Analyzing Changes in the Individual Dimensions of a Behaviorally Anchored Rating Scale

Ms. Maizy Benner, Purdue University
Dr. Daniel M. Ferguson, Purdue University

Daniel M. Ferguson is CATME Managing Director and the recipient of several NSF awards for research in engineering education and a research associate at Purdue University. Prior to coming to Purdue he was Assistant Professor of Entrepreneurship at Ohio Northern University. Before assuming that position he was Associate Director of the Inter-Professional Studies Program [IPRO] and Senior Lecturer at Illinois Institute of Technology and involved in research in service learning, assessment processes and interventions aimed at improving learning objective attainment. Prior to his University assignments he was the Founder and CEO of The EDI Group, Ltd. and The EDI Group Canada, Ltd, independent professional services companies specializing in B2B electronic commerce and electronic data interchange. The EDI Group companies conducted syndicated market research, offered educational seminars and conferences and published The Journal of Electronic Commerce. He was also a Vice President at the First National Bank of Chicago [now J.P. Morgan Chase], where he founded and managed the bank’s market leading professional Cash Management Consulting Group, initiated the bank’s non-credit service product management organization and profit center profitability programs and was instrumental in the breakthrough EDI/EFT payment system implemented by General Motors. Dr. Ferguson is a graduate of Notre Dame, Stanford and Purdue Universities, a special edition editor of the Journal of Engineering Entrepreneurship and a member of Tau Beta Pi.

Dr. Matthew W. Ohland, Purdue University

Matthew W. Ohland is Professor of Engineering Education at Purdue University. He has degrees from Swarthmore College, Rensselaer Polytechnic Institute, and the University of Florida. His research on the longitudinal study of engineering students, team assignment, peer evaluation, and active and collaborative teaching methods has been supported by the National Science Foundation and the Sloan Foundation and his team received Best Paper awards from the Journal of Engineering Education in 2008 and 2011 and from the IEEE Transactions on Education in 2011 and 2015. Dr. Ohland is an ABET Program Evaluator for ASEE. He was the 2002–2006 President of Tau Beta Pi and is a Fellow of the ASEE, IEEE, and AAAS.

Mr. Behzad Beigpourian, Purdue University

Behzad Beigpourian is a Ph.D. student and Research Assistant in Engineering Education at Purdue University. He earned his master’s in Structural Engineering from Shahid Chamran University in Iran, and his bachelor’s in Civil Technical Teacher from Shahid Rajae Teacher Training University in Iran, Tehran. He has been official Technical Teacher at Ministry of Education in Iran from 2007 to 2018, and received many certificate in education such as Educational Planning, Developing Research Report, and Understanding School Culture. During these years, he has taught construction courses in several technical schools. Mr. Beigpourian currently works in the CATME project, which is NSF funding project, on optimizing teamwork skills and assessing the quality of Peer Evaluations.
Analyzing Changes in the Individual Dimensions of a Behaviorally Anchored Scale for Teamwork

Abstract:

Individual Item or Dimension Analysis (IDA) using the Social Relations Model (SRM) has the power to reveal subtle changes in behavior as recorded in a behaviorally anchored scale. These changes cannot be seen in Across Dimension Analysis (ADA). Analysis of mean and standard deviation across all 5 CATME dimensions does not uncover differences at the individual item (or dimension) basis. The purpose of this analysis is to establish the usefulness of IDA SRM as a statistical analysis method.

The Social Relations Model decomposes variance into components. In our case, the variance is identified as rater, target, or relationship variances. Rater variance measures how individuals tend to rate their teammates. A large rater variance would indicate that students give their teammates ratings that are idiosyncratic – they don’t relate to the ratings given by others. Target variance measures how consistently individuals are rated by their teammates. A larger target effect would indicate that students tend to receive similar ratings from various raters. Relationship variance measures how ratings are affected by unique dyadic relationships after other variances have been removed. A larger relationship effect would indicate that the ratings are based more on distinct relationship pairings within the team – that is, that the various team members display different behaviors to each of their teammates and are rated based on those behaviors. The individual dimensions can be defined as follows: Having (H) relevant KSAs, Contributing (C) to the team’s work, Interacting (I) with team members, Keeping (K) the team on track, and Expecting (E) quality.

In this analysis, we examined three different teamwork studies that were based on evaluating average changes in multiple behaviorally anchored items and found additional insights into behavior changes through IDA. All data included in this analysis was obtained through the CATME peer evaluation system and our studies focused on methods of improving the quality of peer evaluations. In an initial study of a large Midwestern university, IDA SRM analysis revealed significant results in dimensions C, I, K, and H that were not seen using an ADA analysis, the average rating scores across items (or dimensions). Our IDA analyses will compare IDA and ADA analyses and intervention groups who received various types of teamwork training and control groups who did not receive any teamwork training.

Introduction:

The purpose of this study is to establish the merits of analysis on an individual dimension (or factor) basis. Results on an individual dimension basis are masked when only looking at results across the dimensions. IDA may reveal changes that are not apparent in across dimension analysis, and has the power to show changes in particular dimensions because of certain training exercises. This information would be of use to instructors who felt their students needed additional teamwork training and wanted details of its effects. This analysis includes description and results of three prior studies of individual dimension analysis.
The basis of the CATME system is a set of five factors, formed using factor analysis that defines teamwork behavior. Factor analysis correlates measurable and observable variables to key factors that share a common variance. Factor analysis can be used in both a confirmatory and exploratory manner. Confirmatory factor analysis is used to confirm hypotheses and uses path analysis to represent factors. Exploratory factor analysis attempts to uncover patterns by testing predictions and exploring the data. Exploratory analysis is used to discover the number of factors and analyze which variables are highly correlated, typically discarding those variables cross-loading highly on multiple factors [1]. A key assumption of factor analysis is that factors are unique and uncorrelated with another. This assumption naturally leads to the idea that variables are only correlated because of their common factor [2].

Detailed explanations of the CATME dimensions and the factor analysis method used can be found in the paper by Loughry, Ohland, and Moore (2007) [3]. Analyzing individual dimensions of teamwork is not unique to the CATME system. Solansky (2010) developed a set of factors to describe teamwork, identifying the five factors of agreeableness, team mean conscientiousness, openness to experience, collectivism, and preference for teamwork [4]. Greguras, Robie, and Born (2001) also developed a five factor system of cooperation, ideas, effort, reliability, and quality, and used this system to prove the usefulness of individual dimension analysis [5]. However, these instruments measure qualities of teamwork, and are not behaviorally focused as the CATME system is.

**Literature Review:**

The five dimensions of CATME are evaluated using a multi-factor behaviorally anchored scale. The more commonly used Likert scale was originally proposed by R Likert in 1932 and provides a quantitative measure of attitude [6]. Likert scales are designed to show level of agreement with a proposed statement. It is theorized that test reliability increases with the number of items used to measure the same concept [7]. Scales can be symmetric and have the level of neutrality be exactly between two extremes, or be asymmetric and have fewer choices on one side. Asymmetric scales can force choices. The key assumptions of a Likert scale are that participants are allowed to choose clearly opposed alternatives and conflicting issues show results that, “constitute an empirical check on degree of success,” [6].

Behaviorally anchored rating scales (BARS) like that used by CATME achieve higher reliability by providing descriptive anchors for the scale points rather than by increasing the number of items. Psychometric studies comparing BARS to other scales have mixed results [8], yet BARS instruments have numerous advantages over Likert-type rating scales, including greater interrater reliability and less leniency error [9], improved user reactions and face validity particularly in collectivist cultures [8], and facilitating the development frame-of-reference training [10]. The benefits of a BARS instrument have also been demonstrated in the specific context of teamwork in a precursor to the CATME instrument [11].

While there is value in measuring how the five CATME dimensions combine to give an overall sense of each student’s overall effectiveness as a team member, Greguras, et al. (2001) used their system to prove the usefulness of individual dimension analysis, and this present work focuses on the advantages of individual dimension analyses. The key analysis tool in this study is
individual dimension analysis supported by the Social Relations Model (SRM) [12]. SRM is a special case of generalizability theory, a two-way, random-effects ANOVA with actor and partner factors, in which relationship is the interaction term. The SRM evaluates three key variances in teamwork data: rater, target, and relationship variance. Rater variance describes how a team member rates their teammates. A larger rater variance indicates that team members are giving multiple targets similar ratings, which is idiosyncratic and undesirable. Target variance measures how consistently an individual is rated by their teammates. A larger target variance would indicate that a student is being rated consistently. Relationship variance measures how the rating is influenced by the unique relationship between the two people (i.e., the interaction between the rater and target). A larger relationship variance is of concern because it would indicate that an individual is rating their teammates based on reasons outside of the team context. Differences found in the individual dimensions may not be present in ADA (an average of all five dimensions).

Research Methodology:
This analysis examines three different teamwork studies that were based on evaluating changes in rating behavior resulting from various training. Study One compares the rating behavior of FYE students in 10 different intervention tests conducted in the fall of 2016. These multiple tests included a student section that received two different rater error trainings to the ratings of several sections of FYE students in the fall of 2016 who received no rater error training. Study Two compares the ratings of FYE students who received video based peer evaluation training in the summer of 2016 on three CATME dimensions [C, I, H] to the ratings of a similar FYE student section in the summer of 2015 who received no rater training. Study Three compares the ratings of students who used the Rater Practice 2.0 Model in the fall of 2016 to the peer ratings of FYE students who used the Rater Practice 3.0 model in the fall of 2017. The goal of the all of these research studies was to increase the quality of peer evaluations within student teams.

The Rater Practice system was used in two of the three analyses described here. Rater Practice is a game like simulation that allows students to practice rating fictional teams in an environment free of consequences. The Rater Practice 2.0 system was a tool used to practice giving a peer evaluation on a fictional team. Users were given a fictional team of 3 and asked to rate their behavior in the five CATME dimensions using a brief description of their behavior. The rating scale used descriptions of the variables involved in the factor being analyzed. They were given feedback after all five dimensions were complete. The Rater Practice 3.0 system was updated to include immediate feedback, ability to play multiple times with unique teams, and instructional videos (details below).

The purpose of Study One was to examine if there was a measurable difference in peer evaluations between a FYE student section of 120 students that was trained using one of more of the 10 rater training interventions and FYE student sections that received no training. Rater practice was one of the rater training interventions used and rater practice is a tool for learning the CATME peer evaluation schema. Rater Practice training occurred before peer evaluation one in the fall of 2016.
Participants in Study Two received lecture and video based training in regards to dimensions C, I, and H between peer evaluations two and three. Each video based training lasted \(\frac{1}{2}\) hour and involved videos, interactive activity within each 4-person team, and interaction with the class as a whole. The purpose of this study was to examine if video based interactive class room training on specific CATME dimensions had a measurable impact on the peer ratings that students provided to themselves and their peers for specific BARS/CATME dimensions.

Study Three examined the quality of peer evaluations for a large sample of FYE students before (fall of 2015) and after (fall of 2017) changes to the Rater Practice tool to make it a more game-like simulation. The changes made to the Rater Practice (RP) 2.0 system to create the 3.0 version can be seen in the paper from Ferguson, Shu, Cao, and Ohland (2017) [13]. One of the important changes was that feedback is given after assessment of each individual CATME dimension (immediate feedback), instead of after the assessment of all dimensions (delayed feedback). This change was justified by Carter (1984) and Kulhavy and Anderson (1972), who agreed that immediate feedback is most useful when the task concerns discrimination learning [14,15]. Discrimination learning is defined as, “learning to respond differently to different stimuli,” [16], and involves differentiation and unitization of stimuli [16]. Immediate feedback is also supported by Gibbs and Simpson (2005), who argued that students need feedback while the information is still relevant and can be used for future work [17]. The Rater Practice tool prompts a student to differentiate between qualities that their peers possess and in theory should increase the quality of peer evaluations.

An additional change made to the Rater Practice 3.0 system was the addition of a suite of instructional videos on how to use the system. These videos were created for both the instructor and the students on how to operate the system. The format of video instruction used is a sequential-step explanation and is a form of observational learning, which is learning through the behavior of others. Observational learning through video allows students to see a “flawless performance” of the task, and can be viewed repeatedly as needed [18]. The instructional videos on Rater Practice should be sufficient in learning how to operate the simulation and can be viewed and practiced until the system is understood. If students comprehend the instructional videos, the barrier to many simulations - the instructions being non-intuitive - would be overcome.

Analysis:

A positive pattern seen in Social Relations Model analysis would be a decrease in rater effect over time and an increase in target effect. Without observing a team to know whether team members manifest different behaviors to different teammates, it is difficult to know whether relationship variance is problematic, but the researchers typically assume that a decrease in relationship effect is a positive trend. Due to the large population size in the studies, any meaningful result will be statistically significant, so we chose to consider a change of 5% or more between two peer evaluations to be of interest. Data was collected from students in engineering courses at a large Midwestern university. Students were required to complete the peer evaluations, which counted for grade points in the course.
Figure 1. Study One ADA Graph

Figure One describes the results of Study One. Across dimension Analysis (ADA) using the SRM model shows the average of each variance over all five dimensions. In the graph, ‘NORP’ describes the group who did not receive rater practice training, and ‘RP’ describes a group who did receive rater error training. The rater training occurred before the first peer evaluation in the RP group. In study one, across dimension analysis (ADA) using SRM revealed a decrease in rater variance of the untrained NORP group of 14.8% and 7.6% for the trained group. Relationship variance increased in both groups, showing an overall increase of 9.9% for the untrained group and an overall increase of 5.1% for the trained group.

Figure 2. Study One-Dimension C SRM Graph

Dimension C showed a decrease in rater variance of 10.1% between evaluations two and three percent for the untrained group. The trained participants did not show any meaningful changes in rater variance. The untrained participants showed an increase of target variance of 7.7% between evaluations two and three, and the trained participants showed an increase between evaluations one and two of 5.6%. Relationship variance increased for the untrained group by 6.3% between evaluations one and two and 8.4% for the trained group between evaluations two and three.
Dimension I saw an increase of 7.4% in target variance between evaluations two and three of the untrained group and an increase of 6.3% between evaluations one and two of the trained group. The trained group saw a decrease in relationship variance of 10.7%.

In dimension K, both the trained (10.6%) and untrained (18.8%) groups showed a decrease in rater variance. The trained group saw an increase in target variance between evaluations one and two of 8%. The untrained group saw an overall increase in relationship variance of 13.1%.

Dimension E revealed an overall decrease in rater variance of 20.4% for the untrained group, and a decrease between evaluations one and two of 12.3%. Relationship variance increased by a total of 14.1% for the untrained group and 7.4% for the trained group.

The results for dimension H showed a decrease in rater variance between evaluations two and three in the untrained group of 16%. Target variance showed an increase of 6.6% for the untrained group, and a decrease in the trained group of 6.8%. Relationship variance increased in the trained group by 9.3% between reviews two and three.

![Across Dimension SRM Analysis](image)

**Figure 3. Study Two Across Dimension SRM Analysis**

Across dimension SRM results for study two revealed a consistent and extreme decrease in rater variance across the four evaluations (53% in total) for students who received training. Target variance saw a consistent increase over the four evaluations (30.3% in total) in the trained group, and relationship variance showed an overall increase of 22.7%. This pattern was not observed in the untrained group, whose results were erratic and did not show a clear pattern of change.
Each of the CATME dimensions in the trained participants displayed an overall decrease in rater variance from evaluations one to four, dimensions C, K, E, and H showed an overall increase in target variance, and dimensions I, K, E and H displayed an increase in relationship variance. Dimension C saw a decrease of 44.3% in rater variance, and an increase of 37.6% in target variance. Dimension I showed a decrease of 29.1% in rater variance, and an increase of 22.2% in relationship variance. Dimension K showed a decrease in rater variance of 55.1%, an increase of 19.5% in target variance, and an increase of 22.8% in relationship variance. Dimension E revealed a decrease in rater variance of 58.3%, an increase of 19.5% in target variance, and an increase of 38.8% in relationship variance. Dimension H showed a decrease of 46.6% in rater variance, an increase of 16.1% in target variance, and an increase of 30.6% in relationship variance.

Study Three ADA compares FYE students in 2017 vs 2016, but there are no significant differences in the demographic or academic profile from one FYE cohort year to the next. In Study Three, the ADA SRM analysis found a significant lowering of rater variance between RP 2.0 compared to RP 2.0 for the second (11.1%) and third (6.3%) peer evaluations. Target variance did not show significant change in the first or second review, but increased significantly.
in review three (6.4%). Relationship variance increased in the first (5.1%) and second (6.5%) reviews, and showed no significant change in the third review.

![Study Three- Dimension C SRM Analysis](image)

**Figure 5. Study 3 Dimension C SRM Graph**

SRM analysis of dimension C (contributing) showed a significant decline in rater variance in review two (7.4%), but not in reviews one or three. Target variance increased significantly in reviews two (10.4%) and three (6.0%). Relationship variance did not show a significant change in any review. Analysis of dimension I (interacting) showed a significant decrease in rater variance in reviews two (9.5%) and three (7.5%). Target variance showed a significant increase in reviews two (6.0%) and three (8.1%). Relationship variance did not show a significant change in any review.

SRM analysis of dimension K (keeping) showed a significant decline in rater variance for reviews two (9.7%) and three (9.5%). Target variance showed a significant increase in review three (10.3%) only. Relationship variance showed a significant increase in review two (6.7%) only. Analysis of dimension E (expecting) showed a significant decrease in rater variance for reviews two (5.7%) and three (5.1%). There were no significant changes in target variance. Review two showed a significant increase in relationship variance of 6.4%. Dimension H (having) showed a significant decrease in rater variance for reviews two (10.9%) and three (5.9%). Review three showed the only significant increase in target variance of 5.1%. Relationship variance increased significantly in review two only (8.2%).

**Discussion:**

Study One across dimension SRM analysis showed an overall decrease in rater variance for both the trained and untrained groups, no significant changes in target variance, and an overall increase for both groups in relationship variance. However, the trained group did see a faster decrease in rater variance than the untrained group and started at lower value than the untrained group.

However, changes were seen on the individual dimension basis. In dimension I, a decrease in relationship variance was observed in the trained group that was not seen in the untrained group. In dimension K, an increase in target variance was observed in the trained
group, but not in the untrained group. Target variance in dimension H showed a significant
decrease, which is unusual. These results indicate that the training had a net effect on certain
aspects of teamwork behavior, such as giving students the ability to understand dimension I
better as to rate their team members more uniquely in regards to that dimension. Additionally,
Rater Practice had an effect of lowering rater variance faster than not using Rater Practice.

The results of Study Two revealed that IDA confirmed the results of ADA SRM model
runs. Across dimension results showed that the trained group saw positive results in rater and
target variance for each dimension, unique from the untrained group. In dimensions K, E, and H
a pattern identical to that of ADA was seen. Dimension C saw a decrease in rater variance and an
increase in target variance. Dimension I showed a decrease in rater variance and increase in
relationship variance. The confirmation of across dimension results by individual dimension
analysis is a positive result because it is possible to see dimension by dimension how training
effected students perceptions of peer evaluations. These results also reveal that the training had a
very positive influence on participant’s ability to properly conduct peer evaluations. While the
untrained group showed no clear patterns, the trained group saw clear positive patterns in each
dimension.

In Study Three, SRM ADA analysis showed a positive result in across dimension
analysis. Rater variance declined in each review from RP 2.0 to RP 3.0, which indicates that
students are differentiating between their teammates more after using RP 3.0, and not giving
each teammate the same rating. Target variance increasing from RP 2.0 to RP 3.0 in reviews two
and three indicates that students were rated more consistently by their teammates after using the
RP 3.0 system. RP 3.0 training made a lasting effect because an improvement in peer evaluations
was seen in both reviews two and three. An increase in relationship variance in reviews one and
two indicates that students were rating their teammates based on interaction outside of the team
context in RP 3.0 more so than in RP 2.0.

SRM analysis of individual dimensions also yielded interesting results. The dimension
contributing (C) showed a significant increase in target variance, but no significant patterns in
rater or relationship variance. This result suggests that students were rated by their peers more
consistently in dimension C after playing RP 3.0 than they were after playing RP 2.0. The
dimension interacting (I) showed significant changes in rater and target variance. This result
suggests that the rater dimension practice influences how consistently students are rated and rate
their peers. The dimension having (H) showed a significant decrease in rater variance from RP
2.0 to RP 3.0 in reviews two and three. This result suggests that students are better able to rate
their peers individually in dimension H after playing the RP 3.0 simulation than they were after
playing the RP 2.0 simulation.

Conclusion:

The positive results across each of the three studies highlights the usefulness of
individual factors or CATME dimension analysis for instructors. In Study One, analysis showed
that IDA confirmed results of across dimension analysis and revealed unique changes in factor or
CATME dimensions. Study Two IDA confirmed the results of across dimension analysis and
showed the importance of training when implementing CATME peer evaluations in the
classroom.
Study Three revealed the intricacies of the individual dimensions by showing how training can influence specific points of understanding a peer evaluation using the CATME factors or dimensions. Specifically, students were better able to understand certain aspects of ratings in dimensions C, I, and H because of the improvements in the rater practice system.

Using IDA, instructors can find changes in a class’ teamwork behavior after their intervention of choice. The details of IDA are often serve as confirmation of ADA SRM, but also may reveal unique changes in dimensions because of training interventions. These results can be used to fill in holes in a class’s teamwork training and create better functioning teams.

**Limitations:**

All three studies were conducted using data from first-year engineering students at one Midwestern University. Students in different disciplines or higher level courses may have a different understanding of peer evaluations and be impacted in unique ways by the training interventions.
References:


