# Exploratory Factor Analysis of a survey on group-exam experiences and subsequent investigation of the role of group familiarity

Joss Ives and Jared B. Stang
*Dept. of Physics & Astronomy, University of British Columbia, 6224 Agricultural Road, Vancouver, BC V6T 1Z1*

José Arias-Bustamente
*Dept. of Forest Resources Management, University of British Columbia, 2424 Main Mall, Vancouver, BC, Canada, V6T 1Z4*

Patrick J. Dubois
*Dept. of Psychology, University of British Columbia, 2136 West Mall, Vancouver, BC, Canada, V6T 1Z4*

Analise Hofmann
*Dept of Cellular & Physiological Sciences, University of British Columbia, 2350 Health Sciences Mall, Vancouver, BC, Canada, V6T 1Z3*

We report on an exploratory study in which we investigate the factor structure of an in-development survey on student experiences during group exams and subsequently examine how these factors can be modelled using performance and self-reported performance measures, while focusing on the role of group familiarity because it is a measure we felt we could bolster through future intervention. We ran an Exploratory Factor Analysis on a suite of survey items that sought to investigate aspects of their group-exam experience, such as participation equity, the prevalence of productive group-work behaviours, and their personal experiences within the group. After stepwise item removal took us from an item pool of twenty-one down to fourteen, our Exploratory Factor Analysis saw a four-factor structure emerge as the preferred option, consistent with the dominant areas of focus of our underlying survey design. The four factors that emerged—*presence of under-contributors*, *presence of dominators*, *productive group-work behaviours* and *personal experience*—and the items that were removed as part of the factor analysis process indicate directions for future item development. These results suggest that our survey could be sensitive to the impact of interventions designed to improve overall student experience in group exams by targeting improvements in sense of academic familiarity with their groupmates, participation equity or productive group-work behaviours.

*2020 PERC Proceedings* edited by Wolf, Bennett, and Frank; Peer-reviewed, doi.org/10.1119/perc.2020.pr.Ives

## I. INTRODUCTION

This study is part of a larger project designed to characterize and improve student experiences in two-phase collaborative group quizzes and exams, where students are first assessed individually via a quiz or exam, and then re-take the assessment in small groups. Students tend to have an overall positive view of this assessment format [1], and as summarized in Reference [2], two-phase group exams have been shown to improve retention, to have a positive impact on class environment and to reduce self-reported test anxiety.

As part of our ongoing development of a survey to help us understand which factors correlate with positive and negative experiences in these group exams, this study reports on an Exploratory Factor Analysis (EFA) of this survey. This study also reports on subsequent modelling of these survey factors, via Multiple Linear Regression techniques, to help us understand the role that performance, demographic and self-reported measures may play in these student experiences. Since the larger project has the ultimate goal of improving student experiences, we focus on group familiarity as a measure that we feel we can bolster.

### A. Survey development

This section details our iterative survey development process and highlights the literature from which we adapted items or otherwise drew inspiration.

The development of our first version of the survey drew from the work of Rieger and Heiner [1], in which they categorized the positive and negative themes that emerged when they asked students "to describe their experience with the group exam in one or two sentences". Additionally, we drew on our own experiences as instructors using this instructional technique to develop items based on our perceptions of what a productive group looks like and what were the common underlying causes that led to students having negative experiences with their group during this type of exam.

After each round of survey deployment, we analyzed the themes emerging from our final open-ended item, "Is there anything else you would like us to know about your group exam experience?" to further develop and refine items. Multiple times throughout the process we conducted focus group interviews with survey participants to ensure items were being interpreted as intended and to provide a venue from which additional themes could emerge.

The literature on student experiences with in-class group-work informed our survey development related to themes of group familiarity [3, 4], participation equity [3], productive group-work behaviours [4, 5], comfort [3], and gender [6] and other social identities [7].

We hypothesized that we might most easily be able to bolster group familiarity, participation equity or productive group-work through interventions or instructional design choices, thus these themes were heavily emphasized in our item development with the intention of being able to measure the impact of future interventions.

Including minor and major revisions, the survey administered in this study represents approximately its tenth iteration.

The EFA reported in this manuscript was performed on twenty-one 6-point-scale survey items, which focused primarily on participation equity, the prevalence of productive group-work behaviours, and personal experiences within the group, such as comfort and overall perception of how the group worked together. Thirteen of these items were framed as asking about participants' observations about the group, with items such as "One or more people were pushy, overbearing or intimidating" and "Everyone agreed on our answer before moving on to the next question". The other eight items were framed as asking about how they felt about their group exam experience, with "I"-statement items such as "I would be happy to work with this group again" and "I felt comfortable offering my opinions". As will be discussed in Section II B, our final EFA model retained only fourteen of these twenty-one items.

Beyond these twenty-one items, other survey items touched on topics such as group size, how many within the group had agreed in advance of the exam to work together, relative participation of all group members, group familiarity and open answer items at the end.

### B. Data collection and participants

Participants were students from a single-instructor introductory calculus-based physics course for non-majors at a large North American R1 university during summer 2019, taught using interactive methods in a small lecture hall and supported by in-class teaching assistants. Although informal group-work was common in the course, the midterm exam was the first time where the students experienced a two-phase collaborative group exam in this course. For the group-phase of this exam, students chose their own groups of 3 or 4 and the course provided no guidance related to effective group-work behaviours. The group phase of the exam was identical to the individual phase of the exam and both phases took place during the same exam sitting. The survey was administered three days after the course's midterm exam was taken by 180 students (69.4% women, 30.6% men). Class time was provided to the students to complete the survey and 138 students completed it during that time and 11 further students completed it during the subsequent five-day period before the survey closed. In addition to the class time provided, draws for gift cards at the campus bookstore were used to further incentivize participation.

Of the 149 survey records, we removed five of them because all twenty-one items related to the factor analysis were incomplete. Two additional records were removed because their submitted answers did not vary among these items. Finally, eight records were removed because it was not possible to associate them with specific students that had taken the

midterm. After these filters had been applied, we found no duplicate records. The median time to complete the survey was four minutes and fifteen seconds for those 134 records (72.8% women, 27.2% men) surviving the filtering process.

## II. FACTOR STRUCTURE OF THE GROUP EXAM EXPERIENCE SURVEY

This section details an EFA used to determine how many dimensions were present in a subset of twenty-one items from this deployment of the survey and to help identify potentially problematic items for future development of the survey. A useful output from an EFA is a set of factor scores associated with each dimension for each survey response, which allows for further analyses using the resulting factor structure of the survey.

Unless otherwise noted, factor analyses (using the `fac` function) and statistical tests were performed using the `psych` package [8] within `R` [9]. Because we expect the factors to correlate with one another, we chose an oblique rotation method ('oblimin') for the factor analysis and used the 'tenBerge' correlation-preserving regression method to determine factor scores.

### A. Descriptive statistics

All records that passed the initial filters (Section I B) had complete responses for all twenty-one items. The descriptive statistics presented in this section are for the fourteen items that remained in the final EFA model.

Items were measured on a 6-point scale, with mean values for positively-coded (negatively-coded) items ranging from $4.15$ to $5.20$ ($1.51$ to $2.72$). All but one of the fourteen items met the skewness $< |2.0|$ guideline [10], with that item having a skewness of $2.32$. Although we chose to retain this item in the model, the skewness being beyond the guideline threshold indicates that this item should be revised in the future to invite larger variation in student answers. All fourteen items met the liberal kurtosis $< |7.0|$ standard [10].

Although Mardia's multivariate normality test indicated our data showed significant multivariate skewness and kurtosis, a factor analysis was still appropriate given that we used Principal Axis Factoring. This factoring method is considered to be one of the more robust methods available for non-normal and ordinal data [11].

To look for outliers in the responses, we used Mahalanobis distance, which is a multi-dimensional measure of how far away a survey record is away from the average. Six survey records were identified as having a significant Mahalanobis distance ($p < .001$), but it was determined that none of them needed to be removed after inspecting each of them in detail.

While items need to correlate with each other so that they can be grouped into factors (factorability), they must also not be overly redundant (multicollinearity). Factorability was found to be satisfied when inter-item correlations were examined and found to range from $|.16|$ to $|.66|$, with a significant majority of these correlations $> |.3|$. This also showed a lack of multicollinearity, with no items showing above the recommended upper threshold of $|.70|$ [11]. Factorability was also measured using the Kaiser-Meyer-Olkin measure of sampling adequacy, where values $\geq .6$ suggest good factorability [12]; values ranged from $.74 - .93$ for our items. Multicollinearity was also tested using the variance inflation factor, which tests to see whether two variables are closely enough related that one will inflate the variance of the other in a regression model. Our maximum factor of $3.86$ was far below the upper threshold of 10 for acceptable values [11].

### B. Exploratory Factor Analysis Methods and Results

This section details multiple rounds of Exploratory Factor Analyses (EFAs), starting with the multiple different factor models on the original twenty-one items and refining these to a four-factor model using fourteen of these items.

We used inspection of a parallel analysis scree plot to determine a range of two to four factors as reasonable for our data set with all twenty-one items, so we initially explored all three of these factor solution options.

In the initial four-factor model, two participation equity factors emerged. *Presence of under-contributors* corresponded to a lack of participation equity due to the presence of people in the group contributing much less than others. *Presence of dominators* corresponded to a lack of participation equity due to the presence one or more people that dominated their group. A third factor, *productive group-work behaviours* was related to items designed to identify groups exhibiting behaviours that we considered to be productive. Finally, the fourth factor, *personal experience*, is related to personal comfort and their perception of how their group worked together.

In the three-factor model, the *presence of dominators* and *personal experience* factors merged into one larger factor. In the two-factor model, the *presence of dominators*, *personal experience* and *productive group-work behaviours* factors merged into one larger factor.

Early on during this process, we identified the four-factor solution as the preferred model because its factors lined up best with the underlying intentions of our item design and the four-factor solution was also the best model according to the degrees-of-freedom-corrected root mean squares of the residuals goodness of fit measure.

The next step was to do a stepwise series of item removals, re-running the EFAs after each removal. These removals were based primarily on factor loadings, which are a measure of how strongly an item is associated with each factor. Our first criterion was to keep only items with factor loadings $> |.4|$. However, we wanted to avoid items that loaded strongly across multiple factors, so our second criterion was to remove items with loadings higher than $|.3|$ across two

| | Loading |
|---|---|
| *Presence of dominators factor* ($R^2 = .80$) | |
| One+ were pushy, overbearing or intimidating | .73 |
| I felt criticized or judged by one or more group members | .68 |
| Some opinions were ignored or not respected | .40 |
| *Presence of under-contributors factor* ($R^2 = .96$) | |
| One+ didn't contribute much | .98 |
| One+ didn't seem well-prepared/didn't know much | .66 |
| The participation-level and contributions from all group members was balanced/equal/the same | $-.51$ |
| *Productive group-work behaviours factor* ($R^2 = .93$) | |
| We took time to explain answers to group members who didn't understand | .81 |
| Answers were discussed & explained so everyone could learn | .78 |
| Everyone agreed on our answer before moving on to the next question | .42 |
| We used our time effectively | .48 |
| *Personal experience factor* ($R^2 = .84$) | |
| Overall, I feel that our group worked well together | .95 |
| I would be happy to work with this group again | .65 |
| I felt comfortable offering my opinions | .70 |
| I felt that I was included in decision-making processes in my group | .59 |

TABLE I. The fourteen survey items in the final EFA model and how they load onto their primary factor. Factor loadings $< .3$ on other factors are not shown. "One+" indicates "One or more". Items with negative loadings indicate they were treated as being reverse-coded with respect to that factor.

or more factors. The stepwise removal of each of the first three items was based on the first criterion and for the next four items it was based on the second criterion, reducing our number of items from twenty-one to fourteen. Items removed for loading $> |.3|$ across two factors include "Everyone contributed relevant opinions, explanations or questions to the group," "Everyone's perspective was considered fairly," and "Everyone was patient, attentive and respectful," which each loaded above this threshold on the *productive group-work behaviours* factor and either the *Presence of dominators* or *Presence of under-contributors* factors.

After each removal, we re-ran the two-, three- and four-factor EFAs, which allowed us to monitor if a model other than the four-factor one might become preferred with fewer items, but the four-factor model remained preferred for the same reasons as described above.

Table I details the final four-factor, fourteen-item model with their primary loading coefficients.

## III. FACTOR SCORE REGRESSION MODELLING

We used a set of four Multiple Linear Regressions (`lm` function [9]) to model each of the four factors from the final EFA model. In this section we describe the demographic and self-reported predictor variables used for our modelling, but focus our attention on group familiarity as the predictor variable we felt we may be most able to bolster through future intervention.

Our measure of familiarity with one's group members, *Know.Avg*, comes from the average of that participant's responses to an item that asked them to rate how well they knew each group member academically on a 6-point scale, from "Not at all/Just met" to "Very well". This item appeared earlier in the survey than the twenty-one items used for the EFA.

In addition to *Know.Avg*, each of the four regression models included these four additional predictors:

- *Solo.Score* and *Group.Score* were included to control for measures of individual and group performance and are their percentage scores on the solo and group phases of the exam, respectively.
- *Particip.Self* is a measure of their perception about their own participation level in the group relative to the rest of the group and was determined from an item asking the participant to rate the relative contributions/level of participation of each of the group members, including themselves, such that the total sums to 100%. To calculate *Particip.Self*, we took the difference between how they scored themselves and the average of the scores they gave the rest of the group.
- *Gender* was taken from from university registration records, where our institution forced students to choose between radio button options for 'male' and 'female' to indicate their 'gender.' We recognize that this forced-binary operationalization of gender is an over-simplification and does not fully represent the complexity with which individuals experience gender [13].

In the first iteration of the regression models we used these five predictor variables, but found that the regression coefficients associated with *Gender* were not statistically significant for any of the four models. This was consistent with findings from Theobald *et al.* [3], who found that binary-gender did not have predictive power with respect to students reporting having a dominator in their group or in their feeling comfortable with their group. We re-ran the regressions without *Gender* and found that this improved the model fit in all cases, as measured by AIC (a relative estimator of fit quality which accounts for goodness of fit and simplicity of the model). The results of the four four-predictor regressions are shown in Table II.

We use the rest of this section to discuss, for completeness, additional alternate sets of models that were considered to test the robustness of the results and to test the effects of alternate measures to those used in the main set of models.

We considered two additional measures of familiarity as replacements for *Know.Avg* in the four regression models. The first, *Know.Max*, extracted the maximum familiarity with group members rating from those used to calculate *Know.Avg*. We hypothesized that a person's strongest relationship in the group could be associated differently with our group exam experience measures than their average familiarity with group

| | EFA Factors | | | |
|---|---|---|---|---|
| | Presence of dominators | Presence of under-contributors | Productive group-work behaviours | Personal experience |
| *Know.Avg* | $-0.17$ | $-0.13$ | $0.26$ ** | $0.32$ *** |
| *Solo.Score* | $0.03$ | $0.14$ | $-0.10$ | $-0.18$ * |
| *Group.Score* | $-0.08$ | $-0.12$ | $0.22$ * | $0.34$ *** |
| *Particip.Self* | $0.13$ | $0.23$ ** | $-0.13$ | $0.02$ |
| $R^2$ | .05 | .10 ** | .14 *** | .21 *** |

TABLE II. The top four rows of numerical cells show the linear regression coefficients for models predicting the factor scores for the four standardized EFA factors (columns) from four standardized predictors (rows). *Gender* was removed from all four models because it provided no statistically significant coefficients and it worsened the quality of all fits, as measured by AIC. Statistical significance reported as *: $p<.05$; **: $p<.01$; ***: $p<.001$.

members. We found that this resulted in only small differences in the results presented in Table II, with the most notable difference being that in predicting the *presence of dominators* factor, *Know.Avg* became a statistically significant ($p<.05$) regression coefficient for *Know.Max*.

The second alternate familiarity measure was "Are you friends with at least one person that was in your group?" (from Ref. [3]), which provides a different and possibly more broad measure of familiarity. Unlike *Know.Avg* and *Know.Max*, this friend item offered no predictive power with respect to our four regression models.

## IV. DISCUSSION AND CONCLUSIONS

We remind the reader that the results presented here are from an exploratory study where we performed many comparisons and considered many models. Thus, statistical significance is likely to be overstated. Instead, we use the results to indicate possible trends and suggest areas of focus for future studies. Additionally, the student composition of summer courses is often different from the composition of the same courses during the regular academic year, so we are cautious of generalizing these finding across terms.

The results from this factor analysis demonstrate a structure consistent with the underlying item design of our group exam experiences survey, with factors emerging for *personal experience*, *productive group-work behaviours*, the *presence of dominators* and the *presence of under-contributors*. The latter two factors target different mechanisms by which student might experience or perceive participation inequity within their group. As shown in Table I, the *personal experience* factor contains all of the surviving "I"-statement items other than "I felt criticized or judged by one or more group members."

Further development of survey items will aim to increase the number of items targeting each individual factor. Items that showed some cross-loading across factors may be retained to investigate how their loadings may change with the stability of results improvements that will come from a larger sample size.

The regression modelling of our factor scores demonstrated that this set of predictor variables captures only a relatively small fraction of the variance, with $R^2$ values ranging from .05 for the *presence of dominators* factor to .21 for the *personal experience* factor. We saw that group familiarity—operationalized as an individual's average response to how well they reported knowing each of the other individuals in their group academically—provided predictive power in the *personal experience* and *productive group-work behaviours* factors, but not in the two participation equity factors, *presence of dominators* and *presence of under-contributors*. Although the lack of predictive power for familiarity on the two participation equity factors is perhaps surprising, it must be noted that high familiarity does not imply intention on the part of an individual student in putting their group together. Thus a student's ability to avoid groups with high participation inequity may only share a small correlation with our familiarity measures.

Since our study was not able to investigate causal relationships, we can only hypothesize that interventions targeted at improving students' sense of academic familiarity with their groupmates could have an overall positive impact on their group exam experience. We have subsequently piloted an in-class pre-midterm group review intervention, where students were encouraged to arrange their groups ahead of time, and were helped to form groups during the review activity if needed. In addition to targeting improvements in intra-group familiarity, we also sought to improve equity in participation and decision-making in this intervention through the sharing sharing of curated responses to a previous survey item: "Provide advice you would give to a future student in this class to get the most out of their group exam experience". Although we have not yet analyzed the results of this pilot intervention, an example from group testing in medical education [4] supports the plausible efficacy of the familiarity component of this intervention.

## ACKNOWLEDGMENTS

We gratefully acknowledge the financial support for this project provided by UBC Vancouver students via the Teaching and Learning Enhancement Fund. We also thank Joy Chen, Maggie Wu and Rosanne Persaud for their contributions to the project.

[1] G. W. Rieger and C. E. Heiner, Examinations that support collaborative learning: The students' perspective, Journal of College Science Teaching **43**, 41 (2014).

[2] C. Rawn, J. Ives, and B. Gilley, Two-stage exams increase learning and laughter on exam day in classes of any size, in *Strategies for Teaching Large Classes Effectively in Higher Education* (Cognella Academic Publishing, San Diego, CA, 2018).

[3] E. J. Theobald, S. L. Eddy, D. Z. Grunspan, B. L. Wiggins, and A. J. Crowe, Student perception of group dynamics predicts individual performance: Comfort and equity matter, PLoS One **12**, e0181336 (2017).

[4] G. J. Sinner, J. C. Briggs, F. T. Stevenson, and S. J. Nazian, Group testing in medical education: An assessment of group dynamics, student acceptance, and effect on student performance, Medical Science Educator **23**, 346 (2013).

[5] B. L. Wiggins, S. L. Eddy, L. Wener-Fligner, K. Freisem, D. Z. Grunspan, E. J. Theobald, J. Timbrook, and A. J. Crowe, ASPECT: A survey to assess student perspective of engagement in an active-learning classroom, CBE Life Sciences Education **16**, 1 (2017).

[6] N. Dasgupta, M. M. Scircle, and M. Hunsinger, Female peers in small work groups enhance women's motivation, verbal participation, and career aspirations in engineering, Proceedings of the National Academy of Sciences **112**, 4988 (2015).

[7] S. L. Eddy, S. E. Brownell, P. Thummaphan, M.-C. Lan, and M. P. Wenderoth, Caution, student experience may vary: social identities impact a studentâs experience in peer discussions, CBE Life Sciences Education **14**, ar45 (2015).

[8] W. Revelle, *psych: Procedures for Psychological, Psychometric, and Personality Research*, Northwestern University, Evanston, Illinois (2019), r package version 1.9.12.

[9] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2013).

[10] D. L. Bandalos and S. J. Finney, Factor analysis: Exploratory and confirmatory, in *The reviewer's guide to quantitative methods in the social sciences* (Routledge, 2018) pp. 98–122.

[11] E. Knekta, C. Runyon, and S. Eddy, One Size Doesn't Fit All: Using Factor Analysis to Gather Validity Evidence When Using Surveys in Your Research, **18**, rm1 (2019).

[12] B. G. Tabachnick and L. S. Fidell, *Using multivariate statistics*, 6th ed. (Pearson Education, Boston, 2013).

[13] A. L. Traxler, X. C. Cid, J. Blue, and R. Barthelemy, Enriching gender in physics education research: A binary past and a complex future, Physical Review Physics Education Research **12**, 020114 (2016).