AC 2010-1665: QUALTIY AND CONSISTENCY IN IDEA PITCH, RESEARCH PROPOSAL AND BUSINESS PLAN COMPETITION JUDGING

Daniel Ferguson, Ohio Northern University

Daniel Ferguson is Assistant Professor of Entrepreneurship at Ohio Northern University. Prior to coming to Ohio Northern University he was Associate Director of the Inter-professional Studies Program at Illinois Institute of Technology and Co-PI on multiple National Science Foundation grants relating to assessment processes and interventions aimed at improving learning objective attainment. Prior to his University assignments he was the Founder and CEO of the The EDI Group, Ltd., an independent professional services company specializing in B2B electronic commerce and electronic data interchange; and a Vice President at the First National Bank of Chicago, where he founded and managed the bank's market leading professional Cash Management Consulting Group and initiated the bank's non credit service product management organization and profit center profitability programs.

Michele Govekar, Ohio Northern University

Michele A. Govekar, Ph.D.

Michele is a Professor of Management at Ohio Northern University.

Her academic research and publications focus on the management and history of US firms' international operations, corporate- nonprofit interactions, nonprofit organizations, and the process, outcomes and assessment of management education.

She currently teaches undergraduate courses in strategy, corporate citizenship, small business and project management. She is the co-program chair of the North American Management Society 2009 conference and past Division Chair of the Management History Division of the Academy of Management for 2005-06.

Amanda Stype, Ohio Northern University

Amanda Stype is a 2009 graduate of the James F. Dicke College of Business Administration at Ohio Northern University with an Honors degree in International Business & Economics and also in applied mathematics. She currently is a graduate student at Bowling Green State University in their Master of Economics program. She plans to pursue a Ph.D. in applied microeconomics in the near future.

Title: Quality and Consistency in Idea Pitch, Research Proposal and Business Plan Competition Judging.

Submitted by

Daniel M. Ferguson, Michele A. Govekar and Amanda C. Stype

Submitted by Daniel M. Ferguson, Michele A. Govekar, and Amanda C. Stype   Thursday, April 01, 2010

Abstract: The results of Entrepreneurial Idea Pitch and Research Proposal Competitions often determine the award of cash prizes [e.g. $100,000 at MIT] and scarce resources. The recipients of these awards are determined by judging processes. These judging processes are rarely audited or evaluated as to quality or consistency. We contend that judging processes will be more fair and perceived as less subjective with a high level of consensus between judges [interrater agreement], especially for those ranked as best. Our research calculates a_{WG} for idea pitch competitions, identifies interventions that improved interrater agreement over time including external factors that may support higher quality judging processes. We seek participation in a collaborative application to continue the research.

2

Introduction:

If a student or innovator has a business plan that is judged to be viable and competitive, this can be seen as a predictor of future success. However, the process to determine viability of plans is uncertain and often perceived as overly subjective. This past year we collected data from Idea Pitch Competitions at Business, Engineering, Pharmacy and Arts and Sciences colleges. We identified interventions or conditions that, either cited in the literature or from practical experience, were present or applied in these judging situations. We used the a_{WG} statistical measurement techniques identified by Brown and Hauenstein (2005) to calculate consensus among the judges in initial and final rounds. We identify interventions that appear to improve consensus over time across assessment or competition situations. Finally we invite collaborative efforts to test individual or combinations of interventions that offer the highest levels of consensus and continuous improvement in a_{WG}. Lack of a high level of interrater agreement can indicate poor judging and a weak judging process, a professional issue of some importance. Very often judges volunteer to be part of the judging process but possess varying degrees of knowledge or expertise regarding the outcomes or knowledge being judged. When organizing assessments or competitions, effectively executing the event is the main goal and there may be no formal attempt to measure interrater agreement or to improve the level of consensus, ie. interrater agreement, in the assessments or competitions over time. Our goal is to provide evidence supporting the use of interventions that do, in fact, improve interrater agreement as well as the perceived fairness of the event.

In competition and proposal assessment situations, it is important to measure consensus between judges. Measuring consensus between judges can expose problems within the judging process such as a difference in expectations between judges. Because many times the outcome of these assessments and competitions greatly affects the participant, whether it is in the form of determining someone's grade, a person who receives an award, or a person who receives further funding, it is important that those organizing and heading these competitions or assessments seek to make the situation as fair as possible. In most competitions, each presenter or group is rated by a different group or panel of judges, which makes a high level of evaluation consensus

3

between judges critical to the outcome of the competition being valid and fair. We believe that measuring consensus between judges on a criterion by criterion basis is the appropriate process to determine the quality of a judging process. Also this method identifies which criteria are causing significant differences between judges which provide an opportunity to improve the judging consensus over time.

Body: Consensus between judges is more commonly known as interrater agreement. Interrater agreement is measured in a variety of ways, depending on the number of judges (2 or more than 2) and the nature of the criteria (dichotomous or categorical). When there are multiple judges using a greater than 2 point scale, there are two commonly used ways to measure interrater agreement.

The first of these is the r_{WG} index developed by James, Demaree and Wolf (1984). Although this method is the more common, it has several drawbacks including scale dependency, the assumption that a uniform distribution models perfect disagreement, the need for a distribution to model disagreement, and dependence on the number of judges (Kozlowski and Hattrup 1992, Brown and Hauenstein 2005). The a_{WG} index developed by Brown and Hauenstein corrects for these problems and measures consensus among judges. It calibrates itself according to the scale and the number of judges. The equation for a_{WG} is:

$$a_{WG} = 1 - 2*S_x^2/\{[(H+L)*X-X^2-(H*L)]*[K/(K-1)]\},$$

where H is the upper bound of the rating scale, L is the lower bound of the rating scale, X is the average rating for the particular contestant for one criteria, S_x^2 is the variance for the results for one particular contestant for one criteria, and K is the number of judges.

Changes made in the judging process to help increase consensus between judges and therefore interrater agreement are called interventions. Some interventions have already been tested and shown in the literature to improve interrater agreement. Other possible interventions to improve interrater agreement have been suggested by judges and participants in post-event feedback sessions for the events observed in this case study.

Interventions

4

Interventions can be divided into three categories. These are: assessment/judging format interventions, judge-related interventions, and other interventions. These categories are not mutually exclusive, and some interventions may both involve the judging format and be judge-related. See Table 1 for the interventions in our case study.

Rubrics

One form of intervention is mandating the use of rubrics. There is significant research into rubric design and rubric construction, as well as many different definitions of what constitutes a rubric. For our purposes, we define a rubric as an assessment tool which defines the characteristics of a behavior that is associated with numerical levels. An example of a rubric that fits our definition of rubric is the meta-rubrics being developed by the Association of American Colleges and Universities for various forms of academic assessments.

There are many different interventions related to rubrics. First, there is a difference between having a rubric and not having a rubric. Sometimes assessors are asked to score a presentation on a scale, say from 1-10; this is not a rubric. In these cases, assessors are not told what criteria to base their resulting assessment on. A rubric explicitly states what behavior or evidence one should be looking for. If criteria are to be weighted uniformly or differently, the rubric should also make this clear. Furthermore, a rubric as we are defining it, will explicitly describe what behaviors or characteristics of the presentation/presenter are associated with each possible numerical outcome.

It has been shown that rubrics on a two point scale (yes/no), also known as check sheets, as opposed to a 3 point scale (yes/no/partial) show higher interrater agreement (Huber, Baroffio, Chamot, Herrmann, Nendaz, & Vu, 2005). Although a two point scale is not practical for many criteria, it may be useful for specific criteria in certain situations. It has also been shown that having the judges or assessors involved in rubric development increases agreement by increasing the judge's understanding of the criteria.(Huber, Baroffio, Chamot, Herrmann, Nendaz, & Vu, 2005). The input of those who will use the rubric allows for the opportunity to decrease ambiguous wording within the rubric and increase the understanding that assessors have of each

5

criterion. Agreement also increases when rubric criteria are separated rather than combined thus having the separate criteria of "speaking skills" and "level of enthusiasm" should result in higher interrater agreement than having a criteria that measures "speaking skills and level of enthusiasm".

Question and Answer Sessions Directed at Presenters:

Another assessment/judging format intervention that may affect interrater agreement is whether or not the presentation involves a question and answer session. It is our hypothesis that allowing for questions at the end of a presentation increases interrater agreement, especially in the context of shorter presentations, such as 'idea pitch' competitions. Idea pitch presentations are typically no longer than two minutes. We believe that allowing questions at the conclusion of these presentations will allow judges to clarify any details they may have missed as well as receive clarification about the idea and the presenter, thus improving their perceptions and agreement.

Asking questions may also force the judges to postpone their rating decisions and add data from the question and answer sessions to the ratings they are giving. Observing actual idea pitch competitions, we note that in situations where the presenter gives an extremely short presentation, judges often assess the presenter just as quickly. This tendency to rate more quickly on potentially less data, we believe leads to lower a_{WG}.

Judge-based Interventions:

Judge-based interventions include judges' training prior to the event, frame of reference training, convergent participation, partner-based consensus, use of a head judge, and the ability for judges to easily confer after they have observed the presentation and/or interacted with the presenter.

Judges' training can mean a variety of things. Trainings occur before the event and vary in length. In general, training includes an overview of the event, the background of the participants, an explanation of the criteria, and a time to allow judges to ask questions about what they are assessing and how they are to complete the assessment process.

6

One specific form of judges' training is Frame of Reference (FOR) training. FOR training provides judges the opportunity to generate a shared understanding of the dimensions being measured. As explained by Jackson. et al (2005) in the literature, this type of training was initially intended only for judges who varied markedly from the 'normal' assessments of their peers. These judges would review and judge case studies and then compare their ratings to what was considered normal for the organization. By providing all raters with a frame of reference and frame of reference training, agreement, that is $a_{WG,}$ increases (Jackson, et al. 2005).

All participants in the study conducted by Jackson, et. al. (2005) had experience judging as well as experience in the subject matter being judged, but no previous FOR or psychological training. They received both a training manual and on-site training, rated a test case and discussed discrepancies in their ratings. They then rated another simulation. Agreement on both behavioral and trait ratings increased after training. This study occurred in the field of human resource management, but we believe it also applies to idea pitch and poster competitions.

Another intervention is convergent participation. With this technique, objects or participants are judged twice. Judges discuss their ratings in a moderated forum at an intermission in the event. The study that used this intervention was rating online learning objects (Vargo, et al. 2003). Once the criteria were assessed by each judge individually, the judges were then brought together to discuss ratings, beginning with the objects which exhibited the highest variance in ratings.

The [convergent participation] conversation is moderated and judges are allowed to adjust their individual evaluation as others present their arguments. At the end of this conference, a review is published based on mean ratings and the comments of the participants. Judges had full knowledge of how their ratings related to the ratings of others on their team, without those others being identified. A day after this discussion, the judges were again asked to rate the sets of learning objects on their own. The results showed that discussing ratings between rating processes can reduce rater bias [a tendency to rate high or low], and decrease rater variance [a tendency to rate differently on an absolute scale than other raters], thus increasing interrater agreement or $a_{WG.}$

Submitted by Daniel M. Ferguson, Michele A. Govekar, and Amanda C. Stype   Thursday, April 01, 2010

We believe that the convergent participation intervention is adaptable to situations involving human participants in addition to learning objects. One possible procedure is that the judges visit with each participant and view their presentation, rating each participant individually. After discussing the ratings with other judges in a closed setting, the judges see each presentation again. The second time judges are more attuned to behaviors other judges observed, or they may be more forgiving of the participant resulting in ratings more similar to those bestowed by other judges and increasing $a_{WG..}$

In other situations, judges have been broken into pairs (Jung 2003)to report one rating. The head judge will then compare the ratings of each pair to those of the others. The benefit of this is that the two judges within a pair are forced to confer and to agree before they turn their single rating in to the head judge. They still retain the option to disagree, in which case they do not submit a rating. This intervention comes from a study which examined the ratings of software processes for new ISO standards (Jung 2003). It also evolved practically in a university-level evaluation of individual college assessment plans within that author's experience.

Designation and use of a head-judge/s is an intervention that we have used within our events. Each head judge is assigned a team of judges. Each team of judges is then assigned a set of participants that a certain number of judges must rate. For example, a team may consist of three to five judges, be assigned to rate 12 contestants, with each contestant rated by a minimum of three judges. The head judge determines the logistics required to fulfill these conditions. Also, the head judge fields questions about the criteria and the judging process. At the end of the event, the head judge collects the scoring sheets and checks them over to make sure the judges filled them out correctly. Head judges often receive special training. In our future practice, the role of the head judge may be expanded to include moderating post-judging or collaborative discussions.

Finally, some judging setups make it easier for judges to confer on ratings than others. In a presentation situation, where all participants are in the room and follow one right after the other, it is not possible for the judges to confer without the other contestants overhearing. In situations where one person presents at a time and then either the presenter or the judges leave the room, it is possible for judges to confer on ratings. It is also

possible for judges to confer after a set of participants present. Another possibility is to have presenters at assigned stations and let judges rotate through the stations. In this case, the judges can step away from a station before completing their rating and also have the opportunity to confer privately among themselves.

Allowing judges to confer about ratings allows a judge to know if their perceptions are different from those of their peers. However, just because one judge has different perceptions, it does not mean that judge's ratings are incorrect. It is possible that a specific judge picked up on something in the presentation or question and answer session that the other judges missed. Rather, we believe that conferring among judges allows each judge to improve their understanding of the rubrics and inceases $a_{WG.}$

External Factors

Several external factors, outside the immediate control of the organizers of an event, may also affect interrater agreement or increase disagreement. Many of these factors have to do with the background of the judges. For example, if one judge has had previous interactions with the person they are judging, whether those be positive or negative, it may skew their marks away from the marks received from judges who have no prior knowledge of that person. It has been shown that judges with the same affect toward a person they are rating, if they previously knew the person, have higher interrater agreement (Tsui and Barry 1986). However, it is very possible, even likely, that a team rating a person may be composed of judges with prior experience with the presenter, plus judges who have had no interaction with and know nothing about the presenter. In many academic situations, it is possible that the project advisor may be a part of the team rating the presentation. In this case, the advisor may view the presentation much more leniently/severely than other judges. Conversely, the project advisor may have much higher expectations than the other judges and will know if the presenters portray any of the facets of their project inadequately or inaccurately. This extra knowledge of presenters and the content of the idea pitch or presentation raises the question as to whether project advisors should be rating their advisees since they may interpret rubrics quite differently from judges without that practical knowledge.

Another factor which may impact interrater agreement is whether or not the judges have previous experience with the rubric. Judges who have previously judged using the same or a similar rubric may be more adept with it. They know the criteria on the rubric better than new judges and have more experience seeking out those criteria when viewing presentations. Some judges will have more experience judging presentation skills than others, whether or not rubrics are used. These judges may be more skilled at assessing the quality of a presentation due to their previous experience seeing other presentations.

Likewise, some judges will have more experience in the subject area of the presentation than others. For example, if the presentation content involves design or construction processes, some judges will have more experience and ability to judge the feasibility of the design than others. In some cases, a judge may have no prior experience with the presentation content they are supposed to rate, or the topic about which they are seeing a presentation. In other cases, such as academic capstone presentations, some or all of the assessors may be at least nominally familiar with the topic or area.

Data, Results, and Discussion

This past year we collected data from idea pitch competitions with eighty-six separate presentations at Business, Engineering, Pharmacy and Arts and Sciences colleges. All judges used the same four criterion rubric (attached as Appendix 1). We identified interventions or conditions that, either cited in the literature or from practical experience, were present or applied in these judging situations. We list these interventions in Table 1.

We collected judges' assessments for each competitor, each round and each competition. We further separated the items or criteria evaluated for each competition and for each round. We used the a_{WG} statistical measurement techniques identified by Brown and Hauenstein (2005) to calculate consensus among the judges by presenter by item and by initial and final rounds for a total of 344 a_{WG} calculations. We present descriptive statistics on these measurements in Table 2. We note a slight improvement in the consensus measures from fall 2008 through spring 2009, but differences also exist between consensus on items/categories. Similarly we note a slight improvement in the consensus measures from Round One to the Final Round, but not for every

10

competition or for every category. We further analyzed the results using some simple t-tests to identify if and where statistically significant differences occurred between rounds, items and/or interventions. We present these results in Table 3.

Using simple t-tests (with unequal variance) we find no statistically significant differences between fall, winter and spring rounds (although we approach significant differences between Winter and Spring, which have more comparable numbers of participants); however the key intervention between Winter and Spring was a simple judges training session directly preceding the competition. Further inspection of results in Table 2 reveals subtle differences between the judges' evaluations over the three competitions, rounds and criteria which provide opportunities for further research. Our results are complicated by the very different numbers of participants, judges and groupings in each season. In each competition we find other intervening factors that may account for the different results. These include different numbers of participants, different numbers of judges, and different composition of final round judging teams. These confounding factors are beyond the control of most competition organizers.

Conclusions:

Although there are many factors outside the control of event organizers which may adversely affect interrater agreement, there are also a variety of interventions which may increase consensus between the often-volunteer judges with diverse backgrounds who judge and assess many different kinds of events. In this second year of quasi-experimental research we have been more systematic at attempting to hold several factors constant to identify the impact of those factors over which we have no control. Our last competition for this year is at the end of April.

We believe that it is important to try to create a judging process that is as fair as possible and to increase consensus when possible without encouraging groupthink or silencing dissenting opinions. A high level of consensus means that the judges are in agreement about the quality of a presentation, candidate, or new product proposal; that is, they deliver a fair and knowledgeable assessment of the best and worst options. A low level of

11

consensus may indicate a large difference in expectations among the various judges and assessors. If this low-level consensus is present in a situation where the resulting reward is highly important (academically or monetarily), it presents a problem. Low-level consensus and varying expectations indicates that the quality of the outcome may not be as high as event organizers would like. Data from three sets of idea pitch competitions conducted last year show how difficult it may be to demonstrate and, more important, evaluate consensus among judges *post hoc*, that is, after the fact. We are conducting further data collection on idea pitches this year using a more controlled, quasi-experimental design, implementing many of the interventions mentioned above in a systematic manner in order to better understand the contributors to higher consensus. We also have other data sets (from another university) with more comparable numbers of participants, which we are analyzing.

Recommendations:

It is our recommendation at this time that institutions begin to retain data from their various competitions and assessments and calculate Brown and Hauenstein's metric for this data. We encourage all to track judging factors as identified in our analysis, plus any other interventions that they believe influence interrater agreement. This will provide a baseline reading for what agreement is typically like for these events. We invite participation in collaboration in continuing this research; we plan to construct a website where data can be shared, a_{WG} can be automatically computed for participants and more evaluations can be investigated in a larger data set. When our future research shows that various interventions significantly increase interrater agreement, we will recommend that assessments and competitions implement these interventions, when cost-effective and practical, to make judging and assessing processes notable for higher quality and less variances due to inadequate preparation for the judging process.

Table 1. Competitions/Assessments and Interventions Observed*

Data set	Date	Head Judge	Rubric	Familiarity with Rubric	Expertise in area of judging	Previous judging experience	Judges Training/ Orientation
Idea Pitch Competitions (individuals)	Fall 2008		Y		Y	Y	
	Winter 2009	Y	Y	Y	Y	Y	
	Spring 2009	Y	Y	Y	Y	Y	Y

- **1. Significant changes in judges occurred each quarter**

- **2. Rubrics were modified each quarter.**

- **3. Participating students are from multiple colleges and grade levels but a majority were required to participate as a class requirement.**

Table 2: Idea Pitch Agreement a_{WG} Descriptive Statistics

Round, Participant n	Rating One Mean/StDev	Rating Two Mean/StDev	Rating Three Mean/StDev	Rating Four Mean/StDev
Fall Round One, n=12	0.649437 / 0.143652	0.767342 / 0.171045	0.714745 / 0.157692	0.760140 / 0.118046
Fall Final Round, n=5	0.611190 / 0.314423	0.789947 / 0.054503	0.725540 / 0.140855	0.879284 / 0.114198
Winter Round One, n=19	0.743688 / 0.166457	0.760464 / 0.192499	0.783599 / 0.160819	0.701029 / 0.246908
Winter Final Round, n=12	0.691104 / 0.143278	0.567903 / 0.135534	0.579745 / 0.156102	0.697065 / 0.106161
Spring Round One, n=25	0.612989 / 0.207536	0.68871 / 0.370396	0.746874 / 0.176080	0.789165 / 0.146375
Spring Final Round, n=12	0.689583 / 0.135073	0.641691 / 0.239871	0.706851 / 0.229789	0.724109 / 0.119272
All Rounds One, n=56	0.665143 / 0.188179	0.729905 / 0.281504	0.75245 / 0.166211	0.753042 / 0.183376
All Rounds Final, n=29	0.676697 / 0.174017	0.636719 / 0.190992	0.657477 / 0.193941	0.739673 / 0.127443

Table 3. Results of Statistical Analyses

Fall Versus Winter Versus Spring

	Fall	Winter	Spring		Fall	Spring
n=	48	76	100		48	100
Mean	0.7229	0.7472	0.7094		0.7229	0.7094
Variance	0.0230	0.0373	0.0608		0.0230	0.0608
Df	116	174			137	
t-Stat	-0.7795	1.1390			0.4088	
P T<t	0.2187	0.1281			0.3417	

Appendix 1: Rating Form Used

IDEA PITCH: Contestant name_____ JUDGES CRITERIA AND RUBRICS: Judge_____					
1. How well does the pitch articulate a specific problem or unmet need and identify the customer	No clear problem statement No clear customer identification, Information confusing		Either the problem statement or the customer identification is well done		Both the problem statement and the target customer are very clearly indentified
	1-Very poorly done	2-poorly done	3-ok	4-well done	5-Very well done
2. How unique and viable is the idea in addressing this specific need?	No uniqueness to the idea. Business idea does not appear viable as presented or address the need as presented		Idea has some good features, and may be viable based on evidence still to be developed.		Unique idea, and idea appears viable as a business, directly addresses the need previously discussed
	1-Very poorly done	2-poorly done	3-ok	4-well done	5-Very well done
3. How effectively and passionately does the presenter articulate the problem and solution?	Stated case is disorganized and not persuasive, No passion no obvious commitment		Some conviction, good evidence included, presentation has interesting if not convincing content		Clearly passionate about opportunity, clearly excited and committed to business idea. Complete and convincing case
	1-Very poorly done	2-poorly done	3-ok	4-well done	5-Very well done
4. How effective or accomplished are the speaker's skills?	No time control, No eye contact, poor articulation, no vocal emphasis,		acceptable delivery, obviously practiced timing, while not overly persuasive-professionally done		Solid eye contact, very persuasive, positive tone and expressions, proper dress and facial expressions, timing great
	1-Very poorly done	2-poorly done	3-ok	4-well done	5-Very well done

16