information existing in the rating structure from Series 1. If Series 2's results accord with those from Series 1, then these improvements create the potential for more efficient testing of the alphabetic model.

Issues considered below are: Will the Series 2 scaling modifications yield valid rating estimates for the new pseudoletters? And, can the Series 2 numerical scale estimates be compared with the estimates of pseudoletters scaled in Series 1?

### *7.2.1. Experimental Series 2 subjects and method*

Differences between Experimental Series 1 and 2 designs are: (a) Series 2 consists of four round-robin experiments rather than five. (b) These four experiments do not employ the overlapping design of Series 1, however, each Series 2 experiment includes some pseudoletters previously used in Series 1. (c) The pseudoletters incorporated in Experiments 2.2, 2.3 and 2.4 are not based on the outcome of prior experiments in Series 2. And (d) Experimental Series 2 assigned pseudoletters to the experiments via a selection heuristic which aimed to both maximize performance-rating estimate diversity, and optimize the diversity of pseudoletters across alphabetic categories.

### *7.2.2. Experimental Series 2 pseudoletter selection*

Pseudoletters were assigned to four booklets of Series 2 as follows: First, new pseudoletters (43–67 in Table 1) were randomly assigned to each experiment. Next, because they were few in number, previously-rated semigrammatical and nongrammati-

[13] The pseudoletters used in Series 2 were created using finer dot-per-inch resolution and smoothing than the characters used in Experimental Series 1. This is seen in different samples of the two experimental booklets available in Appendix A.

cal pseudoletters (i.e., Series 1 items) were randomly assigned in proportions which yielded booklets consisting of approximately equal numbers of items from the three alphabetic categories. Series 1's rated Grammatical pseudoletters were selected on the basis of performance-ratings with the aim being to include in each experiment a representative sample of pseudoletters from the Series 1 rank ordering. Selecting Series 2 pseudoletters in this way pairs Series 2's newcomers against previously-rated players which fairly represent the spread of the Series 1 scale rank-ordering.

Thus, Series 2 consisted of three experimental booklets involving complete 2-AFC pairwise contests between 14 pseudoletter players (previously-rated and newcomers). A forth experimental booklet (Experiment 2.4) otherwise identical to the three described, used seven newcomers rather than six. Thus this fourth booklet involved 15 pseudoletters. Each booklet was adjudicated by ten undergraduate subjects participating in the experiment for partial course credit. Series 2 data total 3780 pairwise datapoints.

### *7.2.3. Scaling the Experimental Series 2 data*

As in Experimental Series 1, the Batchelder and Bershad algorithm was applied to the Series 2 data to derive performance-rating estimates for pseudoletters. Identical scaling methods were used, but the Series 2 rating structure was independently generated from that of Series 1. Deriving an independent rating structure for Series 2 was necessary to gauge the impact of paradigm changes and to determine whether incorporating additional pseudoletters could be successfully achieved using the simpler nonoverlapping design of Series 2.

Also, Series 2 data are scaled separately from that of Series 1 to determine if independently Series 2's rating scale is consistent with the alphabetic model. Such a result would both support the Series 1 findings, and demonstrate that the Series 2 methods can be used to independently introduce newcomers into an existing rating structure and permit comparisons across experimental series. [14] Although the scaling methods used in the two series are the same, the procedures for scaling Series 2 data are somewhat simpler, as is explained below.

As in Series 1, all Series 2 *newcomers* were given initial scale estimates, $r1$'s, using only the datapoints arising from newcomer items matched against previously-rated items. These $r1$'s were derived using Eq. (3), and were computed within experiment as was the case for Series 1 newcomer estimation.

The rerating of Series 2 pseudoletters differed from that of the Series 1 (Appendix B details the Series 1 rerating procedure). In Series 2 all estimates subsequent to $r1$ (e.g., $r2$ through $r7$) were determined recursively using Eq. (B1) in Appendix B. Thus, the $r2$ through $r7$ estimates computed for the pseudoletters *introduced* in Series 2 were computed on a game-by-game basis using all available Series 2 data and across all booklets. [15]

[14] Note that Series 2 and Series 1 have separate scales that each contain uniquely determined rating estimates for pseudoletters incorporated in *both* Series 1 and 2.

[15] To parallel Series 1 as much as possible, the Series 2 recursive estimation was carried out through $r7$. In both Experimental Series 1 and Series 2 the between rating vector variances appeared relatively stable after the first iterative computation (in the case of Series 1 this was $r5$ and Series 2 it was $r2$).

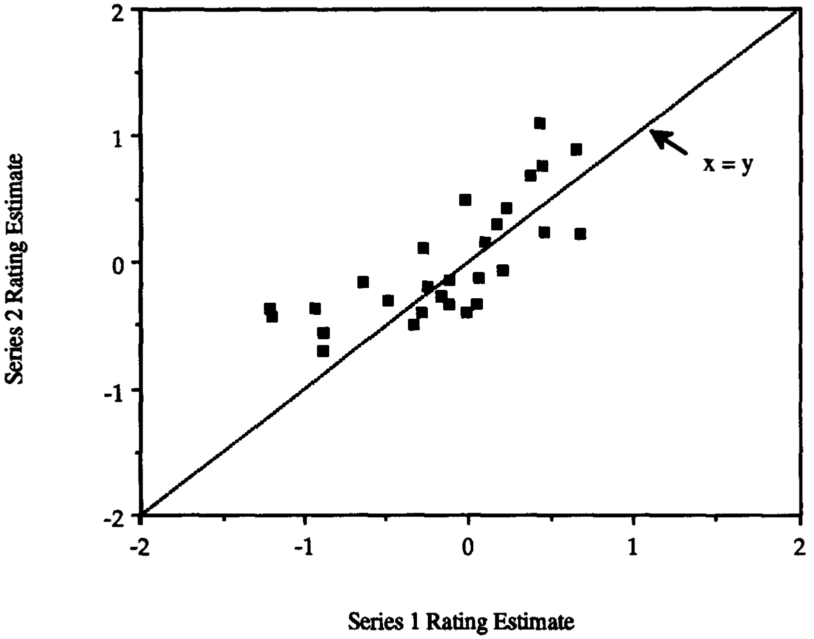


Fig. 2. Plot of performance ratings for pseudoletters involved in both Experimental Series 1 and Experimental Series 2. $N = 28$.

Experimental Series 2's rating structure is related to that of Series 1 only by the 'reference point' of performance ratings of the pseudoletters involved in *both* experimental series. This reference-point refers to both the parameter $Q$ which reflects the Series 1 average scale estimates of pseudoletters paired against new items, and $\sigma$ reflecting the spread of the estimates. Thus, the two rating structures are linked only by the influence of a subset of Series 1 estimates upon the initial computation of Series 2 rating estimates. [16]

### *7.2.4. Experimental Series 2 rating scale*

Initial analyses aim to ascertain whether agreement is found between the Series 1 and Series 2 scales. A comparison of performance ratings for the subset of pseudoletters involved in *both* experiments shows that Experimental Series 2 scale largely matches the Series 1 scale (see Fig. 2). (Series 2 scale is correlated with the Series 1 scale by Pearson's $r = 0.751$, $n = 28$, for the items involved in both series).

Further analyses assessed the measure of agreement between the alphabetic model

[16] If the test paradigm modification did not require an *independent* determination of the Series 2 scale, then Series 2 pairwise data could have easily been incorporated into the Series 1 rating structure through recursive estimates on the aggregate Series 1 and Series 2 data. This would have produced a single scale with estimates for all 67 pseudoletters (the 42 used in Series 1 and the 25 additional newcomers introduced in Series 2) all incorporated into the same rating structure and calibrated on the same scale.

and (1) the Series 1 scale, and (2) the Series 2 scale. Pseudoletters tested in Series 1 and Series 2 were shown by Goodman and Kruskal Gamma measures (hereafter '$\gamma$' or 'Gamma') to be independently scaled in accord with the alphabetic model tested. Gamma is a nonparametric measure of association that makes no scaling assumptions beyond the ordinal level and ignores tied data, and therefore is the preferred measure for the present data. A rationale for using Gamma in this context is given in Jameson (1989, pp. 91-92), and in general for the Gamma statistic in Goodman and Kruskal (1954); and that for Gamma as an ordinal measure is in Freeman (1986).[17]

For pseudoletters scaled in each series: Series 1 scale is correlated with the alphabetic model at $\gamma = 0.74$ ($n = 42$); Series 2 scale is similarly correlated $\gamma = 0.61$ ($n = 53$). Demonstrating that Experimental Series 2 supports the findings of Series 1. These results are encouraging since (a) pseudoletters in Series 2 were separately scaled using a simplified scaling procedure; and (b) the Series 2 scale was based on data garnered by a modified paradigm requiring far fewer observations (only 378 pairwise contests) than either an overlapping round-robin or a complete pairwise design would have required.[18]

However, to optimally compare the ratings of two pseudoletters, each appearing exclusively in a different experimental series, one must consider the issue of calibrating the two series' scales. We turn to that now.

### 7.3. Calibrated performance Ratings Scales from Series 1 and Series 2

According to the Rating System model, it is appropriate to compare the rating estimates of the Series 2 newcomers with those pseudoletters previously estimated in Series 1. However, it is recognized that this is not the most accurate comparison possible because the two separate scales are independently derived and are therefore not calibrated measures. What is needed is a quick method to calibrate them.

The goal is to combine the ratings from the two scales in some principled way that makes use of the scale values of pseudoletters common to the two series to effect the calibration. Two methods which aim to accomplish this (called Methods 1 and 2) are described in Appendix C, and their respective 'Calibrated Rating Scales' are hereafter called *CRS1* and *CRS2*.

Once a properly calibrated scale is obtained one can examine how this scale is associated with an alphabetic model. The CRS1 scale (Table 1 col. 7) was presented by Jameson and Romney (1990) and Jameson (1994) for this purpose.[19] The CRS2 scale and its analysis are introduced here (Table 1 col. 8.).

[17] Gamma is easily interpreted as the proportion of *hits* between two variables by the transformation: $p = (1 + \gamma)/2$. Where $p$ is the proportion of cases in which the two variables are in agreement and $\gamma$ is the observed Gamma statistic between the two variables. Because gamma disregards tied data, the $p$ presented here is a proportion-of-agreement measure for the cases that Gamma considered (i.e., the total number of datapoint comparisons *minus* the number of tied cases).

[18] If implemented as an overlapping round-robin design Series 2 would have required 601 pairwise matches involving 146 additional pairwise observations and an additional scaling step beyond that used in the Series 1 design. Note that for Series 2 each pairwise match was judged by ten subjects and therefore yielded ten observations per pairing.

[19] Jameson and Romney (1990) found the CRS1 scale to be strongly supported in independent empirical studies.

### 7.4. The expected results

Results are presented below for calibrated performance ratings and *two* variants of Watt's alphabetic model, hereafter called 'Model 1' and 'Model 2'; both versions are described in Jameson (1994).

Put briefly, Model 2 can be described as an improved variant of Model 1. Therefore, Model 2 is predicted to perform better as a description of the empirical data than Model 1. In essence, Model 2 is an elaboration of Model 1. That is, it incorporates additional rules which, for a given pseudoletter, often produce distinctive feature descriptions which differ from those of Model 1.

For our purposes, the key difference between these models lies in how they relate to the pseudoletters used in the present experiments. Model 1 makes predictions about pseudoletter acceptability on the assumption that informants would mostly ignore relatively minor departures from the canonical letters. (For instance, Model 1 assumed that informants would accept pseudoletter P as an ordinary 'P' ignoring the angularity of its cusp.) In contrast, Model 2 was developed after pilot experiments demonstrated that informants were in fact attending to what had previously been considered very minor departures from conventional letterforms.

Some differences between Model 1 and Model 2 are seen in Fig. 3 representing pseudoletter 'P' analyzed using both models. Model 1's characterization assigns 'P' to the grammatical set of pseudoletters, whereas Model 2 assigns it to the nongrammatical class. Model 2 reevaluates and improves upon Model 1's analysis at the level of distinctive-feature analysis. For example, Fig. 3 shows that characterizing certain features of 'P' is problematic for Model 1 at the Kinemic level of analysis, because Model 1's assumption is that this kind of criteria would not dramatically enter into subjects wellformedness judgments. Pilot studies, which employed a small subset of the 67 items used here, found such Kinemic criteria to apparently impact subjects' judgments, and Model 2 gives a finer analysis of such feature relations. Model 2 thus predicts that 'P' would prove relatively less grammatical. Models 1 and 2 are further discussed by Jameson (1994, pp. 247–254, 280–285) and are generally considered by Watt (1994, pp. 103–107).

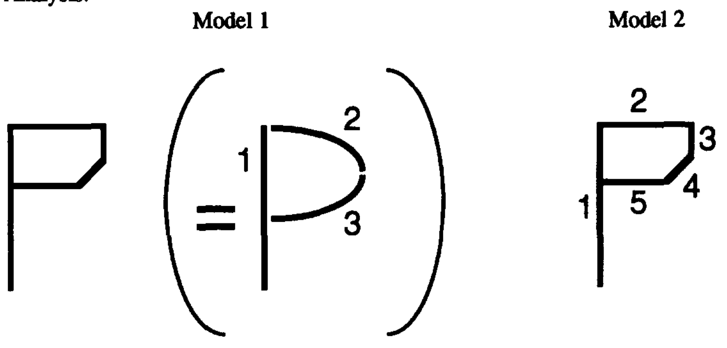
Below two basic questions are explored: (a) Are pseudoletters' performance ratings consistent with the grammatical classification of pseudoletters given by Watt's models?; and (b) are the outcomes of *unobserved* pairwise contests between two pseudoletters reliably predicted by the numerical scale and the Rating System model?

With respect to (a) above, we expect the average performance ratings from the three grammatical classes to be ordered:

$$\text{Grammatical} > \text{Semigrammatical} > \text{Nongrammatical}. \qquad (4)$$

Concerning the predictive capabilities of the Rating System model, mentioned in (b) above, we expect the pairwise outcomes predicted by the model to describe accurately the behavior of new subject samples for both previously tested pairwise comparisons and pairwise comparisons *never before tested.*

Analysis:



| | Model 1 | | | Model 2 | | | | |
|---|---|---|---|---|---|---|---|---|
| Segments: | 1 | 2 | 3 | 1 | 2 | 3 | 4 | 5 |
| VRT | + | - | + | + | ^ | + | + | ^ |
| HRZ | ^ | - | + | ^ | + | ^ | + | + |
| TCE | + | + | + | + | + | + | + | + |
| FLN | + | - | - | 4 | 3 | 1 | 1 | 2 |
| CCV | ^ | + | + | ^ | ^ | ^ | ^ | ^ |
| VSM | + | ( - ) | | + | ( - ) | | | |
| HSM | + | ( + ) | | + | ( - ) | | | |

Fig. 3. Pseudoletter ‘ P ’ characterized by Model 1 and Model 2.

*Note.* The feature assignments above are given the values +, −, or ^. Excepting the feature ‘FLN’, a line-length feature, which in Model 2 is assigned either the value ‘4’ (for full-length), ‘3’, ‘2’ and ‘1’. Lengths ‘2’ and ‘1’, which do not occur in the canonical English letters at all, assign low wellformedness ( = low predicted acceptability) to any letter in which they occur. This and the lack of local symmetry in Model 2’s analysis of ‘ P ’ changes the grammatical classification for this item. Concatenators have been omitted from this illustrative analysis, as they make only a negligible contribution to wellformedness/acceptability. Model 1 assumes that subjects would consider ‘ P ’ as simply a slightly abberant (or poorly-drawn) normal ‘P’; accordingly ‘ P ’ was classified as Grammatical and considered acceptable to subjects. Model 2 presents a reanalysis of ‘ P ’ and classifies the item as illformed, or nongrammatical, and thus predicts the item to be of low acceptability. The author would like to thank W.C. Watt for the analyses in Fig. 3.

### *7.5. The empirical results: Calibrated performance rating scales compared with the alphabetic model*

Initial findings of the experiments described above are presented in Table 2 which give analyses for the two calibrated performance rating scales described above.

Table 2 contains average performance-rating measures for each calibrated scale. As predicted, the data in Table 2 indicates that the performance ratings from both calibrated scales support Model 2 more strongly than Model 1. This is seen in the finding that Model 2 classifications satisfy the relation expressed in Eq. (4) above, whereas Model 1 classifications do not. However, similarities between calibrated scales are not easily

Table 2
Average performance ratings for pseudoletters according to two rating scales and two alphabetic models

| | | Grammatical | Semigrammatical | Nongrammatical |
|---|---|---|---|---|
| | *n*: | 29 | 8 | 30 |
| CRS1 Scale | $\mu$ | 0.250 | −0.259 | 0.002 |
| and Model 1 | $\sigma$ | 0.497 | 0.416 | 0.540 |
| CRS2 Scale | $\mu$ | 0.182 | −0.132 | 0.079 |
| and Model 1 | $\sigma$ | 0.354 | 0.343 | 0.524 |
| | *n*: | 29 | 12 | 26 |
| CRS1 Scale | $\mu$ | 0.474 | −0.038 | −0.309 |
| and Model 2 | $\sigma$ | 0.383 | 0.526 | 0.343 |
| CRS2 Scale | $\mu$ | 0.391 | 0.013 | −0.257 |
| and Model 2 | $\sigma$ | 0.297 | 0.444 | 0.341 |

*Note:* For row headings: *n* denotes the number of ratings in the computation; $\mu$ denotes the average performance-rating; and $\sigma$ denotes the standard deviation; *CRS1* denotes calibrated rating scale 1 and *CRS2* denotes calibrated rating scale 2.

determined by the descriptive measures in Table 2, for this the Goodman and Kruskal Gamma is considered.

The CRS1 and the CRS2 methods produce rating scales which are correlated at $\gamma = 0.88$, $n = 67$. Because the two calibrated scales are very similar only the results for CRS2 are presented hereafter.

Gamma statistics correlating the entire CRS2 rating scale with two versions of Watt's alphabetic model are: $\gamma(\text{CRS2 \& Model 1}) = 0.22$ ($n = 67$) and $\gamma(\text{CRS2 \& Model 2}) = 0.67$ ($n = 67$). These Gammas indicate that Model 2 is clearly much better at predicting the empirical data than is Model 1.[20]

To interpret the finding that the numerical rank-orderings support the pseudoletter classifications of Watt's cognitive model, one might examine possible 'psychological' constituents underlying the decision processes subjects employ in making choices in the 2-AFC experimental task.

Note that the 2-AFC task employed here is not a 'perceptual-discrimination' task. Asking subjects to evaluate pseudoletter acceptability aims to elicit a context dependent 'cognitive' judgment. No criteria are suggested to subjects as a basis for their judgments, and the claim is that the task does not naturally prompt perceptual criteria as a basis.

How can this claim that the task elicits cognitive rather than perceptual judgments be examined? While many alphabetic criteria might impact acceptability judgments, suppose for example that subjects primarily based their judgments on a pseudoletter's

[20] Incidentally, CRS1 is similarly correlated with Model 2. Whether considering all 67 pseudoletters, or a subset that excludes semigrammatical items, measures of association suggest that Model 2 out performs Model 1 at predicting the empirically obtained rating scale rank-order.

confusibility with, or discriminability from, a canonical letter.[21] If this was the primary criterion for acceptance, then the rating scale rank-order of pseudoletters would contain highly-discriminable, nonconfusible pseudoletters in the top most positions of the rank-order, and less-discriminable, confusible pseudoletters in the bottom most ranking positions. The empirically observed preference rank-ordering shows this not to be the case. Rather, a measure of interletter similarity shows top-ranked and bottom-ranked pseudoletters to be equally similar to the closest canonical English alphabet counterpart.[22] This analysis suggests that a simple explanation of pseudoletter rankorder using discrimination-based criteria is not a fitting explanation for our empirically observed preference scale. What is needed is an exploration of possible cognitive criteria underlying subject choices – an examination that is beyond the scope of this paper.

### *7.6. Considering the predictive capabilities of the Rating System*

Additional scaling issues deserving consideration are (1) the general prediction of subsequently observed pairwise preferences, and (2) the prediction of outcomes for pseudoletters never before empirically paired. If the Rating System is a valid model of

[21] This particular criterion (i.e., confusability with existing letters) is but one possible factor underlying subjects' choice behavior. There are several other criteria (e.g., ease of production in specific writing contexts, deviant level of complexity, degree of asymmetry of form, etc.) possibly contributing to the empirically observed preference ordering, but none of these criteria alone come close to explaining the observed rank-order when compared with the explanatory power of Watt's model.

[22] Actually, the top-ten ranked pseudoletters are *more similar* to their canonical counterparts than are the bottom-ten ranked pseudoletters. That is, using a measure of interletter similarity suggested by Townsend (1971b), a 'template' measure of each pseudoletter's similarity with the most similar existing uppercase English letter counterpart was computed. This 'template' measure involves superimposing templates of canonical letters with pseudoletters and quantifying the amount of overlap for the two forms. An Overlap Ratio is thus computed for each pseudoletter defined as the proportion of overlap with a canonical letter relative to the total area of the pseudoletters. (The proportion of overlap provides a similarity index independent of complexity of compared forms). This simple measure of physical similarity was found by Townsend to explain 50% of the variance of the similarity structure in the confusion data as represented by Luce's choice model similarity parameters computed upon scaled distances. Holbrook (1975) found this measure (i.e., Townsend's 'template' measure) to offer a fairly strong prediction of a Luce choice-model similarity measure (Luce, 1963), and reported Townsend's set of similarity parameters to be the most reliable and valid of the available measures (pp. 533–535). For two sets of pseudoletters considered (i.e., top-ten ranked and bottom-ten ranked), an 'Average Overlap Ratio' was computed. It was found that the average overlap-ratio for the top ranking pseudoletters equaled 62%, whereas the average overlap-ratio for the bottom ranking pseudoletters equaled 49%. *If* subjects were basing their choices on 'discrimination' criteria, *then* we would expect to observe *greater* average-overlap for bottom-ranked pseudoletters compared to the average-overlap observed for top-ranked pseudoletters. In fact, the *opposite* relation is observed.
This suggests that confusability and discriminability as measured above are not the primary criteria underlying preferences for pseudoletters in the 2-AFC experimental task. This is further demonstrated by the fact that subjects 'accept' certain pseudoletters despite the fact that these items are more redundant to, or more similar to, canonical letters and are thereby, as predicted by analogous confusion matrix results of Townsend (1971a,b), more easily mistaken for existing uppercase English letters. These results eliminate the possibility that the rank-order performance scale of pseudoletters is simply a measure of pseudoletter acceptance based upon 'perceptual' criteria.

pseudoletter degrees of acceptability, then performance ratings should suggest what is to be expected in subsequently observed pairwise data.

### *7.6.1. Deriving predictions from the calibrated scales*

Properties inherent in the Rating System permit predictions of outcomes for pairwise matches to be computed from performance ratings. Through these predictions, for the set of 67 pseudoletters $A = \{a_1,\ldots,a_i,\ldots,a_j,\ldots,a_{67}\}$, one can examine the *observed* conditional probability with which any given item in $A$ is preferred over any other item, denoted $P_O(a_i|a_i a_j)$, and compare that value with the same conditional probability *predicted* by the Rating System Model, denoted $P_M(a_i|a_i a_j)$ (see Appendix D's procedure for deriving predictions). The expectation is that in considering all possible pairings of items, the probabilities *predicted* by the model will largely resemble the *observed* probabilities. Such a finding would further support use of the Rating System model in the present application. Such a finding would also imply that the model can predict the outcomes for pairwise comparisons of items which have *never* been empirically paired, even when the predictions are based solely upon performance ratings from incomplete pairings of the 67 pseudoletters.

The basic idea underlying the predictions of $P_M(a_i|a_i a_j)$ is that it is a monotonic function of the difference in the true ratings of two pseudoletters. Such an assumption is typical of those frequently used by mathematical psychologists in measurement and scaling settings (Batchelder and Simpson, 1988, p. 295). The formal rationale is given in Batchelder and Simpson (1988, pp. 298–300). Once the model predictions have been derived, the two matrices separately containing $P_O(a_i|a_i a_j)$ and $P_M(a_i|a_i a_j)$ can be compared.

### *7.6.2. Comparing model predictions $P_M(a_i|a_i a_j)$ with empirically observed outcomes $P_O(a_i|a_i a_j)$*

As a measure of similarity between the two matrices $P_M(a_i|a_i a_j)$ and $P_O(a_i|a_i a_j)$ Goodman and Kruskal's Gamma statistic is employed. Gamma is appropriate for this comparison because it permits evaluation of hypotheses concerning variation of the conditional probabilities on the level of a rank-order analysis. This evaluation disregards local variation in the conditional probabilities and supports the prediction model if the rank-order of items is preserved. This rank-order analysis is preferred over other commonly used matrix similarity measures. For example, a Chi-square measure of matrix similarity tests for any level of random variation and linear deviation, and thus is not the desired test of the stated rank-order hypothesis.

The Goodman and Kruskal Gamma measure between the matrices of Series 1 and 2 observed probabilities and CRS2 predicted probabilities is $\gamma = 0.696$, $n = 1440$.[23] This shows that the calibrated rating scale gives rise to predictions that very closely model the empirical data.

Although Gamma is the preferred measure for comparisons between $P_O(a_i|a_i a_j)$ and

[23] CRS2 Gamma computations involve comparisons of two matrices each containing 1440 conditional probability datapoints.

$P_M(a_i|a_i a_j)$, it is conservative in that it only makes use of ordinal information in the data. A second measure, Hubert's QAP (Hubert and Schultz, 1976), is presented below which is more sensitive to variation in comparisons of the conditional probabilities $P_O(a_i|a_i a_j)$ and $P_M(a_i|a_i a_j)$ because it uses an interval scale analysis. This interval scale analysis is presented with Experiment 3 data below.

### 7.7. *Experiment 3: An experiment to test the predictions of the Rating System Model*

While the Goodman and Kruskal Gamma measures presented for Series 1 and 2 data imply that the Rating System predictions accord with subjects' preferences, a stronger test of the Rating System as a model of pseudoletter acceptability is in the prediction of comparisons not previously observed. A third experiment was conducted to test such predictions.

Experiment 3 collected pairwise data like those of Series 1 and 2. These new data are compared against the conditional probability matrices predicted by the Rating System (see Appendix D). If the Rating System model accurately models subjects' preferences for pairwise compared pseudoletters, then these new data should be predicted by the Rating System.

#### 7.7.1. *Experiment 3 subjects and method*

The 2-AFC design previously employed was used in Experiment 3. However, a single nonoverlapping round-robin among 14 pseudoletter players was tested using a single booklet containing 91 2-AFC questions. It involved 58% not-previously-observed pairwise competitions and 42% previously-observed ones. The goal was both to test the predictive capabilities of the Rating System model and to gauge the amount of between-experiment variation in the Rating System models' predictions. Pseudoletters were selected for Experiment 3 to satisfy this 58%-novel/42%-replicated constraint, however pseudoletters were also chosen to be evenly distributed across the CRS2 scale. Forty undergraduate subjects each provided data for 53 *novel* pairwise comparisons and 38 *replicated* pairwise comparisons.

#### 7.7.2. *The results of Experiment 3*

Experiment 3 data are examined only as conditional probabilities observed for pseudoletters pairings, not as a source for rating scale estimation. In addition to Goodman and Kruskal Gamma measures discussed above, alternative analyses between Experiment 3 observations $P_O(a_i|a_i a_j)$ and prediction matrices $P_M(a_i|a_i a_j)$ are now presented.

#### 7.7.3. *Goodman and Kruskal Gammas for Experiment 3 data*

Gammas between Experiment 3's observations $P_O(a_i|a_i a_j)$ and the model's predictions $P_M(a_i|a_i a_j)$ suggest that repeated observations are well predicted. CRS2 was found to predict Experiment 3 choice-data at about the same level as that observed earlier in Experimental Series 1 and Series 2. Moreover, Experiments 3's novel observations are predicted as well as replicated observations. That is, the overall Gamma between CRS2 predictions and Experiment 3 observations equalled $\gamma = 0.662$ for all pairwise comparisons considered ($n = 182$); and, considering only novel comparisons equaled $\gamma = 0.656$ ($n = 106$), whereas a similar Gamma considering only replicated comparisons equaled $\gamma = 0.669$ ($n = 76$).

### 7.7.4. Hubert's QAP analysis of Experiment 3 data

Interval-scale level analysis of these data is given by Hubert's Quadratic Assignment Program, hereafter *QAP* (Hubert and Schultz, 1976). QAP yields an index value, called QAP-Gamma, that measures structural closeness between two matrices. QAP evaluates interval association between two matrices, becoming, in effect, a specialized significance test for the ordinary Pearson correlation coefficient. It also provides Z-scores for the obtained index value for significance tests of the structural similarity of two matrices (Hubert, 1980; Hubert and Baker, 1978). This index is defined as:

$$\text{QAP Gamma} = \Sigma\Sigma(O(i,j) * M(i,j)),$$

where $O(i, j)$ and $M(i, j)$ are values of the two matrices to be compared. The observed value is evaluated relative to a distribution of the QAP Gamma indices based upon a random permutation of rows and columns of the two matrices. The standard normal is assumed as the underlying distribution, and structural similarity between two matrices is measured by the standard Z-score, the null-hypothesis being interpreted as no structural similarity between the two matrices (Hubert and Schultz, 1976).

The QAP test is read backwards from the Chi-square test. In the Chi-square, presented below, a large Chi-square and small probability means that the two matrices are very different. Conversely, a high QAP Gamma and a small probability indicate that the two matrices are very similar.

The results from the QAP analysis of Experiment 3 data are presented below. In addition to the Z-scores from the QAP analysis, Pearson's Correlation Coefficients for the data (taking values of the two matrices as two interval-level random variables) are also presented.[24]

The QAP Gamma measure between Experiment 3's $P_O(a_i|a_i a_j)$ and the CRS2 model's predictions $P_M(a_i|a_i a_j)$ equals 49.95 (Z-score 25.461, $p < 0.001$), and the two matrices are correlated at Pearson's $r = 0.84$. This Z-score ($Z = 25.461$) indicates a very significant structural similarity between the CRS2 prediction matrix and the observed Experiment 3 data matrix. This again confirms that the Rating System Model gives rise to a performance-rating scale that predicts independently observed preference data for pseudoletters. The manner in which the observation and prediction matrices differ, however, is better suggested by the analysis below.

### 7.7.5. Chi-square analysis of Experiment 3 data

Although Chi-square is not an ideal statistical model for goodness-of-fit comparisons between $P_O(a_i|a_i a_j)$ and $P_M(a_i|a_i a_j)$ (it is not a test at the level of the stated rank-order hypothesis mentioned above, and the assumption of observation-independence is not

[24] Correlation coefficients are monotonically related to the QAP Gamma and may be more widely understood than the QAP Gamma might be.

met), it is used here to explore pseudoletter pairings that possibly represent problematic 2-AFC comparisons. Chi-square can be thought of as a kind of distance measure. The probability is interpreted as the chance that the two matrices could be that distant by sampling error alone. The Chi-square values for the 91 comparisons between $P_O(a_i|a_i a_j)$ and CRS2's $P_M(a_i|a_i a_j)$ is $\chi^2(91) = 232.92$, $p < 0.001$. The average Chi-square value for these 91 individual comparisons equals 2.56, and may give a more general impression of the extent to which a predicted outcome typically deviates from the observed outcome.[25]

If we examine the 'fit' of each individual comparison (i.e., with criterion: $\chi^2(1) \geq 3.841$, $p < 0.05$) we find 20 of the 91 are poor and that these are largely comparisons in which 3 specific pseudoletters are involved.[26] Comparisons that exceeded the stated criterion were either: (1) largely due to new-item pairings (these accounted for the most extreme Chi-square values); or (2) were found to be large Chi-square measures from previously observed pairs associated with predictions based on a small number of observations (i.e., 10 compared to the maximum 40 possible). Further analyses did not find commonalities in comparisons exceeding the criterion. For example, no consistent patterns were observed with respect to rating estimates, grammatical classification, possible change in classification from Model 1 to Model 2, old-pairings versus novel-pairings, etc.

These results suggest that the conditional probabilities arising from comparisons involving these pseudoletters are problematic with respect to the Rating System model. The problem might, in some cases, be due to fewer observations for some pairings; or could reflect a high level of intransitivity in subject's paired-comparison choices among distantly ranked pseudoletters. At present these are speculations. Further empirical work and model assessment are needed to determine what factors of rating scales, alphabetic models, or behaviors in the forced-choice task, might depress the predictive capabilities of the Rating System methods.

## *7.8. Summary of the empirical findings*

### *7.8.1. Summary of Experimental Series 1 and 2 findings*

Comparisons between calibrated rating scales and two variants of Watt's alphabetic model were shown to generally support Watt's model. Specifically, statistical analyses found an improved variant, Model 2, was a better model of the empirical phenomena than the Model 1 variant. Methods presented for calibrating independent rating scales produced a combined scale informative in evaluating alphabetic models. Finally, Good-

[25] This Chi-square is given by the following computations: the values in the lower-half matrix of $P_O(a_i|a_i a_j)$ are compared against the corresponding values in $P_M(a_i|a_i a_j)$. These proportions are binomial distributed and are approximated by the Normal Distribution and are thus Chi-square distributed with 1 degree-of-freedom. Individual Chi-square measures are computed using observed and predicted sample frequencies, and these 91 Chi-squares are summed to produce an overall Chi-square value with 91 degrees of freedom.

[26] These three pseudoletters are items 36, 38, and 50 in Table 1. They are involved in 14 of the 20 Chi-square computations which exceed the 3.841 good-fit criterion.

man and Kruskal Gamma statistics comparing observed conditional probabilities with those predicted by the Rating System model found the calibrated scale served well as a basis to predict the observed conditional probabilities.

### *7.8.2. Summary of Experiment 3 findings*

Experiment 3 tested the accuracy of the Rating System model for predicting pairwise conditional probabilities not previously used in constructing the rating scales. Two different analyses (Goodman and Kruskal's Gamma and Hubert's QAP) showed the Rating System model overall to predict accurately. As expected, *replicated* pairwise contests were well predicted by the Rating System model, and, surprisingly, *novel* pairwise contests were predicted at the same level. A cell-by-cell examination of Chi-square measures between Experiment 3's observed conditional probabilities and Rating System predictions proved useful in identifying specific pairwise comparisons that may represent inherently problematic 2-AFC items. The findings of Experiment 3's Gamma analysis parallel that observed for Experimental Series 1 and 2 data.

### *7.8.3. Summary of methodological advances*

The methodological advances given by the present investigations can be summarized as follows: Experimental Series 1 introduced a empirical paradigm and scaling techniques employed for testing Watt's alphabetic model. Experimental Series 2 established that a simplified paradigm and scaling procedures could be alternatively employed to obtain empirical results similar to that found in Experimental Series 1. Both studies support the continued use of the empirical methods in the testing of generative alphabetic models.

Possible methods for combining separate rating scales from independent experiments were presented and were found to yield a calibrated scale which was (a) interpretable with respect to the alphabetic models tested, and (b) useful in distinguishing between two competing variants of Watt's alphabetic model as models of the empirical phenomena. Moreover, predictions of paired-comparison outcomes derived from the calibrated scale and the Rating System model were found to closely approximate independent pairwise observations from Experiment 3, suggesting that the presented scaling techniques can be utilized for predicting paired-comparison outcomes which have yet to be empirically assessed.

But what indicates that a 'good' rating scale is given by the suggested calibration method? As stated earlier a good scale of grammaticality is one which (a) for any given pair of pseudoletters $i$ and $j$, it predicts accurately the empirical preference of $i$ over $j$, (b) reflects the continuous nature of grammaticality through continuous-valued performance ratings, and (c) approximates empirical preferences when both ordinal and interval-scale information is used for predictions. Empirical findings presented here indicate that the obtained rating scale satisfies these criteria.

## 8. General discussion

A major goal stated at the outset of this paper was to provide new methods for testing empirically generative or inductive-like models similar to Watt's model of the alphabet.

To this end the Batchelder and Bershad Rating System was employed in a new paradigm to provide numerical measures for assessing the generative aspects of Watt's cognitive model of the alphabet. Findings from experiments presented here demonstrate that the Rating System serves as a valuable method for testing the empirical viability of generative alphabetic models, and for identifying, in a test between two models of the phenomena, the generative model which best describes the empirical data. To my knowledge these empirical findings are the first instance in which the Batchelder and Bershad Rating System has been successfully employed as an empirical test in the psychological literature.

The methods presented here also represent a significant advance for writing-systems research and other studies of inductive systems, since they permit strong tests of generative models (see earlier discussion in Section 5. Moreover, because these investigations involve the numerical scaling of data derived from judgments for which the *correct* answers are unknown (i.e., are only known by way of a theoretical alphabetic model), one would expect that the presented methods would easily generalize to categorization judgments involving probabilistic category assignments and, therefore, are made by subjects with some uncertainty.

### *8.1. Methodological implications of the present techniques for psychology*

The numerical scaling techniques and experimental methodology presented in this paper are considered useful methods for psychology in general. This is because the procedures can be generalized to many psychological situations involving (a) tests in which 'inductive' or 'generative' processes are examined, (b) tests of these processes in which either within-subject or between-subjects analyses are needed, and (c) tests which simply need to evaluate large numbers of items in complete comparisons without overtaxing subjects by requiring an excessive number of judgments.

Just as independent analyses were employed to confirm the results presented above (see footnote 19), similar Consensus Model analyses can be employed for the purposes of independently validating the Rating System results in other psychological domains, and to determine if subgroups from within the subject sample are found to consistently employ different decision strategies, or criteria, when making choices in a paired-comparison task.

Possible uses of the presented methods are:

(1) Generalizations to other symbol systems: The presented methods could be used to evaluate, and improve, generative models of a variety of writing and symbol systems, including widely employed iconic systems of the kind currently use in computer user-interfaces and air-traffic control systems, and symbolic data representation. Such a test of symbol systems would involve tasks in which subjects evaluated symbols or icons as a signifiers of a particular meaning or concept, or as acceptable extensions of an existing set of items. Similar to the designs presented here, the task would involve forced-choice judgments in which subjects choose items, from a pairwise comparison, that are optimal descriptors of a given semantic value, or those which best extend the existing set of symbols. Psychological studies like these could make important contribu-

tions to the scientific study of symbol systems by providing an empirical basis for implementing 'psychologically optimal' information systems.

(2) Generalizations for categorization studies: A different class of problems in which the present methods could apply are studies involving 'categorization' and prototype-exemplar investigations.

One possible study, which investigates a domain involving continuously varying stimuli, is the investigation of the psychological relationship between color appearance and color semantics. Such a generalization would involve numerically scaling color-appearance samples based upon subjects' preferences for pairwise compared color samples as 'best exemplars' of a given color name. The structure of the scaling of the color samples could be examined for a meaningful correspondence to the organization of a psychophysical color space.[27] This investigation could provide important data concerning the cognitive relationship between perceived colors and color-names. One other psychological domain that has been investigated using these methods is the relation between facial expressions and verbal emotion-labels (Alvarado and Jameson, 1996). These studies suggest that the present methodology can give new insights in otherwise well-investigated domains of psychological phenomena.

A second generalization of the present methods is to investigations of cognitive categories involving stimuli with discrete representations. Such a study could address issues surrounding 'prototype' theory or investigations of 'natural categories' (e.g., Rosch, 1973). Generalizing the present methods to a specified cognitive domain, the experimental task could assess preferences for pairwise compared items in the context of category membership. Numerically scaling subjects' preferences for items' membership in a category, or as 'prototypes' or 'exemplars' of a cognitive domain (i.e., 'living' things), might provide a simply derived, yet informative, scale of category membership or exemplar-prototype status.

## Acknowledgements

This work was supported in part by a National Science Foundation – North Atlantic Treaty Organization Postdoctoral Fellowship in Science and Engineering #RCD-9154448, and by a University of California President's Postdoctoral Fellowship.

Thanks are extended to R.D. Luce, L. Narens and W.C. Watt who offered numerous suggestions on earlier versions of this paper; and to D. Bamber, W. Batchelder, J.-C. Falmagne, L. Freeman, and X. Hu for suggestions concerning data analysis.

## Appendix A. Sample experimental booklets

Experimental Series 1 sample question

[27] Research involving just such a generalization is currently being carried out by the author.

Question #1:

ABCDEFGHIJKLMNOPQRSTUVWXYZ

Which of the following choices is the best "new letter" candidate?

1.
2.

Experimental Series 2 sample question

**Question #1:**

ABCDEFGHIJKLMNOPQRSTUVWXYZ

**Which of the following choices is the best "new letter" candidate?**

**1.**

**2.**

## Appendix B. Refining performance ratings through recursive reestimation

Using all of the data collected from each of the five round-robin experiments $r2$ estimates, a second performance rating, were computed within experiment. A portion of data used in computing $r2$ estimates are *new* data which was not utilized in $r1$ calculations.[28] These $r2$ estimates reflect a more informative $Q$ value. That is, $r2$ estimates were determined using $Q$ values reflecting all 14 representative $r1$ estimates for the 14 pseudoletters incorporated in each of the five experiments, whereas the $r1$ estimates were calculated using $Q = 0$ for the first 14 rating estimates, and only nominally representative $Q$ values (each reflecting the estimates of only 7 pseudoletters rather than 14) were used for the remaining 28 $r1$ estimates. The $r2$ estimates are

[28] For example, considered for the first time are the outcomes of games from experiments 1.2, 1.3, 1.4, and 1.5 involving unrated players against other unrated players. (Contests of this kind total 1680 additional datapoints.)

considered better approximations of the *true* performance ratings of pseudoletters than the preliminary $r1$ estimates.

The $r2$ estimates were established dynamically for each pseudoletter using the formula for previously-rated players where *true* ratings remain unchanged across tournaments (see Batchelder and Simpson, 1988). The estimation formula is [29]:

$$rn_i = \{2\sigma(W_i - L_i)/N\} + Q, \tag{B1}$$

where,

$\sigma$ is the standard deviation of the underlying distribution of true ratings, [30]

$L_i$ and $W_i$ represent the $i$th player's number of observed *losses* and *wins*, respectively (for $i = 1$ to $14$),

$N$ is the number of observations per player incorporated into the rating computation,

And, $Q$ is the mean performance rating of those players engaged in the experiment.

In general, Eq. (B1) is sufficient for recursively rescaling initial scale estimates. In the general, non-overlapping case this equation would be used for rerating scale values until the estimates satisfied a specified stability criterion. However, in overlapping experimental designs, as is the case in Series 1, some additional explanation is required. The remainder of this appendix simply demonstrates the recursive reestimation of overlapping round-robin data using Experimental Series 1 as an example.

Due to the overlapping design and to the fact that $r2$ ratings were computed within experiment, some pseudoletters' $r2$'s were independently estimated twice. For example, the typical player *introduced* in Experiment 1.1 is rated once using that data, and again using Experiment 1.2 or 1.3 data depending upon the particular pseudoletter being considered. (Similarly, a pseudoletter introduced in Experiment 1.2 or 1.3 is $r2$ rated using that data, and again using the data of either Experiment 1.4 or 1.5.) For cases where $r2_i$ is estimated twice the second of these estimates is arbitrarily assigned to $r3_i$. Thus players *introduced* in Experiments 1.1, 1.2, 1.3 (namely pseudoletters 1 to 28) are assigned both $r2$ and $r3$ estimates, while pseudoletters *introduced* in Experiments 1.4 and 1.5 (namely pseudoletters 29–42) are assigned $r2$ estimates only. Thus the vector of $r2$ estimates contains 42 rating estimates, whereas the $r3$ vector is incomplete with only 28 rating assignments (the last 14 values being unassigned).

To utilize the information contained in both $r2$ and $r3$, $r4$ estimates were derived by assigning $r4$ the arithmetic average of $r2$ and $r3$ values when both values were present

[29] Because Eq. (B1) is also later employed in the recursive reestimation of rating estimates a notation change is required: The $n$ in $rn$ will take a value to indicate the iteration of the rating estimate at issue. For example, the fifth iterative recomputation will be denoted $r5$.

[30] Sigma, $\sigma$, replaces $\alpha$ as a scaling constant, eventually permits in later computations an adjustment in scale that reflects the spread of the scale. The assumption is that the distribution underlying pseudoletter rating scale is Standard Normal which follows from the alphabetic model tested. Overall, an application of the model will generate pseudoletter candidates which are normally distributed with respect to wellformedness. The assumption is that the distribution of wellformedness will be reflected in the rating scale. Whatever the true underlying distribution might be, the Normal Distribution assumption has been shown appropriate through empirical validation (Batchelder and Simpson, 1988) and analytic analysis (Yellott, 1977) and is known to be valid in cases where the true underlying distribution is the Double Exponential model or Dawkins Exponential model. Here the distribution is assumed Normal, $N(0, 1)$, therefore $\sigma = 1$.

(i.e., pseudoletters 1 to 28). When $r3$ was not available the rating estimate in $r2$ was simply assigned to $r4$.

Thus the $r4$ estimates can be described as the performance-rating average of $r2$ and $r3$ when a $r3$ value exists, or simply a reassigned $r2$ rating estimate otherwise. Deriving $r4$ in this way is consistent with the dynamic estimation procedures described in Batchelder and Bershad (1979) given the assumption of pseudoletters' unchanging performance abilities over time. [31]

Finally to test whether the $r4$ ratings are good approximations of the *true* performance ratings of pseudoletters a test of stability over recursive recomputation is carried out. To accomplish this Eq. (B1) is employed to successively recompute rating estimates *across* experiments, on a game-by-game basis, for each of 42 pseudoletters until the between rating-vector variances reach, or are close to, zero. In other words the outcomes of 5460 paired-comparison games, are observed randomly, one-at-a-time, and after each observation the performance ratings for the two players relevant to the observations are recomputed. [32] This procedure uses all available data from Experimental Series 1. If after one pass through all the data some pseudoletters were found to continue to exhibit sizeable between-rating variances for the most recent estimates, then the iterative procedure was carried out again. To satisfy this criterion the iterative procedure was carried out three separate iterations which produced a vector of final rating estimates for Experimental Series 1.

Thus stable performance ratings, which theoretically are better than the preceding performance ratings, are provided in the $r7$ estimates of Series 1 (see Table 1 for Series 1 scale). In this way Experimental Series 1 data was used to scale the first 42 pseudoletters in Table 1.

This technique is presented as a general procedure for obtaining a stable rating scale when scaling psychological data. Depending upon the psychological stimuli being scaled one may find that variations in the iterative process also produce a stable scale.

## Appendix C. Two calibration methods

For both methods described below the Series 1 scale (hereafter *Scale 1*) will be calibrated to Series 2 scale's (hereafter *Scale 2*) standard. The rationale for modifying Scale 1 is that it is assumed that Scale 2 is a 'richer' scale (by virtue of more extensive empirical base), therefore Scale 1 should be adjusted to Scale 2's standard.

[31] Holding constant the sample alphabet context and the forces underlying the 'wellformedness continuum' (i.e., the *true* alphabetic model), a pseudoletter's *ability* to perform against another pseudoletter in that context can not *improve* or *change over time* as long as the pseudoletters in question are correctly ordered along the wellformedness continuum, and wellformedness is the corollary of *ability*, then a pseudoletter's ability to win pairwise matches will not improve in a closed system. If the model we are using is close to the correct or *true* model the subject's have in their heads, then the model rank-order will largely resemble the true rank-ordering given by the wellformedness continuum.

[32] The random order in which the $r$'s are recomputed is the same random order in which the questions appeared in the experiments.

A calibrated performance-rating scale is believed to provide a more accurate ordering of the 67 pseudoletters than that given by the ratings reported in Jameson (1989) because each items' rating is individually rescaled to reflect the structure of all items' ratings. In this way, calibrated performance ratings may more closely approximate the *true* performance ratings of the items.

### *C.1. Method I: Calibrating via average rating estimates*

The first method of adjustment aims to provide a rough combining of the information in Scale 1 and Scale 2 by using the scale values of pseudoletters with rating estimates in *both* scales. This calibration method yields a scale hereafter referred to as 'Calibrated Rating Scale 1', or 'CRS1'.

Calibration Method I: Every rating estimate from Scale 1 was adjusted by a constant, calibrating with respect to the overall averages of the two scales and the averages of items with separate estimates in both scales. [33] Then for each pseudoletter involved in *both* experiments the two respective rating estimates (from Scale 1 and Scale 2) were averaged, producing a single rating estimate, CRS1($r_i$), for each item. [34]

This procedure produces a single performance rating for each of the 67 items, calibrated to the same scale, and which together represent *Calibrated Rating Scale 1* for the 67 new-letter forms. [35] Table 1 (col. 7) contains the CRS1 performance-ratings and alphabetic model classification for each pseudoletter.

### *C.2. Method II: Calibrating according to both the average rating and the distribution of rating estimates*

For the purposes of comparison with CRS1, a second method of calibrating the estimates in Scale 1 to Scale 2 was undertaken. Theoretically this method is preferred to that given in Method I because it analytically determines the coefficients for optimally combining the information in the two scales based solely upon the items occurring in both scales. The CRS1 calibration method uses the coefficients $a$ and $b$ of a linear transform represented by the function $f$:

$$f(r_{i1}) = a(r_{i1}) + b = r_{i2}, \qquad \text{(C1)}$$

where, for $i = 1$ to $n$ pseudoletters, and $r_{i1}$ is pseudoletter $i$'s performance-rating from Scale 1 and $r_{i2}$ is that from Scale 2.

[33] This was achieved through two steps: First, every rating estimate from Scale 1 was adjusted by a constant so that the Scale 1 overall average rating-estimate equaled the overall average of Scale 2. Second, all the adjusted Scale 1 estimates were each adjusted a second time by a constant so that the Scale 1 average rating for the set of pseudoletters involved in both experiments was equal to that for Scale 2. This two step adjustment was used because it separately considers the overall average scale estimate and the average estimates of the items common to both scales. However, it is numerically equivalent to adjusting Scale 1's estimates simply once using the sum of the two mentioned constants as the adjustment.

[34] Because Scale 1 contains only ratings for pseudoletters 1 through 42, the ratings for pseudoletters 43 through 67 in CRS1 are simply equal to the corresponding ratings in Scale 2.

[35] Findings for CRS1 are reported in Jameson and Romney (1990) and Jameson (1994).

To find coefficients $a$ and $b$, the expression below is minimized:

$$\sum_{i=1}^{28} (r_{i2} - a(r_{i1}) - b)^2. \tag{C2}$$

The sum in Eq. (C2) is taken over the 28 items which have ratings in both Scale 1 and Scale 2. Minimizing the expression in (C2) yields an estimate of the desired $a$ and $b$ coefficients. Once the coefficients are defined the rating estimates in Scale 1 can be transformed, or calibrated, using the algorithm presented below, thereby permitting comparisons with rating estimates exclusively in Scale 2.

Given Scale 1, let $i$ be any given pseudoletter in Scale 1, and for each $i$, transform the Scale 1 value $r_{i1}$ using

$$\text{CRS2}(r_i) = a(r_{i1}) + b,$$

where CRS2($r_i$) is the calibrated scale value.

The coefficients $a$ and $b$ were determined by finding the derivatives for the function $f$ (see Eq. (C1)) when the slope equaled zero, or at the minimum point in the parabolic curve.[36] The CRS2 performance-ratings are presented in Table 1, column 8.

The calibration methods employed here are only two possible methods. Depending upon the scales and stimuli being assessed, better calibration methods may exist for combining two scales from different experiments. As long as the method used achieves an appropriate linear normalization of the two scales, then it will accord with the Rating System model and will yield results similar to the two methods described above.

## Appendix D. Deriving conditional probability matrices

To compare the model predictions with the empirical observations, the probability with which $a_i$ is preferred given the pairwise comparison of $a_i$ and $a_j$ (denoted $p(a_i|a_i a_j)$) was calculated, for all items which empirically pairwise met, using all data observed in Experimental Series 1 and 2. The calculations produced a 67-by-67 partial matrix of conditional probabilities consisting of 1440 cell values (a bit more than one-third of the values possible in the full matrix).

To derive the conditional probabilities predicted by the Rating System model a generalization of the statistical theory gave the following procedure: Let $R = \{r_1, \ldots r_i, \ldots, r_j \ldots, r_{67}\}$ denote the set of performance ratings for the 67 pseudoletters. For all possible pairs of pseudoletters the differences between the Performance Ratings of each item was determined, denoted $d_{ij}$ for all $i$ and $j$. Next these $d_{ij}$'s were employed in the following piecewise linear equation given in Batchelder and Bershad (1979, p. 44):

$$L(d_{ij}) = \begin{array}{ll} 1 - b, & \text{for } d_{ij} \geq B \\ 1/2 + a^{-1}d_{ij}, & \text{for } -B < d_{ij} < B \\ b, & \text{for } d_{ij} \leq -B \end{array}$$

[36] The coefficients were determined as $a = 0.67133$ and $b = 0.0984$.

where $d_{ij}$ denotes the performance-rating differential, and where $B = 1.75$, $b = 1/16$, and $a = 4$, are given by approximating the cumulative distribution function of a Standard Normal Distribution in the region $\pm 1.75$ by a line segment with slope $1/4$. This approximation while being rough numerically, was shown by Batchelder and Bershad (1979, p. 44) to satisfactorily approximate actual ratings. Constrained to the interval $(-1.75 \leq d_{ij} \leq 1.75)$ the system given above acts like a monotone rating system for rating differences, $d_{ij}$, and is referred to as the Uniform System in Batchelder and Simpson (1988, p. 300).

## References

Aczél, J. and F.S. Roberts, 1989. On the possible merging function. Mathematical Social Sciences 17, 205–243.

Alvarado, N. and K. Jameson, 1996. New findings on the contempt expression. Cognition and Emotion (in press).

Batchelder, W.H. and N.J. Bershad, 1977. The statistical analysis of a Thurstonian model for rating chess players. Social Sciences Working Papers, 128, August, 1977. School of Social Sciences, University of California, Irvine.

Batchelder, W.H. and N.J. Bershad, 1979. The statistical analysis of a Thurstonian model for rating chess players. Journal of Mathematical Psychology 19, 39–60.

Batchelder, W.H. and R.S. Simpson, 1988. Rating systems for human abilities: The case of rating chess skill. UMAP modules in undergraduate mathematics and its applications: Module 698. Reprinted in: P.J. Campbell (ed.), UMAP modules 1988: Tools for teaching (pp. 289–314). Arlington, MA: Consortium for Mathematics and its Applications, Inc.

Bartlett, F.C., 1932. Remembering. Cambridge: Cambridge University Press.

Bostic, R., R.J. Herrnstein and R.D. Luce, 1990. The effect on the preference-reversal phenomenon of using choice indifferences. Journal of Economic Behavior and Organization 13, 193–212.

Bruner, J.S., R.D. Busiek and A.L. Minturn, 1952. Assimilation in the immediate reproduction of visually perceived figures. Journal of Experimental Psychology 44, 151–155.

Carmichael, L., H.P. Hogan and A.A. Walter, 1932. An experimental study of the effect of language on the reproduction of visual perceived forms. Journal of Experimental Psychology 15, 73–86.

Chomsky, N. and G.A. Miller, 1963. 'Introduction to the formal analysis of natural languages'. In: R.D. Luce, R.R. Bush and E. Galanter (eds.), Handbook of mathematical psychology, Vol. II (pp. 269–321). New York: Wiley.

Chomsky, N. and M. Halle, 1968. The sound patterns of English. New York: Harper and Row.

Dunn-Rankin, P., 1968. The similarity of lower-case letters of the English alphabet. Journal of Verbal Learning and Verbal Behavior 7, 990–995.

Eden, M. and M. Halle, 1961. 'The characterization of cursive handwriting'. In: C. Cherry (ed.) Information theory: Fourth London Symposium. Washington, DC: Butterworths.

Elo, A., 1978. The rating of chess players, past and present. New York: Arco Publishing.

Fehrer, E., 1935. An investigation of the learning of visually perceived forms. American Journal of Psychology 47, 187–221.

Freeman, L.C., 1986. Order-based statistics and monotonicity: A family of ordinal measures of association. Journal of Mathematical Sociology 12, 49–69.

Gardin, J.C., 1958. 'On the coding of geometrical shapes and other representations with reference to archaeological documents'. In: Proceeding of the International Conference on Scientific Information (pp. 889–915). Washington, DC: National Academy of Sciences.

Gibson, E.J., H. Osser, W. Schiff and J. Smith, 1963. 'An analysis of critical features of letters tested by a confusion matrix'. In: Cooperative Research Project #639: A Basic Research Program on Reading. Washington, DC: U.S. Office of Education.

Gibson, E.J., 1965. Learning to read. Science 148, 1066–1072.

Gibson, E.J., 1969. Principles of perceptual learning and development. New York: Appleton Century Crofts.

Goodman, L.A. and W.H. Kruskal, 1954. Measures of association for cross classifications. Journal of American Statistical Association 49, 733–764.

Holbrook, M.B., 1975. A comparison of methods for measuring the interletter similarity between capital letters. Perception & Psychophysics 17, 532–536.

Hubel, D.H. and T.N. Wiesel, 1962. Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. Journal of Physiology 160, 106–154.

Hubel, D.H. and T.N. Wiesel, 1965. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. Journal of Neurophysiology 28, 229–289.

Hubert, L.J. and J. Schultz, 1976. Quadratic assignment as a general data analysis strategy. British Journal of Mathematical and Statistical Psychology 29, 190–241.

Hubert, L.J., 1980. Analyzing proximity matrices: the assessment of internal variation in combinatorial structure. Journal of Mathematical Psychology 21, 247–264.

Hubert, L.J. and F.B. Baker, 1978. Evaluating the conformity of sociometric measurements. Psychometrika 43, 31–41.

Jameson, K.A., 1989. An empirical investigation of semiotic characterization of alphabetic systems. Doctoral dissertation, University of California, Irvine. Ann Arbor, MI: University Microfilms International (Microfilm order no. 8923073).

Jameson, K. and A.K. Romney, 1990. Consensus on semiotic models of alphabetic systems. Journal of Quantitative Anthropology 2, 289–304.

Jameson, K., 1994. 'Empirical methods for evaluating generative semiotic models: An application to the roman majuscules'. In: W.C. Watt (ed.), Writing systems and cognition (pp. 247–291). Dordrecht: Kluwer.

Kolers, P.A., 1968. The recognition of geometrically transformed text. Perception & Psychophysics 3, 57–64.

Kolers, P.A., 1969. Clues to a letter's recognition: Implications for the design of characters. Journal of Typographic Research 3, 145–168.

Künnepas, T., 1966. Visual perception of capital letters. Scandinavian Journal of Psychology 7, 189–196.

Luce, R.D., 1963. 'Detection and recognition'. In: R.D. Luce, R.R. Bush and E. Galanter (eds.), Handbook of mathematical psychology, Vol. I (pp. 103–189). New York: Wiley.

Miller, G.A. and S. Isard, 1963. Some perceptual consequences of linguistic rules. Journal of Verbal Learning and Verbal Behavior 2, 217–228.

Narens, L. and R.D. Luce, 1983. How we may have been misled into believing in the intercomparability of utility. Theory and Decision 15, 247–260.

Quirt, R. and J. Svartnik, 1966. Investigating linguistic acceptability. The Hague: Mouton.

Rosch, E.H., 1973. Natural categories. Cognitive Psychology 4, 328–350.

Thurstone, L.L., 1927. A law of comparative judgment. Psychological Review 34, 273–286.

Thurstone, L.L., 1959. The measurement of values. Chicago, IL: University of Chicago Press.

Townsend, J.T., 1971a. Theoretical analysis of an alphabetic confusion matrix. Perception & Psychophysics 9, 40–50.

Townsend, J.T., 1971b. Alphabetic confusions: A test of models for individuals. Perception & Psychophysics 9, 449–454.

Townsend, J.T. and F.G. Ashby, 1982. Experimental test of contemporary mathematical models of visual letter recognition. Journal of Experimental Psychology: Human Perception and Performance 8, 834–864.

Townsend, J.T., G.G. Hu and R.J. Evans, 1984. Modelling feature perception on brief displays with evidence for positive interdependence. Perception and Psychophysics 36, 5–49.

Watt, W.C., 1975. What is the proper characterization of the alphabet? I: Desiderata. Visible Language 9, 293–327.

Watt, W.C., 1980. What is the proper characterization of the alphabet? II: Composition. Ars Semeiotica 3, 3–46.

Watt, W.C., 1981. What is the proper characterization of the alphabet? III: Appearance. Ars Semeiotica 4, 269–313.

Watt, W.C., 1988a. What is the proper characterization of the alphabet? IV: Union. Semiotica 70, 199–241.

Watt, W.C. and D. Jacobs, 1975. 'The child's conception of the alphabet'. In: M.D. Douglass (ed.), 39th Yearbook, Claremont Reading Conference. Claremont, CA: Claremont Graduate School.

Watt, W.C., 1979. Iconic equilibrium. Semiotica 28, 31–62.

Watt, W.C., 1988b. 'Canons of alphabetic change'. In: D. de Kerckhove and C. Lumsden (eds.), The alphabet and the brain: The lateralization of writing (pp. 122–152). Berlin: Springer-Verlag.

Watt, W.C. (ed.), 1994. Writing systems and cognition: Perspectives from psychology, physiology, linguistics, and semiotics. Dordrecht/Boston/London: Kluwer Academic Publishers.

Yellott, J.I., Jr., 1977. The relationship beteen Luce's Choice Axiom, Thurstone's theory of Comparative Judgment, and the double exponential distribution. Journal of Mathematical Psychology 15, 109–144.