Developments in Business Simulation and Experiential Learning, Volume 28, 2001

AN ALTERNATIVE VIEW TO SCORING SELECTED SIMULATION GAMES AND EXPERIENTIAL EXERCISES

John R. Dickinson, University of Windsor
bjd@uwindsor.ca

ABSTRACT

Evaluating performance of participants in simulation games and experiential exercises is an ubiquitous need. A milieu of criteria may be employed, some of them, e.g., earnings, by nature being metric measures, others of them being more qualitative originally, e.g., evaluation of written plans or evaluation of role plays, and requiring assignment of metric scores by the instructor. This paper describes an alternative approach requiring original performance measures at an ordinal level, actually reducing the data scale level of inherently metric measures or making for less demanding evaluations by instructors.

INTRODUCTION

Measuring the performance of participants is an integral component of simulation gaming and experiential exercises generally. In application, where participants are students, performance measurement is, of course, required in order to assign grades (Anderson and Lawton 1992). Performance is also a common criterion in basic research. Example fields include comparing simulation gaming with other methods of learning (Miles, Biggs, and Schubert 1986), manager/student characteristics (Gosenpud and Washbush 1996), manager/student styles and practices (Teach 1993), external validity (Norris 1986; Wolfe and Roberts 1986), internal validity (Dickinson and Faria 1997; Wellington, Dickinson, and Faria 1990), administration parameters (Patz 2000), and so on.

Research on performance measurement includes the work of Pogossian (1999, 1998, 1997). During discussion of his 1999 paper, Professor Pogossian described a conceptual approach to measurement drawn from the realm of international chess competition. The underlying condition of that idealized approach would have each competitor playing each other competitor numerous times. This paper integrates the rationale of the paradigm described by Pogossian with a well-developed model of "comparative judgments" toward an alternative approach to measuring the performance of participants in simulation games and experiential exercises.

FRAMEWORK

Again, the idealized competition format would have each chess player compete against each other chess player numerous times. It may not be practically feasible to do this, but perhaps that condition can be approached, if not fully realized. Basic results, then, would be the proportions of times each competitor prevailed over each other competitor. A conceptually similar schema underlies round-robin tournaments where each participant competes against each other participant, though where each participant competes against each other participant only once, the analogous proportions are the extremes of either 0 or 1. A key characteristic of these competition formats is that no account is taken of by how much a competitor wins or loses a round, only whether or not he or she wins or loses.

In competitive simulation games and experiential exercises that comprise a series of periods of competition or a series of trials, a corresponding proportions matrix is readily obtainable. For example, in business games, companies (i.e., participants) commonly compete against other companies within their industries over a succession of several simulated quarters or years. Based on some indicator of performance, companies/participants may be rank ordered. Common indicators include earnings or, often in the case of total enterprise games, an index comprising various facets, e.g., stock price, return on investment, etc.. Though not in the guise of the paired comparisons of chess competitions, each company each period does compete against each other company. The repetition of this competition each period allows for the determination of a proportions matrix, P. The P matrix is n x n, where n is the number of companies/participants (within a given industry). P_{jk} is the proportion of periods participant k ranked higher than participant j over the course of the entire competition. That is, the P matrix is column-dominated. P_{kj} equals $1-P_{jk}$ and the entries on the diagonal are left vacant. The P matrix essentially comprises paired comparisons proportions.

WHY?

Original Nonmetric Measures

Depending on the game or exercise situation, basic performances of participants might only be dichotomous measures of won-lost or better-worse originally or, at best, rank ordered. (Note that the dichotomous measures are simple rank orders.) A series of experiential exercises, for instance, may be such that the instructor finds dichotomous or rank ordered determinations more meaningful than metric scores. This may occur, for example, where the exercise is not scored objectively, but is evaluated by the instructor and he or she does have sufficient basis for making "how much more or less" types of evaluations, but does have sufficient basis for making "more or less" types of evaluations. A second example is where exercises are of varying difficulty. Numerically, a second-place score of 70 on a relatively difficult exercise might be at a disadvantage compared to a second-place score of 80 on a relatively easy exercise. A rank of second in both cases may be deemed more meaningful. Circumstances such as these lead naturally to the proportions matrix described earlier.

Original Metric Measures

Other game or exercise situations, though, employ metric indicators of participant performances, such as the composite indices as described above or earnings. It is the cumulative value of the indicator at the end of the competition that is the basis for evaluation of participants.

These indicators are of interval or ratio data types and the question arises as to why one would wish to transform more informative interval or ratio measures into less informative rank orders.

An essential reason is that the single cumulative indicator approach may not be the most philosophically desirable. Under that approach, performances during each period of the competition or on each exercise are weighted equally. Simply enough, in that cumulation a disastrous performance in a single period or on a single exercise may be never overcome by superior performances in all other periods or on all other exercises. (There is certainly a school of thought that holds this scenario to be acceptable, as discussed under "Issues" below.) Parallel accommodations of this phenomenon are plentiful. The practices of disregarding the lowest of a series of quiz scores, weighting a first case analysis less than a second case analysis, and grading on improvement are of the same underlying philosophical ilk.

A second, though similar reason, is that the tacit assumption that performances in each period of a simulation competition or each experiential exercise are comparable. This may be questionable. The competition environment, i.e., the playing field, may have been altered by the instructor by design. Examples might be the imposition of events such as a strike and varying exchange rates. The instructor may redefine the parameters of the competition to effect, say, growth of the market or to alter the strategy decision mix. Factors other than by instructor design may also make periods structurally different such as a radical strategy by a competitor, a company dropping out of the competition, and so on. While the playing field remains level across the competitors, it changes from period to period. A given metric indicator value of performance in one period may not be comparable to the same numeric indicator value in a different period. (Note that a scheme that weights performance in earlier periods less than performance in later periods does not address many of these factors.)

The rationale for using the P matrix as the basis for participant evaluation, then, may stem from the original nature of performance being dichotomous or from a reasoned philosophical desire to transform higher level data measures into the lower level rank orders of paired comparisons.

ANALYTICAL PROCEDURE

The basis for analysis, i.e., determination of each participant's final performance measure value, is the P matrix as described above. A simple approach is to sum the respective columns of the P matrix. The performance value for participant k, then, would be the sum of the proportions in column k. A variety of more sophisticated models have been put forth, however, for transforming the proportions matrix into scale values (Greenberg 1965; Guttman 1946; Nishisato 1978). A seminal method is Thurstone's law of comparative judgment (Thurstone 1927a, 1927b). Following, Thurstone's law is described briefly and then its application is illustrated.

Thurstone's Law of Comparative Judgment

Thurstone's law is founded on a postulated discriminal process. Essentially when presented with a stimulus j, some value, d_j' for the stimulus is "excited" in the subject. Repeated exposures to the stimulus would result in a distribution of such values, the mean of which, s_j, is taken to be the scale value of the stimulus and the standard deviation of which is called the discriminal dispersion. Variation in the values is attributed to random effects. For example, hypothetically a person rating the sweetness of a (disguised) food product on a scale from 0 to 100 might not give the exact same rating in repeated trials due to fluctuations in sensory receptors and other transitory physical and psychological effects. The distribution of these discriminal processes is postulated to be normal.

A similar process would apply for a second stimulus k. Repeated exposures to the second stimulus would yield a distribution of discriminal process values, d_k', and a mean

value, s_k, for it. The extent to which the distributions would overlap would be indicative of the differences in the sweetness of the two products. Generally, the sweeter product would be rated higher than the less sweet product, but not necessarily always. The greater the preponderance of one product being rated higher than the other, the further apart their distributions, and the further apart their sweetness scale values, s_j and s_k. Thurstone's law, then, relates scale values to this preponderance or proportion.

In the context of competitive simulation games and experiential exercises, the corresponding conceptual view is that performance in a given period or on a given exercise is a sample drawn from a distribution, the mean of which is the true measure of student performance. Thurstone's law provides a model by which the P matrix may be used to estimate these true measures. More precisely, Thurstone's law estimates differences in these true measures between students.

A complete presentation of Thurstone's law is beyond the scope of this paper. For such a presentation, the reader is referred to Thurstone's original works (1927a, 1927b) and to Torgerson (1958). Nine versions of Thurstone's law have been developed, representing combinations of data collection modes and simplifying conditions that allow estimation of Thurstone's model. Suffice to say that it is the Condition C version of the law that is applied here.

ILLUSTRATION

For the sake of illustration, consider an hypothetical simulation game competition in which five companies, comprising an industry, compete against each other over ten competition periods. The traditional company performance criterion is total earnings at the end of the competition, i.e., the sum of the ten period earnings amounts. The example may be easily applied to a series of experiential exercises. Too, the performance criterion may just as readily be a composite index, as with many total enterprise games.

Total earnings for each of the five companies are presented in Table 2, ranging from a low of $185.44 for Company 5 to a high of $395.82 for Company 2. Single period earnings were examined to determine the percent of periods in which Company k earned more than Company j for all pairs of companies, yielding the P matrix presented in Table 1. With rows denoted by j and columns denoted by k, Table 1 presents the proportions of periods in which the earnings of Company k exceeded the earnings of Company j. For example, Company 2 earned more than Company 1 in eight of the ten competition periods.

TABLE 1

PERIODS COMPANY k OUTPERFORMED COMPANY j (%)

Company	k = 1	2	3	4	5
j = 1	0	.8	.6	.4	.5
2	.2	0	.6	.1	.2
3	.4	.4	0	.2	.4
4	.6	.9	.8	0	.4
5	.5	.8	.6	.6	0

The P matrix was analyzed using the Condition C version of Thurstone's law, with results presented in Table 2. Also presented in Table 2 are the sums of the column percents for each company, a simple approach to deriving a metric score for each company from the P matrix.

TABLE 2

ORIGINAL, THURSTONE, AND COLUMN %s PERFORMANCE MEASURES

	ORIGINAL			TRANSFORMED		
Company	Earnings	Thurstone	Sum of Col. %s	Earnings	Thurstone	Sum of Col. %s
2	395.82	.967	2.9	100.00	100.00	100.00
3	367.42	.745	2.6	91.88	89.75	91.67
1	363.02	.256	1.7	90.62	67.24	66.67
4	230.40	0.00	1.3	52.68	55.43	55.56
5	185.44	.155	1.5	39.82	62.58	61.11

Transformation of Scores

Thurstone scale values are unitless, the minimum value traditionally and arbitrarily being fixed at zero. To facilitate comparisons across the original earnings, Thurstone scale, and sums of the column percents performance scores, each measure was linearly transformed to have a maximum of 100 and a mean of 75.

RESULTS

Consider the final scores yielded by the traditional earnings criterion and by Thurstone's law applied to the P matrix, each transformed to a maximum of 100 and a mean of 75 (Table 2). For three of the companies (Companies 2, 3, and 4) differences between the two approaches are less than 3 points. However, for Companies 1 and 5, the differences are dramatic. Based on earnings, the score for Company 1 is 90.62, while based on the P matrix its score is 67.24, a

decrease of 23.38 points. In contrast, the difference for Company 5 is an increase of 22.76 points, from 39.82 based on earnings to 62.58 based on the P matrix.

Company 1 ($363.02) earned $177.58 more than did Company 5 ($185.44), resulting in a difference in their transformed earnings-based final scores of 50.8 points (90.62-39.82). Yet their final scores derived from the P matrix differ by only 4.66 points (67.24-62.58).

Underpinnings of the P Matrix Measures

To better understand the philosophical basis for the more equal scores derived from the P matrix, it is informative to examine some of the matrix elements. Over the ten competition periods, Company 1 earned more than Company 5 in five periods, while Company 5 earned more than Company 1 in the other five periods. Compared to the top earning company, Company 2, both Companies 1 and 5 earned more than Company 2 in two of the ten periods. Compared to the second highest earning company, Company 3, both Companies 1 and 5 earned more than Company 3 in four of the ten periods. In paired comparisons with other companies, the only distinction between Companies 1 and 5 is that the former earned more than Company 4 in six periods, while the latter earned more than Company 4 in four periods. On the basis of head-to-head, better-worse comparisons, it is easy to understand the nearly equal Thurstone final measure values for Companies 1 and 5.

Absent from this examination, of course, is consideration of *by how much* one company earned more than another.

Results for the sums of the column percents are very close to those derived using Thurstone's law (Pearson correlation=.999). For this example, it is the use of the P matrix as a basis for determining final scores that is critical, not which of the two analysis methods is applied to that matrix.

ISSUES

Philosophical Basis

This study has developed an approach by which participants *might* be evaluated on the basis of a succession of trials. Philosophically, there is the issue of whether participants *should* be evaluated on an accumulation of individual period performances, i.e., trials, as opposed to the cumulative final state of the enterprise at the end of the competition. In support of the latter, are "real world" arguments such as cumulative performance being the basis for valuations by investors and managers having to live with past failures and successes. Similarly, numerous games and exercises are designed to be on-going. They are expressly designed to provide that competitive environment and, thus, participants should be evaluated *vis-à-vis* that environment.

Such arguments are not clearly conclusive, however. There does exist the scenario in which a participant's superior performance in all periods but one or two is not sufficient to overcome poor performance in those periods. The participant prevailed in all periods but a few and that basis for evaluation has some substance. Simply seeing an assignment through to successful completion, in a success-failure dichotomy, is certainly a basis for evaluation of managers in the real world.

Empirical Validity

Deciding on the philosophical basis on which students are to be evaluated is the initial step toward an operational evaluation method. That decision, though, should also be supported by empirical validation. Investigations of the validity of the approach developed here are presently underway, using both Monte Carlo and real data banks and alternative analytical methods.

Feasibility

Another type of issue is the feasibility of implementing the approach developed here. To students and other interested parties, it may be difficult to achieve understanding and acceptance of the approach.

There also exist practical implementation issues. This approach to performance measurement is ideally suited to competitive simulations and experiential exercises involving repeated (independent) trials, particularly those in which participants either win or lose or in which participants may be rank ordered only. In these two scenarios, the basic paired comparison data are readily obtainable. For some types of games and exercises, though, a paired comparison approach is more problematic, both practically and philosophically.

Practically, games such as total enterprise typically progress on a period by period basis, with the state of the enterprise at the end of one period being the state of the enterprise at the beginning of the next period. That is, participants manage an on-going company and benefit from or are burdened by the effects of strategies of preceding periods. In this condition, one participant's strategy in a given period may be superior to that of a second participant's strategy, yet that superiority may not be reflected in enterprise evaluation criteria for that period. In this type of simulation, other than the first period, participants do not enter a given competition period under identical conditions and those non-identical starting conditions may dominate superior performance in an individual period. To an extent, game designers may address this by attempting to isolate individual period performance. For example, interest expense to service previously incurred debt may be removed from enterprise performance criteria. Other such adjustments, though, may not be feasible. For

instance, strategy decisions such as advertising may have carry-over effects. It may not be possible to remove these effects without altering the fundamental nature of the game.

Diagnosis and Basic Research

The feasibility of the evaluation approach developed here is a material concern, with practical hurdles possibly being prohibitive. At the same time, there is no reason to accept these hurdles as fatal *a priori* nor, as a policy matter, does it seem tenable to dismiss what might be a more valid approach than traditional evaluation.

Also, should this approach prove conceptually more appealing than traditional evaluation and should it prove to accomplish greater discrimination among students, it may be useful for diagnostic and basic research purposes.

REFERENCES

Anderson, Philip H. and Lawton, Leigh (1992), "A Survey of Methods Used for Evaluating Student Performance on Business Simulations" *Simulation & Gaming*, Vol. 23 (December), 490-498.

Dickinson, John R. and Faria, A. J. (1997), "A Random Strategy Criterion for Validity of Simulation Game Participation" *Simulation & Gaming*, Vol. 28 (September), 263-275.

Gosenpud, Jerry and Washbush, John (1996), "Total Enterprise Simulation Performance as a Function of Myers-Briggs Personality Type" *Simulation & Gaming*, Vol. 27 (June), 184-205.

Greenberg, M. G. (1965), "A Method of Successive Cumulations for the Scaling of Pair-Comparison Preference Judgments" *Psychometrika*, Vol. 30 (December), 441-448.

Guilford, J. P. (1954), *Psychometric Methods*, Second Edition (New York: McGraw-Hill Book Company).

Guttman, L. (1946), "An Approach for Quantifying Paired Comparisons and Rank Orders" *Annals of Mathematical Statistics*, Vol. 17, 144-163.

Miles, Wilford G., Jr., Biggs, William D., and Schubert, James N. (1986), "Student Perceptions of Skill Acquisition through Cases and a General Management Simulation" *Simulation & Games*, Vol. 17 (March), 7-24.

Nishisato, Shizuhiko (1978), "Optimal Scaling of Paired Comparisons and Rank Order Data: An Alternative to Guttman's Formulation" *Psychometrika*, Vol. 43, 263-271.

Norris, Dwight R. (1986), "External Validity of Business Games" *Simulation & Games*, Vol. 17 (December), 447-459.

Patz, Alan L. (2000), "One More Time: Overall Performance in Total Enterprise Simulation Performance" in Page, Diana and Snyder, LT (editors), *Developments in Business Simulation and Experiential Learning*, Volume 27 (Statesboro, GA: Association for Business Simulation and Experiential Learning), 254-258.

Pogossian, Edward M. (1999), "Increasing Efficiency of Management Skill Assessment" in Morgan, Sandra and Page, Diana (editors), *Developments in Business Simulation and Experiential Learning*, Volume 26, 80-81.

Pogossian, Edward M. (1998), "Development of Management Skill Assessment" in Butler, John K., Jr, Leonard, Nancy H., and Morgan, Sandra W. (editors), *Developments in Business Simulation and Experiential Learning*, Volume 25 (Statesboro, GA: Association for Business Simulation and Experiential Learning), 29, 30.

Pogossian, Edward M. (1997), "Ability of Efficient Evaluation of Knowledge-Based Management Strategies" in Butler, John K., Jr. and Leonard, Nancy H. (editors), *Developments in Business Simulation and Experiential Learning*, Volume 24 (Statesboro, GA: Association for Business Simulation and Experiential Learning), 128, 129.

Teach, Richard D. (1993), "Forecasting Accuracy as a Performance Measure in Business Simulations" *Simulation & Gaming*, Vol. 24 (December), 476-490.

Thurstone, L. L. (1927a), "A Law of Comparative Judgment" *Psychological Review*, Vol. 4 (July), 273-286.

Thurstone, L. L. (1927b), "Psychophysical Analysis" *American Journal of Psychology*, Vol. 38, 368-389.

Torgerson, Warren S. (1958), *Theory and Methods of Scaling* (New York: John Wiley & Sons, Inc.)

Wellington, William J., Dickinson, John R., and Faria, A. J. (1990), "The Impact of a Market Leader on Simulation Competitors' Strategies" in Weinroth, Jay and Hilber, Joe E. (editors), *Simulation in Business and Management, 1991* (San Diego, CA: The Society for Computer Simulation), 33-40.

Wolfe, Joseph and Roberts, C. Richard (1986), "The External Validity of a Business Management Game: A Five-Year Longitudinal Study" *Simulation & Games*, Vol. 17 (March), 45-59.