You Name It: Comparing Holistic and Analytical Rating Methods of Eliciting Preferences in Naming an Online Program Using Ranks as a Concurrent Validity Criterion

Michael J. Roszkowski, La Salle University, USA
Scott Spreat, Woods Services, Inc., USA

ABSTRACT

Current and prospective students (n = 133) were surveyed about their preferences for a name for a new online series of courses to be launched by a university. Preferences for each of five names were solicited by means of analytical ratings, holistic ratings, and rankings. All three techniques were employed to assure that the most appropriate name for the program was selected, but this also afforded us the opportunity to study several theoretical issues: (a) Do the different methods lead to discrepant decisions at the aggregate level? (b) Is the holistic rating or the analytical rating approach more closely related to the rankings? (c) To what extent is lack of agreement between ratings and rankings due to lack of differentiation in ratings? The authors find that at the aggregate level all three methods suggest the same name for the program; the holistic rating is slightly more highly correlated with the ranking; and the lack of differentiation in ratings is one reason producing inconsistencies between ratings and rankings.

Keywords: Analytical Rating, Holistic Rating, Name Preferences, Ranking, Rating

INTRODUCTION

The elicitation of preferences typically involves asking the respondent to indicate a choice by either “rating” or “ranking” a set of stimuli. Not only do the relative merits of ratings and rankings continue to be debated, but there is also the ongoing controversy as to whether holistic or analytic ratings work best. One issue that does not seem to have been adequately addressed in the literature, which we explore further, is whether holistic or analytical ratings are more strongly related to rankings. Furthermore, we examine the degree to which non-differentiation in ratings accounts for the lack of agreement between each type of rating and ranking. Finally, we examine whether there is homogeneity of variance in ratings across ranks.
ADVANTAGES AND DISADVANTAGES OF RANKING VERSUS RATING

Inappropriately, the terms rate and rank are sometimes used interchangeably as if they were synonymous, disregarding a fundamental difference. That is, a rating requires one to assign a value to a stimulus using a common scale, whereas a ranking asks one to compare different objects directly to one another by arranging them in some order with respect to some attribute (such as importance, agreement, quality or preference, etc.). Paulhus (1991) identified three types of potential response biases with rating scales: social desirability bias, acquiescence bias, and extreme response bias (i.e., stringency and leniency). The chief virtue of ranking is that the procedure prevents the respondent from failing to differentiate between stimuli due to response styles bias such as acquiescence or extreme response (Baumgartner & Steenkamp, 2001; Berkowitz & Wolkon, 1964; Douceur, 2009; Harzing et al., 2009; Shuman & Presser, 1981; Toner, 1987), but the drawback is that it may force the respondent to artificially differentiate between items that may in fact be viewed as equivalent. Likewise, ranking does not allow for determination of the degree of difference between the objects being compared. Ranking is also a more time-consuming procedure; on average, it takes three times longer to answer a ranking than a rating question (Munson & McIntyre, 1979), although it is argued that the process thereby produces better quality data. According to a review by Krosnick (1999), the improvement in data quality occurs because ranking demands a greater degree of attention and respondents thereby make fewer mistakes when using this answer format.

Overall, Krosnick considers ranks to generally be more reliable and have higher validity with criterion measures in a variety of contexts. Comparisons of the merits of absolute performance appraisals (various rating formats) and relative (various ranking formats) have been the focus of much research in industrial psychology. Generally, relative formats are more valid measures of actual job performance when a “hard” criterion exists, such as sales volume (Goffin et al., 1996; Heneman, 1986; Nathan & Alexander, 1988). Moreover, Hartzig et al. (2009) found rankings to superior over ratings in cross-cultural studies. O’Mahony, Garske, and Klapman (1980) used a signal detection index of difference to determine whether rating or ranking is preferable for identifying differences in food flavors, and report that ranking is superior.

Although ranking is not subject to the acquiescence bias and extreme response bias from which ratings can suffer, ranking is subject to other errors. For one, there is the so called terminal error whereby items appearing first and last on a list are over-ranked in relation to items in the middle of a display (Wagner & Hoover, 1974a, 1974b). Moreover, ranking is context dependent and the ranks assigned to a given stimulus can shift dramatically depending on how many elements are being considered (Krosnick, Thomas, & Shaeffer, 2003), although that criticism may also be true of ratings (cf. Hsee, 1996). If too many items are ranked, low test-retest reliability can result (Krosnick, Thomas, & Shaeffer, 2003; Peng, Nisbett, & Wong, 1997), especially for the lower ranked items (Ben-Akiva, Morikawa, & Shiroishi, 1991). From a statistical perspective, rankings are problematic because they are ipsative scores, meaning that they lack independence since the prior rank determines the possible ranks of remaining ones (Bean & Papadakis, 1994; Dunlap & Cornwell, 1994; Van Deth, 1983). Therefore, conducting a conventional factor analysis on rankings is controversial, but it is possible to do it with other models (Jackson & Alwin, 1980; Cheung, 2004; Hino & Imai, 2008).

AGREEMENT BETWEEN RATINGS AND RANKINGS

Typically, at the aggregate level, rankings and ratings lead to the same conclusion (e.g., Barnard & Ehrenberg, 1990; Driesener & Ro-
maniuk, 2002; Russell & Gray, 1999; Rankin & Grube, 1980; Stillwell et al., 1981) although Herk and van de Velden (2007) have reported that the degree of correlation varies across countries due to culturally-based response tendencies (e.g., Germany=.77, France=.75, Spain=.65, Italy=.64, UK=.62), which other researchers (e.g., Harzing, 2006) have also observed. Moreover, examining the relationship between the holistic rating and the ranking of each of the nine values in Kahle’s (1983) List of Values (LOV) in five EU countries (UK, France, Germany, Italy, and Spain), van Herk and van de Velden (2007) determined that the relationship between ranks and holistic ratings is heteroscedastic, with smaller variance in ratings for the highest and lowest ranked items than for the items with middle ranks. They reported (p. 1102): “Across all countries, people seem to have the most difficulty in assigning scores to the values that they consider to be of medium importance.” For example, when assessing nine values, those ranked at ‘5’, ‘6’ or ‘7’, were given ratings of between ‘4’ and ‘9’. By contrast, respondents appear quite capable of separating the most important values from those considered to be of least importance.”

It has been proposed that disagreement between ranks and ratings may reflect either the respondent’s lack of real preference or it may be due to an unwillingness to differentiate (DeCarlo & Luthar, 2000; Klein, Dülmer, Ohr, Quandt, & Rosar, 2004; Mills, 1991). There is some empirical support for the latter supposition, but the evidence is equivocal. Russell and Gray (2004) found that respondents with a strong concordance between their ranks and their ratings exhibited a larger spread in their ratings, suggesting greater differentiation. Likewise, Krosnick and Alwin (1988) reported that when participants who rated a set of stimuli (values) similarly—termed low differentiators—were excluded from the sample, the ratings correlated more strongly with a conceptually related variable.

In contrast, Lee, Soutar, and Louviere (2007) report that removing low differentiators only had a minimal impact on improving the quality of ratings. They reported that when examining ratings about the importance of nine different values, certain values that conceptually should be negatively correlated were in fact showing positive correlations in the ratings but not in the rankings. For example, the contradictory statements “having security in life” and “having an exciting life” correlated +.37 (p < .01) with each other when rated on a five-point Likert scale, but as logically should be the case, the correlation for these two items was -.22 (p <.05) when the values were ranked. Deleting several classes of respondents who failed to differentiate in their ratings did not change the positive correlation to a negative one, although the magnitude of the positive correlation was lowered. That is, after removing respondents who used the same rating for all nine values, the positive correlation between “having security in life” and “having an exciting life” dropped from +.37 to +.31 (p <.01). Eliminating from the sample the raters who assigned the same rating to eight of the nine values reduced the size of the correlation further, to +.26 (p <.01). Finally, excluding respondents who rated seven of the nine values with the same rating resulted in a correlation of +.24 (p <.01). So while the magnitude of the correlation dropped after removing non-differentiators, the direction of the correlation remained incorrect (i.e., positive rather than negative).

Thomas (2011) studied the impact that non-differentiation in ratings had on validity coefficients in three different Web-based surveys. In the one survey, eliminating non-differentiators did not affect the validity coefficients, and in the second survey it increased validity, but in the third survey it actually lowered validity. Moreover, a study by Maio, Roese, Seligman, and Katz (1996) raises some serious questions about the merits of forcing low differentiators to make the forced distinctions inherent in ranking. This team of researchers asked undergraduates from the University of Western Ontario to rate and rank “values” and the results of these two preference formats were then correlated to conceptually related “attitudes” towards controversial issues (e.g.,
“sanctity of life—legalized abortion” was correlated with “equality-feminism”). Overall, it was reported that ratings produced stronger correlations than rankings between corresponding values and the attitudes. Surprisingly, when the participants were split into high, moderate, and low differentiators (based on the spread in their values ratings) and the correlations of value ranks-to-attitudes were compared to the correlations of value ratings-to-attitudes in each group separately, the validity coefficients of ratings were greater than the ones for rankings in moderate- and low-differentiating groups whereas there was no difference between the validity of ratings and rankings in the highly differentiating group. The conclusion reached by these researchers was that “…value ranking forces people to make distinctions that they would not otherwise make” (p. 179).

Disputes about whether ranking or rating is the more valid technique are unlikely to be resolved any time soon. The search for consensus on whether rating or ranking is the better technique has proven to be elusive. Some researchers espouse the merits of a combined rating-ranking approach (Chiu & Alliger, 1990; Seth, 2004; van Herk & van de Velden, 2007; Lacey et al., 2008; Green, Reid, Passante, & Canipe, 2008; Schubert et al., 2008). For example, if respondents first pick their most and least important values, and then rate them, this leads to greater differentiation between items than a simple rating procedure (McCarty & Shrum, 2000). Lacey et al. shows how differences in rating scale calibration between different groups can be determined by comparing rankings to ratings, and Schubert et al. show the value of using both approaches to validate a scale.

HOLISTIC RATINGS VERSUS ANALYTIC RATINGS

Within the rating approach, a distinction can be drawn between holistic and analytic ratings. With holistic ratings, a single number is assigned to represent one’s overall opinion of a given stimulus rather than considering each feature of the stimulus individually. The holistic approach to rating is sometimes called global, unitary, impressionistic, or intuitive. In contrast, using an analytical approach, one would first rate each of the features, characteristics or components of a stimulus separately and then add or average the individual scores on each element to arrive at the overall impression. Other less frequently used names for the analytical approach include aggregated, disaggregated, atomistic, decomposed, componential, and discrete. The relative merits and drawbacks of these two approaches to rating are debated in contexts such as language assessment (Carr, 2000), essay evaluation (Baca, 2001), second-language writing and speaking assessment (Sawaki, 2007), job analysis (Gibson, Harvey, & Harris, 2007), and physician performance (Daelmans et al., 2005; Williams et al., 2003).

ADVANTAGES AND DISADVANTAGES OF HOLISTIC RATING SCALES

Holistic scales, due to their simplicity, are very efficient in terms of both time and cost (White, 1984), and generally respondents seem to prefer completing holistic instead of analytical rating scales (Arkes et al., 2010). The primary weakness of holistic scales is that they cannot offer any type of diagnostic information, such as an analysis of the strengths and weaknesses of a product, person or service (Bachman, Lynch, & Mason, 1995; Hamp-Lyons, 1995; Roid, 1994). In program evaluation, a related concern is that holistic scales may not be fine-grained enough to capture small changes (Haswell, 1988). Analytic scores are most useful if each sub-score provides some unique information, but in applications such as writing assessment (Barkaoui, 2010; Sawaki, 2007) and the assessment of the English speaking performance of foreign born persons (Xi, 2007), there exist large inter-correlations among the analytic rating dimensions so that little distinct information emerges from the process. Therefore, in some applications, this
supposed advantage of analytical scales may be illusionary.

White (1984) maintains that a holistic rating is typically more reliable than a rating that is derived from the composite of several subscales. While instances of better or equal reliability for holistic ratings have been documented (e.g., Cornelius & Lyness, 1980; Goulden, 1994; Johnson, Fisher, Willeke, & McDaniel, 2003; Lyness & Cornelius, 1982; Nicholson et al., 2009; Olson, 1988; Sanchez & Levine, 1994), generally it is the decomposed measures that tend to be more reliable than their holistic counterparts in a variety of contexts (Arkes et al., 2006; Brown & Bailey, 1984; Butler & Harvey, 1988; Goulden, 1994; Hamp-Lyons, 1995; Klein et al., 1998; Morera & Budescu, 1998, 2001; Murray et al., 2002; Ravinder, 1992; Wiegle, 2002). Morera and Budescu (1998, 2001) attribute the greater reliability to the fact that analytic scales follow the “divide and conquer” principle, taking a complex issue and breaking it down into smaller more manageable parts, which can provide very specific criteria for a rating (Brown & Bailey, 1984).

When the psychometrics of holistic scales is studied, the focus is generally on reliability, especially inter-rater reliability. As noted by Huot (1990, p. 204): “Holistic scoring literature has been dominated by reliability, the procedures used to attain it, and the influences which detract from it.” Holistic methods vary in level of guidance on scoring from impressionistic (no guidelines) to rubric-based wherein a common set of criteria and examples of each type of holistic score are offered (e.g., “anchor papers” in writing assessment). The better the rubric and the training in the use of the rubric, the greater is the expected reliability, especially in terms of inter-rater agreement (Cherry & Meyer, 1993; Elder, Knoch, Barkhuizen, & von Randow, 2005; Huot, 1993; Mishler & Hogan, 1982; McCloy, 1970; Weigle, 1994, 1998, 2002; North & Schneider, 1998). Generally, increasing the number of qualified raters can also improve holistic score reliability (Swartz et al., 1999).

**VALIDITY IS THE BOTTOM LINE CRITERION**

However, the ultimate criterion of whether holistic or analytical ratings are better is not their reliability but rather their validity (Arkes et al., 2010; Charney, 1984; Huot, 1990). In many fields where ratings are used, it is hard to identify a “gold standard” criterion for assessing validity (Morgeson & Campion, 2000; Sanchez & Levine, 2000; Howie, Heaney, & Maxwell, 2004). Consequently, reliability often serves as the primary proxy for validity, under the assumption that reliability places an upper limit on the value of a validity coefficient (Arkes et al., 2006). Although theoretically this is the case, in practice one can find exceptions, such as in study by Johnson, Sallis, and Hovell (1999), where the ranking format for assessing health values showed higher one-week test-retest reliability than the ratings of these same health values, yet the rankings had lower validities than the ratings when compared to self-reported health behaviors. This finding indicates that exceptions can occur to the principle that reliability places a limit on validity and shows the need for validity assessment.

In the absence of a gold standard, agreement between different methods of assessment also serves as a basis for establishing validity (Harvey & Wilson, 2000). Pointing to the difficulty of establishing an appropriate criterion measure, Arkes et al. noted:

“Demonstrating the superiority of one rating procedure over another is particularly difficult in a preference task because there is no obvious criterion. If the disaggregated rating procedure leads me to rank order chocolate over vanilla and vanilla over strawberry, but the holistic procedure leads to reverse ordering, how would it be possible to ascertain which is the ‘correct’ or ‘gold standard’ ordering?” (p. 252)

Their solution to this quandary is to compare the holistic and analytical ratings
against rankings based on a paired comparison procedure:

“We acknowledge that the paired comparison procedure is not a gold standard of preference ordering that the holistic and disaggregated procedures must match in order to be deemed valid. Instead we suggest that the method—holistic or disaggregated—that more closely matches the ordering derived from the paired comparison procedure thereby incurs degree of convergent validity.” (p. 252)

STUDIES USING RANKS AS A VALIDATION CRITERION FOR RATINGS

Thus, one possible criterion for convergent validity is ranking, which some consider to be a superior procedure. Although studies exist using ranking as the criterion for either holistic ratings (e.g., Stellmack et al., 2009; Boor, Wartman, & Reuben, 1983; Castel, Miró, & Rull, 2005) or analytic ratings (e.g., DeCarlo & Luthar, 2000; Obenchain, Abernathy, & Wiest, 2001; LeBlanc, Matson, Cherry, & Bamberg, 1999; DeCarlo & Luthar, 2000; Smith & Barnes, 2007), surprisingly few studies could be found comparing whether analytic or holistic ratings are more strongly associated with rankings.

The study by Arkes et al. (2010) pitted holistic ratings and analytical (which they call disaggregated) ratings against each other in the context of a college preference decision. Arkes et al. solicited ratings about the attractiveness of various colleges, which were described in terms of a standardized set of features. The holistic rating was based on a scale of 1 (not at all attractive) to 100 (very attractive). Two types of analytical ratings were used. The first was a simple (unit weighted) sum score of four features describing each college, where each feature was rated on a ten-point scale on which 1 = poor and 10 = very high. The second analytical rating score was a weighted sum of these four characteristics, with the weight being the importance assigned to each feature; the importance weights were made on a scale of 1 to 100 with the sum of the weights given to the four features being required to add to 100.

To determine the criterion scores, the undergraduates (n = 84) serving as the participants in this study were requested to complete a paired comparison procedure in which eight schools were pitted against each other in the 28 possible combinations. The schools were then ranked for each participant based on the results of the paired comparisons, and these ranks for each person were correlated with the individual’s (a) holistic rating, (b) unweighted analytical rating, and (c) importance-weighted analytical score. The average correlations between the ranks and these three ratings were .71, .76, and .76 respectively. In other words, both types of analytical ratings showed a slight advantage over the holistic ratings, although as in most cases, the weighting did not contribute anything, as is most often the case (Roszkowski & Spreat, 2010).

Johnson, Fisher, Willeke, and McDaniel (2003) conducted an assessment of a program aiming to improve family literacy using portfolios (n = 42) and compared holistic and analytic ratings with ranking. In the analytic scoring system, the six program goals were each rated on a five-point scale (with verbal descriptors at the extremes and the mid-point of the scale) and were then summed. The holistic ratings were based on rubrics describing four levels of family literacy: proficient (presence of most of the family literacy skills), developing (development of some, but not all, of those skills), emerging (development of a few family literacy skills), and not yet (no evidence of development of such skills). Prior to completing the holistic and analytical ratings, the family educators were also asked to rank order the participants according to their perceived family literacy levels. The inter-rater reliability (index of dependability) was .74 for the analytic method and .79 for the holistic method when the estimates were based on the average of two raters’ scores. The correlation between the analytic and holistic methods of scoring was high (.93) and their correlations...
with the ranks were .59 and .50 for the analytic and holistic methods, respectively.

Mackeigan, O’Brien, and Oh (1999) compared the congruence of composite and holistic preference measures for four methods of treating diabetes over the course of a 30 year period with the rank-order preference of these four options. Treatment of diabetes is generally sequential, starting with just a diet and exercise regimen, progressing to oral administration of a single drug, then to a combination of oral drugs, and finally to injection of insulin. The four methods varied in terms of the drugs used and the duration of oral drug monotherapy, dual therapy, triple therapy, and insulin. For each method, information was given about the required frequency of drug administration and blood glucose testing, the drug’s ability to control glucose, and its side effects.

The analytical rating for each of the four treatment option involved eliciting a preference weight for each of the components of the given treatment, multiplying the preference by its duration in years, and summing these products. Under the holistic approach, the respondent was asked to rate the value of entire sequence of health-state paths through time specified under each of the four treatment options. The correlations between the analytical and holistic approaches to rating the four options resulted in the following correlations: .74, .76, .75, and .82. However, the agreement was low between the rankings of the four treatment options and the ratings of these same options using either the holistic or the analytical approach. In terms of a weighted kappa, the associations between the holistic ratings and the rankings were .06, .08, .07, and .15. The corresponding analytical rating-to-ranks associations were also meager: .17, .06, -.01, and .17.

In a study by Marzano (1975), six essays were ranked by three university professors who taught composition, and their ranks were averaged and correlated with ratings of the same essays provided by 8 college seniors studying to become secondary English teachers. Half made holistic ratings and half analytical ratings. The students’ holistic ratings correlated .80 with the professors’ ranks but the analytical ratings correlated .47 with these same ranks. The higher correlation for the holistic method occurred even though it has a lower inter-rater reliability than the analytical method (.70 vs. .59). Marzano considered the professor’s rankings as the criterion (Kendall’s Coefficient of Concordance =.85), and therefore concluded that the holistic method was more valid despite being less reliable.

CONVERGENCE BETWEEN HOLISTIC AND ANALYTICAL RATINGS

Good convergence between holistic ratings and analytical ratings has been reported on tasks such as the ratings of the skills of medical students and residents, which resulted in a correlation of .71 (Murray et al., 2002). Olson (1988) found a correlation of .66 between the analytic and holistic methods of scoring students’ computer programs in an introductory computer science class, and Klein et al. (1998) reported correlations of .71 at grade 5 and .80 at grade 8 between analytic and holistic ratings of elementary school science projects. Vacc (1989) computed correlations ranging between .56 and .80 in the analytical and holistic evaluation of elementary school writing. Goulden (1994), who examined the holistic and analytic ratings of speech by classroom teachers, also found evidence for concurrent validity. However, according to some reports (Butler & Harvey, 1988; Harvey et al., 1994; Gibson, Harvey, & Harris, 2007) there is poor convergence between holistic and analytical methods of job analysis.

REASONS FOR LACK OF AGREEMENT BETWEEN HOLISTIC RATINGS AND AGGREGATED RATINGS

A good deal of the disagreement between analytical ratings and holistic ratings is random measurement error. Aggregated ratings generally have higher reliabilities than single
item holistic ratings, which would mean that other things being equal, they should correlate more highly with a given criterion. However, a potential problem with analytical measures is that the elements used to create the aggregated scores, when imposed, may not be relevant to the rater. As Karapanos, Martens, and Hassenzahl (2009, p. 640) observe, “…different individuals may form different evaluative judgments even while having no disagreement on the perceived quality of the product, e.g. both might think of it as a novel and hard-to-use product, but they disagree on the relative importance of each quality. In extreme cases, individuals might even use entirely different attributes to evaluate a product, reflecting the qualities they consider important for the specific product being evaluated.” A similar problem can occur with holistic ratings since the criteria for the rating are implicit, so idiosyncrasies in the selection of criteria and their mental weighting are bound to exist (Barkaoui, 2010; Brown, 1995; Eckes, 2008; Goulden, 1994). Moreover, Arkes et al. (2010) indicate that holistic ratings are subject to “gaming”, whereby raters use inappropriate criteria for the rating, despite instructions to not base their ratings on these criteria.

RATIONALE FOR PRESENT STUDY

The selection of an appropriate name for a product or service is a critical step in the branding process, particularly in the net economy (Kollman & Suckow, 2007). The present study compares the relationship of holistic ratings and analytical ratings to rankings in the context of measuring a preference for a name of an online educational product. It further examines the degree to which non-differentiation is a factor in attenuating the correlation between rankings and ratings, looking at both analytical and holistic ratings from the same sample. Lastly, we seek to corroborate van Herk and van de Velden (2007) finding that the relationship between rankings and ratings is heteroscedastic, such that ratings for middle ranked items have greater variance than the ratings for the highest and lowest ranked items. The study by van Herk and van de Velden (2007) considered only holistic ratings whereas we look at both holistic and analytical ratings.

METHOD

Participants

The participants were respondents to an Internet-based survey. Participation was solicited from current and prospective part-time undergraduate students and graduate students at one university. An e-mail invitation containing the address to a general (non-individualized) link to an Internet-based survey (Survey Monkey) was sent to 2,619 potential respondents. There were 167 respondents, which translates to a response rate of about 6.3%, but only 133 answered all 8 questions of relevance to this study. The respondents were about equally split between prospective (43%) and current students (57%), and were mainly at the graduate (84%) level.

Instrumentation

The data were drawn from the survey. The introduction to the survey was as follows: “La Salle University is selecting a name for identifying and marketing its online academic programs. The name will be applied to a new approach to online education that offers a richer multimedia experience within online courses. The University is interested in current and prospective student feedback on a list of proposed names for this new online program hub. The names under consideration are: La Salle World Class, La Salle University Online, Electric Exploring, iLearn.La Salle, Online Explorer. We would like your feedback on each of these potential names. Thank you in advance for your participation in this survey.”

Among other things the survey contained the following questions in the sequence shown.

1. Please evaluate each name on how easy it is to remember: 1=very hard, 2= somewhat hard, 3= somewhat easy, 4= very easy.
Table 1. Mean preference for the names as assessed by holistic ratings, analytical ratings, and ranks

<table>
<thead>
<tr>
<th>Name</th>
<th>Holistic</th>
<th>Analytical</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>World Class</td>
<td>2.47</td>
<td>48.58</td>
<td>3.02</td>
</tr>
<tr>
<td>iLearn.La Salle</td>
<td>3.25</td>
<td>70.09</td>
<td>1.96</td>
</tr>
<tr>
<td>Electric Exploring</td>
<td>1.74</td>
<td>35.74</td>
<td>4.18</td>
</tr>
<tr>
<td>La Salle University</td>
<td>2.85</td>
<td>64.22</td>
<td>2.38</td>
</tr>
<tr>
<td>Online Explorer</td>
<td>2.13</td>
<td>42.04</td>
<td>3.46</td>
</tr>
</tbody>
</table>

2. How distinctive is each name (i.e., compared to online programs offered by other colleges and universities)?: 1=not at all, 2= somewhat, 3= quite

3. In your opinion, to what extent does each name convey what the product is all about?: 1=not at all, 2= somewhat, 3= quite a bit

4. How innovative (original) is each name ?: 1=not at all, 2= somewhat, 3= very

5. How self-explanatory is each proposed name?: 1=not at all, 2= somewhat, 3= quite

6. How well does the name fit with the La Salle brand ?: 1= not all, 2= somewhat, 3= very

7. What is your overall opinion of each name?: 1= dislike very much, 2= somewhat dislike, 3= somewhat like, 4= like very much

8. Please rank the names in order of your preference, from first choice to last choice.

The sum of questions 1 to 6 constitutes the Analytical Rating, which was calculated as a simple sum score but in percent of maximum units, as recommended by Cohen, Cohen, Aiken, and West (1999). Question 7 is the Holistic Rating and question 8 captures the Rank.

RESULTS

The internal consistency reliability of the sum of the analytical ratings for each name was: World Class=.83, iLearn.La Salle =.81, Online Explorer =.80, Electric Exploring =.72, La Salle University Online=.60.

The average holistic ratings, analytical ratings and ranks for each name are shown in Table 1. It is clear that on the basis of grouped data, the order of preference for the names is identical under all three methods of eliciting a preference: La Salle Online, iLearn, World Class, Online Explorer, and Electric Exploring. In fact, at the group level, the holistic rating –rank correlation is 1.00, the holistic rating-analytical rating correlation is .99, and the analytical rating-rank correlation is .98.

Table 2 deals with outcomes of analyses at the individual level, reporting the correlations between (a) the rank of the name and its holistic rating, and (b) the rank of the name and its analytical rating. Also shown are the results of the test for the statistical significance of the difference between the Rank-Holistic Rating correlation versus the Rank-Analytical Rating correlation for each of the five proposed names. In view of the controversy about whether it is proper to use the Pearson correlation with ordinal data, for each relationship, two correlations are reported: Spearman and Pearson. Descriptively, the holistic rating has the higher correlation with the rank in all five instances using the both the Pearson and the Spearman correlation procedures. These differences reach statistical significance in the case of “La Salle University Online” and “Online Exploring.” (While correlations are reported to the second decimal place, values to the third place were used in the testing for differences between correlations in order to increase precision of the t-test.)

The correlations for each name were averaged using Fisher z transformations. The average Pearson correlation for the rank-holistic rating association is .69 and for the rank-analytical rating association it equals .63. In terms of the Spearman procedure, the average rank-holistic rating correlation is again .69 whereas the average rank-analytical rating correlation...
is .59. Thus, on average, the holistic rating can explain about 8% more of the variance in name ranks based on the Pearson procedure and nearly 13% more based on the Spearman.

The reason why the analytic ratings of the names are not as highly related to the ranks could be due to several factors: (1) the analytical rating constitutes a sum score of facets that do not enter into the formulation of either the ranking or the holistic rating, or (2) other features besides the ones specified in the analytical sum score are the basis for the holistic ratings and ranks. To check for the first possibility the rating of each facet for each of the five names was correlated to its (a) analytical sum rating, (b) the holistic rating, and (c) the rank of that name. To avoid inflated analytical correlations, the analytical sum excluded the particular facet (i.e., corrected item-total correlations were computed). Based on the similar results produced by the Pearson and Spearman procedures, Table 3 only reports on the Pearson correlations. The average facet-based correlations were .51, .52, and .46 for the analytical ratings, holistic ratings, and the ranks, respectively. In other words, the explanation that the specified facets were not a consideration in the choice is not a strong possibility, although something else may have been used as well. The second one is more likely.

Next, we assessed the extent to which non-differentiation accounts for the attenuation of the correlation between the ratings and rankings. To do this, we computed an intra-individual standard deviation of the analytical scores (sums) that each respondent gave to each of the five names. In the sample of 133, the intra-individual standard deviation had a mean of 25.04 and a median of 24.92. The sample was then split on the median into low differentiation ($n = 66$) and high differentiation ($n = 67$) groups. For example, suppose an individual had the following analytical rating scores for the five names: 18, 42, 16, 83, and 91; the standard deviation of these scores would be: 34.04, which would place that individual into the high differentiator group. On the other hand, if the scores were 38, 22, 30, 15, and 35 then the standard deviation would be 9.28 and that person would fall into the low differentiator category.

We next correlated the ranks and analytical sum scores in each of the two groups separately,
Table 3. Pearson correlations of facets to analytical rating, holistic rating, and rank for each name preference

<table>
<thead>
<tr>
<th>World Class</th>
<th>Analytical</th>
<th>Holistic</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>easy it is to remember</td>
<td>.58</td>
<td>.62</td>
<td>.56</td>
</tr>
<tr>
<td>distinctive</td>
<td>.59</td>
<td>.52</td>
<td>.52</td>
</tr>
<tr>
<td>conveys what product is about</td>
<td>.62</td>
<td>.67</td>
<td>.60</td>
</tr>
<tr>
<td>innovative (original)</td>
<td>.51</td>
<td>.59</td>
<td>.44</td>
</tr>
<tr>
<td>self-explanatory</td>
<td>.67</td>
<td>.59</td>
<td>.56</td>
</tr>
<tr>
<td>fit with brand</td>
<td>.64</td>
<td>.65</td>
<td>.56</td>
</tr>
<tr>
<td>La Salle University Online</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>easy it is to remember</td>
<td>.30</td>
<td>.52</td>
<td>.26</td>
</tr>
<tr>
<td>distinctive</td>
<td>.37</td>
<td>.32</td>
<td>.31</td>
</tr>
<tr>
<td>conveys what product is about</td>
<td>.29</td>
<td>.29</td>
<td>.21</td>
</tr>
<tr>
<td>innovative (original)</td>
<td>.52</td>
<td>.37</td>
<td>.23</td>
</tr>
<tr>
<td>self-explanatory</td>
<td>.14</td>
<td>.16</td>
<td>.04</td>
</tr>
<tr>
<td>fit with brand</td>
<td>.42</td>
<td>.54</td>
<td>.33</td>
</tr>
<tr>
<td>Electric Exploring</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>easy it is to remember</td>
<td>.34</td>
<td>.46</td>
<td>.57</td>
</tr>
<tr>
<td>distinctive</td>
<td>.40</td>
<td>.34</td>
<td>.42</td>
</tr>
<tr>
<td>conveys what product is about</td>
<td>.54</td>
<td>.42</td>
<td>.62</td>
</tr>
<tr>
<td>innovative (original)</td>
<td>.42</td>
<td>.36</td>
<td>.45</td>
</tr>
<tr>
<td>self-explanatory</td>
<td>.52</td>
<td>.48</td>
<td>.57</td>
</tr>
<tr>
<td>fit with brand</td>
<td>.59</td>
<td>.68</td>
<td>.59</td>
</tr>
<tr>
<td>iLearn.La Salle</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>easy it is to remember</td>
<td>.51</td>
<td>.67</td>
<td>.57</td>
</tr>
<tr>
<td>distinctive</td>
<td>.48</td>
<td>.41</td>
<td>.42</td>
</tr>
<tr>
<td>conveys what product is about</td>
<td>.68</td>
<td>.61</td>
<td>.62</td>
</tr>
<tr>
<td>innovative (original)</td>
<td>.53</td>
<td>.57</td>
<td>.45</td>
</tr>
<tr>
<td>self-explanatory</td>
<td>.62</td>
<td>.67</td>
<td>.57</td>
</tr>
<tr>
<td>fit with brand</td>
<td>.68</td>
<td>.70</td>
<td>.59</td>
</tr>
<tr>
<td>Online Explorer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>easy it is to remember</td>
<td>.39</td>
<td>.46</td>
<td>.33</td>
</tr>
<tr>
<td>distinctive</td>
<td>.44</td>
<td>.45</td>
<td>.45</td>
</tr>
<tr>
<td>conveys what product is about</td>
<td>.73</td>
<td>.63</td>
<td>.50</td>
</tr>
<tr>
<td>innovative (original)</td>
<td>.48</td>
<td>.51</td>
<td>.50</td>
</tr>
<tr>
<td>self-explanatory</td>
<td>.63</td>
<td>.54</td>
<td>.43</td>
</tr>
<tr>
<td>fit with brand</td>
<td>.70</td>
<td>.77</td>
<td>.60</td>
</tr>
</tbody>
</table>
using both the Pearson and Spearman techniques. The resulting correlation coefficients are shown in Table 4. Descriptively, for each of the five names, the correlation of the rank with the analytical rating was higher in the high differentiation group than the corresponding correlation in the low differentiation group. For one of the names (iLearn.La Salle) the difference in both the Pearson and Spearman correlations in these two groups was large enough to reach statistical significance. Also, statistical significance was reached on the difference between the two Spearman correlations in the case of a second name, and the difference on the basis of the Pearson was close to significance. Furthermore, the difference in the correlations of rank with analytical rating among low versus high differentiators was near significance in the case of a third name. More telling may be the average correlations in the two groups across the five names: (a) Pearson r of .55 in the low differentiator group and .73 in the high differentiator group, (b) Spearman rho of .51 for the low differentiators and .69 for the high differentiators.

We also conducted the same dichotomization into low and high differentiators based on the holistic ratings. The intra-individual standard deviation of the five holistic ratings ranged from .45 to 1.64, with a median of 1.14 and a mean of 1.11. The split on 1.14 resulted in a somewhat unequal grouping, with 57 cases in the low differentiator group and 77 cases in the high differentiator group. The Pearson and Spearman correlations between ranks and holistic ratings in these two groups appear in Table 5.

In every instance, the value of the correlation coefficient in the high differentiator group exceeded that of the corresponding one in the low differentiator group, although not all differences achieved conventional levels of statistical significance. Averaging across the correlations for the five names, the Pearson correlation between rank and holistic rating is approximately .53 for the low differentiators and .78 for the high differentiators. On the

<table>
<thead>
<tr>
<th>Name</th>
<th>Type of Correlation</th>
<th>Low Intra-individual SD (n=66)</th>
<th>High Intra-individual SD (n=67)</th>
<th>z-test for Difference between Independent Correlations</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>World Class</td>
<td>Pearson r</td>
<td>.65</td>
<td>.81</td>
<td>1.89</td>
<td>.059</td>
</tr>
<tr>
<td></td>
<td>Spearman rho</td>
<td>.62</td>
<td>.81</td>
<td>2.28</td>
<td>.023</td>
</tr>
<tr>
<td>La Salle University Online</td>
<td>Pearson r</td>
<td>.30</td>
<td>.56</td>
<td>1.89</td>
<td>.059</td>
</tr>
<tr>
<td></td>
<td>Spearman rho</td>
<td>.32</td>
<td>.58</td>
<td>1.83</td>
<td>.067</td>
</tr>
<tr>
<td>Electric Exploring</td>
<td>Pearson r</td>
<td>.55</td>
<td>.66</td>
<td>0.99</td>
<td>.321</td>
</tr>
<tr>
<td></td>
<td>Spearman rho</td>
<td>.37</td>
<td>.44</td>
<td>0.51</td>
<td>.608</td>
</tr>
<tr>
<td>iLearn.La Salle</td>
<td>Pearson r</td>
<td>.60</td>
<td>.81</td>
<td>2.46</td>
<td>.014</td>
</tr>
<tr>
<td></td>
<td>Spearman rho</td>
<td>.62</td>
<td>.85</td>
<td>2.94</td>
<td>.003</td>
</tr>
<tr>
<td>Online Explorer</td>
<td>Pearson r</td>
<td>.61</td>
<td>.73</td>
<td>1.29</td>
<td>.198</td>
</tr>
<tr>
<td></td>
<td>Spearman rho</td>
<td>.55</td>
<td>.65</td>
<td>0.86</td>
<td>.391</td>
</tr>
<tr>
<td>Average</td>
<td>Pearson r</td>
<td>.55</td>
<td>.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spearman rho</td>
<td>.51</td>
<td>.69</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. POMP based (item level) correlation between name preference rank and analytical rating of name preferences as a function of dichotomization (median-split) on intra-individual SD.
basis of the Spearman, the average correlations are .51 and .78, respectively. Comparing the advantage of the high differentiator group over the low differentiator group in terms of the rank-holistic rating correlation versus the rank-analytical rating correlation, one finds the difference to be somewhat greater in the case of the holistic ratings.

Finally, we tested the proposition that variance in ratings between subjects is greater in the middle ranks than in the extreme ranks. As may be observed in Table 6, the lowest variance on both the holistic and analytical ratings occurred when the name was ranked as first choice (rank of 1). Descriptively, the variance seems to increase progressively until one reaches the least preferred choice (rank of 5), where it drops. The procedure developed by Pittman (1939) was used to evaluate pairwise differences in homogeneity of variance as a function of rank. The variance in holistic ratings differed significantly by rank in four comparisons: rank 1 vs. rank 4 \([t(2)=4.65, p =.000]\), rank 1 vs. rank 4 \([t(2)=2.88, p =.005]\). None of the differences in variance by rank on the analytical ratings reached significance, although the difference between rank 1 and rank 3 had a fairly low probability on a two-tailed test \([t(2)=1.65, p =.101]\).

### DISCUSSION AND IMPLICATIONS

A primary purpose of any method of ordering items along some dimension is to support decision making. In this particular case, the decision to be made was the selection of a name for a new online program at a university. Selecting a proper name for a product or service is crucial since it an essential ingredient for branding, which many colleges are considering as an approach to marketing their services (Chapleo, 2011). Five names were under consideration. To elicit preferences for these, we used three approaches: (a) holistic rating, (b) analytical rating, and (c) ranking.

### Table 5. Correlation between name preference rank and holistic rating of name preferences as a function of dichotomization (median-split) on intra-individual SD

<table>
<thead>
<tr>
<th>Name</th>
<th>Type of Correlation</th>
<th>Low Intra-individual SD ((n=56))</th>
<th>High Intra-individual SD ((n=77))</th>
<th>z-test for Difference between Independent Correlations</th>
<th>(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>World Class</td>
<td>Pearson r</td>
<td>.52</td>
<td>.85</td>
<td>3.72</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>Spearman rho</td>
<td>.51</td>
<td>.84</td>
<td>3.65</td>
<td>.000</td>
</tr>
<tr>
<td>La Salle University Online</td>
<td>Pearson r</td>
<td>.60</td>
<td>.74</td>
<td>1.45</td>
<td>.146</td>
</tr>
<tr>
<td></td>
<td>Spearman rho</td>
<td>.63</td>
<td>.84</td>
<td>2.65</td>
<td>.008</td>
</tr>
<tr>
<td>Electric Exploring</td>
<td>Pearson r</td>
<td>.48</td>
<td>.61</td>
<td>1.07</td>
<td>.283</td>
</tr>
<tr>
<td></td>
<td>Spearman rho</td>
<td>.33</td>
<td>.41</td>
<td>0.51</td>
<td>.612</td>
</tr>
<tr>
<td>iLearn.La Salle</td>
<td>Pearson r</td>
<td>.72</td>
<td>.85</td>
<td>1.95</td>
<td>.051</td>
</tr>
<tr>
<td></td>
<td>Spearman rho</td>
<td>.77</td>
<td>.87</td>
<td>1.86</td>
<td>.063</td>
</tr>
<tr>
<td>Online Explorer</td>
<td>Pearson r</td>
<td>.60</td>
<td>.78</td>
<td>1.87</td>
<td>.062</td>
</tr>
<tr>
<td></td>
<td>Spearman rho</td>
<td>.60</td>
<td>.76</td>
<td>1.67</td>
<td>.096</td>
</tr>
<tr>
<td>Average</td>
<td>Pearson r</td>
<td>.53</td>
<td>.78</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spearman rho</td>
<td>.51</td>
<td>.78</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The practical reason for using multiple approaches was to make certain that the correct decision was reached about what to name the online educational program, but this methodology afforded us the opportunity to study whether analytical or holistic ratings were more congruent with rankings, an issue that has not been addressed adequately in the literature, especially in the context of esthetic preferences, where a “gold standard” criterion for establishing validity does not exist. The primary limitation of our study is that the analytical and holistic ratings did not use the same scales. That is, the aggregated ratings were based on whether the name possessed pre-ordained desirable features whereas the holistic ratings solicited opinion about whether the respondent liked or disliked the name. Also, the response rate was disappointing, but there is no reason to believe that it represents some type of bias.

The ranking procedure was selected as the criterion measure in this validity study. One could question whether ranking is an appropriate yardstick against which ratings should be tested because it forces people to make choices. However, as noted by Olsen et al. (2005), in zero sum contexts such as priority setting, decision makers must make a choice, even if the differences between options are minor. Although some researchers contend that ranking is the more valid approach to measuring preferences, this premise does not need to be accepted in order to make sense of our results. We are simply treating ranking as a third measure against which to compare the two types of ratings (holistic versus analytical).

For the reader, the major implication to be drawn is that one method will suffice since the three different approaches led to a common final outcome. Whether selected via the holistic method, the analytical method, or the ranking method, the same name was identified as the number one choice. Not only was the same name selected as the first choice by each method, but each method also identified exactly the same ordering of the five potential names by preference. Equal outcomes, in this case 100% agreement, strongly suggest an equivalence across the three methods of ordering items at the aggregate level.

Lending further credence to the concurrent validity of three approaches to subjectively ordering items along a common dimension are the intercorrelations among the three approaches at the individual level. These values suggest acceptable levels of concurrent validity and indicate that the three approaches share considerable amounts of variance. While discrepancies exist between choices at the individual level based on method, given the concordance at the aggregate level, the methods may be interchangeable when decisions are to be made on the basis of aggregated (group) data.

Although both holistic and analytic approaches to rating demonstrated acceptable levels of concurrence with the ranking procedure, the holistic approach generally yielded slightly stronger correlations. Comparing the magnitude of the relationship between analytic ratings and the rankings to the one between holistic ratings and rankings, we found that the correlations based on holistic ratings exceeded

<table>
<thead>
<tr>
<th>Rank</th>
<th>Holistic Rating</th>
<th>Analytical Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>1</td>
<td>3.83</td>
<td>0.39</td>
</tr>
<tr>
<td>2</td>
<td>3.11</td>
<td>0.71</td>
</tr>
<tr>
<td>3</td>
<td>2.26</td>
<td>0.84</td>
</tr>
<tr>
<td>4</td>
<td>1.83</td>
<td>0.85</td>
</tr>
<tr>
<td>5</td>
<td>1.41</td>
<td>0.64</td>
</tr>
</tbody>
</table>
those of analytic ratings with all five names under both Pearson and Spearman procedures (although statistically significant differences were detectable on only three instances of the 10 correlations).

The reason why the analytic ratings of the names are not as highly related to the ranks as are the holistic ratings could be due to a couple of factors. One might first suspect that the correlations involving the holistic ratings exceeded the ones involving the analytic ratings because the analytic rating procedure created a sum score using attributes that did not enter into the formulation of either the holistic rating or the rank. However, this argument is somewhat tempered by the observation that our study revealed a similar pattern of association among the facets used to construct the analytical sum score and each of the three approaches to determining a preference.

It thus appears more likely that additional features besides the specifically delineated facets went into the both holistic rating and the determination of rank, but we do not know what these are. Nonetheless it seems reasonable to recommend that for the type of task used in our study, the parsimonious procedure for assessing preferences is the holistic rating method, unless (a) information is desired about the specific facets that comprise an analytic rating, and (b) all factors used in the decision are part of the analytic sum score.

Making subjective holistic ratings is easy and economical, and our study suggests that the holistic approach has greater concurrent validity with a ranking procedure, which some consider to be the best basis for eliciting preferences. Given the discrepancies across studies, however, perhaps the type of task needs to be considered closely when deciding on which technique is optimal for eliciting a preference. The redundancy in our project did not create that much more work for the respondent, and the congruence of results gave us reassurance about our decision on what to name the online program. To the benefits identified by others of employing both ratings and rankings, we would add a greater sense of confidence in one’s decision if all techniques point to the same conclusion. In other contexts, there may be considerable differences in cost between the three approaches and the redundancy may create some frustration and fatigue, which could result in lower response rates if the topic salience is low (Roszkowski & Bean, 1990). Moreover, our study was based on an aesthetic preference. It would be of interest to assess the concurrent validity of a holistic rating procedure with a more objective quantification of behavior.

Another possibility is that the correlation of the holistic rating with the rank is higher because ranking is conceptually more similar to holistic rating than to analytic rating (cf. Huot, 1990, p. 208). Mentally, ranking in reality may be a two step process. First, the objects to be ranked need to be rated or evaluated on some basis (implicitly if not explicitly) and then they need to be arranged in order on the basis of each stimulus having the least to the most of the given characteristic. The evaluation preceding the ranking is probably more global than analytical in nature.

A secondary research issue we addressed has to do with the impact of responder ability or willingness to make distinctions among their ratings. The literature shows that the impact of differentiation varies across studies, suggesting that perhaps it depends on some other factors, which remain unknown at present. In this particular study of naming preferences, we found that across each of the five studied name options, the correlation derived from the group with high differentiation in ratings exceeded the correlation derived from the low differentiation group. Our results therefore lend support to the notion that lack of differentiation on the part of some raters is a reason for attenuated correlations between ranks and ratings.

REFERENCES


