Interprofessional Education Perceptions: Intra-rater Reliability of Longitudinal Retrospective Pre-Test Design Scores

Tina Patel Gunaldo PhD, DPT, MHS *Louisiana State University Health Sciences Center at New Orleans*
Kari Brisolara ScD, MSPH *Louisiana State University Health Sciences Center at New Orleans*
Sonia Gasparini PhD *Louisiana State University Health Sciences Center at New Orleans*
Donald Mercante PhD *Louisiana State University Health Sciences Center at New Orleans*
Cornelius Rosenbaum MA *Louisiana State University Health Sciences Center at New Orleans*
Chrissie Alving-Trinh BS *Louisiana State University Health Sciences Center at New Orleans*

Abstract

INTRODUCTION The purpose of the study was to investigate student intra-rater reliability of their retrospective pre-response after completing both year one and year two of an IPE curriculum. Determining intra-rater reliability of student retrospective responses is fundamental when reporting student learning outcomes over a longitudinal experience.

METHODS First and second year health professional students completed a two year IPE curriculum. The Student Perceptions of Interprofessional Clinical Education-Revised instrument, version 2 (SPICE-R2) was administered after year one and two of the curriculum using a retrospective pre-/post-test design. Students were asked to retrospectively reflect on the period of time prior to entering the two year curriculum. The weighted kappa statistic was used to measure agreement between the two retrospective pre-responses.

RESULTS There was no statistically significant ($p>0.05$) difference found between the two retrospective pre-time point means for all students for the SPICE-R2 total score and three factors. However, the results of the weighted kappa statistic for each question showed fair agreement and each weighted kappa statistic was found to be significant ($p< 0.05$).

DISCUSSION There was a lack of intra-rater reliability of student retrospective pre-response scores when measuring after year one and year two of the curriculum. Given this finding, when using a retrospective pre-/post-test design, longitudinal changes in IPE perceptions should be reported independently.

Received: 12/07/2021 *Accepted*: 02/28/2022

Implications for Interprofessional Practice

- It is important to consider the use of traditional pre-/post-test design versus retrospective pre-/post-test design when measuring interprofessional education perceptions or attitudes using quantitative instruments.
- There is a lack of intra-rater reliability of student retrospective pre-response scores when measuring after year one and year two of the curriculum.

INTRODUCTION

Measuring student learning is essential in the development and assessment of an effective curriculum. A commonly used method to measure interprofessional education (IPE) learning outcomes is through the use of validated perception or attitudinal questionnaires. The IPE literature includes the use of numerous validated perception or attitudinal quantitative surveys. These survey instruments have been used to measure change after single short-term experiences or longitudinally. The framework most frequently cited to measure the impact of IPE training is the expanded Kirkpatrick Model (Barr et al., 2000). The expanded Kirkpatrick Model includes four levels: 1) reaction, 2) learning, 3) behavior, and 4) results (Barr et al., 2000). Level 2 and 4 are further subdivided: 2a-modification of attitudes/perceptions, 2b-acquisition of knowledge/skills, 4a-change in organizational practice and 4b-benefits to patients/clients (Barr et al., 2000). There is encouragement from the IPE community to move beyond level 2 measurements. However, there are gaps in the education literature at level 2. Specific to this study is research related to the longitudinal use of a retrospective pre-/post-test design (RPP) to measure IPE perceptions.

Perception change can be measured using a traditional pre-/post-test design (TPP) or a retrospective pre-/post-test design (RPP). In TPP, participants are asked to self-evaluate perception and attitude in real time, once before an intervention and again after. Use of TPP is limited by response shift bias. Participants' post-test results might not reflect a shift in perception because of a shift in internal standards (knowledge) due to the intervention (Howard et al., 1979). RPP eliminates response shift bias by administering a survey at a single point after the intervention and asking participants to reflect on a time point before the experience, thus keeping the frame of reference constant (Howard, 1980). While RPP eliminates drawbacks such as response shift bias, it is not without limitations. Some limitations include social desirability bias, effort justification bias, impression management, maturation, consistency theory, and self-inflation bias (Hill, 2020). These limitations address basic human nature but do not account for recall bias.

Recall bias is a form of information bias that can shift data sets and analysis towards or away from the null hypothesis when data collected is purely based on memory, as are retrospective pre-responses in RPP (Howard, 1980). Individuals may incorrectly remember attitudes, perceptions, or depth of knowledge at a given time point. Findings from a case-controlled study investigating parental occupational exposure and leukemia in children indicated parental responses were influenced by recall bias. Parents with a child diagnosed with leukemia recalled specific one-time or insignificant events that could have contributed to the diagnosis, while parents with a child not diagnosed with leukemia did not report these types of events (Schüz, Spector, & Ross, 2003). In addition, as more time passed between the child's birthday and administration of the survey, parents with a child diagnosed with leukemia reported less exposure events, while reporting by parents in the control group remained constant (Schüz et al., 2003).

The effect of recall bias in retrospective pre-responses over time has not been studied in IPE. It is important to know the impacts of recall bias when using retrospective pre-responses in a repeated measure design. If retrospective pre-response scores are not the same, then each data collection point is independent of any other collection point, which limits comparisons over the time being investigated. Investigating the impacts of recall bias on retrospective pre-responses was important to Louisiana State University Health Sciences Center at New Orleans (LSUHSCNO) as related to reporting student IPE learning outcomes.

LSUHSCNO requires an IPE curriculum for first and second year students. The curriculum is two years in length spanning both fall and spring semesters. Student learning is measured through quantitative surveys, reflections and capstone projects. One of the quantitative surveys utilized the RPP test design. This survey was administered to students at the end of year one and repeated at the end of year two. It is unknown if students can reliably assess their retrospective pre-response over time. The purpose of the study was to investigate student intra-rater reliability of their retrospective pre-response after completing both year one and year two of the IPE curriculum. The research question "can students reliably assess retrospective pre-responses at the end of year one and year two of the curriculum?" was investigated. Answering this question is important in understanding how to interpret student learning outcomes of a longitudinal curriculum when using a RPP survey design.

METHODS

The large-scale IPE curriculum at LSUHSCNO is known as Team Up: Compassion, Communication and Collaboration™. Team Up™ is a two-year required program integrated within 6 Schools: Allied Health Professions, Dentistry, Graduate Studies, Medicine, Nursing, and Public Health (LSUHSCNO, 2017). Students were randomly assigned to 60 teams of 13-15 students representing various professions. Student teams met on a monthly basis, three times during the fall semester and three times during the spring semester. The last session in the spring semester of both years was a presentation of the final project. After the final presentations for year one and year two, students were given a week to complete the Student Perceptions of Interprofessional Clinical Education-Revised instrument, version 2 (SPICE-R2). The survey was created using a RPP design. A link to the SPICE-R2 survey was placed in the electronic learning platform used at the institution. For both time points, students were asked to retrospectively reflect on their perceptions prior to beginning Team Up™.

Participants

Students were first and second year students, enrolled in one of eighteen health professional programs. Students were grouped by School for analysis: Allied Health Professions (Audiology, Cardiovascular Sonography, Physical Therapy, Physician Assistant, Respiratory Therapy, Speech-Language Pathology); Dentistry (Dental Hygiene, Dentistry); Medicine; Nursing (undergraduate); Public Health (Behavioral and Community Health Sciences, Environmental and Occupational Health Sciences, Epidemiology, and Health Policy and Systems Management). Students completed both year one (2018-2019) and year two (2019-2020) of the curriculum.

Survey Instrument

The SPICE-R2 (Table 1) is a validated IPE instrument assessing student attitudes in three domains or factors (Interprofessional Teamwork and Team-based Practice, Roles/Responsibilities for Collaborative Practice, and Patient Outcomes from Collaborative Practice) (Zorek et al., 2017). The ten-item survey uses a five point Likert-type scale (1=strongly disagree, 2=disagree, 3=neutral, 4=agree, 5=strongly agree). Research studies using the SPICE-R2 have reported change in perception outcomes using TPP and RPP (Gunaldo et al., 2021; McGregor, Lanning, & Lockeman, 2018).

Analysis

The analysis was completed using the R programming language (version 4.0.2) and included only those students who completed both surveys at the end of year one and the end of year two. Data was analyzed at the all student level and school level. Significance of mean scores between time points was calculated using a paired t-test due to the sufficient sample size. A weighted kappa statistic, which is a reliability statistics, was also calculated to compare with the results of the t-test.

The weighted kappa statistic, denoted k_w, is the generalized kappa statistic for ordinal data (Cohen, 1960; Sim & Wright, 2005). Cohen devised both the normal kappa statistic, for nominal data, and the weighted kappa statistic, most often used for ordinal data. A kappa statistic measures agreement adjusting for chance agreement (Cohen, 1960). The kappa statistic ranges from -1 to +1, where -1 indicates disagreement beyond chance, 0 means neither agreement/disagreement beyond chance, and +1 means agreement beyond chance (Cohen, 1960). We selected the weighted kappa statistic because the Likert scale data is ordinal and this allows us to penalize those answers that are one Likert scale off from the correct response to a lesser extent, and to penalize those answers that are further away from the correct response to a greater extent (Cohen; Sim & Wright, 2005). For the study, the equal weighting scale was used. That is, the difference between subsequent Likert scale responses is just 1. Equal weights were used, meaning that the difference between any sequential Likert answers, for example 2 and 3 is just 1. Kappa statistics are often reported with an interpretation based on a range in which the kappa statistic itself falls into (Landis & Koch, 1977).

RESULTS

Five hundred fifty-two students completed the SPICE-R2 at the end of year one and year two (Allied Health n=99; Dentistry n=104; Medicine n=187; Nursing n=148; Public Health n=14). Table 2 includes the retrospective pre-test mean at both time points for each of the ten questions, each of the three factors and the total SPICE-R2. There was no statistically significant (p>0.05) difference found between the two time point means for all students and students in each of the Schools, except the Patient Outcomes factor for the School of Public Health.

	Question	Factor
1.	Working with students from different disciplines enhances my education.	Teamwork
2.	My role within an interprofessional team is clearly defined.	Roles/Responsibilities
3.	Patient/client satisfaction is improved when care is delivered by an interprofessional team.	Patient Outcomes
4.	Participating in educational experiences with students from different disciplines enhances my ability to work on an interprofessional team.	Teamwork
5.	I have an understanding of the courses taken by, and training requirements of, other health professionals.	Roles/Responsibilities
6.	Healthcare costs are reduced when patients/clients are treated by an interprofessional team.	Patient Outcomes
7.	Health professional students from different disciplines should be educated to establish collaborative relationships with one another.	Teamwork
8.	I understand the roles of other health professionals within an interprofessional team.	Roles/Responsibilities
9.	Patient/client-centeredness increases when care is delivered by an interprofessional team.	Patient Outcomes
10.	During their education, health professional students should be involved in teamwork with students from different disciplines in order to understand their respective roles.	Teamwork

Table 1. *SPICE-R2 Items and Respective Factors*

	End of Year One		End of Year Two			
SPICE-R2 Question (Q)	Mean	SD	Mean	SD	P-value	Cohen's D
Q1	3.92	0.891	3.88	0.791	0.312	0.043
Q2	3.67	0.995	3.69	0.934	0.779	0.012
Q3	4.34	0.688	4.25	0.695	0.003	0.127
Q4	3.84	0.912	3.77	0.874	0.116	0.067
Q5	2.93	1.021	3.00	1.037	0.163	0.060
Q6	3.45	0.924	3.65	0.901	<0.001	0.191
Q7	3.98	0.799	3.98	0.786	0.881	0.006
Q8	3.43	0.954	3.39	0.924	0.374	0.038
Q9	4.09	0.714	4.06	0.748	0.328	0.042
Q10	3.71	0.936	3.74	0.618	0.618	
Factors						
Teamwork	15.45	2.906	15.37	2.740	0.502	0.029
Roles	10.03	2.292	10.08	2.313	0.695	0.017
Patient Outcomes	11.89	1.774	11.96	1.903	0.432	0.033
Total	37.37	5.373	37.40	5.577	0.901	0.005

Table 2. *SPICE-R2 descriptive statistics, t-test and Cohen's D results for retrospective pre-responses at the end of year one and year two of the IPE curriculum for all students*

While no differences were found in paired means, except for one factor and one School, the weighted kappa statistic revealed disagreement between the two time points for the same rater. Table 3 reports the results of the weighted kappa statistic for each question. All ten questions showed fair agreement. Fair agreement is defined as any kappa statistic between the bounds of .2 and .4. Most of the data fell between 0.2 and 0.3 with a mean of 0.294, a median of 0.289, and minimum of 0.259 and maximum of 0.338. Further, each weighted kappa statistic was found to be significant (p< 0.010).

SPICE-R2 Question (Q)		P-value	Interpretation
Q1	0.335	<0.001	Fair Agreement
Q2	0.309	<0.001	Fair Agreement
Q3	0.301	<0.001	Fair Agreement
Q4	0.314	<0.001	Fair Agreement
Q5	0.276	<0.001	Fair Agreement
Q6	0.267	<0.001	Fair Agreement
Q7	0.338	<0.001	Fair Agreement
Q8	0.259	<0.001	Fair Agreement
Q9	0.264	<0.001	Fair Agreement
Q10	0.278	<0.001	Fair Agreement

Table 3. *Weighted kappa statistic values for SPICE-R2 questions comparing retrospective pre-responses at the end of year one and year two of the IPE curriculum for all students*

DISCUSSION

Longitudinal studies measure a variable(s) over time. According to the Theory of Planned Behavior, a positive change in IPE perceptions after engagement in a longitudinal IPE curriculum could support a change in collaborative behaviors (Fishbein & Ajzen, 1975). Being able to measure this change can support IPE integration in curricula. Based upon the findings from this study, when students were asked to reflect upon the retrospective pre-time point of prior to Team Up^{TM}, student responses after year one and year two varied. Therefore, students could not reliably recall their retrospective pre-response at two different time points. RPP is often cited as more reliable than TPP due to lack of response shift bias (Little et al., 2020). However, the lack of reliability in consistently reporting the same retrospective pre-test time point questions the reliability of this point for comparison. This finding illustrates a drawback of the RPP method, which should be considered when evaluating and reporting longitudinal outcomes. In this case, the change in perception after year one is independent of the change in perception after year two.

When quantitative measurements are taken in real time, one could evaluate change in scores between any two points. However, when using a RPP test design, the change occurring within a single time frame is independent of the change occurring during a second consecutive time frame. For example, if the change after engaging in Team Up^{TM} year one is 2.0 and the change after year two is 3.0 using the same retrospective pretest reference point, the change in scores should be reported independently. Based upon the findings from this study, one should not interpret that 2.0 points of change occurred in year one and 1.0 points of change occurred in year two. In addition, when assessing a change after year two of the curriculum using the same example, it would be difficult to indicate when the 3.0 point change occurred over the two year period. We could only state that there was a 2.0 point change that occurred in year one and a 3.0 point change that occurred over a period of two years, so it is more difficult to quantify the change that occurred during year two only.

Although not reported in the results section of the paper, the post-test total SPICE-R2 score for year one was 39.21 and for year two, 42.13. The change in perceptions over the first year of the curriculum was 1.84 and the change over both years was 4.73. While we know a positive change in perceptions occurred over year one and both years, using the RPP design limits our ability to specifically measure a change over year two of the curriculum. Based upon the differences in means, there was greater change in perceptions over two years, as compared to one year. For educators, this result provides support for a longitudinal two-year curriculum regardless of the change occurring each year.

Theoretically, if reflection and memory were perfect, the retrospective pre-test mean after year one would be identical to the mean after year two. However, these values differed. This difference in retrospective pre-test perceptions could be attributed to recall bias since the data was completely dependent on students' memories.

Memory is not always reliable, and as time goes on, memory of knowledge and specific attitudes can fade. Research indicates that up to 20% of specific detail about a personal event is forgotten after one year (Bradburn et al., 1987). Reflection after two years is thus more difficult to ascertain previous perception and level of knowledge. Since details of events fade over time, people depend on inferences to fill in the gaps (Bradburn et al., 1987). Inferences can sway autobiographical data thus influencing the retrospective pre-test perception data.

The difference could also be attributed to the vague choices and descriptors used on the Likert scale, making it harder for participants to pick choices that accurately described their perception at that time and harder to reliably report. Developing more objective measurement choices could possibly decrease the difference seen between retrospective pre-test scores after year one and year two.

The finding from this research is new to the IPE literature and was researched at a single institution. Other institutions using a repeated RPP design have the opportunity to contribute to the body of research in this area. Additional research at other institutions would allow for findings to be compared even if different instruments and time frames were utilized. This study measured students post-year one and post-year two. However, it would be beneficial to know if reliability of the retrospective pre-test score improves if the time frame was shortened to six months and twelve months, for example.

Kappa statistics come with limitations. Many papers have been written discussing such limitations. These arguments concern the prevalence, bias, symmetry (or asymmetry), and balance (or imbalance) of the categories and their responses (Byrt, Bishop & Carlin, 1993; Cicchetti & Feinstein, 1990; Feinstein & Cicchetti, 1990; Flight & Julious, 2015; Sim & Wright, 2005). To adjust for these concerns, there exist the prevalence and bias adjusted kappa (PABAK) (Byrt, Bishop, & Carlin, 1993). However, no such adjustment exists for the weighted kappa statistics. Graham and Jackson (1993) discussed issues specific to the weighted kappa, and consider use of the ordinal logistic regression in the place of the weighted kappa statistic. However, we found that such concerns were not relevant here.

CONCLUSION

As noted before, the findings of this study are new to the IPE literature. The focus of this paper was to discover if students could reliably report retrospective pre-scores at two different time points over a period of two years. The results indicate fair agreement. More research is needed in this area to assist educators in selecting the most appropriate measurement design to evaluate curricula.

References

Barr, H., Freeth, D., Hammock, M., Koppel, I., & Reeves, S. (2000, August). Evaluations of interprofessional education. https://www.caipe.org/resources/publications/barr-h-freethd-hammick-m-koppel-i-reeves-s-2000-evaluations-of-interprofessional-education

Bradburn, N. M., Rips, L. J., & Shevell, S. K. (1987). Answering autobiographical questions: the impact of memory and inference on surveys. *Science*, *236*, 157-161. https://doi.org/10.1126/science.3563494

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology,46*(5), 423-429. https://doi.org/10.1016/0895-4356(93)90018-V

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, *43*(6),551-558. https://doi.org/10.1016/0895-4356(90)90159-M

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213-220. https://doi.org/10.1037/h0026256

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1):37-46. https://doi.org/10.1177/001316446002000104

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low Kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*;43(6),543-549. https://doi.org/10.1016/0895-4356(90)90158-L

Fishbein, M., & Ajzen, I. (1975). *Beliefs, attitudes, intentions, and behavior*. Boston: Adison-Wesley.

Flight, L., & Julious, S. A. (2015). The disagreeable behaviour of the kappa statistic. *Pharm Stat*, *14*(1),74-78. https://doi.org/10.1002/pst.1659

Graham, P., & Jackson, R. (1993). The analysis of ordinal agreement data: beyond weighted kappa. *Journal Clinical Epidemiology, 46*(9),1055-1062. https://doi.org/10.1016/0895-4356(93)90173-X

Gunaldo, T. P., Lockeman, K., Pardue, K., Breitbach, A., Eliot, K., Goumas, A., Kettenbach, G., Lanning, S., & Mills, B. (2021). An exploratory, cross-sectional and multi-institutional study using three instruments to examine student perceptions of interprofessional education. *Journal of Interprofessional Care*, *36*(2), 1–8. https://doi.org/10.1080/13561820.2021.1892614

Hill, L. G. (2020). Back to the future: Considerations in use and reporting of the retrospective pretest. I*nternational Journal of Behavioral Development*, *44*(2), 184–191. https://doi.org/10.1177/0165025419870245

Howard, G. S. (1980). Response-shift bias: A problem in evaluating interventions with pre/post self-reports. *Evaluation Review, 4*(1), 93–106.https://doi.org/10.1177/0193841X8000400105

Howard, G. S., Ralph, K. M., Gulanick, N. A., Maxwell, S. E., Nance, D. W., Gerber, S. K. (1979). Internal invalidity in pretest-posttest self-report evaluations and a re-evaluation of retrospective pretests. *Applied Psychological Measurement, 3*(1),1-23. https://doi.org/10.1177/014662167900300101

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*,*33*(1),159. https://doi.org/10.2307/2529310

LSUHSCNO. (2017). Team Up: Commit to Compassion, Communication, and Collaboration. http://www.lsuhsc.edu/administration/academic/cipecp/team_up_overview.aspx. Accessed: March 15, 2021.

McGregor, M. R., Lanning, S. K., & Lockeman, K. S. (2018). Dental and dental hygiene student perceptions of interprofessional education. *Journal of Dental Hygiene, 92*(6), 6–15.

Schüz J., Spector, L. G., & Ross, J. A. (2003). Bias in studies of parental self-reported occupational exposure and childhood cancer. *American Journal of Epidemiology*, *158*(7), 710-6. https://doi.org/10.1093/aje/kwg192. PMID: 14507608

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*,*8 5*(3),257-268.https://doi.org/10.1093/ptj/85.3.257

Zorek, J. A., Lockeman, K. S., Eickhoff, J. C., & Gunaldo, T. P. (2017). Multi-institutional validation of the Student Perceptions of Interprofessional Clinical Education-Revised instrument, version 2 (SPICE-R2). American Interprofessional Health Collaborative and Canadian Interprofessional Health Collaborative, Collaborating Across Borders Meeting, Banff, Canada, October 2, 2017.

Corresponding Author

Tina Patel Gunaldo, PhD, DPT, MHS

Center for Interprofessional Education
and Collaborative Practice
LSU Health Sciences Center New Orleans
433 Bolivar Street
New Orleans, LA 70112

tgunal@lsuhsc.edu