Examining the efficacy of a professional development assessment tool

Alistair McInerny
Physics, North Dakota State University, 1340 Administration Ave, Fargo, North Dakota, 58105

Mila Kryjevskaia[1], Alexey Leontyev[2]
[1]*Physics, North Dakota State University, 1340 Administration Ave, Fargo, North Dakota, 58105*
[2]*Chemistry, North Dakota State University, 1340 Administration Ave, Fargo, North Dakota, 58105*

Professional development is an effective way to diffuse evidence-based instructional practices and support their implementation. At the same time, assessing the impacts of professional development opportunities in physics education is still an emerging field. In this paper, we examine the efficacy of an assessment instrument based on the Theory of Planned Behavior, a theoretical framework originating from psychology. The theory has been used in other fields to understand and predict behaviors, but it has not been utilized in STEM Education research. The theory posits that participants' attitudes, norms, and perceived behavioral control beliefs toward a particular behavior (e.g., implementation of active learning) predict intentions which, in turn, determine and explain that behavior. Previous work presented empirical data that show that a link exists between intentions (as measured by the instrument) and adoption of active learning strategies (as measured by the COPUS observation protocol). In this paper, we focus on examining evidence for the validity and reliability of measurements produced by the instrument, which is necessary to provide further validity evidence to the empirical results of the prior study. We also discuss the retrospective pre-test methodology for data collection which helps minimize the effects of biases associated with self-reported data collected via traditional pre-post survey administrations. We used Arjoon, Xu, and Lewis's framework for validity and provide evidence for the reliability, internal structure, and temporal stability of our measures. We will also compare results from our instrument to those obtained using the Approaches to Teaching Inventory.

2021 PERC Proceedings edited by Bennett, Frank, and Vieyra; Peer-reviewed, doi.org/10.1119/perc.2021.pr.McInerny

I. INTRODUCTION

Benefits of active learning strategies for students in STEM courses are well documented [1,2,3]. Despite that, lecture remains the most prominent mode of instruction throughout all STEM disciplines [4]. Research into the diffusion of effective active learning strategies (ALS) revealed that simply making potential adopters aware of the existence of pedagogical innovations is not likely to enact changes in instructors' behaviors. Henderson and colleagues argue that effective strategies for facilitating change "are aligned with or seek to change the beliefs of the individuals involved; involve long-term interventions, lasting at least one semester; require understanding a college or university as a complex system and designing a strategy that is compatible with this system." [5] Currently, dissemination efforts are shifted toward longer-term professional development (PD) programs such as workshops and faculty learning communities (FLC) [5-8]. However, assessment of their effectiveness remains particularly challenging.

Ideally, observable changes in behavior should provide evidence for the impacts of a PD program. Despite increased attention to the development of teaching observation protocols, their implementations are typically limited by practical constraints [4] (e.g., training and paying observers, logistical issues, etc.). Instead, it is common to rely on self-report measures (e.g., surveys, interviews) to collect data on participants' teaching practices and their perceptions of the effectiveness of PD [9, 10]. These efforts are critical for moving PD research forward. We argue, however, that it is also valuable to explore a different assessment methodology that uses self-reported data to *predict* changes in behavior.

Our team drew on PD practices in other fields and psychology research to design, implement, and evaluate a PD assessment methodology based on the *Theory of Planned Behavior (TPB),* which has been successful at explaining and predicting a range of behaviors, such as smoking and exercising [11,12]. The TPB framework was used to develop an instrument that relies on participants' self-reported data to predict their intentions toward an implementation of ALS. Those intentions determine whether this behavior will be performed. A previous study [13] demonstrated that a link exists between intentions to use ALS (as measured by the instrument) and the adoption of that behavior (as measured by COPUS observations [4, 14]). The data suggest that instructors with higher post-PD intentions to adopt ALS spend more time using peer instruction techniques (e.g., clicker questions) and/or group work, and less time lecturing. This empirical result could be considered "proof of concept" for the utility of the instrument for PD; however, the study does not discuss evidence for the validity and reliability of measurements, which is the primary goal of this paper. We use the framework presented by Arjoon *et al* to examine evidence for reliability, temporal stability, internal structure, and relation to other variables [15]. We also discuss an important element of our assessment methodology, the *retrospective* pre-test, which addresses a phenomenon characteristic of self-report measures, response-shift bias, which is not addressed by traditional pre-/post-test methodology [16, 17].

II. DESIGN AND METHODOLOGY

A. *Theoretical framework and instrument design*

The Theory of Planned Behavior [11, 12] suggests that an individual's intent to perform an action (*e.g.,* implement ALS) can be predicted by a combination of three factors: 1) positive or negative feelings toward that action (attitudes), 2) beliefs about important others' (peers, coworkers) attitudes towards that action (norms in the community), and 3) perception of affordances and the individual's abilities to perform that action (perceived behavioral control or PBC). The intentions to perform an action then determine actual behavior. In this framework, the effects of other important aspects, such as demographic characteristics or experiences, are already accounted for by one of the three factors above. For example, if a participant experienced being stereotyped by race or gender in educational settings, they might believe that more inclusive teaching methodologies are important, thus exhibiting a more positive attitude towards ALS. They may also feel less supported in their department, which would affect their normative beliefs. If they are a victim of the stereotype threat, they may express doubts about their abilities to create an engaging learning environment thus underestimating their PBC.

The development of the instrument (*the Beliefs and Intentions to Use Active Learning Strategies survey,* BIUALS) followed standard practices of psychological measurements as prescribed by Ajzen and Fishbein [11,12]. A pilot study elicited salient beliefs about behavioral outcomes, normative referents, and control factors. The results were used to create 7-point Likert scale items. Most items utilize response scales ranging from "strongly agree" to "strongly disagree," as shown in Table I. These scales are common for self-report measures developed in PER [18]. Response scales for attitude items use the standard technique which asks a respondent to evaluate a statement on a set of bipolar adjectives. The adjectives chosen should reveal a person's evaluation or feeling of "favorableness or unfavorabliness" toward the behavior in question. To discourage repetitive answers, reverse scales are employed on several items.

It was also important to develop items that directly correspond to participants' *own* attitudes toward and intentions to perform a *specific* action under defined circumstances as opposed to general behavior. As such, all items (1) specify that the statements concern participants' own evaluations (e.g., "for me," "my department,") and (2) include a common descriptor of the expected behavior (*i.e.,*

TABLE I. Three and four factor component analysis of BIUALS survey data

Item	3 Factors			4 Factors			
	C 1	C 2	C 3	C 1	C 2	C 3	C 4
Attitude 1: "For me, using an ALS in the classroom at some time during the next month is [Extremely Pleasing … Extremely Annoying]"	-	-	-.91	-	-	-.89	-
Attitude 2: "For me, using an ALS in the classroom at some time during the next month is [Extremely Negative … Extremely Positive]"	-	-	-.86	-	-	-.90	-
Attitude 3: "For me, using an ALS in the classroom at some time during the next month is something I [Extremely like … Extremely dislike]"	-	-	-.88	-	-	-.91	-
Attitude 4: "For me, using an ALS in the classroom at some time during the next month is [Extremely worthless … Extremely valuable]"	-	-	-.74	-	-	-.76	-
Attitude 5: "For me, using ALS is in the classroom at some time during the next month is [Extremely punishing … Extremely rewarding]"	-	-	-.88	-	-	-.89	-
Norm 1: "My department chair/head thinks I should use an ALS in the classroom at some time during the next month" [Strongly disagree … Strongly agree]	-	-.79	-	-	-.72	-	-
Norm 2: "My department chair/head approves of my using an ALS in the classroom at some time during the next month" [Strongly disagree … Strongly agree]	-	-.81	-	-	-	-	-.85
Norm 3: "My department chair/head wants me to use an ALS in the classroom at some time during the next month" [Strongly disagree … Strongly agree]	-	-.84	-	-	-.78	-	-
Norm 4: "My department chair/head would support me using an ALS in the classroom at some time during the next month" [Strongly disagree … Strongly agree]	-	-.77	-	-	-	-	-.90
Norm 5: "My department colleagues think I should use an ALS in the classroom at some time during the next month" [Strongly disagree … Strongly agree]	-	-.81	-	-	-.91	-	-
Norm 6: "My department colleagues approve of my using an ALS in the classroom at some time during the next month" [Strongly disagree … Strongly agree]	-	-.85	-	-	-.21	-	-.75
Norm 7: My department colleagues want me to use an ALS in the classroom at some time during the next month" [Strongly disagree … Strongly agree]	-	-.80	-	-	-.92	-	-
Norm 8: "Most of my department colleagues will use an ALS in the classroom at some time during the next month" [Strongly disagree … Strongly agree]	-	-.71	-	-	-.59	-	-.22
Norm 9: "My department colleagues would support me using an ALS in the classroom at some time during the next month" [Strongly disagree … Strongly agree]	-	-.76	-	-	-	-	-.77
PBC 1: "It would be difficult for me to use an ALS in the classroom at some time during the next month" [Strongly disagree … Strongly agree]"	.61	-	-	.54	-	-	-.30
PBC 2: "How much control do you have over whether you use an ALS in the classroom at some time during the next month" [Completely No Control … Complete Control]	.89	-	-	.89	-	-	-
PBC 3: "It is mostly up to me if I use an ALS in the classroom at some time during the next month" [Strongly disagree … Strongly agree]"	.81	-	-	.84	-	-	-
PBC 4: "I feel confident that I could use an ALS" [Strongly disagree … Strongly agree]	.61	-	-.32	.59	-	-.32	-
PBC 5: "I feel confident that I could use an ALS in the classroom at some time during the next month even if I was very busy." [Strongly disagree … Strongly agree]	.59	-	-	.60	-	-	-
PBC 6: "I feel confident that I could use an ALS in the classroom at some time during the next month even if I was in a bad mood" [Strongly disagree … Strongly agree]	.68	-	-	.67	-	-	-
PBC 7: "I feel confident that I could use an ALS in the classroom at some time during the next month even if the weather was bad" [Strongly disagree … Strongly agree]	.70	-	-	.70	-	-	-
Cronbach's α	.90	.93	.93	.90	.91	.93	.88

the behavior will be performed "in the classroom at some time during the next month."). While seemingly repetitive, these descriptors are necessary. Table 1 contains all attitudes, norms, and PBC items as well as results of the factor analysis discussed in Section III.

B. *Addressing the response shift bias*

Traditionally, self-reported data obtained through surveys administered pre- and post-intervention are used to gauge changes resulting form that intervention. In the context of PD, it is expected that the intervention will lead to positive (or desired) shifts in survey responses. At the same time, it is expected that the participants' understanding of affordances and limitations of ALS will become more sophisticated [16, 17]. This shift in understanding may, in turn, result in a shift (or recalibration) in the internal reference frames participants use for assessing their *current* attitudes, beliefs, and perceptions regarding ALS. Because of this recalibration, self-reported pre- and post-test responses often suffer from the phenomenon known as the *response shift bias*. In other words, participants' attitudes and beliefs regarding ALS that were originally reported with high marks, in retrospect, are likely to be over-estimated. A common practice employed to control for the effect of response shift bias is the *retrospective*-pre-test methodology.

The retrospective-pre-test is identical to the pre-test and post-test but asks participants to evaluate themselves as they were prior to the PD. It is administered after the PD, at the same time as the post-test. The goal is for participants to re-assess their pre-intervention beliefs about ALS using the same internal reference frame that informed their post-intervention responses. Prior study has shown that response shift bias exists in the data from our population [15]: pre- and post-test comparisons do not accurately capture the impacts of PD suggesting that retrospective pre-test should be used in place of a traditional pre-test. In this study, we provide further evidence for the accuracy of retrospective-pre-tests.

C. *Population and data collection*

The instrument was administered to cohorts of STEM faculty enrolled in a university-wide PD program intended to support faculty adoption of ALS [19]. The instrument was given before the start of the program as a pre-test (Pre). It was also given after the first 2-day workshop as both a post-test (Post-test 1) and as a retrospective pre-test (Retro pre-test 1). It was administered again 6 months later after a semester of FLC and the second 2-day workshop (Post-test 2 and Retro pre-test 2). Retrospective surveys ask participants about beliefs they had before the start of the program (e.g., first day of the program). The multi-year, multi-cohort nature of this study imposed some limitations on the data collection. All participants (N=80) provided responses used in the analysis of the structure and reliability, but only 53 participants responded to both retrospective pre-tests that were matched to probe temporal stability.

III. DATA AND ANALYSIS

A. Internal structure

Analysis of the internal structure was performed to examine the relationships among items. An exploratory factor analysis was done to establish the degree to which the instrument structure is aligned with the theoretical basis for its design, namely items corresponding to the same construct load on the same factor. This analysis is critical for validating the results of the previous, "proof of concept" study [13] since in the absence of evidence from factor analysis, it would be invalid to average item scores to find a single factor score – an approach used to calculate differences in pre- and post-PD scores to determine change.

Eigenvalues suggest the existence of 4 factors. The scree plot indicates that retaining 3 factors was also reasonable. Extracting 4 factors produced curious results. Two of the four factors loaded exactly as expected (Attitudes and PBC), but the Norms items split into two factors. By inspection, it appears that one sub-set of these items concerns whether the department wants a participant to implement ALS, while the other subset relates to whether the department will actually support such an action. This result illustrates a dichotomy that often occurs and presents a significant obstacle to instructional innovations: the department recognizes advantages of ALS but may not necessarily support an initiative to adopt these strategies because of the perceived constraints, such as concerns of reduced content coverage, lack of TA support, poor student evaluations, etc. The detection of such a pattern could serve as a red flag indicating the dangerous misalignment between intentions and practices, similar to issues discussed by Henderson et al. [5, 20, 21].

The 4-factor model explains an additional 4% of variance compared to the 3-factor structure. The Norms items, however, do not significantly overlap with other categories suggesting the viability of the 3-factor structure as well. Extracting 3 variables reveals that all items align with their presumed categories as designed. While it may be advantageous to break Norms items into 2 factors for reasons above, taken together, they still measure "peers' attitudes."

B. Reliability

Reliability was established by exploring a) internal consistency using Cronbach's α and b) temporal stability using a modified approach to test-retest analysis [16].

Internal consistency was explored by probing the degree to which participants responding to certain items in a particular way were likely to respond to other items corresponding to the same construct in the similar way. Items designed to measure a particular construct should correlate with each other; if the correlation is weak or not present the items should either be removed or redesigned. Analysis revealed that the reliability of each factor is quite high (for both 3- and 4- factor structures). Further investigation into the "reliability if deleted" shows that removing any given item still leaves Cronbach's α above 0.8 for all factors.

One can argue that, for practical considerations, the number of items per construct could be reduced because of the apparent redundancy. While practicality is key for adoption of the instrument, a significant reduction in the number of items per construct will likely affect precision of the measurements. Ajzen and Fishbein recommend using at least 3 items per construct. A common rule of thumb for Likert scale is to include at least 4 items per construct to treat average scores as quasi-continuous. If users wish to adopt this instrument as a research tool, retaining 4 items per construct will likely be acceptable. In this case, the practical utility of the instrument may be improved while maintaining precision. If a ceiling effect is suspected, more precise measurements may be needed to detect smaller changes that could be missed via a reduced version.

Temporal stability is a measure of the instruments' reliability over time. While certain experiences are expected to cause participants' responses to change (e.g., instructional interventions), the simple passage of time should not. Typically, temporal stability is established by administering an instrument twice to the same population in a short period of time, eliminating the possibility that any changes in the

test-retest measurements, if observed, are due to intervention. In our study, we used data from the retrospective pre-tests rather than a more traditional test-retest. Because we were limited to the population of instructors who applied to participate in the PD program, we were not able to identify a period of time during which the participants were not exposed (at least to some degree) to opportunities to think about, consider adopting, or even practice ALS. However, the retrospective pre-test methodology may allow for probing the temporal stability by examining how participants' retrospective self-evaluations of their views of ALS change over time, if at all. Specifically, the retrospective pre-test askes participants to re-evaluate beliefs about ALS that they had had prior to PD through the lens of their current understanding. While it is expected that the participants' understanding of ALS will change due to PD, their retrospective self-evaluations should be affected to a much lesser degree, if at all, over the same period of time.

TABLE II. Correlation coefficients between retrospective scores collected 6 months apart (Retro 1 - Retro 2) and Cohen's *d* of shifts in retrospective and post-test scores collected over the same time period (Retro 1 - Post 2).

Construct	Retro 1 - Retro 2 Corr. Coeff.	Retro1-Post 2 Effect Size
Intentions	0.68	0.78
Attitudes	0.55	1.11
Norms	0.74	0.36
PBC	0.66	0.87

Data analysis revealed that, indeed, scores on retrospective surveys administered 6 months apart remain relatively stable over time. The correlation coefficients between scores are shown in Table II. Paired *t*-tests did not detect any significant differences in retrospective responses for any construct between Retro 1 and Retro 2. Because these retrospective scores attempt to measure participants Pre-PD beliefs and are collected 6 months apart we believe this provides evidence for the stability of retrospective scores.

Paired *t*-tests applied to retrospective pre-test 1 and post-test 2 indicate desirable and statistically significant shifts in beliefs about and intentions to adopt ALS, as indicated by the effect sizes shown in Table II. (See Section II.B for an explanation why *retro*-pre-test was used to probe the impact of PD.) We note that the changes in participants' perceptions of norms trail behind changes in other constructs. This result is somewhat expected since changes in community norms are notoriously slow. While participants may start to perceive their community in a more favorable way (perhaps due to support through PD or FLC), it is likely that the changes in personal beliefs will outpace perceived changes in the norms.

The examination of the temporal stability of retrospective pre-test responses provides evidence that the instrument functions as intended, detecting changes due to the PD over time (comparison of Retro 1 and Post 2 scores) while also showing the stability of retrospective self-evaluations measured 6 months apart. This result adds evidence to the quality of inference provided by the BIUALS instrument.

C. Relation to other variables

Building empirical relationships between the constructs measured by the instrument and other theoretically relevant constructs could provide additional evidence for the validity. The TPB posits that intentions predict behavior. Thus, empirical results that establish the link between the measures of intentions and actual observable behavior constitute strong evidence for the validity of inference (i.e., high intentions will likely lead to desirable behavior). As discussed above, such a link was already established in a prior study. However, further investigations are necessary to explore the functioning of the instrument with different populations and various types of PD.

For completeness and transparency, we note that we also administered the *Approaches to Teaching Inventory* (ATI) [22, 23] to assess individual's pedagogical practices along two dimensions: instructor-focused approach and student-focused. The ATI was administered 6 months apart (as a pre-test and after workshop 2) to some cohorts of participants. The instrument did not detect any measurable changes in pre-/post- responses. This result is inconsistent with the data from COPUS observations over the same time period which revealed desirable shifts in adoption of ALS. We speculate that ATI items are not sensitive enough (at least for our population and PD). Response shift bias may also be a factor since the ATI was not given in the retrospective format.

IV. CONCLUSIONS

We made an argument for the utility of a new methodology for assessing effectiveness of PD programs. The TPB informed the development of an instrument to measure intentions which, in turn, determine behavior. While the link between the instrument's measures of intentions toward adoption of ALS and the actual behavior was already established empirically in a pilot study, this paper provides evidence for the validity and reliability of these measures. We also advocate for the adoption of the retrospective pre-test methodology for mitigating effects of response shift bias in self-report measures.

We speculate that potential users may be concerned about the length of the instrument and the seeming redundancy of some items. We do not dismiss these concerns but remain optimistic that if the adoption of the instrument gains momentum in the community, the community will merge (over time) on a version that balances usability with the quality of measurements.

IV. ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under grant no. DUE-1525056.

[1] National Research Council, *Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering* (National Academies Press, Washington, DC, 2012).

[2] National Research Council Committee on Undergraduate Physics Education Research and Implementation, *Adapting to a Changing World: Challenges and Opportunities in Undergraduate Physics Education* (National Academies Press, Washington, DC, 2013).

[3] S. Freeman et al, Proceedings of the National Academy of Sciences, **111**, 8410-8415 (2014).

[4] M. Stains, et al. *Science*, American Association for the Advancement of Science, **30** 1468-1470 (2018),

[5] C. Henderson, et al. *Journal of Research in Science Teaching* **48**, 952–984 (2011).

[6] R. Khatri et al. Physics Education Research Conference Proceedings, (2015).

[7] M. Borrego, and C. Henderson. *Journal of Engineering Education*, **25**, (2014).

[8] C. Henderson. *American Journal of Physics*, **76**, 179–187 (2008).

[9] S. Chasteen., et al. *Physics Education Research Conference Proceedings*, (2016).

[10] S. Chasteen, R. Chattergoon, *Physical Review Physics Education Research* **16**, (2020).

[11] I. Ajzen, and M. Fishbein, *Understanding Attitudes and Predicting Social Behavior.* (Prentice-Hall, 1980).

[12] I. Ajzen, and M Fishbein, *Predicting and changing behavior: The reasoned action approach*. (Psychology Press, 2011).

[13] A.M. Semanko, J.L. Ladbury, STEM Courses. *Journal for STEM Educ Res* **3,** 387–402 (2020).

[14] M. K. Smith, et al. *Cell Biology Education*, **12**, 618 (2013).

[15] J. Arjoon, X. Xu, and J. Lewis, *Journal of Chemical Education* **90**, 536-545 (2013).

[16] J. Drennan and A. Hyde, Assessment & Evaluation in Higher Education, **33**, 699 (2008).

[17] G.S. Howard, et al, Applied Psychological Measurement, **3**, 1 (1979).

[18] W. K. Adams, et al. *Physical Review Special Topics - Physics Education Research*, **2**, (2006).

[19] M. Vossen Callens, et al. *Journal of College Science Teaching*, **49**, (2019).

[20] C. Henderson, and M. H. Dancy. *Physical Review Special Topics - Physics Education Research*, **3**, (2007).

[21] C. Henderson, and M. H. Dancy. *American Journal of Physics*, **76**, 79–91. (2008).

[22] K. Trigwell, and M. Prosser. *Educational Psychology Review*, **16**, 409–424 (2004).

[23] K. Trigwell, et al. *Higher Education Research & Development*, **24**, 349–360 (2005).