Improving measurement methods for behavior change interventions: opportunities for innovation

Andrea L. Dunn¹*, Kenneth Resnicow² and Lisa M. Klesges³

Theoretically based behavioral interventions have demonstrated effectiveness in stopping smoking, increasing physical activity and improving nutrition in order to prevent major chronic diseases such as cardiovascular disease, cancer and diabetes [1–7]. Despite the development of effective behavior change programs, the magnitude of change in these behaviors has been relatively modest [8, 9]. The rising rates of diabetes and obesity, for example, indicate there is still much to learn about the mechanisms of behavior change and how to maintain newly acquired behavioral skills. One problem that has slowed behavioral intervention research is that the validity and reliability of our measures has sometimes lagged other innovations such as the development of effective tailored interventions [10, 11] or analytical techniques for assessing moderators and mediators of behavior change, hence making it difficult to understand the mechanisms of behavior change and making it difficult to improve our interventions [12–15].

This special issue of Health Education Research provides an opportunity to consider an important advancement in our behavioral measurement methods, specifically, how we can apply item response models to improve our psychometric methods in health education and health behavior research and practice. Although item response modeling (IRM) has been used in educational testing over the last three decades [16], it is an emerging method in health education and health behavior research [17]. As, for example, in health research, IRM is being widely adopted to improve and revise quality of life questionnaires [18–20]. There are many other innovative applications of IRM and as the nine papers in this issue show us, these applications can aid our understanding of the psychometric properties of scales beyond making questionnaires shorter and beyond what most of us learned in our survey development courses based on classical test theory (CTT). In this brief afterword, we highlight current and future applications of IRM and discuss how these methods might help us improve the efficiency of our research. We believe these methods could lead to considerable improvements in intervention methods, understanding the mechanisms of behavior change and developing and refining theoretical models to make them more parsimonious as well as provide a foundation for considering the feasibility of computerized adaptive testing for measures of behavioral constructs.

For those who are not familiar with IRM methods, the two papers by Wilson colleagues [21, 22] provide an excellent tutorial by presenting an example using the Rasch one-parameter model to examine a measure of self-efficacy using both dichotomous and polytomous models and then comparing IRM and CTT methods. In the first paper, readers learn the basics of IRM, including the relationship of individual items as well as the role of item difficulty response patterns. In other words, item response models estimate a result of the underlying construct...
for the individual and a difficulty estimate of the item on the underlying construct. The implications of these features of IRM are likely to be far reaching for health promotion research and practice. Researchers will be able to determine whether items contribute to an overall understanding of a construct and determine if there are gaps in item difficulty which could enable a more informed decision about development of new items. We see the beginnings of work that builds toward precise measures that yield equal accuracy at all levels [23]. For example, in the quality of life measures, the goal is to have equally precise measures across the continuum of chronically ill to healthy individuals. A similar goal could be established for behavioral constructs like self-efficacy, barriers, benefits, enjoyment and many others. The goal is to increase the reliability of test scores and to improve the amount of information obtained along the continuum of the construct measured by the scale and potentially improving the scale’s ability to detect behavior change.

Another potential application of IRM is in the design of tailored interventions. Behavioral scientists have recognized the promise of tailored interventions aimed at discriminating key motivational strategies for individuals but we have lacked methods to effectively differentiate individuals on constructs targeted by these strategies [11, 24, 25]. For example, it is unlikely that current measures of self-efficacy for physical activity clearly differentiate levels among individuals with either high or low self-efficacy. For example, in the case of two people who might score 90 out of 100 on a self-efficacy scale for physical activity, we might want to tailor the self-efficacy advice somewhat differently depending on how they achieved that level of self-efficacy, e.g did confidence increase because of a change of circumstances or because they have been able to work through problems over a long period of time and have developed a lifelong habit. Tailoring interventions is not a new concept and it can be done with CTT, but one of the greatest advantages of IRM is that both the items and total scores are on the same metric means that we know what items a person with low self-efficacy is likely to endorse which strengthen our ability to develop tailored interventions. Therefore, by applying IRM methods, we can build databases to better understand the properties of individual items and the location of these items along a difficulty or frequency of endorsement continuum. Improved precision of items can lead to improved tailoring of interventions. Furthermore, this process can serve as the foundation to develop ‘item banks’ that could be a shared resource among researchers and is instrumental for the development of computerized adaptive testing.

The assembly of item banks is an important first step in computerized adaptive testing which has become the norm in educational testing such as the national nursing licensing examinations and Graduate Record Examination. To illustrate, items on each of these exams are ranked by difficulty. Individuals who answer ‘easy’ items correctly move to next levels of difficulty and do not waste time on ‘easier’ questions. The outcome is a reduction in testing time and increased precision (higher reliability and lower standard error of measurement). Computerized adaptive testing is under development in the quality of life assessment and could also be used for the assessment of many constructs used in health promotion research and practice. This long-term goal would require an iterative process of developing an initial item bank of the different theoretical constructs to be measured. Gaps in the item bank would then be determined using IRM methods and new items developed where needed. IRM could then be used again to determine difficulty ratings of banked items in order to be able to begin to develop more efficient construct measures [23]. For example, individuals who show high levels of efficacy may be administered only items for the more difficult end of the efficacy distribution. A key issue here is whether behavioral health represents skills or abilities that map onto concepts of ‘difficulty’ and whether the field is able to move forward in exploring innovations in measurement methodology.

The papers by Heesch, Mâsse and Dunn [26] Watson, Baranowski and Thompson [27] provide examples of the IRM method applied to several
different scales and demonstrate how we might use IRM to evaluate the psychometric properties of our measures. For example, in the paper by Heesch, three scales are evaluated, namely, physical activity enjoyment [28], barriers to physical activity and benefits of physical activity [29]. The Rasch analyses of these scales demonstrated the unidimensionality of the enjoyment and benefits scales but the barriers scale had two separate factors that seem to be related to time demands and barriers internal to the person. Further, the analyses indicated problematic items within the scales that seemed to tap other minor dimensions. These items could be dropped in a future scale revision. While this finding could have also been determined with CTT, IRM analyses also revealed that the perceived benefits scale was limited in its ability to measure the underlying construct in a large sample of sedentary adults because items were not providing adequate content coverage for those having moderately high to high levels of perceived benefits. The enjoyment scale also did not indicate good item content coverage for those on the low or high end of physical activity enjoyment. As all of these papers have shown, IRM provides detailed information about the performance of the items by identifying ‘difficulty gaps’, which provides an important first step toward improving testing in health promotion research and practice by increasing our understanding of the performance of particular items. In addition to identifying difficulty gaps, IRM also helps to identify ‘response gaps’. For example, the Rasch model not only provides information about whether the distances between responses are equal or not (using the rating scale versus partial credit models) but also tells us about the functioning of the response format using item characteristic curves. By having this information in addition to what CTT provides, we can greatly enhance the efficiencies and understanding of our measurements. This lesson is made even clearer in the paper by Masse, Heesch and Eason [30] comparing confirmatory factor analyses (CFA) with CTT and IRM. Although all of these methods provide useful information on psychometric properties of scales, IRM provides additional but complementary information that cannot be provided by CFA or CTT. IRM extends our capabilities but does not replace the need to consider CFA or CTT approaches.

IRM also provides a method for enhancing our understanding of how various scales purporting to measure comparable constructs might be similar or different. Currently, we have difficulty comparing intervention outcomes (or surveys for that matter) that have administered different scales. Methods of scale linkage are highly desirable when trying to evaluate and contrast the results of several studies. The paper by Masse, Allen and Wilson [31] demonstrates how two eight-item self-regulatory scales can be linked and found that this is possible if there were at least four items in common between the measures. Behavioral science researchers often use different scales of the same construct and frequently items within the scale are similar. The IRM linking method is likely to be especially useful in meta-analytic reviews of studies to be able to evaluate the comparability of results as additional scale linkage papers become available. This is relevant to adaptive testing, as this linking could allow us to analyze variables measured with different versions of instruments.

We are enthusiastic about the potential of IRM to advance the behavioral science measurement field. The implications of IRM analyses hold exciting potential to improve our interventions, our understanding of the mechanisms of behavior change and our evaluation of the theoretical basis for behavior change research. We are grateful to the leadership of Drs. Masse, Baranowski and Wilson for providing a stimulating set of papers and we urge all who are interested in furthering the field of behavioral science to consider IRM to advance our knowledge of the psychometric properties of our measurements. In the near future, there are many existing data sets that could be used to continue to build our knowledge base and we would urge the use of complementary approaches that employ CTT, CFA and IRM methods to gain a more complete picture of the performance of these scales in different populations. In the future, it will be possible to develop item banks and to develop computerized adaptive tests for research. Such an effort will require...
leadership to build the necessary infrastructure for such an undertaking. Given that behavior underlies much of the chronic disease in the world, it is easy to understand the necessity of reaching these intermediate and distant goals.

References

8. Rothman AJ. “Is there nothing more practical than a good theory?”: why innovations and advances in health behavior change will arise if interventions are used to test and refine theory. *Int J Behav Nutr Phys Act* 2004; 1: 11.